Harry Halpin

Social Semantics

The Search for Meaning on the Web

September 12, 2011

Springer

I dedicate this thesis to my father and mother, Harry Halpin Sr. and Rebecca Halpin. One must always remember that our parents knew us before we knew even ourselves.

Foreword

@@FOREWARD BY HENRY S. THOMPSON

Place, month year *Firstname Surname*

Preface

While there may seem to be no more abstract and theoretical pursuit that the study of *meaning* itself, the calling of such a simultaneously overly dramatic and likely ill-fated project nonetheless emerges out of a very concrete struggle with a particular subject matter - the World Wide Web, which is already is now becoming almost a transparent part of our everyday activity. In fact, at this juncture I would hold the Web to part of the very social and cognitive fabric that maintain our very being. As has been noticed by - ofcourse! - Wittgenstein, the aspects of things that are most important for us are hidden because of their simplicity and familiarity. Far from feeling alone and isolated in a lonely world devoid of meaning, I take it for granted that we strive to live in a meaningful world, a world bursting apart at the seams with undiscovered facets; the Web being a particular popular manner of intertwingling these facets together at this moment. Even though there's no a priori reason why that our individual human 'minds' can escape whatever framework they are inhabit to understand the process, respectably called *semantics*, by which meaning somehow exists in a world that is, at least according to the more mature science of physics, ultimately atomic in nature. Yet I do not exist in a world of atoms (or even bytes) but a world rich in of full-blooded coffee, tables, chairs, web-pages, trees, family, and friends. *Representations* are key to my world, from the warmth in my heart the mental image of parents invokes to a tangible relaxation looking upon the Mediterranean from my window.

A friend once said that the world is not composed of atoms, it is composed of stories. Across the Mediterranean, I find the courage of Egypt contagious and I follow their stories, one by one, as external digital photos and text in my Web browser. So this book can be considered the trace of my particular story. Writing a thesis on the Web was sternly looked down upon at my University, and I even remember the Principal distinctly asking me why one would ever want to write a thesis about this newfangled and quite hacked together thing called "the Web." Yet the kindness and support of those mentioned in my acknowledgments, ranging from my Ph.D. advisors Henry S. Thompson and Andy Clark, to my parents, and even to Tim Berners-Lee, who kept my sanity even when it appeared most of the rest of the graduate school at the University of Edinburgh thought I had clearly lost the plot.

There are memories that make the bleak moments of writing endless chapters in my little stone villa in Old Burdiehouse Road worthwhile. One in particular was when I presented the Representational State Transfer architecture to a group of theoretical computer scientists and so explained why the 'back' button on a browser did not work. When Phil Wadler agreed that that inventor of the Web had been correct about how the Web should work, I felt I was perhaps on to something in this study of the Web.

The story I am telling in this book is not exactly the story I had hoped to tell as a graduate student. This book is to a large extent a reworked and re-edited version of my thesis, and as such suffers from the problems that any thesis has, namely that the studies it compromises were done as small shots in the dark in order to reveal some aspect of a much larger and more sophisticated question. Of these studies, the ones I did most quickly such as my study on tagging have so far received the most acclaim, while the ones I was most intellectually attached to have so far only garnered a small but eager group of fellow travelers in the 'philosophy of the Web' around me. Looking back on this book, I can only notice that it is essentially a preface to a much larger work that can properly do justice to the question of the Web means not only to our notion of representation and semantics, but to larger questions of cognition and intelligence, which ultimately always are always profoundly ethical questions. This larger endeavor will no doubt be another book in due course. However, in this book we point to the hypothesis that all of these questions are social, and take a stab at what that entails. At the time of writing these studies I did not have access to the massive Web-scale data-sets or processing power needed to formulate a testable theory of social computation, and as I sit here in Yahoo! Research, I cannot but be amazed at the seemingly unlimited amount of computational power I have and the fact that I have an entire copy of the Web accessible from my desktop. So I will simply deliver this book to clear space for this future theoretical framework and sketch the components of an adequate computational operationalization of social semantics. The idea came upon me in a visit to Amsterdam when I first arrived in Europe: Meaning is not something we possess alone, but something we create together. To this day, I can still not think of a better way to phrase the hypothesis of social semantics.

Barcelona, April 2011 *Harry Halpin*

Acknowledgements

The majority of this book was written as my thesis *Sense and Reference on the Web* at the University of Edinburgh, so this book would not have been possible without the support of my community of friends and colleagues across the globe and in Edinburgh. In particular, this thesis would not have been possible had it not been for the unwavering support of my advisor, Henry S. Thompson, who encouraged me to to pursue considering the Web architecture a first-rate citizen of inquiry, a brave act few advisors would have been willing to do. I would also like to thank Andy Clark for philosophical inspiration and Victor Lavrenko for his invaluable help on the empirical evaluation. Conversations and support from other colleagues at Edinburgh, in particular Ewan Klein and Kavita Thomas, has been important. However, even more support has come from the global community of Semantic Web hackers and researchers. I have been particularly privileged to have had numerous discussions with Pat Hayes and Tim Berners-Lee on these subjects, and I hope I have accurately given an exegesis of their debate. My time at Duke, where I have was fortunate enough to study under Fredric Jameson and Michael Hardt, has had a decisive if subterranean influence on this thesis. Various friends and co-authors deserve my gratitude. In particular, I would like to single out Rob Didham, Kavita Thomas, Dan Connolly, Brandon Jourdan, Jochen Leidner, Maciej Zurawski, Priya Reddy, Malamo Korbetis, Claire Grover, Richard Tobin, Peter Buneman, Phil Wadler, Valentin Robu, Jonathan Oppenheimer, Michael Wheeler, Laura Gomez, Piotr Bultoc, Dan Brickley, Orit Halpern, Paolo Bouquet, Nicholas and Rita Tishuk, Ras Al-Majnuun, Alexandre Monnin, and everyone in Bilston, the Forest Cafe, and Carrboro. Others shall not be named to protect the innocent. Lastly, I have found intellectually invaluable my time at the Santa Fe Institute, the Oxford Internet Institute, the Island seminar with Brian Cantwell Smith, and the Interface Seminar at Duke University - and more recently, my time with Peter Mika at Yahoo! Research Barcelona. Special gratitude must go to the late Karen Sparck Jones, who called me out of the blue and encouraged this philosophical approach to information retrieval and semantics when I was first beginning.

Contents

Chapter 1
Introduction

You have abandoned the old domain, the old concepts. Here you are in a new domain, for which new concepts will give you the knowledge. The sign that a real change in locus and problematic has occurred, and that a new adventure is beginning, the adventure of science in development. **Louis Althusser** (1963)

This book is an inquiry into representation. Given the almost impossibly wide scope of possible kinds of questions pertaining to representations, we will deploy an analysis that is simultaneously both historical and scientific by restricting our inquiry to the an investigation of representations on the World Wide Web. Yet regardless of our careful scoping, we will nonetheless be blindly driven into the realm of *semantics*, the hard question of how meaning is assigned to representation - a question that is as hard it seems as that of the more popular hard problem of consciousness Chalmers (1995). The nature of representation is no longer fashionable to even pursue in philosophy or even in artificial intelligence; it is a problem whose immensity overwhelms us. As a subject matter the apparent phenomenon of *reference*, the suspiciously mysterious - and so perhaps even non-existent! - connection between a representation and that which it represents, verges upon the totality of our social relationship with the world. From Plato's theory of Forms to the evolution of representation in artificial life Halpin (2006), science is littered with theories of the semantics, all of which equally purport to solve this thorny matter in one way or another. One would be forgiven in given the lack of clear success of any theory so far that perhaps the question is unscientific or simply intractable in nature, yet that compels us with only a more irresistible attraction.

At first glance, representation and semantics seems strangely old-fashioned, particularly given the current enthusiasm for embodiment in cognitive science, which in its more extreme versions leads to claims of "intelligence without representation" Brooks (1991). Yet this fetish for embodiment may be strangely disciplinary and - although radical on the surface - actually end up being a reactionary fad when viewed within context of a larger landscape outside academic philosophy and cognitive science. In particular, computer science - with the exception of the peculiarly anthropomorphic line of research of AI in robotics - does not seem to care about

embodiment. In his *One Hundred Billion Lines of C++*, computer scientist-turned-philosopher Brian Cantwell Smith notes that in artificial intelligence debates over representation tend to frame the debate as if it were between "classical" logic-based symbolic reasoners and some "connectionist" and "embodied" alternative ranging from neural networks to epigenetic robotics (1997). Smith then goes on to aptly state that the kinds of computational systems discussed in artificial intelligence and philosophy tend to ignore the vast majority of existing systems, for "it is impossible to make an exact estimate, but there are probably something on the order of 10, or one hundred billion lines of C++ in the world. And we are barely started. In sum: symbolic AI systems constitute approximately 0.01% of written software" (1997). What Smith fails to mention is that the same small fraction likely holds true of "non-symbolic AI" computational systems such as robots, artificial life, and old-fashioned connectionist networks (an exception may soon be made for the machine-learning that runs phenomena such as advertising and search on the Web). As raw statistics of deployed systems by themselves hold little intellectual weight, no doubt a philosopher could argue that the vast majority of computational systems may have no impact on our understanding of representation and intelligence. In other words, what the vast majority of the planet is doing with computation and representation - which is increasingly focused on the World Wide Web - is simply intellectually uninteresting. In this book we argue otherwise.

Yet while one can easily deny if anything resembling digital representations exists 'inside the brain,' it is much harder to argue that there are no digital representations on the Web. As one clicks from web-page to web-page, it seems that the Web is nothing but a vast network of digital representations. The thesis of this book is that the wide class of computational systems outside of those traditionally considered by artificial intelligence or philosophy present what Cantwell Smith calls a "middle distance" where questions of representation (and perhaps even intelligence) come to the forefront in a peculiarly obvious manner and are likely more tractable than in the case for humans, given the relative complexity of computers and humans (Smith, 1995). At the present moment, with all the totalizing attraction of a black hole, computational systems the world over are becoming part and parcel of the World Wide Web, described by Tim Berners-Lee - the person widely acclaimed to be the 'inventor' of the Web - as "a universal information space"(1992). We further argue that not only may the Web may not only reveal some general insights about representation, but its unique historical status as the first *actual* universal information space may prompt an entire re-thinking of semantics. When asked to consider this hypothesis, Michael Wheeler - a philosopher who is well-known for his Heideggerian defense of embodiment - surmises that while "the power of the Web as a technological innovation is now beyond doubt" but "what is less well appreciated is the potential power of the Web to have a conceptual impact on cognitive science" and so may provide a new "fourth way" in addition to the "three kinds of cognitive science or artificial intelligence: classical, connectionist, and (something like) embodied-embedded" (2008). While countless papers have been produced on the technical aspects of the Web, very little has been done explicitly on the Web *qua* Web as a subject matter of interest to philosophy. This does not mean there

has not been interest, although the interest has come in particular more from the side of those engineers working on developing the Web rather than those already entrenched in philosophy, linguistics, and artificial intelligence (Halpin et al, 2006; Bouquet et al, 2007, 2008). In this spirit, what we will undertake in this thesis as a whole is to apply many well-known philosophical theories of reference and representation to the phenomenon of the Web, and see which theory survives - and lastly, if the Web points a way to a *new* theory of semantics, which we surmise may be a social semantics.

1.1 Scope

The World Wide Web is without a doubt one of the most significant computational phenomena to date. Yet there are some questions that cannot be answered without a *theoretical* understanding of the Web. Although the Web is impressive as a practical success story, there has been little in the way of developing a theoretical framework to understand what – if anything – is different about the Web from the standpoint of long-standing questions of representation and semantics in philosophy. While this situation may have been tolerable so far, serving as no real barrier to the further growth of the Web, with the development of the Semantic Web, a next generation of the Web "in which information is given well-defined meaning, better enabling computers and people to work in cooperation," these philosophical questions come to the forefront, and only a practical solution to them can help the Semantic Web repeat the success of the hypertext Web (Berners-Lee et al, 2001). At this moment, there is little doubt that the Semantic Web faces gloomy prospects - and perhaps for good reason. On first inspection, the Semantic Web appears to be a close cousin to another intellectual project, known politely as 'classical artificial intelligence' (also known as 'Good-Old Fashioned AI') an ambitious project whose progress has been relatively glacial and whose assumptions have been found to be cognitively questionable (Clark, 1997). The initial bet of the Semantic Web was that somehow the *Web* part of the Semantic Web would somehow overcome whatever problems the Semantic Web inherited from classical artificial intelligence, in particular, its reliance on logic and inference as the basis of meaning (Halpin, 2004).

This thesis is explicitly limited in scope, concentrating only on the terminology necessary to phrase a single, if broad, question: How can we determine the meaning of a URI on the Web? Although the thesis is interdisciplinary, as it involves elements as diverse as the philosophy of language and machine-learning, these elements are only harnessed insofar as they are necessary to phrase our central thesis and present a possible solution. Due to constraining ourselves to the scope of the Web and the topic of representation, this thesis is not an attempt to develop a philosophy of computation (Smith, 2002), or a philosophy of information (Floridi, 2004), or even a comprehensive "philosophy of the Web" (Halpin, 2008b). These are much larger projects outside the scope of a single book, and even a single individual's life-long calling. However, in combination with more fully-formed work in the philosophy,

we hope that at least this book provides a starting point for future work in these areas. So we use notions from philosophy selectively, and then define the terms in lieu of our goal of articulating the principles of Web architecture and the Semantic Web, rather than attempting to articulate or define the terms of a systematic philosophy or with reference to the many arguments over these terms in analytic philosophy. Many of the terms in this thesis could be explored much further, but by virtue of our scoping not explored, as to constrain the book to a reasonable size. Unlike a philosophical work, in this book counter-arguments and arguments are generally not given for terminological definitions, but instead references are given to the key works that explicate these terms further.

This thesis does not inspect every single possible answer to the question of *What is the meaning of a URI?*, but only three distinct positions. An inspection of every possible theory of meaning and reference is beyond the scope of the thesis, as is an inspection of the tremendous secondary literature that has accrued over the years. Instead, we will focus only on theories of meaning and representation that have been brought up explicitly in the various arguments over this question in the Web by the primary architects of the Web and the Semantic Web. Our proposed solution of social semantics rests on a theory of meaning, a neo-Wittgensteinian theory, that is one of the most infamously dense and infuriatingly obscure theories of meaning.

Finally, while the experimental components of this book has done its best to be realistic, they are in no way complete. Pains have been taken to ensure that the experiments, unlike much work in the Semantic Web, at least uses real data, feedback from real users, and is properly evaluated over a range of algorithms and parameters. Work on tagging systems takes its data from a real system, *del.icio.us*, as well. While various parts of the experiments could no doubt be optimized and scaled up still further, these experiment should be sufficient to motivate our movement towards social semantics, although a full formalization of such a theory and testing it of would require access to the data of a large-scale search engine such as Google, which for the time being it outside of scope. For future work, we would like to pursue the formalization and large-scale testing of social semantics.

1.2 Summary

The thesis of this book must be stated in a two-fold fashion, first to analyze the problem and then to propose a solution. To analyze the problem of representation on the Web, one must ask the question: **What is the meaning of a URI?**. First, we will must clarify the problem that the Web is a kind of new language that can be defined by its engineering conformance to the principles of Web architecture, but nonetheless inherits the problems regarding sense and reference from the philosophy of natural language. So there is no easy way out of the hard question of representation. Our proposed answer is then that only a theory of representation and semantics that takes into account the socially grounded use of a multiplicity of representations is sufficient to provide the meaning of a representation on the

Web, from which the meaning of a peculiar URI can be derived. In essence, we turn the question on its head; instead of saying that a URI can have its meaning only by virtue of what representations can be accessed from it, we instead say that the network of representations and their use provides the meaning of a URI.

In order to orient the reader to the Web, we give an extended introduction to its history and its architecture in Chapter 2, while introducing the philosophical terminology in concert with examples from the Web that undergirds the rest of the book. In Chapter 3 we propose that the Semantic Web, as embodied by the Resource Description Framework (RDF), is a kind of URI-based knowledge representation language for data integration based on the principles of Web architecture, and illustrate it by providing the elements of Web architecture in terms of a formal Semantic Web ontology. These works have in earlier forms been published as *An Ontology of Resources: Solving the Identity Crisis* Halpin and Presutti (2009) with Valentina Presutti and my very early essay *The Semantic Web: The Origins of Artificial Intelligence Redux* Halpin (2004).

In 4 we illustrate the crisis of the Semantic Web: There is no answer to the aforementioned question of how to assign meaning to a URI. There are at least two distinct positions to this question on the Semantic Web, each corresponding to a distinct philosophical theory of semantics. The first response is the ***logicist position***, which states that *the meaning of a URI is determined by whatever model(s) satisfy the formal semantics of the Semantic Web* (Hayes, 2004). This answer is identified with both the formal semantics of the Semantic Web itself and the traditional Russellian theory of names and its descriptivist descendants (Russell, 1905). While this answer may be sufficient for automated inference engines, this answer is insufficient for humans, as it often crucially under-determines what kind of things the URI identifies. As the prevailing position in early Semantic Web research, this position has borne little fruit. Another response is the ***direct reference position*** for the Web, which states that *the meaning of a URI is whatever was intended by the owner.* This answer is identified with the intuitive understanding of many of the original Web architects like Berners-Lee and a special case of Putnam's 'natural kind' theory of meaning. This position is also a near relative to Kripke's famous response to Russell (Kripke, 1972; Putnam, 1975). Further positions that have been marginal to the debate on the Web, such as that of semiotics, are not explored. An earlier version of this work has been previously published as *Sense and Reference on the Web* in the journal *Minds and Machines* Halpin (2011).

Then we dive from the heights of theory to the depths of experimental work. In Chapter **??**, we begin the exploration of an alternative form of discovering the meaning of a representation, namely that of 'bottom-up' collaborative tagging systems, where users simply 'tag' a resource with a term they find useful or descriptive and so define the 'sense' of a URI as a set of terms. We commit a number of experiments to determine if these tags converge over time and over a diversity of resources. Then in Chapter **??** we extend this exploration to search engines, considering the 'bag-of-words' produced by a document to be equivalent to a set of tags, and so, the sense of the URI. In particular, we explore this using documents from both the Semantic Web and the hypertext Web, and use relevance models to combine them. The

study of tagging was previously published as *The Complex Dynamics of Collobora-tive Tagging* in *ACM Transactions on the Web* with Valentin Robu and Hana Shepard Halpin et al (2007); Robu et al (2009), while the short user study was published with Dirk Bollen as *An Experimental Analysis of Suggestions in Collaborative Tagging* Bollen and Halpin (2009). The study of search engines and relevance feedback was previously published as *Relevance Feedback between Web Search and the Semantic Web* with Victor Lavrenko Halpin and Lavrenko (2011).

We finally turn to formulate a third position in Chapter **??**, the ***social semantics***, which states that since the Web *is a form of language, and as language exists as a public mechanism among multiple agents, then the meaning of a URI is determined by the socially-grounded use of networks of representations on the Web by ordinary users.* As vague as this position seems at first glance, we argue this analysis of meaning and representation is the best fit to how natural language works, and it supersedes and even subsumes the two other positions. Furthermore, it goes beyond a certain quietism about natural language attributed to Wittgenstein as well as a certain belief in the occult powers of some 'mental' lexicon. Ideas in this version were previously published with Andy Clark and Michael Wheeler as *Towards a Philosophy of the Web: Representation, Enaction, Collective Intelligence* Halpin et al (2010). The entire Ph.D. thesis was submitted and approved to University of Edinburgh, with Yorick Wilks being the external examiner, as *Sense and Reference on the Web* Halpin (2009b), with the precis being published with Henry S. Thompson as *Social Meaning on the Web: From Wittgenstein to Search Engines* in *IEEE Intelligent Systems* Halpin and Thompson (2009). Note that all previously published versions of work in this book have been edited, amended, and otherwise expanded.

As Wittgenstein would say, one must remember that every "language game" comes with, a "form of life" (Wittgenstein, 1953), and the Web is a new form of life that goes beyond the philosophy of natural language, and leads us straight into a new philosophy of dynamic machinic and human assemblages, a philosophy-to-come of collective intelligence.

1.3 Notational Conventions

In order to aid the reader, this book employs a number of notational conventions. In particular, we only use "double" quotes to quote a particular author or other work. When a new word is introduced and used in an unusual manner to be clarified later, we use 'single' quotes. The use of 'single' quotes is also used when a word is supposed to be understood as the word *qua* word, a mention of the word, rather than a use of the word. When a term is defined, the word is first labeled using ***bold and italic*** fonts, and either immediately followed or preceded by the definition given in *italics*. Mathematical or formal terms are *italicized*, as is the use of *emphasis* in any sentence. Finally, the names of books and other works are often italicized. In general, technical terms like HyperText Transport Protocol (HTTP) are often abbreviated by their capitalized initials. The World Wide Web is usually referred to by

the web. One of the largest problems of this whole area historically has had a rather ad-hoc use of terms, and we hope this fairly rigorous notational convention helps separate the use, mention, definition, and direct quotations of words.

Chapter 2
Architecture of the World Wide Web

All the important revolutions that leap into view must be preceded in the spirit of the era by a secret revolution that is not visible to everyone, and still less observable by contemporaries, and that is as difficult to express in words as it is to understand.
G.W. F. Hegel (1959)

In order to establish the relative autonomy of the Web as a subject matter, we recount its origins and so its relationship to other projects, both intellectual such as Engelbart's Human Augmentation Project, as well as more purely technical projects such as the Internet (1962). It may seem odd to begin this book, which involves very specific questions about representation and meaning on the Web, with a historical analysis of the Web. To understand these questions we must first have an understanding of the boundaries of the Web and the normative documents that define the Web. The Web is a fuzzy and ill-defined subject matter - often considered a ill-defined 'hack' by both academic philosophers and computer scientists - whose precise boundaries and even definition are unclear. Unlike some subject matters like chemistry, the subject matter of the Web is not necessarily very stable, like a 'natural kind,' as it is a technical artifact subject to constant change. So we will take the advice of the philosopher of technology Gilbert Simondon, "Instead of starting from the individuality of the technical object, or even from its specificity, which is very unstable, try to define the laws of its genesis in the framework of this individuality or specificity, it is better to invert the problem: it is from the criterion of the genesis that we can define the individuality and the specificity of the technical object: the technical object is not this or that thing, given *hic et nunc* but that which is generated" (1958). In other words, we must first trace the creation of the Web before attempting to define it, imposing on the Web what Fredric Jameson calls "the one absolute and we may even say 'transhistorical' imperative, that is: Always historicize!" (1981). Only once we understand the history and signficance of the Web, will we then proceed to dissect its components one-by-one, and attempt to align them with certain still-subterranean notions from philosophy.

2.1 The History of the Web

What is the Web, and what is its significance? At first, it appears to be a relative upstart upon the historical scene, with little connection to anything before it, an ahistorical and unprincipled 'hack' that came unto the world unforeseen and with dubious academic credentials. The intellectual trajectory of the Web is a fascinating, if unknown, revolution whose impact has yet to be historically comprehended, perhaps even by its creators. Although it is well-known that the Web bears some striking similarity to Vannevar Bush's 'Memex' idea from 1945, the Web is itself usually thought more of as a technological innovation rather than an intellectually rich subject matter such as artificial intelligence or cognitive science (1945). However, the Web's heritage is just as rich as artificial intelligence and cognitive science, and can be traced back to the selfsame root, namely the 'Man-Machine Symbiosis' project of Licklider (1960).

2.1.1 The Man-Machine Symbiosis Project

The first precursor to the Web was glimpsed, although never implemented, by Vannevar Bush, chief architect of the military-industrial complex of the United States of America. For Bush, the primary barrier to increased productivity was the lack of an ability to easily recall and create records, and Bush saw in microfiche the basic element needed to create what he termed the "Memex," a system that lets any information be stored, recalled, and annotated through a series of "associative trails" (1945). The Memex would lead to "wholly new forms of encyclopedias with a mesh of associative trails," a feature that became the inspiration for "linking" in hypertext (Bush, 1945). However, Bush could not implement his vision on the analogue computers of his day.

The Web had to wait for the invention of digital computers and the Internet, the latter of which bears no small manner to debt to the work of J.C.R. Licklider, a disciple of Norbert Wiener (Licklider, 1960). Wiener thought of feedback as an overarching principle of organization in any science, one that was equally universal among humans and machines (1948). Licklider expanded this notion of feedback loops to that of feedback between humans and digital computers. This vision of 'Man-Machine Symbiosis' is distinct and prior to cognitive science and artificial intelligence, both of which were very infantile disciplines at the time of Licklider, and both of which are conjoined at the hip by hypothesizing that the human mind can be construed as either computational itself or even implemented on a computer. Licklider was not a true believer in the computational mind, but held that while the human mind itself might not be computational (Licklider cleverly remained agnostic on that particular gambit), the human mind was definitely *complemented* by computers. As Licklider himself put it, "The fig tree is pollinated only by the insect Blastophaga grossorun. The larva of the insect lives in the ovary of the fig tree, and there it gets its food. The tree and the insect are thus heavily interdependent:

the tree cannot reproduce without the insect; the insect cannot eat without the tree; together, they constitute not only a viable but a productive and thriving partnership. This cooperative 'living together in intimate association, or even close union, of two dissimilar organisms' is called symbiosis. The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today" (1960). The goal of 'Man-Machine Symbiosis' is then the enabling of reliable coupling between the humans and their 'external' information as given in digital computers. To obtain this coupling, the barriers of time and space needed to be overcome so that the symbiosis could operate as a single process. This required the invention of ever decreasing low latency feedback loops between humans and their machines.

In pursuit of that goal, the 'Man-Machine Symbiosis' project was not merely a hypothetical theoretical project, but an concrete engineering project. In order to provide the funding needed to assemble what Licklider termed his "galactic network" of researchers to implement the first step of the project, Licklider became the institutional architect of the Information Processing Techniques Office at the Advanced Research Projects Agency (ARPA) (Waldrop, 2001). Licklider first tackled the barrier of time. Early computers had large time lags in between the input of a program to a computer on a medium such as punch-cards and the reception of the program's output. This lag could then be overcome via the use of time-sharing, taking advantage of the fact that the computer, despite its centralized single processor, could run multiple programs in a non-linear fashion. Instead of idling while waiting for the next program or human interaction, in moments nearly imperceptible to the human eye, a computer would share its time among multiple humans (McCarthy, 1992).

In further pursuit of its goal of human-machine symbiosis, in which some over-enthusiastic science-fiction fans or academics with a penchance for the literal might see the idea of a cyborg, the 'Man-Machine Symbiosis' project gave funding to two streams of research: artificial intelligence and another lesser-known strand, the work on 'human augmentation' exemplified by the Human Augmentation Project of Engelbart (1962). Human augmentation, instead of hoping to replicate human intelligence as artificial intelligence did, only thought to enhance it. At the same time Licklider was beginning his 'Man-Machine Symbiosis' project, Douglas Engelbart had independently generated a proposal for a "Human Augmentation Framework' that shared the same goal as the 'Man-Machine Symbiosis' idea of Licklider, although it differed by placing the human at the centre of the system, focusing on the ability of the machine to extend to the human user. In contrast, Licklider imagined a more egalitarian partnership between humans and digital computers, more akin to having a somewhat intelligence machine as a conversational partner for the human (1962). This focus on human factors led Engelbart to the realization that the primary reason for the high latency between the human and the machine was the interface of the human user to the machine itself, as a keyboard was at best a limited channel even compared to punchcards. After extensive testing of what devices enabled the lowest latency between humans and machines, Engelbart invented the mouse and

other, less successful interfaces, like the one-handed ‘chord’ keyboard (Waldrop, 2001). By employing these interfaces, the temporal latency between humans and computers was decreased even further. Strangely enough, we have not - despite all the hyperbole around tactile or haptic interfaces from various media-labs - gone far beyond keyboards, mice, and touch-screens in fifty years.

2.1.2 The Internet

The second barrier to be overcome was space, so that any computer should be accessible regardless of its physical location. The Internet “came out of our frustration that there were only a limited number of large, powerful research computers in the country, and that many research investigators who should have access to them were geographically separated from them” (Leiner et al, 2003). Licklider’s lieutenant Bob Taylor and his successor Larry Roberts contracted out Bolt, Beranek, and Newman (BBN) to create the Interface Message Processor, the hardware needed to connect the various time-sharing computers of Licklider’s “galactic network” that evolved into the ARPANet Waldrop (2001). While BBN provided the hardware for the ARPANet, the software was left undetermined, so an informal group of graduate students constituted the Internet Engineering Task Force (IETF) to create software to run the Internet (Waldrop, 2001).

The IETF has historically been the main standardization body that creates the protocols that run the Internet. It still maintains the informal nature of its foundation, with no formal structure such as a board of directors, although it is officially overseen by the Internet Society. The IETF informally credits as their main organizing principle the credo “We reject kings, presidents, and voting. We believe in rough consensus and running code” (Hafner and Lyons, 1996). Decisions do not have to be ratified by consensus or even majority voting, but require only a rough measure of agreement on an idea. The most important product of these list-serv discussions and meetings are IETF RFCs (Request for Comments) which differ in their degree of reliability, from the unstable ‘Experimental’ to the most stable ‘Standards Track.’ The RFCs define Internet standards such as URIs and HTTP (Berners-Lee et al, 1996, January 2005). RFCs, while not strictly academic publications, have a *de facto* normative force on the Internet and therefore on the Web, and so they will be referenced considerably throughout this book.

Before the Internet, networks were assumed to be static and closed systems, so one either communicated with a network or not. However, early network researchers determined that there could be “open architecture networking” where a meta-level “internetworking architecture” would allow diverse networks to connect to each other, so that “they required that one be used as a component of the other, rather than acting as a peer of the other in offering end-to-end service” (Leiner et al, 2003). In the IETF, Robert Kahn and Vint Cerf devised a protocol that took into account, among others, four key factors, as cited below (Leiner et al, 2003):

1. Each distinct network would have to stand on its own and no internal changes could be required to any such network to connect it to the Internet.
2. Communications would be on a best effort basis. If a packet didn't make it to the final destination, it would shortly be retransmitted from the source.
3. Black boxes would be used to connect the networks; these would later be called gateways and routers. There would be no information retained by the gateways about the individual flows of packets passing through them, thereby keeping them simple and avoiding complicated adaptation and recovery from various failure modes.
4. There would be no global control at the operations level.

In this protocol, data is subdivided into 'packets' that are all treated independently by the network. Data is first divided into relatively equal sized packets by TCP (Transmission Control Protocol), which then sends the packets over the network using IP (Internet Protocol). Together, these two protocols form a single protocol, TCP/IP (Cerf and Kahn, 1974). Each computer is named by an Internet Number, a four byte destination address such as *152.2.210.122*, and IP routes the system through various black-boxes, like gateways and routers, that do not try to reconstruct the original data from the packet. At the recipients end, TCP collects the incoming packets and then reconstructs the data.

The Internet connects computers over space, and so provides the physical layer over which the universal information space of the Web is implemented. However, it was a number of decades before the latency of space and time became low enough for something like the Web to become not only universalizing in theory, but universalizing in practice, and so actually come into being rather than being merely a glimpse in a researcher's eye. An historical example of attempting a Web-like system before the latency was acceptable would be the NLS (oNLine System) of Engelbart (1962). The NLS was literally built as the second node of the Internet, the Network Information Centre, the ancestor of the domain name system. The NLS allowed any text to be hierarchically organized in a series of outlines, with summaries, giving the user freedom to move through various levels of information and link information together. The most innovative feature of the NLS was a journal for users to publish information in and a journal for others to *link* and comment upon, a precursor of blogs and wikis (Waldrop, 2001). However, Engelbart's vision could not be realized on the slow computers of his day. Although time-sharing computers reduced temporal latency on single machines, too many users sharing a single machine made the latency unacceptably high, especially when using an application like NLS. Furthermore, his zeal for reducing latency made the NLS far too difficult to use, as it depended on obscure commands that were far too complex for the average user to master within a reasonable amount of time. It was only after the failure of the NLS that researchers at Xerox PARC developed the personal computer, which by providing each user their own computer reduced the temporal latency to an acceptable amount (Waldrop, 2001). When these computers were connected with the Internet and given easy-to-use interfaces as developed at Xerox PARC, both temporal and spatial latencies were made low enough for ordinary users to access the Internet. This convergence of technologies, the personal computer and the Internet,

is what allowed the Web to be implemented successfully and enabled its wildfire growth, while previous attempts like NLS were doomed to failure as they were conceived before the technological infrastructure to support them had matured.

2.1.3 The Modern World Wide Web

Perhaps due to its own anarchic nature, the IETF had produced a multitude of incompatible protocols such as FTP (File Transfer Protocol) and Gopher (Postel and Reynolds, October 1985; Anklesaria et al, 1993). While protocols could each communicate with other computers over the Internet, there was no universal format to identify information regardless of protocol. One IETF participant, Tim Berners-Lee, had the concept of a "universal information space" which he dubbed the "World Wide Web" (1992). His original proposal to his employer CERN brings his belief in universality to the forefront, "We should work towards a universal linked information system, in which generality and portability are more important than fancy graphics and complex extra facilities" (Berners-Lee, 1989). The practical reason for Berners-Lee's proposal was to connect the tremendous amounts of data generated by physicists at CERN together. Later as he developed his ideas he came into direct contact with Engelbart, who encouraged him to continue his work despite his work being rejected at conferences like ACM Hypertext 1991.[1]

In the IETF, Berners-Lee, Fielding, Connolly, Masinter, and others spear-headed the development of URIs (Universal Resource Identifiers), HTML (HyperText Markup Language) and HTTP (HyperText Transfer Protocol). Since by being able to reference anything with equal ease due to URIs, a web of information would form based on "the few basic, common rules of 'protocol' that would allow one computer to talk to another, in such a way that when all computers everywhere did it, the system would thrive, not break down" (Berners-Lee, 2000). The Web is a *virtual space for naming information* built on top of the physical infrastructure of the Internet that could move bits around, and it was built through specifications that could be implemented by anyone, "What was often difficult for people to understand about the design was that there was nothing else beyond URIs, HTTP, and HTML. There was no central computer 'controlling' the Web, no single network on which these protocols worked, not even an organization anywhere that 'ran' the Web. The Web was not a physical 'thing' that existed in a certain 'place.' It was a 'space' in which information could exist" (Berners-Lee, 2000).

The very idea of a *universal* information space seemed at least ambitious, if not *de facto* impossible, to many. The IETF rejected Berners-Lee's idea that any identification scheme could be universal. In order to get the initiative of the Web off the ground, Berners-Lee surrendered to the IETF and renamed URIs from *Universal Resource Identifiers* (URIs) to *Uniform Resource Locators* (URLs) (Berners-Lee, 2000). The Web begin growing at a prodigious rate once the employer of Berners-

[1] Personal communication with Berners-Lee.

Lee, CERN, released any intellectual property rights they had to the Web and after Mosaic, the first graphical browser, was released. However, browser vendors started adding supposed 'new features' that soon led to a 'lock-in' where certain sites could only be viewed by one particular corporate browser. These 'browser wars' began to fracture the rapidly growing Web into incompatible information spaces, thus nearly defeating the proposed universality of the Web (Berners-Lee, 2000).

Berners-Lee in particular realized it was in the long-term interest of the Web to have a new form of standards body that would preserve its universality by allowing corporations and others to have a more structured contribution than possible with the IETF. With the informal position of merit Berners-Lee had as the supposed inventor of the Web (although he freely admits that the invention of the Web was a collective endeavour), he and others constituted the World Wide Web Consortium (W3C); a non-profit dedicated to "leading the Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web" (Jacobs, 1999). In the W3C, membership was open to any organization, commercial or non-profit organization. Unlike the IETF, W3C membership came at a considerable membership fee. The W3C is organized as a strict representative democracy, with each member organization sending one member to the Advisory Committee of the W3C, although decisions technically are always made by the Director, Berners-Lee himself. By opening up a "vendor neutral" space, companies who previously were interested primarily in advancing the technology for their own benefit could be brought to the table. The primary product of the World Wide Web Consortium is a W3C Recommendation, a standard for the Web that is explicitly voted on and endorsed by the W3C membership. W3C Recommendations are thought to be similar to IETF RFCs, with normative force due to the degree of formal verification given via voting by the W3C Membership and a set number of implementations to prove interoperability. A number of W3C Recommendations have become very well known technologies, ranging from the vendor-neutral later versions of HTML (Raggett et al, 1999), which stopped the fracture of the universal information space, to XML, which has become a prominent transfer syntax for many types of data (Bray et al, 1998).

This book will cite W3C Recommendations when appropriate, as these are one of the main normative documents that define the Web. With IETF RFCs, these normative standards collectively define the foundations of the Web. It is by agreement on these standards that the Web functions as a whole. However, the rough-and-ready process of the IETF and the more bureaucratic process of the W3C has led to a terminological confusion that must be sorted in order to grasp the nature of representations on the Web, causing even the most well-meaning of souls to fall into a conceptual swamp of undefined and fuzzy terms. This is true in spades in particular over the hotly-contested term 'representation.'

2.2 The Terminology of the Web

Can the various technologies that go under the rubric of the World Wide Web be found to have common principles and terminology? This question would at first seem to be shallow, for one could say that any technology that is described by its creators, or even the public at large, can be considered trivially 'part of the Web.' To further complicate the matter, the terms like the 'Web' and the 'Internet' are elided together in common parlance, and so are often deployed as synonyms. In a single broad stroke, we can distinguish the Web and the Internet. The Internet is a type of packet-switching network as defined by its use of the TCP/IP protocol. The purpose of the Internet is to get bits from one computer to another. In contrast, the Web is a space of names defined by its usage of URIs. So, the purpose of the Web is the use of URIs for accessing and referring to information. The Web and the Internet are then strictly separable, for the Web, as a space of URIs, could be realized on top of other types of networks that move bits around, much as the same virtual machine can be realized on top of differing physical computers. For example, one could imagine the Web being built on top of a network built on principles different from TCP/IP, such as OSI, an early competitor to the TCP/IP stack of networking protocols (Zimmerman, 1980). Likewise, before the Web, there were a number of different protocols with their own naming schemes built upon the Internet like Gopher (Anklesaria et al, 1993).

Is it not presumptuous of us to even hope that such an unruly phenomenon such as the Web even has guiding principles? Again we must appeal to the fact that unlike natural language or chemistry, the Web is like other engineered artifacts, created by particular individuals with a purpose, and designed with this purpose in mind. Unlike the case of the proper function of natural language, where natural selection itself will forever remain silent to our questions, the principal designers of the Web are still alive to be questioned in person, and their design rationale is overtly written down on various notes, often scribbled on some of the earliest web-pages of the Web itself. It is generally thought of that the core of the Web consists of the following standards, given in their earliest incarnation, HTTP (Berners-Lee et al, 1996), URI (Berners-Lee, 1994a), and HTML (Berners-Lee and Connolly, June 1993). So the basic protocols and data formats that proved to be successful were the creation of a fairly small number of people, such as Tim Berners-Lee, Roy Fielding, and Dan Connolly.

The primary source for our terminology and principles of Web architecture is a document entitled *The Architecture of the World Wide Web* (AWWW), a W3C Recommendation edited by Ian Jacobs and Norm Walsh to "describe the properties we desire of the Web and the design choices that have been made to achieve them" (Jacobs and Walsh, 2004). The AWWW is an attempt to systematize the thinking that went into the design of the Web by some of its primary architects, and

as such is both close to our project and an inspiration..[2] In particular, AWWW is an exegesis of Tim Berners-Lee's notes on "Design Issues: Architectural and philosophical points"[3] and Roy Fielding's dissertation "Architectural Styles and the Design of Network-based Software Architectures" (Fielding, 2010), often abbreviated as REST. The rationale for the creation of such a document of principles developed organically over the existence of the W3C, as new proposed technologies were sometimes considered to be either informally compliant or non-compliant with Web architecture. When the proponents of some technology were told that their particular technology was not compliant with Web architecture, they would often demand that somewhere there be a description of this elusive Web architecture. The W3C in response set up the Technical Architecture Group (TAG) to "document and build consensus" upon "the underlying principles that should be adhered to by all Web components, whether developed inside or outside W3C," as stated in its charter.[4] The TAG also maintains a numbered list of problems (although the numbers are in no way sequential) that attempts to resolve issues in Web architecture by consensus, with the results released as notes called 'W3C TAG findings,' which are also referred to in this discussion. The TAG's only Recommendation at the time of writing is the aforementioned *Architecture of the Web: Volume 1* but it is reasonable to assume that more volumes of *Architecture of the Web* may be produced after enough findings have been accumulated. The W3C TAG's AWWW is a blend of common-sense and sometimes surprising conclusions about Web architecture that attempts to unify diverse web technologies with a finite set of core design principles, constraints, and good practices (Jacobs and Walsh, 2004). However, the terminology is AWWW is often thought to be too informal and ungrounded to use by many, and we attempt to remedy this in the next few chapters by fusing the terminology of Web architecture with our own peculiar brand of philosophical terminology.

To begin our reconstruction of Web architecture, the first task is the definition of terms, as otherwise the technical terminology of the Web can lead to as much misunderstanding as understanding. To cite an extreme example, people coming from communities like the artificial intelligence community use terms like 'representation' in a way that is different from those involved in Web architecture. We begin with the terms commonly associated with a typical exemplary Web interaction. For an agent to learn about the *resource* known as the Eiffel Tower in Paris, a person can access its *representation* using its *Uniform Resource Identifier (URI)* `http://www.tour-eiffel.fr/` and retrieve a webpage in the HTML *encoding* whose *content* is the Eiffel Tower using the HTTP *protocol*.

---

[2] Although to what extent the Web as it actually exists follows these design choices is still a matter for debate, and it is very clear some of the more important parts of the Web such as the ubiquity of scripting languages, and thus HTML as mobile code, are left unmentioned.

[3] There exist a collection of unordered personal notes available at: `http://www.w3.org/DesignIssues/`, which we also refer directly to in the course of this chapter.

[4] Quoted from their charter, available on the Web at: `http://www.w3.org/2001/07/19-tag` (last accessed April 20th, 2007).

2.2.1 Protocols

A ***protocol*** is *a convention for transmitting information between two or more agents*, a broad definition that encompasses everything from computer protocols like TCP/IP to conventions in natural language like those employed in diplomacy. A protocol often specifies more than just the particular encoding, but also may attempt to specify the interpretation of this encoding and the meaningful behaviour that the sense of the information should engender in an agent. An ***agent*** is *any thing capable of interacting via a protocol*. These are often called a 'user agent' on the Web, and the term covers both web-browsers, humans, web spiders, and even combinations such as humans operating web-browsers. A ***payload*** is *the information transmitted by a protocol*. Galloway notes that protocols are "the principle of organization native to computers in distributed networks" and that agreement on protocols are necessary for any sort of network to succeed in the acts of communication (2004).[5] The paradigmatic case of a protocol is TCP/IP, where the payload transmitted is just bits in the body of the message, with the header being used by TCP to ensure the lossless delivery of said bits. TCP/IP transmits strictly an encoding of data as bits and does not force any particular interpretation on the bits; the payload could be a picture of the Eiffel Tower, web-pages about the Eiffel Tower, or just meaningless random bits. All TCP/IP does is move some particular bits from one individual computer to another, and any language that is built on top of the bit-level are strictly outside the bounds of TCP/IP. Since these bits are usually communication with some purpose, the payload of the protocol is almost always an encoding on a level of abstraction above and beyond that of the raw bits themselves.

The Web is based on a ***client-server architecture***, meaning that *protocols take the form of a request for information and a response with information*. The ***client*** is defined as *the agent that is requesting information* and the ***server*** is defined as *the agent that is responding to the request*. In a protocol, an ***endpoint*** is *any process that either requests or responds to a protocol*, and so includes both client and servers. The client is often called a ***user-agent*** since it is the user of the Web. A user-agent may be anything from a web-browser to some sort of automated reasoning engine that is working on behalf of another agent, often the specifically human user. The main protocol in this exposition will be the ***HyperText Transfer Protocol*** (HTTP), as most recently defined by IETF RFC 2616 (Fielding et al, 1999). HTTP is a protocol originally intended for the transfer of hypertext documents, although its now ubiquitous nature often lets it be used for the transfer of almost any encoding over the Web, such as its use to transfer XML-based SOAP (originally the *Simple Object Access Protocol*) messages in Web Services (Box et al, 2000). HTTP consists of sending a ***method***, *a request for a certain type of response from a user-agent to the server*, including information that may change the state of the server. These methods have a list of ***headers*** that *specify some information that may be of used by the*

[5] Although unlike Galloway, instead of descending into a sort of postmodern paranoia of protocols, we recognize them as the very conditions of collectivity.

server to determine the response. The ***request*** is *the method used by the agent and the headers, along with a blank line and an optional message body*.

The methods in HTTP are HEAD, GET, POST, PUT, DELETE, TRACE, OPTIONS, and CONNECT. We will only be concerned with the most frequently used HTTP method, GET. GET is informally considered 'commitment-free,' which means that the method has no side effects for either the user-agent or the server, besides the receiving of the response (Berners-Lee et al, 1996). So a GET method should not be used to change the state of a user-agent, such as charging someone for buying a plane ticket to Paris. To change the state of the information on the server or the user-agent, either PUT (for uploading data directly to the server) or POST (for transferring data to the server that will require additional processing, such as when one fills in a HTML form) should be used. A sample request to `http:///www.example.org` from a Web browser user-agent is given in Figure 2.1.

```
GET /index.html HTTP/1.0
User-Agent: Mozilla/5.0
Accept: */*
Host: www.example.org
Connection: Keep-Alive
```

Fig. 2.1 An HTTP Request from a client

The first part of an HTTP response from the server then consists of an HTTP ***status code*** which is *one of a finite number of codes which gives the user-agent information about the server's HTTP response itself.* The two most known status codes are HTTP 200, which means that the request was successful, or 404, which means the user-agent asked for data that was not found on the server. The first digit of the status code indicates what general class of response it is. For example, the two hundred series (`2xx`) response codes mean a successful request, although 206 means partial success. The `4xx` codes indicate that the user-agent asked for a request that the server could not fulfill, while `1xx` is informational, `3xx` is redirectional, and `5xx` means server error. After the status codes there is an ***HTTP entity*** which is "*the information transferred as the payload of a request or response*" (Fielding et al, 1999). This technical use of the word 'entity' should be distinguished from our earlier use of the term 'entity' like the Eiffel Tower who can only be realized by the thing itself, not in another realization. In order to do so, we will take care to preface the protocol name 'HTTP' before any 'HTTP entity,' while the term 'entity' by itself refers to the philosophical notion of an entity. An HTTP entity consists of "entity-header fields and... an entity-body" (Fielding et al, 1999) An ***HTTP response*** consists of *the combination of the status code and the HTTP entity*. These responses from the server can include an additional header, which specifies the date and last modified date as well as optional information that can determine if the desired representation is in the cache and the content-type of the representation. A sample HTTP

response to the previous example request, excluding the HTTP entity-body, is given in Figure 2.2.

```
HTTP/1.1 200 OK
Date: Wed, 16 Apr 2008 14:12:09 GMT
Server: Apache/2.2.4 (Fedora)
Accept-Ranges: bytes
Connection: close
Content-Type: text/html; charset=ISO-8859-1
Content-Language: fr
```

Fig. 2.2 An HTTP Response from a server

In the HTTP response, an HTTP entity body is returned. The encoding of the HTTP entity body is given by the HTTP entity header fields that specify its `Content-type` and `Content-language`. These are both considered different languages, as a single webpage can be composed in multiple languages, such as the text being given in English with various formatting given in HTML. Every HTTP entity body should have its particular encoding specified by the `Content-type`. *The formal languages that can be explicitly given in a response or request in HTTP* are called ***content types***. In the example response, based on the header that the content type is `text/html` a user-agent can interpret ('display as a web-page') the encoding of the HTTP entity body as HTML. Since the same encoding can theoretically represent many different languages besides HTML, a user-agent can only know definitely how to process a message through the content type. If no content type is provided, the agent can guess the content type through various heuristics including looking at the bytes themselves, a process informally called *sniffing*. A user-agent can specify what media types they (can) prefer, so that a web-server that can only present JPEG images can specify this by also asking for the content type `image/jpeg` in the request.

Content-types in HTTP were later generalized as 'Internet Media Types' so they could be applied with any Internet protocol, not just HTTP and MIME (*Multimedia Internet Message Extensions*, an e-mail protocol) (Postel, March 1994). A ***media type*** consists of *a two-part scheme that separates the type and a subtype of an encoding*, with a slash indicating the distinction. Internet media types are centrally registered with IANA at `http://www.iana.org/assignments/media-types/`, although certain 'experimental' media types (those beginning with 'x-') can be created in a decentralized manner (Postel, March 1994). A central registry of media types guarantees the interoperability of the Web, although increasingly new media-types are dependent on extensions to specific applications (plug-ins) in order to run. Support for everything from new markup languages to programming languages such as Javascript can be declared via support of its media type.

To move from concrete bits to abstract definitions, a protocol can be defined and implemented in many different types of way. In the early ARPANet, the first wide-area network and foundation of the Internet, the protocol was 'hard-wired' in the

hardware of the Interface Message Processor (IMP), a separate machine attached to computers in order to interface them with ARPANet (Hafner and Lyons, 1996). As more and more networks multiplied, these heterogeneous networks began using different protocols. While the invention of TCP/IP let these heterogeneous networks communicate, TCP/IP does not interpret messages beyond bits. Further protocols built on top of TCP/IP, such as FTP (File Transfer Protocol) for the retrieval of files (Postel and Reynolds, October 1985), Gopher for the retrieval of documents (Anklesaria et al, 1993), and SMTP (Simple Mail Transfer Protocol) for the transfer of mail (Postel, August 1982). Since one computer might hold many different kinds of information, IP addresses were not enough as they only identified where a particular device was on the network. Thus each protocol created its own naming scheme to allow it to identify and access things on a more fine-grained level than IP addresses. Furthermore, each of these protocols was often associated (via registration with a governing body like IANA, the *Internet Assigned Numbers Authority*) with particular ports, such that port 25 was used by SMTP and port 70 by Gopher. With this explosion of protocols and naming schemes, each Internet application was its own 'walled garden.' Names created using a particular protocol were incapable of being used outside the original protocol, until the advent of the naming scheme of the Web (Berners-Lee, 2000).

2.2.2 Information Encoding and Content

There is a relationship between a server sending a message - such as a web-page about the Eiffel Tower - to a client in response to an HTTP request and certain notions from information theory, however hazy and qualitative. To phrase informally, ***information*** is *whatever regularities held in common between a* ***source*** *and a* ***receiver*** (Shannon and Weaver, 1963). Note that the source and receiver do not have to be spatially separate, but can also be temporally separate, and thus the notion of a self-contained 'message' resembling a postcard being sent between sender and receiver is incomplete if not incorrect.[6] To have something in common means to share the same regularities, e.g. parcels of time and space that cannot be distinguished at a given level of abstraction. This definition correlates with information being the inverse of the amount of 'noise' or randomness in a system, and the amount of information being equivalent to a reduction in uncertainty. It is precisely this preservation or failure to preserve information that can be thought of the as sending of a *message* between the source and the receiver over a channel, where the channel is over time, space, and - most likely - both. *Whether or not the information is preserved over time or space is due to the properties of a physical substrate* known as the ***channel***. So in our example, the channel is the fiber-optic or copper wires that must accurately carry the voltages which the bits consist of. The ***message*** is *the physical thing that realizes the regularities of the information due to its local characteristics*, which in

[6] Imagine that your eye color not changing is a message from yourself at ten years old to yourself at seventy!

this case would be particular patterns of bits being preserved over multiple channels as they are popped from an electro-magnetic hard-disk on a server to fibre-optic then over the air via wireless and finally back to the electric charges stored in memory chips in a client device, such as a web browser on a mobile phone. These messages are often called the *realization* of some abstract informational content.

Already, information reveals itself to be not just a singular thing, but something that exists at multiple levels: How do the bits become a message in HTTP? In particular, we are interested in the distinction in information between content and encoding. Here our vague analogy with Shannon's information theory fails, as Shannon's theory deals with finding the optimal encoding and size of channel so that the message can be guaranteed to get from the sender to the receiver, which in our case is taken care of by the clever behavior of the TCP/IP protocol operating over a variety of computational devices (Shannon and Weaver, 1963). Yet, how can an encoding be distinguished from the content of information itself in a particular HTTP message? Let's go back to bits by leaning on aesthetic theory of all things; art critic and philosopher Nelson Goodman defines a ***mark*** *as a physical characteristic* ranging from marks on paper one can use to discern alphabetic characters to ranges of voltage that can be thought of as bits (1968). To be reliable in conveying information, an encoding should be physically 'differentiable' and thus maintain what Goodman calls 'character indifference' so that (at least within some context) each character (as in 'characteristic') can not be mistaken for another character. One cannot reconstruct a message in bits if one cannot tell apart 1 and 0, much as one cannot reconstruct a HTML web-page if one cannot tell the various characters in text apart. So, an ***encoding*** is *a set of precise regularities that can be realized by the message*. Thus, one can think of multiple levels of encoding, with the very basic encoding of bits being handled by the protocol TCP/IP, and then the protocol HTTP handing higher-level encodings in textual encodings such as HTML.

Unforunately, we are not out of the conceptual thicket yet; there is more to information than encoding. Shannon's theory does not explain the notion of information fully, since giving someone the number of bits that a message contains does not tell the receiver *what* information is encoded. Shannon explicitly states, "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem" (1963). He is correct, at least for his particular engineering problem. However, Shannon's use of the term 'information' is for our purposes the same as the 'encoding' of information, but a more fully-fledged notion of information is needed. Many intuitions about the notion of information have to deal with not only how the information is encoded or how to encode it, but what a particular message is about, the *content* of an information-bearing message.[7]

[7] An example of the distinguishment between content and encoding: Imagine Daniel sending Amy a secret message about which one of her co-employees won a trip to the Eiffel Tower. Just determining that a single employee out of eight won the lottery requires at least a three bit encoding and does not tell Amy (the receiver) which employee in particular won the lottery. Shannon's theory

'Content' is a term we adopt from Israel and Perry, as opposed to the more confusing term 'semantic information' as employed by Floridi and Dretske (Israel and Perry, 1990; Dretske, 1981; Floridi, 2004). One of the first attempts to formulate a theory of informational content was due to Carnap and Bar-Hillel (1952). Their theory attempted to bind a theory of content closely to first-order predicate logic, and so while their "theory lies explicitly and wholly within semantics" they explicitly do not address "the information which the sender intended to convey by transmitting a certain message nor about the information a receiver obtained with a certain message," since they believed these notions could eventually be derived from their formal apparatus (Carnap and Bar-Hillel, 1952). Their overly restrictive notion of the content of information as logic did not gain widespread traction, and neither did other attempts to develop alternative theories of information such as that of Donald McKay (1955). In contrast, Dretske's *semantic theory of information* defines the notion of content to be compatible with Shannon's information theory, and his notions have gained some traction within the philosophical community (Dretske, 1981). To him, the content of a message and the amount of information – the number of bits an encoding would require – are different, for "saying 'There is a gnu in my backyard' does not have more content than the utterance 'There is a dog in my backyard' since the former is, statistically, less probable" (Dretske, 1981). According to Shannon, there is more information in the former case precisely because it is less likely than the latter (Dretske, 1981). So while information that is less frequent may require a larger number of bits in encoding, the content of information should be viewed as to some extent separable if compatible with Shannon's information theory, since otherwise one is led to the "absurd view that among competent speakers of language, gibberish has more meaning than semantic discourse because it is much more less frequent" (Dretske, 1981). Simply put, Shannon and Dretkse are talking about distinct notions that should be separated, the notions of encoding and content respectively.

Is there a way to precisely define the content of a message? Dretske defines the content of information as "a signal r carries the information that s is F when the conditional probability of s's being F, given r (and k) is 1 (but, given k alone, less than 1). k is the knowledge of the receiver" (1981). To simplify, the ***content*** of any information-bearing message is *whatever is held in common between the source*

only measures how many bits are needed to tell Amy precisely who won. After all, the false message that her office-mate Sandro won a trip to Paris is also three bits. Yet content is not independent of the encoding, for content is conveyed by virtue of a particular encoding and a particular encoding imposes constraints on what content can be sent (Shannon and Weaver, 1963). Let's imagine that Daniel is using a code of bits specially designed for this problem, rather than natural language, to tell Amy who won the free plane ticket to Paris. The content of the encoding `001` could be yet another co-employee Ralph while the content of the encoding `010` could be Sandro. If there are only two possible bits of information and all eight employees need one unique encoding, Daniel cannot send a message specifying which friend got the trip since there aren't enough options in the encodings to go round. An encoding of at least three bits is needed to give each employee a unique encoding. If `01` has the content that 'either Sandro or Ralph won the ticket' the message has not been successfully transferred if the purpose of the message is to tell Amy *precisely* which employee won the ticket.

and the receiver as a result of the conveyance of a particular message. While this is similar to our definition of information itself, it is different. The content is whatever is shared in common as a result of a *particular* message, such as the conveyance of sentence 'The Eiffel Tower is 300 meters high.' The content of a message is called the "facts" by Dretske, (F). This content is conveyed from the source (s) successfully to the receiver (r) when the content can be used by the receiver with certainty, *and* that before the receipt of the message the receiver was not certain of that particular content. Daniel can only successfully convey the content that 'Ralph won a trip to Paris' if before receiving the message Amy does not know 'Ralph won the trip to Paris' and after receiving the message Amy does know that fact. Dretkse himself notes that information "does not mean that a signal must tell us everything about a source to tell us something," it just has to tell enough so that the receiver is now certain about the content within the domain (1981). Millikan rightfully notes that Dretske states his definition too strongly, for this probability of 1 is just an approximation of a statistically "good bet" indexed to some domain where the information was learned to be recognized (2004). For example, lightening carries the content that "a thunderstorm is nearby" in rainy climes but in an arid prairie lightning can convey a dust-storm. However, often the reverse is true, as the same content is carried by messages in different encodings, like a web-page about the Eiffel Tower being encoded in either English or French. These notions of encoding and content are not strictly separable, which is why they together compose the notion of information. An updated famous maxim of Hegel could be applied: For information, there is no encoding without content, and no content without encoding (1959).

The relationship of an encoding to its content, is an ***interpretation***. The interpretation 'fills' in the necessary background left out of the encoding, and maps the encoding to some content. In our previous example using binary digits as an encoding scheme, a mapping could be made between the encoding `001` to the content of the Eiffel Tower while the encoding `010` could be mapped to the content of the Washington Monument. When the word 'interpretation' is used as a noun, we mean the content given by a particular relationship between an agent and an encoding, i.e. the interpretation. Usual definitions of "interpretation" tend to conflate these issues. In formal semantics, the word "interpretation" often can be used either in the sense of "an interpretation structure, which is a 'possible world' considered as something independent of any particular vocabulary" (and so any agent) or "an interpretation mapping from a vocabulary into the structure" or as shorthand for both (Hayes, 2004). The difference in use of the term seems somewhat divided by fields. For example, computational linguistics often use "interpretation" to mean what Hayes called the "interpretation structure." In contrast, we use the term 'interpretation' to mean what Hayes called the "interpretation mapping," reserving the word 'content' for the "interpretation structure" or structures selected by a particular agent in relationship to some encoding. Also, this quick aside into matters of interpretation does not explicitly take on a formal definition of interpretation as done in model theory, although our general definition has been designed to be compatible with model-theoretic and other formal approaches to interpretation.

These terms are all illustrated in Figure 2.3. A source is sending a receiver a message. The information-bearing message realizes some particular encoding such as a few sentences in English and a picture of the Eiffel Tower, and the content of the message can be interpreted to be about the Eiffel Tower.

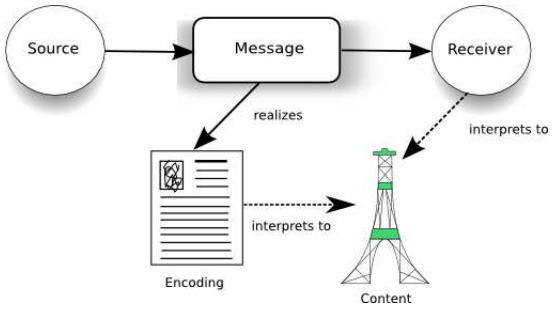

Fig. 2.3 Information, Encoding, Content

The encodings and content of information do not in general come in self-contained bundles, with each encoding being interpreted to some free-standing propositional content. Instead, encodingevs and content come in entire interlocking informational systems. One feature of these systems is that encodings are layered inside of each other and content is also layered upon other content. The perfect example would be an English sentence in an e-mail message, where a series of bits are used to encode the letters of the alphabet, and the alphabet is then used to encode words. Likewise, the content of a sentence may depend on the content of the words in the sentence. When this happens, one is no longer dealing with a simple message, but some form of language. A ***language*** can be defined as *a system in which information is related to other information systematically*. In a language, this is a relationship between how the encoding of some information can change the interpretation of other encodings. Messages always have encodings, and usually these encodings are part of languages. To be more brief, information is *encoded in* languages. The relationships between encodings and content are usually taken to be based on some form of (not necessarily formalizable or even understood) rules. If one is referring to *a system in which the encoding of information is related to each other systematically*, then one is talking about the ***syntax*** of a language. If one is referring to *a system in which the content of information is related to each other systematically*, then one is referring to the ***semantics*** of the language.The lower-level of a language can be ***terms***, *regularities in marks*, that may or may not have their own interpretation, such as the words or alphabet. *Any combination of terms that is valid according to the language's syntax* is a ***sentence*** (sometimes an 'expression') in the language, and *any combination of terms that has an interpretation to content according to the language's semantics* is a ***statement*** in the language.

Particular encodings and content then are accepted by or considered vlaid by the syntax and semantics of a language respectively (and thus the normative importance

of standardization on the Web in determining these criteria. Also, we do not restrict our use of the word ‘language’ to primarily linguistic forms, but use the term ‘language’ for anything where there is a systematic relationship between syntax and (even an informal) semantics. For example HTML is a language for mapping a set of textual tags to renderings of bits on a screen in a web browser. One principle used in the study of languages, attributed to Frege, is the principle of ***compositionality***, where *the content of a sentence is related systematically to terms in which it is composed*. Indeed, while the debate is still out if human languages are truly compositional (Dowty, 2007), computer languages almost always are compositional. In English, the content of the sentence such as ‘Tim has a plane ticket to Paris so he should go to the airport!’ can then be composed from the more elementary content of the sub-statements, such as ‘Tim has a plane ticket’ which in turn can have its content impacted by words such as ‘Paris’ and ‘ticket.’ The argument about whether sentences, words, or clauses are the minimal building block of content is beyond our scope. Do note one result of the distinction between encoding and content is that sentences that are accepted by the syntax (encoding) of a language, such as Chomsky’s famous “Colourless green ideas sleep furiously’ may have no obvious interpretation (to content) outside of the pragmatics of Chomsky’s particular exposition (1957).

2.2.3 Uniform Resource Identifiers

The World Wide Web is defined by the AWWW as “an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URI)” (Jacobs and Walsh, 2004). This naming scheme, not any particular language like HTML, is the primary identifying characteristic of the Web. URIs arose from a need to organize the “many protocols and systems for document search and retrieval” that were in use on the Internet, especially considering that “many more protocols or refinements of existing protocols are to be expected in a field whose expansion is explosive” (Berners-Lee, 1994a). Despite the “plethora of protocols and data formats,” if any system was “to achieve global search and readership of documents across differing computing platforms,” gateways that can “allow global access” should “remain possible” (Berners-Lee, 1994a). The obvious answer was to consider all data on the Internet to be a single space of names with global scope.

URIs accomplish their universality over protocols by moving *all the information used by the protocol within the name itself*. The information needed to identify any protocol-specific information is all specified in the name itself: the name of the protocol, the port used by the protocol, any queries the protocol is responding to, and the hierarchical structure used by the protocol. The Web is then first and foremost a naming initiative “to encode the names and addresses of objects on the Internet” rather than anything to do with hypertext (Berners-Lee, 1994a). The notion of a URI can be viewed as a “meta-name,” a name which takes the existing protocol-

specific Internet addresses and wraps them in the name itself, a process analogous to reflection in programming languages (Smith, 1984). Instead of limiting itself to only existing protocols, the URI scheme also abstracts away from any particular set of protocols, so that even protocols in the future or non-Internet protocols can be given a URI; "the web is considered to include objects accessed using an extendable number of protocols, existing, invented for the web itself, or to be invented in the future" (Berners-Lee, 1994a).

One could question why one would want to name information outside the context of a particular protocol. The benefit is that the use of URIs "allows different types of resource identifiers to be used in the same context, even when the mechanisms used to access those resources may differ" (Berners-Lee et al, January 2005). This is an advantage precisely because it "allows the identifiers to be reused in many different contexts, thus permitting new applications or protocols to leverage a pre-existing, large, and widely used set of resource identifiers" (Berners-Lee et al, January 2005). This ability to access with a single naming convention the immense amount of data on the entire Internet gives an application such as the ubiquitous Web browser a vast advantage over an application that can only consume application-specific information.

Although the full syntax in Backus-Naur form is given in IETF RFC 3986 (Berners-Lee et al, January 2005), a URI can be given as the regular expression `URI= [scheme ":"] [hierarchical component]* [ "?" query ]? [ "#" fragment]?`. First, a ***scheme*** is *a name of the protocol or other naming convention used in the URI.* Note that the scheme of a URI does not determine the protocol that a user-agent has to employ to use the URI. For example, a `HTTP` request may be used on `ftp://www.example.org`. The scheme of a URI merely indicates a preferred protocol for use with the URI. A ***hierarchical component*** is *the left to right dominant component of the URI that syntactically identifies the resource.* URIs are federated, insofar as each scheme identifies the syntax of its hierarchical component. For example, with HTTP the hierarchical component is given by `[authority] [//] [":" port]? ["/" path component]*`. The ***authority*** is *a name that is usually a domain name, naming authority, or a raw IP address, and so is often the name of the server*. However, in URI schemes like `tel` for telephone numbers, there is no notion of an authority in the scheme. The hierarchical component contains special reserved characters that are in HTTP characters such as the backslash for locations as in a file system. For ***absolute URIs***, *there must be a single scheme and the scheme and the hierarchical component must together identify a resource* such as `http://www.example.com:80/monument/EiffelTower` in HTTP, which signals port 80 of the authority `www.example.com` with the path component `/monument/EiffelTower`. The port authority is usually left out, and assumed to be 80 by HTTP-enabled clients. Interestingly enough there are also ***relative URIs*** *in some schemes like HTTP, where the path component itself is enough to identify a resource within certain contexts,* like that of a web-page. This is because the scheme and authority itself may have substituted some special characters that serve as indexical expressions, such as '.' for the current place in the path compo-

nent and '..' as the previous level in the path component. So, `../EiffelTower` is a perfectly acceptable relative URI. Relative URIs have a straightforward translation into absolute URIs, and it is trivial to compare absolute URIs for equality (Berners-Lee et al, January 2005).

The 'hash' (#) and 'question mark' (?) are special characters at the end of URI. The question mark denotes 'query string.' The 'query string' allows for the parameterization of the HTTP request, typically in the cases where the HTTP response is created dynamically in response to specifics in the HTTP request. The 'hash' traditionally declares a ***fragment identifier***, which *identifies fragments of a hypertext document* but according to the TAG, it can also identify a "secondary resource," which is defined as "some portion or subset of the primary resource, some view on representations of the primary resource, or some other resource defined or described by those representations" where the "primary resource" is the resource identified by the URI without reference to either a hash or question mark (Jacobs and Walsh, 2004). The fragment identifier (specified by a 'hash' followed by some string of characters) is stripped off for the request to the server, and handled on the client side. Often the fragment identifier causes the local client to go to a particular part of the accessed HTTP entity. If there was a web-page about Gustave Eiffel, its introductory paragraph could be identified with the URI `http://www.example.com/EiffelTower#intro`. Figure 2.4 examines a sample URI, `http://www.example.org/EiffelTower#intro`:

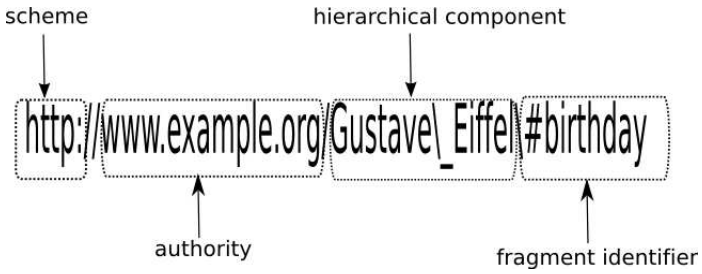

Fig. 2.4 An example URI, with components labelled.

The first feature of URIs, the most noticeable in comparison to IP addresses, is that they can be human-readable, although they do not have to be. As an idiom goes, URIs can be 'written on the side of a bus.' URIs can then have an interpretation due to their use of terms from natural language, such as `www.whitehouse.gov` referring to the White House or the entire executive branch of the United States government. Yet it is considered by the W3C TAG to be bad practice for any agent to depend on whatever information they can glean from the URI itself, since to a machine the natural language terms used by the URI have no interpretation. For an agent, all URIs are opaque, with each URI being just a string of characters that can be used to either refer to or access information, and so syntactically it can only be checked for equality with other URIs and nothing more. This is captured well

by the good practice of ***URI opacity***, which states that "agents making use of URIs should not attempt to infer properties of the referenced resource" (Jacobs and Walsh, 2004). So, just because a URI says *http://www.eiffel-tower.com* does not mean it will not lead one to a web-page trying to sell one cheap trinkets and snake oil, as most users of the Web know. Second, a URI has an owner. The ***owner*** is *the agent that is accountable for a URI*. Interestingly enough, the domain name system that assigns control of domain names in URIs is a legally-binding techno-social system, and thus to some extent a complex notion of accountability for the name is built into URIs. Usually for URIs schemes such as HTTP, where the hierarchical component begins with an authority, the owner of the URI is simply whoever controls that authority. In HTTP, since URIs can delegate their relative components to other users, the owner can also be considered the agent that has the ability to create and alter the information accessible from the URI, not just the owner of the authority. Each scheme should in theory specify what ownership of a URI means in context of the particular scheme.

2.2.4 Resources

While we have explained how a URI is formed, we have yet to define what a URI is. To inspect the acronym itself, a Uniform Resource Identifier (URI) is an identifier for a 'resource.' Yet this does not solve any terminological woes, for the term 'resource' is undefined in the earliest specification for "Universal Resource Identifiers" (Berners-Lee, 1994a). Berners-Lee has remarked that one of the best things about resources is that for so long he never had to define them (Berners-Lee, 2000). Eventually Berners-Lee attempted to define a resource as "anything that has an identity" (Berners-Lee et al, 1998). Other specifications were slightly more detailed, with Roy Fielding, one of the editors of HTTP, defining (apparently without the notice of Berners-Lee) a resource as "a network data object or service" (Fielding et al, 1999). However, at some later point Berners-Lee decided to generalize this notion, and in some of his later works on defining this slippery notion of 'resource,' Berners-Lee was careful not to define a resource only as information that is accessible via the Web, since not only may resources be "electronic documents" and "images" but also "not all resources are network retrievable; e.g., human beings, corporations, and bound books in a library" (Berners-Lee et al, 1998). Also, resources do not have to be singular but can be a "collection of other resources" (Berners-Lee et al, 1998).

Resources are not only a concrete messages or sets of possible messages at a given temporal junction, but are a looser category that includes individuals changing over time, as "resources are further carefully defined to be information that may change over time, such as a service for today's weather report for Los Angeles"(Berners-Lee et al, 1998). Obviously, a web-page with "today's weather report" is going to change its content over time, so what is it that unites the notion of a resource over time? The URI specification defines this tentatively as a 'concep-

tual mapping' (presumably located in the head of an individual creating the representations for the resource) such that "the resource is the conceptual mapping to an entity or set of entities, not necessarily the entity which corresponds to that mapping at any particular instance in time. Thus, a resource can remain constant even when its content – the entities to which it currently corresponds – changes over time, provided that the conceptual mapping is not changed in the process" (Berners-Lee et al, 1998). This obviously begs an important question: If resources are identified as conceptual mappings in the head of an individual(s), then how does an agent know, given a URI, what the resource is? Is it our conceptual mapping, or the conceptual mapping of the owner, or some consensus conceptual mapping? This question and further questions of identity come to centre stage in Chapter **??**. The latest version of the URI specification deletes the confusing jargon of "conceptual mappings" and instead re-iterates that URIs can also be things above and beyond concrete individuals, for "abstract concepts can be resources, such as the operators and operands of a mathematical equation" (Berners-Lee et al, January 2005). After providing a few telling examples of precisely how wide the notion of a resource is, the URI specification finally ties the notion of resource directly to the act of identification given by a URI, for "this specification does not limit the scope of what might be a resource; rather, the term 'resource' is used in a general sense for whatever might be identified by a URI" (Berners-Lee et al, January 2005). Although this definition seems at best tautological, the intent should be clear. A ***resource*** is *any thing capable of having a sense* (content), or in other words, an 'identity' in a language. Since a sense is not bound to particular encoding, in practice within certain protocols that allow access to information, *a resource is typically not a particular encoding of some content but some content that can be given by many encodings*. To rephrase in terms of sense, *the URI identifies content on a level of abstraction, not the encoding of the content*. So, a URI identifies the 'content' of the Eiffel Tower, not just a particular webpage which is subject to change. However, there is nothing to forbid someone from identifying a particular encoding of information with its own URI and resource. For example, one could also have a distinct URI for a webpage about the Eiffel Tower in English, or a webpage about the Eiffel Tower in English in HTML. In other words, a resource can be given *multiple URIs*, each corresponding to a different encoding or even different levels of abstraction. Furthermore, due to the decentralized nature of URIs, often different agents create *multiple URIs for the same content*, which are then called in Web architecture ***co-referential URIs***.

We illustrate these distinctions in a typical HTTP interaction in Figure 2.5, where an agent via a web browser wants to access some information about the Eiffel Tower via its URI. While on a level of abstraction a protocol allows a user-agent to identify some resource, what the user-agent usually accesses concretely is some realization of that resource in a particular encoding, such as a webpage in HTML or a picture in the JPEG language (Pennebaker and Mitchell, 1992). In our example, the URI is resolved using the domain name system to an IP address of a concrete server, which then transmits to the user-agent some concrete bits that realizes the resource, i.e. that can be interpreted to the sense identified by the URI. In this example, all the interactions are local, since the webpage *encodes* the content of the resource. This

HTTP entity can then be interpreted by a browser as a rendering on the screen of Ralph's browser. Note this is a simplified example, as some status codes like `307` may cause a redirection to yet another URI and so another server, and so on possibly multiple times, until an HTTP entity may finally be retrieved.

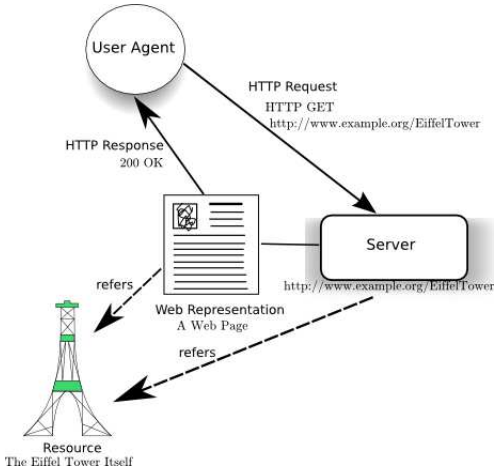

Fig. 2.5 A user agent accessing a resource

One of the most confusing issues of the Web is that a URI does not necessarily retrieve a single HTTP entity, but can retrieve multiple HTTP entities. This leads to a surprising and little-known aspect of Web architecture known as content negotiation. ***Content Negotiation*** is *a mechanism defined in a protocol that makes it possible to respond to a request with different Web representations of the same resource depending on the preference of the user-agent*. This is because information may have multiple encodings in different languages that all encode the same sense, and thus the same resource which should have a singular URI. A "representation" on the Web is then just "an entity that is subject to content negotiation" (Fielding et al, 1999). Historically, the term "representation" on the Web was originally defined in HTML as "the encoding of information for interchange" (Berners-Lee and Connolly, June 1993). A later definition given by the W3C did not mention content negotiation explicitly, defining a representation on the Web as just "data that encodes information about resource state" (Jacobs and Walsh, 2004). To descend further into a conceptual swamp, "representation" is one of the most confusing terms in Web architecture, as the term "representation" is used differently across philosophy. In order to distinguish the technical use of the term "representation" within Web architecture from the standard philosophical use of the term "representation," we shall use the term "Web representation" to distinguish it from the ordinary use of the term "representation" as given earlier in Section 2.2.6. A ***Web representation*** is *the encoding of the content given by a resource given in response to a request that*

is subject to content negotiation, which must then include any headers that specify an interpretation, such as character encoding and media type. So a Web representation can be considered to have *two* distinct components, and the headers such as the media type that lets us interpret the encoding, and the payload itself, which is the encoding of the state of the resource at a given point in time (i.e. the HTML itself). So, ***web-pages*** are *web representations given in HTML*. Web resources can be considered resources that under 'normal' conditions result in the delivery of web-pages.

Our typical Web transaction, as given earlier in Figure 2.5, can become more complex due to this possible separation between content and encoding on the Web. Different kinds of Web representations can be specified by user-agents as preferred or acceptable, based on the preferences of its users or its capabilities, as given in HTTP. The owner of a web-site about the Eiffel Tower decides to host a resource for images of the Eiffel Tower. The owner creates a URI for this resource, `http://www.eiffeltower.example.org/image`. Since a single URI is used, the sense (the depiction) that is encoded in either SVG or JPEG is the same, namely that of an image of the Eiffel Tower, that is, there are two distinct encodings of the image of the Eiffel Tower available on a server in two different iconic languages, one in a vector graphic language known as SVG and one in a bitmap language known as JPEG (Ferraiolo, 2002; Pennebaker and Mitchell, 1992). These encodings are rendered identically on the screen for the user. If a web-browser only accepted JPEG images and not SVG images, the browser could request a JPEG by sending a request for `Accept: image/jpeg` in the headers. Ideally, the server would then return the JPEG-encoded image with the HTTP entity header `Content-Type: image/jpeg`. Had the browser wished to accept the SVG picture as well, it could have put `Accept: image/jpeg, image/svg+xml` and received the SVG version. In Figure 2.6, the user agent specifies its preferred media type as `image/jpeg`. So, both the SVG and JPEG images are Web representations of the same resource, an image of the Eiffel Tower, since both the SVG and JPEG information realize the same information, albeit using different languages for encoding. Since while a single resource is identified by the same URI `http://www.example.org/EiffelTower/image`, different user-agents can get a Web representation of the resource in a language they can interpret, even if they cannot all interpret the same language. In Web architecture, content negotiation can also be deployed over not only differing computational languages such as JPG or SVG, but differing natural languages, as the same content can be encoded in different natural languages such as French and English. An agent could request the description about the Eiffel Tower from its URI and set the preferred media type to '`Accept-Language: fr`' so that they receive a French version of the webpage as opposed to an English version. Or they could set their preferred language as English but by using '`Accept-Language: en`.' The preferences specified in the headers are not mandatory for the server to follow, the server may only have a French version of the resource available, and so send the agent a French version of the description, encoded in HTML or some other formal

language, regardless of their preference.[8] Figure 2.6 shows is that the Web representations are distinct from the resource, even if the Web representations are bound together by realizing the same information given by a resource, since accessing a resource via a single URI can return *different* Web representations depending on content negotiation.

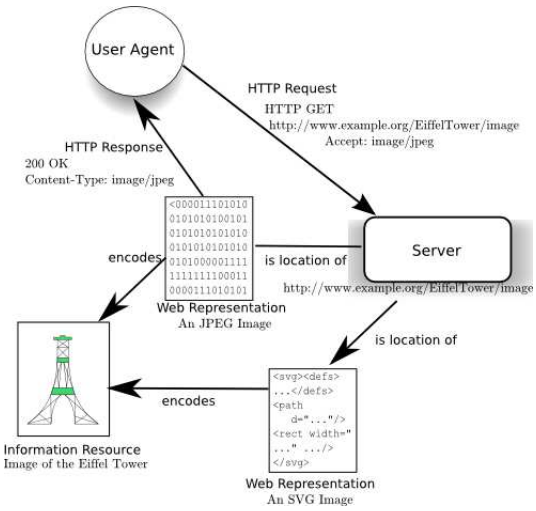

Fig. 2.6 A user agent accessing a resource using content negotiation

The only architectural constraint that connects Web representations to resources is that they are retrieved by the same URI. So one could imagine a resource with a URI called `http://www.example.org/Moon`, that upon accessing using English as the preferred language would provide a web-page with a picture of the moon, and upon accessing with something other than English as the preferred language would provide a picture of blue cheese. While this seems odd, this situation is definitely possible. What binds Web representations to a resource? Is a resource *really* just a random bag of Web representations? Remember that the answer is that the Web representations should have the same *content* regardless of their particular encoding if it is accessible from the same URI. This notion depends on our notion of informational content (sense) as given in Section **??**, which we define by an appeal to Dretske's semantic theory of information (Dretske, 1981). To recall, Dretske's

[8] It is well-known there are some words in French that are difficult if not impossible to translate into English, such as 'frileusement.' Indeed, saying that one natural language encodes the same content as another natural language is akin to hubris in the general case. If this is the case, then it is perfectly reasonable to establish different resources and so URIs for the French and English language encodings of the resource, such as `http://www.eiffeltower.example.org/francais` and `http://www.eiffeltower.example.org/english`. In fact, if one believes the same image cannot be truly expressed by both SVG and JPEG image formats, one could give them distinct URIs as well.

definition of semantic information, "a signal *r* carries the information that *s* is *F* when the conditional probability of *s*'s being *F*, given *r* (and *k*) is 1 (but, given *k* alone, less than 1). *k* is the knowledge of the receiver" (Dretske, 1981). We can then consider the signal *r* to be a Web representation, with *s* being a resource and the receiver being the user-agent. However, instead of some fact *F* about the resource, we want an interpretation of the Web representation by *different* user-agents to be to the same content.[9] From a purely normative viewpoint in terms of relevant IETF and W3C standards, it is left to the owner to determine whether or not two Web representations are equivalent and so can be hosted using content negotiation at the same URI. The key to content negotiation is that the owner of a URI never knows what the capabilities of the user-agent are, what natural and formal languages are supported by it. This is analogous to what Dretske calls the "knowledge" or *k* of the receiver (1981). The responsibility of the owner of a URI should be, in order to share their resource by as many user-agents as possible, to provide as many Web representations in a variety of formats as they believe are reasonably necessary. So, the owner of the URI for a website about the Eiffel Tower may wish to have a number of Web representations in a wide variety of languages and formats. By failing to provide a Web representation in Spanish, they prevent speakers of only Spanish from accessing their resource. Since the maintainer of a resource cannot reasonably be expected to predict the capabilities of all possible user-agents, the maintainer of the resource should try their best to communicate their interpretation within their finite means. The reason URIs identify resources, and not individual Web representations, is that Web representations are too ephemeral to want to identify in of themselves, being by definition the response of a server to a *particular* response and request for information. While one could imagine wanting to access a particular Web representation, in reality what is usually wanted by the user-agent is the content of the resource, which may be present in a wide variety of languages. What is important is that the sense gets transferred and interpreted by the user agent, not the individual bytes of a particular encoding in a particular language at a particular time.

2.2.5 Digitality

The Web is composed of not just representations, but digital representations. One of the defining characteristics of information on the Web is that this information is

[9] Of course, one cannot control the interpretations of yet unknown agents, so all sorts of absurdities are possible in theory. As the interpretation of the same encoding can differ among agents, there is a possibility that the owner of the URI `http://www.example.org/Moon` really thinks that for French speakers a picture of blue cheese has the same sense as a picture of the Moon for English speakers, even if users of the resource disagree. However, it should be remembered that the Web is a space of communication, and that for communication to be successful over the Web using URIs, it is in the interest of the owner of the resource to deploy Web representations that they believe the users will share their interpretation of. So content negotiation between a picture of blue cheese and a picture of the moon for a resource that depicts the Moon is, under normal circumstances, the Web equivalent of insanity at worse, or bad manners at best.

digital, bits and bytes being shipped around by various protocols. Yet there is no clear notion of what 'being' digital consists of, and a working notion of digitality is necessary to understand what can and can not be shipped around as bytes on the Web. Much like the Web itself, we can know something digital when we spot it, and we can build digital devices, but developing an encompassing notion of digitality is a difficult task, one that we only characterize briefly here.

Goodman defined marks as "***finitely differentiable***" *when it is possible to determine for any given mark whether it is identical to another mark or marks* Goodman (1968). This can be considered equivalent to how in categorical perception, despite variation in handwriting, a person perceives hand-written letters as being from a finite alphabet. So, *equivalence classes of marks can be thought of as an application of the philosophical notion of types*. This seems close to 'digital,' so that given a number of types of content in a language, a system is digital if any mark of the encoding can be interpreted to a one and only one type of content. Therefore, in between any two types of content or encoding there can not be an infinite number of other types. Digital systems are the opposite of Bateson's famous definition of information: Being digital is simply having a difference that does not make difference (Bateson, 2001). This is not to say there are characteristics of a mark which do not reflect its assignment in a type, and these are precisely the characteristics which are lost in digital systems. So in an analogue system, every difference in some mark makes a difference, since between any two types there is another type that subsumes a unique characteristic of the token. In this manner, the prototypical digital system is the discrete distribution of integers, while the continuous numbers are the analogue system par excellence, since between any real number there is another real number.

Lewis took aim at Goodman's interpretation of digitality in terms of determinism by arguing that digitality was actually a way to represent possibly continuous systems using the combinatorics of discrete digital states (1971). To take a less literal example, discrete mathematics can represent continuous subject matters. This insight caused Haugeland to point out that digital systems are always abstractions built on top of analog systems (1981). The reason we build these abstractions is because digital systems allow perfect reliability, so that once a system is in a digital type (also called a 'digital state'), it does not change unless it is explicitly made to change, allowing both flawless copying and perfect reliability. Haugeland reveals the purpose of digitality to be "a mundane engineering notion, root and branch. It only makes sense as a practical means to cope with the vagarities and vicissitudes, the noise and drift, of earthy existence" (Haugeland, 1981). Yet Haugeland does not tell us what digitality actually is, although he tells us what it does, and so it is unclear why certain systems like computers have been wildly successful due to their digitally (as in the success of analogue computers was not so widespread), while others like 'integer personality ratings' have not been as successful. Without a coherent definition of digitality, it is impossible to even in principle answer questions like whether or not digitality is *purely* subjective (Mueller, 2007). Any information is ***digital*** when *the boundaries in a particular encoding can converge with a regularity in a physical realization*. This would include sentences in a language that can be realized by sound-waves or the text in an e-mail message that can be re-encoded

as bits, and then this encoding realized by a series of voltages. Since the encoding of the information can be captured perfectly by a digital system, it can be copied safely and effectively, just as an e-mail message can be sent many times or a digital image reproduced countlessly.

To implement a digital system, there must be a small chance that the information realization can be considered to be in a state that is not part of the discrete types given by the encoding. The regularities that compose the physical boundary allows within a margin of error a discrete boundary decision to be made in the interpretation of the encoding. So, anything is capable of upholding digitality if that buffer created by the margin of error has an infinitesimal chance at any given time of being in a state that is not part of the encoding's discrete state. For example, the hands on a clock can be on the precise boundary between the markings on the clock, just not for very long. In a digital system, on a given level of abstraction, the margin of error does not propagate upwards to other levels of abstraction that rest on the earlier level of abstractions. Since we can create physical systems through engineering, we can create physical substrata that have low probabilities of being in states that do not map to digital at a given level of abstraction. As put by Turing, "The digital computers ... may be classified amongst the 'discrete state machines,' these are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously" (Turing, 1950). ***Analogue*** is the rather large and heterogeneous set of *everything that is not digital*. This would include people, such as Tim Berners-Lee himself, who can be represented but not realized as a message, as well as places, like Mount Everest, whose precise boundaries are rather indeterminate. While, according to Hayles, "the world as we sense it on the human scale is basically analogue," and the Web is yet another development in a long-line of biological modifications and technological prostheses to impose digitalization on an analogue world (2005). The vast proliferation of digital technologies is possible because there are physical substrata, some more so than others, which support the realization of digital information and give us the advantages that Haugeland rightfully points out is the purpose of the digital: flawless copying and perfect reliability in a flawed and imperfect world (1981).

2.2.6 Representations

A web-page about the Eiffel Tower seems to be an obvious representation. One can sit at home on one's computer far away from Paris and access a web-page that features a clear picture of - a representation! - of the Eiffel Tower. Furthermore, others from Japan to Egypt should be able to access the exact same representation by accessing the same URI. By claiming to be a "universal space of information," the Web is asserting to be a space where any encoding can be transferred about any content (Berners-Lee et al, 1992). However, there are some distinct differences be-

tween kinds of content, for some content can be distal and other content can be local. *Things that are separated by time and space* are ***distal*** while *those things that are not separated by time and space are* ***proximal***. As synonyms for distal and proximal, we will use ***non-local*** and ***local***, or just ***disconnected*** and ***connected***. Although this may seem to be an excess of adjectives to describe a simple distinction, this aforementioned distinction will underpin our notions of representation. In a message between two computers, if the content is a set of commands to 'display these bytes on the screen' then the client can translate these bytes to the screen directly without any worry about what those bytes represent to a human user. However, the content of the message may involve some distal components, such as the string "The Eiffel Tower is in Paris," which refers to many things outside of the computer. Differences between receivers allow the self-same content of a message to be both distal and local, depending on the interpreting agent. The message to 'display these bytes on the screen' could cause a rendering of a depiction of the Eiffel Tower to be displayed on the screen, so the self-same message causes not only a computer to display some bytes but also causes a human agent to receive information about what the Eiffel Tower in Paris looks like.

Any *encoding of information that has distal content* is called a ***representation***, regardless of the particular encoding of the information. Representations are then a subset of information, and inherit the characteristics outlined of all information, such as having one or more possible encodings and often a purpose and the ability to evoke normative behaviour from agents. To have some relationship to a thing that one is disconnected from is to be *about* something else. Generally, *the relationship of a thing to another thing to which one is immediately causally disconnected* is a relationship of ***reference*** to a ***referent*** or ***referents***, *the distal thing or things referred to by a representation*. The thing which refers to the referent(s) we call the 'representation,' and take this to be equivalent to being a *symbol*. *Linguistic expressions of an natural or formal language* are called ***descriptions*** while *the expressions of a iconic language* is called ***depictions***. To refer to something is to *denote* something, so the content of a representation is its *denotation*. In the tradition of Bretano, the reference relation is considered *intentional* due to its apparent physical spookiness. After all, it appears there is some great looming contradiction: if the content is whatever is held in common between the source and the receiver as a result of the conveyance of a particular message, then how can the source and receiver share some information they are disconnected from?

On the surface this aspect of 'representation' seems to be what Brian Cantwell Smith calls "physically spooky," since a representation can refer to something with which it is not in physical contact (Smith, 1995). This spookiness is a consequence of a violation of *common-sense* physics, since representations allow us to have some sort of what appears to be a non-physical relationship with things that are far away in time and space. This relationship of 'aboutness' or *intentionality* is often called 'reference.' While it would be premature to define 'reference,' a few examples will illustrate its usage: someone can think about the Eiffel Tower in Paris without being in Paris, or even having ever set foot in France; a human can imagine what the Eiffel Tower would look like if it were painted blue, and one can even think of a situation

where the Eiffel Tower wasn't called the Eiffel Tower. Furthermore, a human can dream about the Eiffel Tower, make a plan to visit it, all while being distant from the Eiffel Tower. Reference also works temporally as well as distally, for one can talk about someone who is no longer living such as Gustave Eiffel. Despite appearances, reference is not epiphenomenal, for reference has real effects on the behaviour of agents. Specifically, one can remember what one had for dinner yesterday, and this may impact on what one wants for dinner today, and one can book a plane ticket to visit the Eiffel Tower after making a plan to visit it.

We will have to make a somewhat convoluted trek to resolve this paradox. The very idea of representation is usually left under-defined as a "standing-in" intuition, that a representation is a representation by virtue of "standing-in" for its referent (Haugeland, 1991). The classic definition of a symbol from the Physical Symbol Systems Hypothesis is the genesis of this intuition regarding representations (Newell, 1980): "An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behaviour depends on Y." There are two subtleties to Newell's definition. Firstly, the notion of a representation is grounded in the behaviour of an agent. So, what precisely counts as a representation is never context-free, but dependent upon the agent completing some purpose with the representation. Secondly, the representation *simulates* its referent, and so the representation must be local to an agent while the referent may be non-local: "This is the symbolic aspect, that having X (the symbol) is tantamount to having Y (the thing designated) for the purposes of process P" (Newell, 1980). We will call X a representation, Y the *referent* of the representation, a process P the representation-using *agent*. This definition does not seem to help us in our goal of avoiding physical spookiness, since it pre-supposes a strangely Cartesian dichotomy between the referent and its representation. To the extent that this distinction is held a priori, then it is physically spooky, as it seems to require the referent and representation to somehow magically line up in order for the representation to serve as a substitute for its missing referent.

The only way to escape this trap is to give a non-spooky theory of how representations arise from referents. Brian Cantwell Smith tackles this challenge by developing a theory of representations that explains how they arise temporally (1995). Imagine Ralph, the owner of a URI for that he wants to host a picture of the Eiffel Tower, finally gets to Paris and is trying to get to the Eiffel Tower in order to take a digital photo. In the distance, Ralph sees the Eiffel Tower. At that very moment, Ralph and the Eiffel Tower are both physically connected via light-rays. At the moment of tracking, connected as they are by light, Ralph, its light cone, and the Eiffel Tower are a system, not distinct individuals. An alien visitor might even think they were a single individual, a 'Ralph-Eiffel Tower' system. While walking towards the Eiffel Tower, when the Eiffel Tower disappears from view (such as from being too close to it and having the view blocked by other buildings), Ralph keeps staring into the horizon, focused not on the point the Eiffel Tower was at before it went out of view, but the point where he thinks the Eiffel Tower would be, given his own walking towards it. Only when parts of the physical world, Ralph and the Eiffel Tower, are now physically separated can the agent then use a representation, such as the case of Ralph using an internal "mental image" of the Eiffel Tower or the external

digital photo to direct his walking towards it, even though he cannot see it. The agent is distinguished from the referent of its representation by virtue of not only disconnection but by the agent's attempt to track the referent, "a long-distance coupling against all the laws of physics" (Smith, 1995). The local physical processes used to track the object by the subject are the representation, be they 'inside' a human in terms of a memory or 'outside' the agent like a photo in a digital camera.

This notion of representation is independent of the representation being either internal or external to the particular agent, regardless of how one defines these boundaries.[10] Imagine that Ralph had been to the Eiffel Tower once before. He could have marked its location on a piece of paper by scribbling a small map. Then, the marking on the map could help guide him back as the Eiffel Tower disappears behind other buildings in the distance. This characteristic of the definition of representation being capable of including 'external' representations is especially important for any definition of a representation to be suitable for the Web, since the Web is composed of information that is considered to be external to its human users.

However fuzzy the details of Smith's story about representations may be, what is clear is that instead of positing a connection between a referent and a representation a priori, they are introduced as products of a temporal process. This process is at least theoretically non-spooky since the entire process is capable of being grounded out in physics without any spooky action at a distance. To be grounded out in physics, all changes must be given in terms of connection in space and time, or in other words, via effective reach. Representations are "a way of exploiting local freedom or slop in order to establish coordination with what is beyond effective reach" (Smith, 1996). In order to clarify Smith's story and improve the definition of the Physical Symbol Systems Hypothesis, we consider Smith's theory of the "origin of objects" to be a ***referential chain*** with distinct stages (Halpin, 2006):

- **Presentation**: Process *S* is connected with process *O*.
- **Input**: The process *S* is connected with *R*. Some local connection of *S* puts *R* in some causal relationship with process *O* via an encoding. This is entirely non-spooky since *S* and *O* are both connected with *R*. *R* eventually becomes the representation.
- **Separation**: Processes *O* and *S* change in such a way that the processes are disconnected.
- **Output**: Due to some local change in process *S*, *S* uses its connection with *R* to initiate local meaningful behaviour that is in part caused by *R*.[11]

In the 'input' stage, the *referent* is the cause of some characteristic(s) of the information. The relationship of *reference* is the relationship between the encoding of the information (the representation) and the referent. The relationship of interpretation becomes one of reference when the distal aspects of the content are crucial for the meaningful behaviour of the agent, as given by the 'output' stage. So we have

[10] The defining of "external" and "internal" boundaries is actually non-trivial, as shown in (Halpin, 2008a).

[11] In terms of Newell's earlier definition, 0 is X while S is P and R is Y.

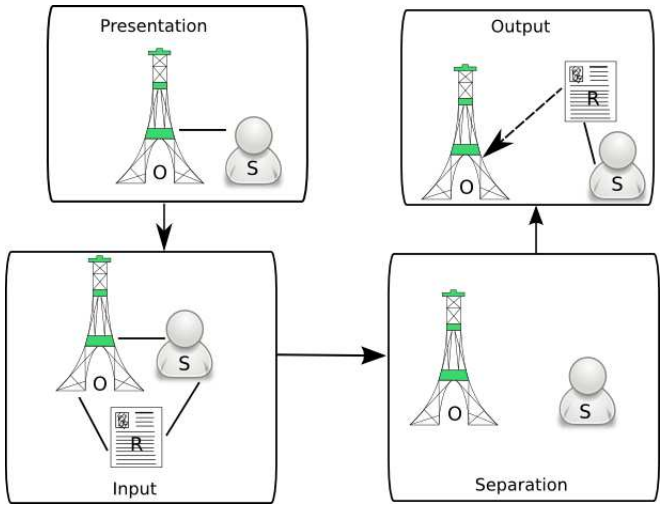

Fig. 2.7 The Referential Chain

constructed an ability to talk about representations and reference while not presupposing that behaviour depends on internal representations or that representations exist a priori at all. Representations are only needed when the relevant intelligent behaviour requires some sort of distal co-ordination with a disconnected thing.

So the interpretation of a representation – a particular kind of encoding of content – results in behavior by the user-agent that is dependent on a distal referent via the referential chain. In this manner, the act of reference can then be defined as the interpretation of a representation. This would make our notion of representation susceptible to being labelled a *correspondence theory of truth* (Smith, 1986), where a representation refers by some sort of structural correspondence to some referent. However, our notion of representation is much weaker, requiring only a causation between the referent and the representation - and not just any causal relationship, but one that is meaningful for the interpreting agent - as opposed to some tighter notion of correspondence such as some structural 'isomorphism' between a representation and its "target," the term used by Cummins to describe what we have called the "referent" of a representation (1996). So an interpretation or an act of reference should therefore not be viewed as mapping to referents, but a mapping to some content where that content leads to meaningful behaviour precisely because of some referential chain. This leads to the notion of a Fregean 'objective' sense, which we turn to later.

Up until now, it has been implicitly assumed that the referent is some physical entity that is non-local to the representation, but the physical entity was still existent, such as the Eiffel Tower. However, remember that the definition of non-local includes *anything* the representation is disconnected from, and so includes physical entities that may exist in the past or the future. The existence of a representation

does not imply the existence of the referent or the direct acquaintance of the referent by the agent using a representation – a representation only implies that some aspect of the content is non-local. However, this seems to contradict our 'input' stage in the representational cycle, which implies that part of our definition of representation is historical: for every *re*-presentation there must be a presentation, an encounter with the thing presented. By these conditions, the famous example of Putnam's example of an ant tracing a picture of Winston Churchill by sheer accident in the sand would not count as a representation (1975). If a tourist didn't know where the Eiffel Tower was, but navigated the streets of Paris and found the Eiffel Tower by reference to a tracing of a Kandinsky painting in his notebook, then the tourist would not then be engaged in any representation-dependent meaningful behaviour, since the Kandinsky painting lacks the initial presentation with the Eiffel Tower. The presentation does not have to be done by the subject that encountered the thing directly. However, the definition of a representation does not mean that the *same* agent using the representation had to be the agent with the original presentation. A representation that is created by one agent in the presence of a referent can be used by another agent as a 'stand-in' for that referent if the second agent shares the same interpretation from encoding to distal content. So, instead of relying on his own vision, a tourist buys a map and so relies on the 'second-order' representation of the mapmaker, who has some historical connection to someone who actually travelled the streets of Paris and figured out where the Eiffel Tower was. In this regard, our definition of representation is very much historical, and the original presentation of the referent can be far back in time, even evolutionary time, as given by accounts like those of Millikan (1984). One can obviously refer to Gustave Eiffel even though he is long dead and buried, and so no longer exists.

Also, the referent of a representation may be to what we think of as real-world patches of space and time like people and places, to abstractions like the concept of a horse, to unicorns and other imaginary things, to future states such as 'see you next year,' and descriptive phrases whose supposed *exact* referent is unknown, such as 'the longest hair on your head on your next birthday.' While all these types of concepts are quite diverse, they are united by the fact that they cannot be completely realized by local information, as they depend on partial aspects of an agent's local information, the future, or things that do not exist. Concepts that are constructed by definition, including imaginary referents, also have a type of 'presence,' it is just that the 'presentation' of the referent is created via the initial description of the referent. Just because a referent is a concept – as opposed to a physical entity – does not mean the content of the representation cannot have an meaningful effect on the interpreter. For example, exchanging representations of 'ghosts' - even if they do not quite identify a coherent class of referents - can govern the behaviour of ghost-hunters. Indeed, it is the power and flexibility of representations of these sorts that provide humans the capability to escape the causal prison of their local environment, to plan and imagine the future.

2.3 The Principles of Web Architecture

It is now possible to show how the various Web terms are related to each other in a more systematic way. These relationships are phrased as five finite principles that serve as the normative Principles of Web architecture: The Principles of Universality, Linking, Self-Description, the Open World, and Least Power. In practice many applications violate these principles, and by virtue of their use of URIs and the HTTP protocol, many of these applications would be in some sense 'on the Web.' However, these principles are normative insofar as they define what could be considered as compliance with Web architecture, and so an application that embodies them is compliant with Web architecture.

2.3.1 Principle of Universality

The ***Principle of Universality*** can be defined as that *any resource that can be identified by a URI*. The notion of both a resource and a URI was from their onset universal in its ambition, as Berners-Lee said, "a common feature of almost all the data models of past and proposed systems is something which can be mapped onto a concept of 'object' and some kind of name, address, or identifier for that object. One can therefore define a set of name spaces in which these objects can be said to exist. In order to abstract the idea of a generic object, the web needs the concepts of the universal set of objects, and of the universal set of names or addresses of objects" (1994a). The more informal notes of Berners-Lee are even more startling in their claims for universality, stating that the first 'axiom' of Web architecture is "Universality" where "by universal' I mean that the Web is declared to be able to contain in principle every bit of information accessible by networks" (1996b). Although it appears he may be constraining himself to only talk about digital 'objects' that are accessible over the Internet in this early IETF RFCs, in later IETF RFCs the principle quickly ran amok, as users of the Web wanted to use URIs to refer to "human beings, corporations, and bound books in a library" (Berners-Lee et al, 1998).

There seems to be a certain way that web-pages are 'on the Web' in a way that human beings, corporations, unicorns, and the Eiffel Tower are not. Accessing a web-page in a browser means to receive some bits, while one cannot easily imagine what accessing the Eiffel Tower itself or the concept of a unicorn in a browser even means. This property of being 'on the Web' is a common-sense distinction that separates things like a web-page about the Eiffel Tower from things like the Eiffel Tower itself. This distinction is a matter of between the use of URIs to *access* and *reference*, between the local and the distal. The early notes of Berners-Lee that pre-date the notion of URIs itself address this distinction between access and reference, phrasing it as a distinction between locations and names. As Berners-Lee states, "conventionally, a 'name' has tended to mean a logical way of referring to an object in some abstract name space, while the term 'address' has been used for

something which specifies the physical location" (1991). So, a ***location*** is *a term that can be used to access the thing*, while a ***name*** is *a term that can be used to refer to a thing*. Unlike access, reference is the use of an identifier for a thing *to which one is immediately causally disconnected*. ***Access*** is *the use of an identifier to create immediately a causal connection to the thing identified* (Hayes and Halpin, 2008). The difference between the use of a URI to access a hypertext web-page or other sort of information-based resource and the use of a URI to refer to some non-Web accessible entity or concept ends up being quite important, as this ability to representationally use URIs as 'stands-in' for referents forms the basis of the distinction between the hypertext Web and the Semantic Web.

Names can serve as identifiers and even representations for distal things. However, Berners-Lee immediately puts forward the hypothesis that "with wide-area distributed systems, this distinction blurs" so that "things which at first look like physical addresses...cease to give the actual location of the object. At the same time, a logical name...must contain some information which allows the name server to know where to start looking" (1991). He posits a third neutral term, "identifier" that was "generally referred to a name which was guaranteed to be unique but had little significance as regards the logical name or physical address" (Berners-Lee, 1991). In other words, an ***identifier*** is a *term that can be used to either access or refer, or both access and refer to, a thing*. The problem at hand for Berners-Lee was how to provide a name for his distributed hypertext system that could get "over the problem of documents being physically moved" (1991). Using simple IP addresses or any scheme that was tied to a single server would be a mistake, as the thing that was identified on the Web should be able to move from server to server without having to change identifier.

For at least the first generation of the Web, the way to overcome this problem was to provide a translation mechanism for the Web that could provide a methodology for transforming "unique identifiers into addresses" (Berners-Lee, 1991). Mechanisms for translating unique identifiers into addresses already existed in the form of the domain name system that was instituted by the IETF in the early days of the expansion of ARPANet (Mockapetris, Novemeber 1983). Before the advent of the domain name system, the ARPANet contained one large mapping of identifiers to IP addresses that was accessed through the Network Information Centre, created and maintained by Engelbart (Hafner and Lyons, 1996). However, this centralized table of identifier-to-address mappings became too unwieldy for a single machine as ARPANet grew, so a decentralized version was conceived based on ***domain names***, where each domain name is *a specification for a tree structured name space, where each component of the domain name (part of the name separated by a period) could direct the user-agent to a more specific "domain name server" until the translation from an identifier to the name to IP address was complete.*

Many participants in the IETF felt like the blurring of this distinction that Berners-Lee made was incorrect, so URIs were bifurcated into two distinct specifications. *A scheme for locations that allowed user-agents via an Internet protocol to access information* were called ***Uniform Resource Locations*** (URLs) (Berners-Lee et al, 1994) while *a scheme whose names that could refer to things outside of the*

causal reach of the Internet were called ***Uniform Resource Names*** (URNs) (Sollins and Masinter, 1994). Analogue things like concepts and entities naturally had to be given URNs, and digital information that can be transmitted over the Internet, like web-pages, were given URLs. Interestingly enough, URNs count *only* as a naming scheme, as opposed to a protocol like HTTP, because they cannot access any information. While one could imagine a particular Web-accessible realization, like a web-page, disappearing from the Web, it was felt that identifiers for things that were not accessible over the Web should "be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name" (Mealling and Daniel, 1999).

Precisely because of their lack of ability to access information, URNs never gained much traction, while URLs to access web-pages became the norm. Building on this observation about the "blurring of identifiers," the notion of URIs implodes the distinction between identifiers used only for access (URLs) and the identifiers used for reference (URNs). A ***Uniform Resource Identifier*** is *a unique identifier whose syntax is given in (Berners-Lee et al, January 2005), that may be used to either or both refer to or access a resource*. URIs subsume both URIs and URNs, as shown in Figure 2.8. Berners-Lee and others were only able to push this standard through the IETF process years after the take-off of the Web. Indeed, early proposals for universal names, ranging from Raymond Lull to Engelbart's 'Every Object Addressable' principle (1990), all missed the crucial advantage of the Web; while classically names in natural language are used for reference, on the Web names can be used to access information. In a decentralized environment this is crucial for discovering the sense of a URI, as illustrated by the notions of 'linking' and 'self-description' detailed next in Section 2.3.2 and Section 2.3.3.

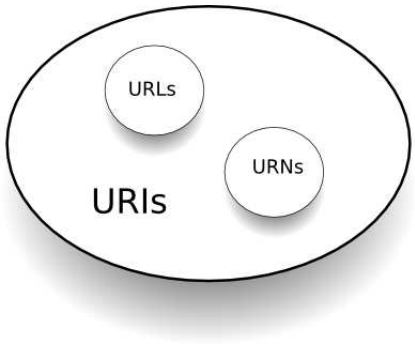

Fig. 2.8 A Venn Diagram describing the relationships between URIs, URNs, and URLs

2.3.2 Principle of Linking

The ***Principle of Linking*** states that *any resource can be linked to another resource identified by a URI*. No resource is an island, and the relationships between resources are captured by the linking, transforming lone resources into a Web. A ***link*** is *a connection between resources*. The *resource that the link is directed from* is called its ***starting resource*** while the *resource a link is directed to* is the ***ending resource*** (DeRose et al, 2001).

What are links for? Just as URIs links may be used for either access or reference, or even both. In particular, in HTML the purpose of links is for access to additional hypertext documents, and so they are sometimes called hyperlinks. This access is often called *following* the link, a transversal from one Web representation to another, that results in access to Web representations of the ending resource. A unidirectional link that allows access of one resource from another is the predominant kind of link in hypertext. Furthermore, access by linking is transitive, for if a user-agent can access a Web representation of the ending resource from the starting resource, then it can access any links present in the Web representation, and thereby access a Web representation of an ending resource. It is precisely this ability to transitively access documents by following links that led the original Web to be a seamless Web of hypertext. While links can start in Web representations, the main motivation for using URIs as the ending resource of a link as opposed to a specific Web representation is to prevent *broken links*, where a user-agent follows a link to a resource that is no longer there, due to the Web representation itself changing. As put by the TAG, "Resource state may evolve over time. Requiring a URI owner to publish a new URI for each change in resource state would lead to a significant number of broken references. For robustness, Web architecture promotes independence between an identifier and the state of the identified resource" (Jacobs and Walsh, 2004).

However, one of the distinguishing features of the Web is that links may be broken by having any access to a Web representation disappear, due to simply the lack of hosting a Web representation, loss of ownership of the domain name, or some other reason. These reasons are given in HTTP status codes, such as the infamous `404 Not Found` that signals that while there is communication with a server, the server does not host the resource. Further kinds of broken links are possible, such as `301 Moved Permanently` or a `5xx` server error, or an inability to even connect with the server leading to a time-out error. This ability of links to be 'broken' contrasts to previous hypertext systems. Links were not invented by the Web, but by the hypertext research community. Constructs similar to links were enshrined in the earliest of pre-Web systems, such as Engelbart's *oNLine System* (NLS) (1962), and were given as part of the early hypertext work by Theodor Nelson (1965). The plethora of pre-Web hypertext systems were systematized into the Dexter Reference Model (Halasz and Schwartz, 1994). According to the Dexter Reference Model, the Web would not even qualify as hypertext, but as "proto-hypertext," since the Web did not fulfill the criteria of "consistency," which requires "in creating a link, we must ensure that all of its component specifiers resolve to existing components" (Halasz and Schwartz, 1994). To ensure a link must resolve and therefore not be

broken, this mechanism requires a centralized link index that could maintain the state of each resource and not allow links to be created to non-existent or non-accessible resources. Many early competitors to the Web like HyperG had a centralized link index (Andrews et al, 1995). As an interesting historical aside, it appears that the violation of this principle of maintaining a centralized link index was the main reason why the World Wide Web was rejected from its first academic conference, ACM Hypertext 1991, although Engelbart did encourage Berners-Lee and Connolly to pursue the Web further.[12] While a centralized link index would have the benefit of not allowing a link to be broken, the lack of a centralized link index removes a bottleneck to growth by allowing the owners of resources to link to other resources without updating any index besides their own Web representations. This was doubtless important in enabling the explosive growth of linking. The lack of any centralized link index, and index of Web representations, is also precisely what search engines like Google create post-hoc through spidering, in order to have an index of links and web-pages that enable their keyword search and page ranking algorithms. As put by Dan Connolly in response to Engelbart, "the design of the Web trades link consistency guarantees for global scalability" (2002). So, broken links and `404 Not Found` status codes are purposeful *features*, not defects, of the Web.

2.3.3 Principle of Self-Description

One of the goals of the Web is for resources to be 'self-describing,' currently defined as "individual documents become self-describing, in the sense that only widely available information is necessary for understanding them" (Mendelsohn, 2006). While it is unclear what "widely-available" means, one way for information to be widely-available is for it to be linked to from the Web representation itself. The ***Principle of Self Description*** states that *the information an agent needs to have an interpretation of a Web Representation (resource) should be accessible from the Web representation itself (URI)*.

How many and what sort of links are necessary to adequately describe a resource? A resource is successfully described if an interpretation of a sense is a possible. Any representation can have links to other resources which in turn can determine valid interpretations for the original resource. This process of following whatever data is linked in order to determine the interpretation of a URI is informally called 'following your nose' in Web architecture.

The ***Follow-Your-Nose algorithm*** states that if a user-agent encounters a representation in a language that the user-agent cannot interpret, the user-agent should, in order:

1. **Dispose of Fragment Identifiers:** As mandated by the URI specification (Berners-Lee et al, January 2005), user-agents can dispose of the fragment identifier in

[12] Personal communication with Tim Berners-Lee.

order to retrieve whatever Web representations are available from the racine (the URI without fragment identifier). For example, in HTML the fragment identifier of the URI is stripped off when retrieving the webpage, and then when the browser retrieves a Web representation, the fragment identifier can be used to locate a particular place within the Web representation.

2. **Inspect the Media Type:** The media type of a Web representation provides a normative declaration of how to interpret a Web representation. Since the number of IETF media-types is finite and controlled by the IETF, a user-agent should be able to interpret these media types.[13]
3. **Follow any Namespace Declarations:** Many Web representations use a generic format like XML to in turn specify a customized dialect. In this case, a language or dialect is itself given a URI, called a ***namespace URI***, *a URI that identifies that particular dialect*. A namespace URI then in turn allows access to a ***namespace document***, *a Web representation that provides more information about the dialect*. In a Web representation using this dialect, a ***namespace declaration*** then *specifies the namespace URI*. In this case, the user-agent may follow these namespace declarations in order to get the extra information needed to interpret the Web representation. As a single Web representation may be encoded in multiple languages, it may have multiple namespace URIs to follow.
4. **Follow any links:** The user-agent can follow any links. There are some links in particular languages that may be preferred, such as the ending resource of a `link` header in HTML or in RDF Schema links such as *rdfs:isDefinedBy* links, or links like OWL by the *owl:imports* (See Chapter **??** for the definition of RDF and OWL). If links are typed in some fashion, each language may define or recommend links that have the normative status, and normative links should be preferred. However, for many kinds of links, their normative status is unclear, so the user-agent may have to follow any sort of link as a last resort.

Using this algorithm, the user-agent can begin searching for some information that allows it to interpret the Web representation. It can follow the first three guidelines and then follow the fourth, applying the above guidelines recursively. Eventually, this recursive search should bottom out either in a program that allows an interpretation of the Web representation (such as a rendering of a web-page or inferences given by a Semantic Web language) or specifications given by the IETF in plain, human-readable text, the natural bottoming point of self-description. This final fact brings up the point that the information that gets one an interpretation is not necessarily a program, but could be a human-readable specification that requires a human to make the mapping from the names to the intended sense.

[13] The finite list is available at *http://www.iana.org/assignments/media-types/*, and a mapping from media types to URIs has been proposed at *http://www.w3.org/2001/tag/2002/01-uriMediaType-9*.

2.3.4 The Open World Principle

The ***Open World Principle*** states that *the number of resources on the Web can always increase*. There can always be new acts of identification, carving out a new resource from the world and identifying it with a URI. At any given moment, a new webpage may appear on the Web, and it may or may not be linked to. This is a consequence of the relatively decentralized creation of URIs for resources given by the Principle of Universality and the decentralized creation of links by the Principle of Linking. Without any centralized link index, there is no central repository of the state of the *entire* Web. While approximations of the state of the entire Web are created by indexing and caching web-pages by search engines like Google, due to the Open World Principle, none of these alternatives will necessarily ever be guaranteed to be complete. Imagine a web-spider updating a search engine index. At any given moment, a new resource could be added to the Web that the web-spider may not have crawled. So to assume that any collection of resources of the Web can be a complete picture of the whole Web is at best impudent.

The ramifications of the Open World Principle are surprising, and most clear in terms of judging whether a statement is true or false. This repercussions transform the Open World Principle into its logical counterpart, the ***Open World Assumption***, which logically states that *statements that cannot be proven to be true cannot be assumed to be false*. Intuitively, this means that the world cannot be bound. On the Web, the Open World Principle holds that since the Web can always be made larger, with any given set of statements that allows an inference, a new statement relevant to that inference may be found. So any agent's knowledge of the Web is always partial and incomplete, and thus the Open World Assumption is a safe bet for agents on the Web. The Open World Principle is one of the most influential yet challenging principles of the Web, the one that arguably separates the Web from traditional research in artificial intelligence and databases in practice. In these fields, systems tend to make the opposite of the Open World Assumption, the Closed World Assumption. The ***Closed World Assumption*** states that logically *statements that cannot be proven to be true can be assumed to be false*. Intuitively, this means that somehow the world can be bounded. The Closed World Assumption has been formalized on a number of different occasions, with the first formalization being due to Reiter (1978). This assumption has often been phrased as an appeal to the Law of the Excluded Middle ($\forall p.p \vee \neg p$) in classical logic (Detlefsen, 1990). *Negation as failure* is an implementation of the Closed World assumption in both logic programming and databases, where failure for the program to prove a statement is true implies the statement is false (Clark, 1978).

2.3.5 Principle of Least Power

The Principle of Least Power states that a *Web representation given by a resource should be described in the least powerful but adequate language*. This principle

is also normative, for if there are multiple possible Web representations for a resource, the owner should chose the Web representation that is given in the 'least powerful' language. The Principle of Least Power seems odd, but it is motivated by Berners-Lee's observation that "we have to appreciate the reasons for picking not the most powerful solution but the least powerful language" (1996b). The reasons for this principle are rather subtle. The receiver of the information accessible from a URI has to be able to decode the language that the information is encoded in so the receiver can determine the sense of the encoding. Furthermore, an agent may be able to decode multiple languages, but the owner of the URI does not know what languages an agent wanting to access their URI may possess. Also, the same agent may be able to interpret multiple languages that can express the same sense. So, the question always facing any agent trying to communicate is what language to use? In closed and centralized systems, this is ordinarily not a problem, since each agent can be guaranteed to use the same language. In an open system like the Web, where one may wish to communicate a resource to an unknown number of agents, each of which may have different language capabilities, the question of which language to deploy becomes nearly insurmountable. Obviously, if an agent is trying to convey some sense, then it should minimally choose a language to encode that sense which is capable of conveying that sense. Yet as the same sense can be conveyed by different languages, what language to choose?

The Principle of Least-Power is a common-sense engineering solution to this problem of language choice. The solution is simply to build first a common core language that fulfills the minimal requirements to communicate whatever sense one wishes to communicate, and then extend this core language. Using HTML as an example, one builds first a common core of useful features such as the ability to have text be bold and have images inserted in general areas of the text, and then as the technology matures, to slowly add features such as the precise positioning of images and the ability to specify font size. The Principle of Least Power allows a straightforward story about compatibility to be built to honor the"be strict when sending and tolerant when receiving" maxim of the Internet, since it makes the design of a new version an exercise in strictly extending the previous version of the language (Carpenter, June 1996). A gaping hole in the middle of the Principle of Least Power is no consistent definition of the concept of 'power,' and the W3C TAG seems to conflate power with the Chomsky Hierarchy, the problem of defining 'power' formally must be left as an open research question.

2.4 Conclusions

The Web, while to a large extent being an undisciplined and poorly-defined space, does contain a set of defining terms and principles. While previously these terms and principles have been scattered throughout various informal notes, IETF RFCs, and W3C Recommendations, in this chapter we have systematized both the terminology and the principles in a way that reveals how they internally build of each other. In

general, when we are referring to the ***hypertext Web***, we are referring *to the use of URIs and links to access hypertext web-pages using HTTP*. Yet there is more to the Web than hypertext. The next question is how can these principles be applied to domains outside the hypertext Web, and this will be the topic of Chapter 3, as we apply these principles to the Semantic Web, a knowledge representation language for the Web.

Chapter 3
The Semantic Web

The task of classifying all the words of language, or what's the same thing, all the ideas that seek expression, is the most stupendous of logical tasks. Anybody but the most accomplished logician must break down in it utterly; and even for the strongest man, it is the severest possible tax on the logical equipment and faculty. **Charles Sanders Peirce, letter to editor B. E. Smith of the Century Dictionary**

The Web is a universal information space, but so far it has been one composed entirely of hypertext documents. As said by Berners-Lee at the World Wide Web conference in 1994, "to a computer, then, the web is a flat, boring world devoid of meaning...this is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them" (1994b). The heart of this particular insight is the realization that it is the content of the information, not its encoding in hypertext, that is of central importance to the Web. The purpose of the architecture of the Web is to connect information of any kind in a decentralized manner, and this architecture can be applied beyond the hypertext documents of its initial incarnation.

The next step in Berners-Lee's programme to expand the Web beyond hypertext is called the *Semantic Web*, a term first used by Foucault in *The Order of Things* (Foucault, 1970). The most cited definition of the Semantic Web is given by Berners-Lee et al. as "the Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (2001). How can information be added to the Web without encoding it in hypertext? The answer is to find a language capable of representing the information about the aforementioned real objects and imaginary concepts. This requires a ***knowledge representation language***, *a language whose primary purpose is the representation of non-digital content in a digital encoding*. So instead of the Eiffel Tower, we will have a number of facts about the Semantic Web, ranging from pictures to its height, encoded in a knowledge representation language available via a URI for the Eiffel Tower.

As the previous exposition of Web architecture explained in detail, resources on the Web are given by a URI that identifies the same content on the Web across dif-

ferent encodings. What drives the Semantic Web is the realization that at least some of the information on the Web is representational, i.e. information about distal content. Then instead of HTML, which is mainly concerned with the presentation and linking of natural language for humans, the Web needs a knowledge representation language which describes the represented content as fully as possible without regard to presentation for humans. The mixture of content and encodings for presentation forces web-spiders to "scrape" valuable content out of hypertext. In theory, encoding information directly in a knowledge representation language gives a spider more reliable and direct access to the information. As Berners-Lee puts it, "most information on the Web is designed for human consumption, and even if it was derived from a database with well defined meanings (in at least some terms) for its columns, that the structure of the data is not evident to a robot browsing the web" (1998b). This has led him to consider the Semantic Web to a Web "for expressing information in a machine processable form" and so making the Web "machine-understandable" (Berners-Lee, 1998b). This leads to the contrast between the Semantic Web as a 'web of data' as opposed to the hypertext 'web of documents.' W3C standards such as XML were originally created, albeit rarely used, precisely in order to separate content and presentation (Connolly, 1998).

Furthermore, the purpose of the Semantic Web is to expand the scope of the Web itself. Most of the world's digital information is not natively stored in hypertext. Instead, it is stored in databases and other non-hypertext documents and spreadsheets. While this information is slowly but surely migrating towards the Web, as more and more of this information is being exposed to the Web via scripts that automatically and dynamically convert data from databases into HTML, the Semantic Web imagines that by having a common knowledge representation language across the entire Web, all sorts of information that previously were not on the Web can become part of the Web. This makes the Semantic Web not a different and parallel Web to the hypertext Web, but an extension of the current Web, where hypertext serves as just one possible language.

3.1 A Brief History of Knowledge Representation

The creation of the Semantic Web then depends on the creation of at least one (if not multiple!) knowledge representation language for the Web, and so the Semantic Web inherits both the successes and failures of previous efforts to create knowledge representation languages in artificial intelligence. The earliest work in digital knowledge representations was spear-headed by John McCarthy's attempts to formalize elements of human knowledge in first-order predicate logic, where the primary vehicle of intelligence was to be considered some form of inference (1959). These efforts reached their apex in Hayes's "Naive Physics Manifesto," which called for parts of human understanding to be formalized as first-order logic. Although actual physics was best understood using mathematical techniques such as differential equations, Hayes conjectured that most of the human knowledge of physics, such

as “water must be in a container for it not to spill” could be conceptualized better in first-order logic (1979). Hayes took formalization as a grand long-term challenge for the entire AI community to pursue, “we are never going to get an adequate formalization of common sense by making short forays into small areas, no matter how many of them we make” (Hayes, 1979). While many researchers took up the grand challenge of Hayes in various domains, soon a large number of insidious problems were encountered, primarily in terms of the expressivity of first-order logic and its undecidability of inference. In particular, first-order logic formalizations were viewed as not expressive enough, being unable to cope with temporal reasoning as shown by the Frame Problem, and so had to be extended with fluents and other techniques (McCarthy and Hayes, 1969). Since the goal of artificial intelligence was to create an autonomous human-level intelligence, another central concern was that predicate calculus did not match very well with how humans actually reasoned. For example, humans often use default reasoning, and various amendments must be made for predicate calculus to support this (McCarthy, 1980). Further efforts were made to improve first-order logic with temporal reasoning to overcome the Frame Problem, as well as the use of fuzzy and probabilistic logic to overcome issues brought up by default reasoning and the uncertain nature of some knowledge (Koller and Pfeffer, 1998). Yet as predicted by Hubert Dreyfus, it seemed none of these formal solutions could solve the fundamental epistemological problem that all knowledge was in front of an immense background of a world that *itself* seemed to resist formalization (Dreyfus, 1979).

Under increasing criticism from its own former champions like McDermott, first-order predicate calculus was increasingly abandoned by those in the field of knowledge representation (1987). McDermott pointed out that formalizing knowledge in logic requires that all knowledge be formalized as a set of axioms and that “it must be the case that a significant portion of the inferences we want...are deductions, or it will simply be irrelevant how many theorems follow deductively from a given axiom set” (1987). McDermott found that in practice neither can all knowledge be formalized and that even given some fragment of formalized knowledge, the inferences drawn are usually trivial or irrelevant (1987). Moving away from first-order logic, the debate focused on what was the most appropriate manner for AI to model human intelligence. Some researchers championed a *procedural* view of intelligence that regarded the representation as itself irrelevant if the program could successfully solve some task given some input and output. This contrasted heavily with earlier attempts to formalize human knowledge that it was called the *declarative versus procedural* debate. Champion of procedural semantics Terry Winograd stated that “the operations on symbol structures in a procedural semantics need not correspond to valid logical inferences about the entities they represent” since “the symbol manipulation processes themselves are primary, and the rules of logic and mathematics are seen as an abstraction from a limited set of them” (1976). While the procedural view of semantics first delivered impressive results through programs like SHRDLU (Winograd, 1972), since the ‘semantics’ were ad-hoc and task-dependent, procedural semantics could not be used outside the limited domain in which they were created. Furthermore, there became a series of intense debates on whether these pro-

grams often purported to do what they wanted even within their domain, as Dreyfus critiqued that it was ridiculous that just because a program was labelled 'understand' that it did actually in any way understand (1979). Interestingly enough, the debate between declarative and procedural semantics is, under the right formal conditions, a red herring since the Curry-Howard Isomorphism states that given the right programming language, there is a tight coupling between logical proofs and programs so that the simplification of proofs can be equivalent to steps of computation (Wadler, 2001).

Within AI, research began into other forms of declarative knowledge representation languages besides first-order logic that were supposed to be in greater concordance with human intelligence and that could serve as more stable substrates for procedural knowledge-based systems. Most prominent among these alternatives were ***semantic networks***, "a graphic notation for representing knowledge in patterns of interconnected nodes and arcs" (1987). Semantic networks are as old as classical logic, dating back to Porphyry's explanation of Aristotelian categories (Sowa, 1987), although their first self-described usage was as a common knowledge-representation system for machine-translation systems by Masterman (1961). Motivated by a correspondence with natural language, semantic networks were used by many systems in natural language processing, such as the work of Wilks in resolving ambiguities using preference semantics and the work of Schank using conceptual dependency graphs to discover identical sentences regardless of their syntactic form (Schank, 1972; Wilks, 1975). Soon semantic networks were being used to represent everything from human memory to first-order logic itself (Quillian, 1968; Sowa, 1976). The approach of semantic networks was given some credence by the fact that often when attempting to make diagrams of 'knowledge,' humans often start by drawing circles connected by lines, with each component labelled with some human-readable description. A semantic network about 'The architect of the Eiffel Tower was Gustave Eiffel' is given in Figure 3.1. Note that it refers declaratively to things in the world, but uses 'natural-language-like' labels on its nodes and edges.

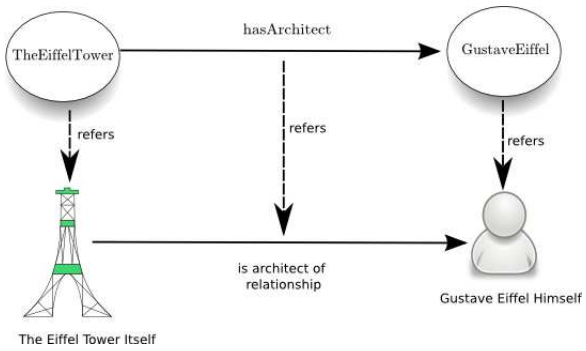

Fig. 3.1 An example semantic network

When researchers attempted to communicate or combine their knowledge representation schemes, no-one really knew what the natural language description 'meant' except the author, even when semantic networks were used as a formal language. The 'link' in semantic networks was interpreted in at least three different ways (Woods, 1975) and no widespread agreement existed on the most common sort-of link, the `IS-A` link, which could represent both subclassing, instantiation, close similarity, and more. This led to an assault on semantic networks by champions of first-order logic like Hayes, who believed that by providing a formal semantics that defined 'meaning', first-order logic at least allowed knowledge representations to be transportable across domains, and that many alternative knowledge representations could be re-expressed in first order-logic (Hayes, 1977a). In response, the field of knowledge representation bifurcated into separate disciplines. Many of the former champions of logic currently do not believe that human intelligence can be construed as logical inference, but researchers still actively pursue the field as it is of crucial importance to many systems such as mathematical proof-proving and it is still used in many less ambitious knowledge-reasoning systems such as ISO Common Logic (Delugach, 2007).

The classical artificial intelligence programme, while fixated on finding a formal language capable of expressing human knowledge, had ignored the problem of tractable inference. This problem came to attention abruptly when KRL, one of the most flexible knowledge representation languages pioneered by Winograd was found to have intractable inference even on simple problems of cryptarithmetic, despite its representational richness.[1] Furthermore, while highly optimized inference mechanisms existed for first-order logic, first-order predicate logic was proven to be undecidable. These disadvantages of alternative representational formats and first-order logic led many researchers, particularly those interested in *an alternative "slot and value" knowledge representation language* known as ***frames*** to begin researching the decidability of their inference mechanisms (Minsky, 1975). This research into frames then evolved into research on ***description logics***, where the trade-offs between the tractability and expressivity where carefully studied (Levensque and Brachman, 1987). The goal of the field was to produce a logic with decidable inference while maintaining maximum expressivity. Although the first description-logic system, KL-ONE, was proven to have undecidable inference for even subsumption, later research produced a vast proliferation of description logics with carefully categorized decidability and features (Brachman and Schmolze, 171-216; Schmidt-Schauss, 1989).

Ultimately, the project of artificial intelligence to design a single knowledge representation system suitable for creating human-level intelligence has not yet succeeded and progress, despite occassional bursts of enthusiasm, is doubtful at best. With no unifying framework, the field of artificial intelligence itself fragmented into many different diverse communities, each with its own family of languages and techniques. Researchers into natural language embraced statistical techniques and went back to practical language processing tasks, while logicians have produced

[1] Personal communication with Henry S. Thompson.

an astounding variety of different knowledge representation languages, and cognitive scientists moved their interests towards dynamical systems and specialized biologically-inspired simulations. The lone hold-out seemed to be the Cyc project, which continued to pursue the task of formalizing all 'common-sense' knowledge in a single knowledge representation language (Lenat, 1990). In one critique of Cyc, Smith instead asked what lessons knowledge representation languages could learn from hypertext, "Forget intelligence completely, in other words; take the project as one of constructing the world's largest hypertext system, with Cyc functioning as a radically improved (and active) counterpart for the Dewey decimal system. Such a system might facilitate what numerous projects are struggling to implement: reliable, content-based searching and indexing schemes for massive textual databases," a statement that strangely prefigures not only search engines, but the revitalization of knowledge representation languages due to the Semantic Web (1991).

3.2 The Resource Description Framework (RDF)

What makes knowledge representation language on the Web *different* from classical knowledge representation? Berners-Lee's early thoughts, as given in the first World Wide Web Conference in Geneva in 1994, were that "adding semantics to the Web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values" (Berners-Lee, 1994b). Having information in "machine-readable forms" requires a knowledge representation language that has some sort of relatively content-neutral language for encoding (Berners-Lee, 1994b). The parallel to knowledge representation in artificial intelligence is striking, as it also sought to find one universal encoding, albeit encoding human-intelligence. The second point, of "allowing links," means that the basic model of the Semantic Web will be a reflection of the Web itself: the Semantic Web consists of connecting resources by links. The Semantic Web is then easily construed as a descendant of semantic networks from classical artificial intelligence, where nodes are resources and arcs are links. Under the aegis of the W3C, the first knowledge representation language for the Semantic Web, the ***Resource Description Language*** (RDF) was made a W3C Recommendation, and it is clearly inspired by work in AI on semantic networks. This should come as no surprise, for RDF was heavily inspired by the work of R.V. Guha on the Meta-Content Framework (Guha, 1996). Before working on MCF, Guha was chief lieutenant of the Cyc project, the last-ditch Manhattan project of classical artificial intelligence (R.V.Guha and D.Lenat, 1993). There are nonetheless some key differences between semantic networks and RDF, as RDF was built in accordance with the Principles of Web Architecture as given in Chapter **??**, as detailed in the next subsections.

3.2.1 RDF and the Principle of Universality

Semantic networks fell out of favour because of their use of ambiguous natural language terms to identify their nodes and arcs, which became a problem when semantic networks were transported between domains and different users, a problem that would be fatal in the decentralized and multi-lingual environment of the Web Woods (1975). According to the Principle of Universality, since a resource can be *anything*, then a component of the knowledge representation language should be considered a resource, and thus can be given a URI. Instead of labelling the arcs and nodes with natural language terms, in RDF all the arcs and nodes can be labelled with URIs. Although few applications had ever taken advantage of the fact before RDF, URIs could be minted for things like the Eiffel Tower *qua* Eiffel-Tower, an absolute necessity for knowledge representation. Since the sense of statements in knowledge representation is usually about content in the world outside the Web, this means that the Semantic Web crucially depends on the rather strange fact that URIs can refer to things outside the Web.

This does not restrict the knowledge-representation language to merely refer to things that we would normally consider outside of the Web, since normal web-pages use URIs as well, and so the Semantic Web can easily be used to refer to normal web-pages. This has some advantages, as it allows RDF to be used to model the relationships between web-accessible resources, and even mix kinds of relationships. This sort of "meta-data" is exemplified by the relationship between a web-page and its human author, which in with RDF would both be denoted by URIs. Lastly, this ability to describe everything with URIs leads to some unusual features, for RDF can then model its own language constructs using URIs, and make statements about its own core language constructs. However, just as all components of RDF may be considered resources, just as all resources may not have URIs, all components of RDF may not have URIs. For example, a string of text or a number may be a component of RDF, and these are called ***literals*** by RDF. In RDF specified anonymous resources are not given a URI, and these are called ***blank nodes***. Yet it would be premature to declare that the deployment of URIs in RDF signal a major improvement over the natural language labels, for URIs can be just as ambiguous as natural language labels by themselves. However, various theories of semantics as well as engineering like the 'follow-your-nose' principle were theorized to solve the problem of ambiguity.

3.2.2 RDF and the Principle of Linking

The second step in Berners-Lee's vision for the Semantic Web, "allowing links to be created with relationship values," follows straightforwardly from the application of the Principle of Universality to knowledge representation. Since RDF is composed of resources, and any resource may link to another resource, then any term in RDF may be linked to another term. This linking forms the heart of RDF, as it allows

disparate URIs to be linked together in order for statements in RDF to be made. The precise form of a statement in RDF is a ***triple***, which consists of two resources connected by a link, as shown in Figure 3.2. This use of RDF shows off the flexibility of using URIs and links for reference instead of access. Lastly, this use of URIs and links *outside* Web representations like those of hypertext web-pages shows the flexibility of the linking paradigm, as RDF is an example of the use of the idea of a *linkbase* that was developed in the hypertext community, in particular in the *Microcosm* hypertext system (a pre-Web forebear that failed due to not being based on open standards and also not being based on the Internet) (Fountain et al, 1990).

Any Web representation *in some form of Semantic Web language* such as RDF are called ***Semantic Web documents***. There are several options for encoding Semantic Web documents. The W3C standardized an encoding of RDF is in a verbose XML format called 'RDF/XML' and a simpler encoding called *Turtle* for triples. In Turtle, a triple is three space-delimited terms (the subject, predicate, and object) ended in a period: `http://www.example.org/EiffelTower http://www.example.org/hasArchitect http://www.example.org/Gustave_Eiffel.` Using namespaces, with `ex="http://www.example.`one abbreviates the example triple to `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel`. As compared to Figure 3.1, the *only* noticeable difference between RDF and a classical semantic network is the use of URIs.

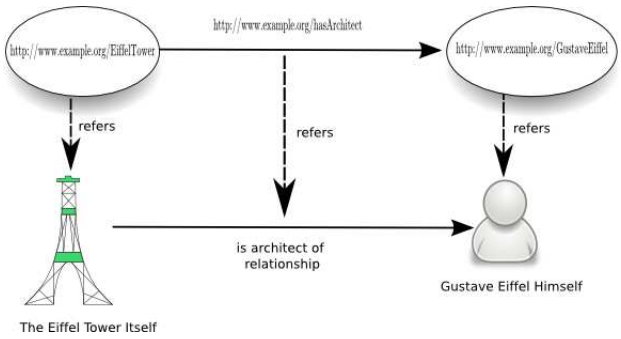

Fig. 3.2 An example RDF statement

There are some restrictions to linking on the Semantic Web. As opposed to the vast numbers and kinds of links possible in XLink, linking on the Semantic Web is directed, like hyperlinks (DeRose et al, 2001) . *The starting resource in the triple* is called the ***subject***, while *the link itself* is called the ***predicate***, and *the ending resource in the triple* is the ***object***. The predicate is usually a role as opposed to an arc role. The major restriction on the Semantic Web is that the subject must be a URI or a blank node, and the predicate must also be a URI. The object, on the other hand, is given the most flexibility, as it may either be a URI, a blank node, or a literal. This predicate-argument structure is a well-known and familiar structure from logic, linguistics, and cognitive science. Triples resemble the binary predicates

in propositional logic needed to express facts, relationships, and the properties of individuals. Furthermore, triples seem similar to simple natural language sentences, where the subject and objects are nouns and the predicate is a verb.

From the perspective of the traditional Web, the main feature of RDF is that links in RDF themselves have a required role URI. It is through this role that URIs are given to relationships outside the Web in RDF. For example, the relationship of 'is architect of' between Gustave Eiffel and the Eiffel Tower could be formalized as a link (as shown in Figure 3.2), as could the relationship between Tim Berners-Lee and the creation of his web-page. Since the relationships are abstract, these URIs then refer to these relationships, the URIs may not be accessible, and RDF predicates are unlike links in traditional hypertext systems. Similarly, a triple by itself can only state a simple assertion, but webs of links may be made between triples to explain complex statements. A set of triples that share resources is called a ***graph***, as illustrated in Figure 3.3 by two triples having the same subject, namely that 'The Eiffel Tower in Paris has as an architect called Gustave Eiffel.'

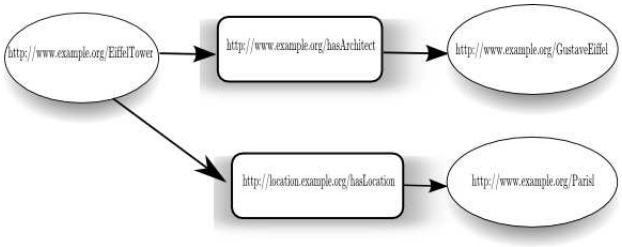

Fig. 3.3 Merging RDF triples

With the ability to make separate statements using URIs, the main purpose of RDF is revealed to be *information integration*. Due to their reliance on URIs, RDF graphs can ***graph merge***, when *two formerly separate graphs combine with each other when they use any of the same URIs*. The central purpose of URIs is to allow independent agents to make statements about the same referent. With a common language of URIs, agents can merge information about the referents of the URIs in a decentralized manner. This is one of the most important applications of the Semantic Web, and it will be further explored in Chapter **??**.

3.2.3 RDF and the Principle of Self-Description

Once the Principle of Universality and the Principle of Linking are obeyed, the Principle of Self-Description naturally follows, and RDF is no exception. Self-description is a crucial advantage of RDF in decentralized environments, since an agent by following links can discover the context of a triple needed for its interpreta-

tion. As witnessed by the Brachman and Smith survey of knowledge representation systems, a bugbear of semantic networks was their inability to be transferred outside of the closed domain and centralized research group that designed them (Brachman and Smith, 1980). The crucial context for usage of a particular semantic network was always lost in transfer, so that what precisely "IS-A" means could vary immensely between contexts, such as the difference between a sub-class relationship or individual identity (Brachman, 1983). By providing self-description, RDF triples can be transported from one context to another, at least in an ideal world where normal conditions, such as when the URIs in the triple can be used to access a web-page describing its content, and correct media types are used.

The hypertext Web, when every resource is linked together, provides a seamless space of linked documents. For example, the W3C tries to deploy its own internal infrastructure in a manner compatible with the principles of Web architecture. Its e-mail lists are archived to the Web, and each e-mail is given a URI, so an agent may follow links seamlessly from one e-mail message to another, and by following links can launch applications to send e-mail, discover more about the group, and in new e-mails reference previous topics. Likewise, an initiative called "Linked Data" attempts to deploy massive public data-sets as RDF, and its main tenet is to follow the Principle of Self Description (Bizer et al, 2008). The hope is that the Semantic Web can be thought of as a seamless web of linked data, so that an agent can discover the interpretation of Semantic Web data by just following links. These links will then go to more data which may host formal definitions or informal natural language descriptions and multimedia depictions. For example, if one finds an RDF triple such as `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel` and discover more information about the Eiffel Tower, like a picture of it or the fact that construction was finished in 1889 by accessing `http://www.example.org/EiffelTower`.

Since RDF is supposed to be an all-purpose knowledge representation system for the Web, RDF statements themselves can also be described using RDF. RDF itself has a namespace document at `http://www.w3.org/1999/02/22-rdf-syntax-ns#`, which provides a description of RDF in RDF itself. In other words, RDF can be meta-modeled using RDF itself, in a similar manner to the use of reflection in knowledge representation and programming languages (Smith, 1984). For example, the notion of a RDF predicate is `http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate`, and is defined there as "the predicate of the subject RDF statement." The same holds for almost all RDF constructs, and a conformant RDF processor can derive from any RDF triple a set of axiomatic triples that define RDF itself, such as `rdf:predicate rdf:type rdf:Property` (all RDF predicates are of the type property). For any RDF statement like `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel`, an RDF-aware agent can then infer that `ex:hasArchitect rdf:type rdf:predicate`, which states in RDF that an architect relationship is a predicate in a RDF triple. However, usually RDF is not hosted according to the Principle of Self-Description. Use of the media type `application/rdf+xml` is not consistent usually, and the namespaces URI of specifications like the RDF Syntax namespace just allow access of to some RDF triples, which is useless to a

machine incapable of understanding RDF in the first place, instead of a more useful RDDL document Borden and Bray (2002). A version of RDDL in RDF (Walsh and Thompson, 2007) with an associated GRDDL transform in order to make it even easier for Semantic Web agents to follow namespace documents to associated resources (Connolly, 2007).

3.2.4 RDF and the Open World Principle

The Principle of the Open World is the fundamental principle of inference on the Semantic Web. A relatively simple language for declaring sub-classes and sub-properties, RDF Schema, abbreviated as RDF(S), was from the beginning part of the vision of the Semantic Web and developed simultaneously with RDF. Yet determining how to specify exactly what other triples may be inferred from a given RDF triple is a non-trivial design problem, since it required adding an inference mechanism to a semantic network, which historically in AI featured little or no inference. Those that do not remember the history of artificial intelligence are bound to repeat it, and the process of specifying inference in RDF led to an almost complete repeat of the 'procedural versus declarative' semantics debate. As originally as defined, the original RDF specification defined its inference procedure by natural language and examples. Yet differing interpretations of the original RDF specification led to decidedly different inference results, and so incompatible RDF processors. This being unacceptable for a Web standards organization, the original defender of formal semantics in artificial intelligence, Pat Hayes, oversaw the creation of a declarative, formal semantics for RDF and RDF(S) in order to give them a principled inference mechanism.

The Open World principle was considered to be a consequence of the lack of centralized knowledge implied by the decentralized creation of URIs and links as given by the Principles of Universality and Linking. The parallel to the removal of centralized link indexes is that on the Semantic Web, "we remove the centralized concepts of absolute truth, total knowledge, and total provability, and see what we can do with limited knowledge" (1998c). Hayes argued, in a similar fashion as he had argued in the original 'procedural versus declarative' semantics debate in AI, that the Semantic Web should just use standard first-order predicate logic. Yet while Berners-Lee accepted the need for a logic-based semantics, he argued against Hayes for the Principle of Open World and monotonicity, and the formal semantics of RDF was designed to obey the Open World Assumption (Hayes, 2002). The reason for maintaining the Open World Assumption was that adding triples in a graph merge should never change the meaning of a graph so one could never retract information by simply adding more triples, or invalidate previously-made conclusions. This monotonicity is considered key, since otherwise every time a RDF triple was merged into a graph the interpretation of the graph could change and so the entire graph might have to be re-interpreted, a potentially computationally expensive operation. By having a design that allows only monotonic reasoning, RDF allows

interpretations to be changed incrementally in order to scale well in the potentially unbounded partial information of the Web. Hayes himself eventually came to agree with Berners-Lee on the issue, noting that reasoning on the Semantic Web "needs to always take place in a potentially open-ended situation: there is always the possibility that new information might arise from some other source, so one is never justified in assuming that one has 'all' the facts about some topic" (2002).

RDF Schema is on the surface a very simple modeling and inference language (Brickley and Guha, 2004). Due to the Open World assumption, unlike schemas in relational databases or XML Schemas, RDF Schemas are not prescriptive, but merely descriptive, and so an agent cannot validate RDF triples as being either consistent or inconsistent with an RDF Schema (Thompson et al, 2004). They cannot make the information given by a triple itself change, but only enrich the description of an existing triple. RDF Schema adds two main features to RDF. First, RDF(S) provides a notion of *class*, or a set of resources. Then RDF(S) allows any resource to be given membership in classes and declare sub-classes (or subsets) of a class that inherit all the triples created to describe the class. Second, RDF(S) also allows properties to have sub-properties and for properties to have types for domains and ranges, such that in for a triple the subject is the domain and the object is the range of a property. Imagine that the property `ex:hasArchitect` has the range `ex:Person` and domain `ex:Building`. Note that RDF Schemas are not automatically applied to triples even if they are mentioned in a triple, such that for a statement like `ex:Eiffel_Tower ex:hasArchitect ex:Gustave_Eiffel`, the fact that the domain of `ex:hasArchitect` is buildings and the range is people, is not known unless the RDF Schema is automatically imported and so merged with the triple itself. An RDF(S)-aware agent that has retrieved the RDF Schema can deduce from the triple that `ex:Gustave_Eiffel rdf:type ex:Person`, namely that Gustave Eiffel is indeed a person. This sort of simple reasoning is again encoded as a set of axiomatic triples and rules for inference and semantic conditions for applying these axioms to infer more triples. See the RDF Formal Semantics for full details (Hayes, 2004). From here on out, the acronym 'RDF' refers to both RDF and RDF(S), whose formal semantics are given together (Hayes, 2004).

In practice, the Principle of the Open World has surprising results. One of the ramifications in RDF is that there is no proper notion of false, but only the notion that something is either inferred or not, and if it is not inferred, it may simply be undefined. Although it seems straightforward, in practice this leads to surprising results. Take the following example: "Gustave is the father of Valentine," which in RDF is `ex:Gustave ex:fatherOf ex:Valentine_Eiffel`. Is George also the father of Valentine (`ex:George ex:fatherOf ex:Valentine`? Operating under the closed world assumption, the answer would be no. Yet operating under the Open World Principle, that statement would be possible, for there is no restriction that the there someone can only have a single father, and in RDF(S) stating such a restriction is impossible. This restriction is possible in the ***Web Ontology Language*** (abbreviated OWL, in an obscure reference to A.A. Milne), an open-world extension of RDF that allows restrictions, such as cardinality, to be placed on predicates. However, even if one set the cardinality of the `ex:fatherOf` predicate to one (so

that one could have at most one father), the results will be surprising: the reasoner will conclude that `ex:George` and `ex:Gustave` refer to the same individual. In contrast to the expected behaviour of many other inference engines, including people, there is no ***Unique Name Assumption***, *the assumption is that each unique name refers to a unique individual*, due to the Open World Principle. The Unique Name Assumption, while very useful for counting, makes an implicit assumption about each name referring to only one individual, and if an individual cannot be found that satisfies the name then that individual must not exist. This further reinforces the tendency of URIs on the Semantic Web, despite their global scope, to be ambiguous, a point we shall return to.

3.2.5 RDF and the Principle of Least Power

Insofar as it is applied to the Semantic Web, the Principle of Least Power is strangely counter-intuitive: traditionally knowledge representation languages were always striving for greater power, yet the Semantic Web begins with RDF, a language purposefully designed to be the least powerful language. The true bet of the Semantic Web is then on triples as the most basic language upon which other languages can be based. The challenge for the Principle of Least Power is how to build the rest of the Semantic Web by expanding on the language of triples.

Inspired by the Principle of Least Power, he envisaged that each language would extend and build upon lower-level languages. On top of RDF, Berners-Lee envisaged a whole stack of more expressive languages being constructed. Although the vagarities of the standardization process have caused various changes in the 'Semantic Web stack' and numerous conflicting versions exist, the original and most popular version of the Semantic Web stack is given in Figure 3.4 (Gerber et al, 2008). The W3C has commenced standardization efforts in a number of these areas, and research in almost all levels of the stack has begun. The majority of the research has focused on extending the Semantic Web with "ontologies" based on description logic like OWL. As should be suspected given their heritage in artificial intelligence, most of the work in description logic applied to OWL has focused on determining the most expressive possible language that preserves decidable inference. OWL itself works well with the Open World Principle, since it only makes an inference by adding inferred statements and classifications, and so remains monotonic. While almost any possible triple is acceptable in RDF, OWL allows users to design ontologies that can even add constraints, such as cardinality and datatyping, that can make some RDF triples inconsistent with a given OWL ontology. Another part of the Semantic Web, originally unforeseen, is the query language ***SPARQL***, a query language for RDF similar to the popular database query language SQL (Prud'hommeaux and Seaborne, 2008). Current work is focused on ***Rule Interchange Format***) (RIF), a rule-language similar to Prolog for both serializing normal rules and operating over RDF data (Boley and Kifer, 2008). Other higher-levels on

the Semantic Web stack such as ‘Unifying Logic’ remain mysterious, if poetic and evocative.

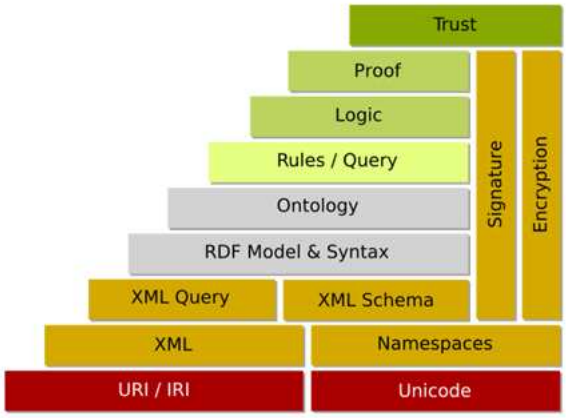

Fig. 3.4 The Semantic Web stack

3.3 Information and Non-Information Resources

One question is whether or not there should be some way to distinguish between URIs used to access web-pages and Semantic Web documents, and URIs used as names for things like physical entities and abstract concepts that are not ‘on the Web.’ This latter class of URIs, URIs that are used as names for entities and abstract concepts, are called ***Semantic Web URIs***. Should a URI be able to both name a non-Web accessible thing in addition to accessing a representation of the thing? This is a difficult question, as it seems the class of web-pages and physical people should be disjointed (Connolly, 2006). The W3C TAG took on this question, calling it the *httpRange-14* issue, which was phrased as the question: what is the range of the HTTP dereference function? (Connolly, 2006)

The TAG defined a class of resources on the Web called an *information resource*, which is a resource “whose essential characteristics can be conveyed in a message” (Jacobs and Walsh, 2004). In particular, this means that an ***information resource*** is a *resource that can be realized as an information-bearing message, even with multiple encodings.* A resource is defined by its sense (content), not the encoding of its Web representations. So information resources would naturally include web-pages and so resources on the hypertext Web, as well as most digital things. However, there are *things that cannot be realized digitally by a message*, but only described or depicted by digital information. These things are ***non-information resources***. Their

only realization is themselves. Many analogue things therefore are non-information resources. It appears that this distinction between information resources and non-information resources is trying to get at the heart of the distinction between a resource being a web-page *about* the Eiffel Tower and a resource *for* the Eiffel Tower itself. A web-page is an information resource, but the Eiffel Tower itself is a non-information resource, as is the text of *Moby Dick* or the concept of red.

The distinction is more subtle than it first appears. The question is not whether something *is* accessible on the Web, but whether it *can be* accessible on the Web by being *in theory* transmitted as an encoding, and therefore Web representation, in a message. For example, imagine a possible world where the Eiffel Tower does not have a web-page. In this world, it would seem counter-intuitive to claim that the web-page of the Eiffel Tower is then not an information resource just because it happens not to *exist* at this moment. This is not as implausible as it sounds, for imagine if the Eiffel Tower's web server went down, so that `http://www.tour-eiffel.fr` returned a `404` status code. A more intuitive case is that of the text of *Moby Dick*. Is the text of *Moby Dick* an information resource? If the complete text of Moby Dick isn't on the Web, one day it might be. However, a particular collector's edition of *Moby Dick* could not be an information resource, since the part of that resource isn't the text, but the physical book itself. Are ordinary web developers expected to have remarkably scholastic discussions about whether or not something is *essentially* information before creating a Semantic Web URI?

Both a web-page about the Eiffel Tower and the text of *Moby Dick* are, on some level of abstraction, carrying information about some content in some encoding. So, if any information resource is any resource which can have its content realized as a Web representation, then information resources *must* be on some level digital so that they can be encoded as Web representations. Then both the text of *Moby Dick* and a web-page about the Eiffel Tower are information resources, even if they are not currently Web-accessible. Digital information can be transmitted via digital encodings, and so *can* in theory be on the Web by being realized as Web representations, even if the resource does not allow access to Web representations at a given time. Lastly, a particular edition of Moby Dick, or Moby Dick in French, or even some RDF triples about *Moby Dick*, are all information resources, with various encodings specified at certain levels of abstraction. It appears that the best story we have to tell about the rather clumsy term 'non-information resource' is that a non-information resource is a thing that is *analogue* and so resists direct digital encoding, but can only be indirectly encoded via representations of the thing in a suitable language. This would then at least be the rather odd combination of physical entities and abstract concepts. So the Eiffel Tower itself, Tim Berners-Lee himself, the integers, and a particular book at a given point in space-time (i.e. on a particular shelf!) are all non-information resources.

Should there be a class to which a web-page about the Eiffel Tower belongs but the text of some as-of-yet unwritten novel does not? In other words, it seems that the class of ***information resources*** is too large, and we need a term for things that are actually accessible over the Web at a given time. We call this kind of thing a

Web resource**, an information resource that has accessible Web representations that realize its information.* A Web resource can then be thought of as a mapping from time of request to a series of Web representation responses, where the information realized by those Web representations *are* the Web resource. This definition is close in spirit to the original pre-Semantic Web thinking behind resources in IETF 1630, as well as in IETF RFC 2616 where a 'resource' is defined as "a network data object or service " and coherent with Engelbart's original use of the term 'resource' (Engelbart and Ruilifson, 1999; Fielding et al, 1999). A ***Semantic Web resource is *a resource that allows access to Semantic Web documents.*

The distinction between information resources and non-information resources has real effects. When the average hacker on the streets wants to add some information to the Semantic Web, the first task is to mint a new URI for the resource at hand, and the second task is to make some of this new information available as a Web representation. However, should a Web representation be accessible from a URI for a non-information resource? If not, should Web representations be accessed from such a non-information resource, as it might confuse the non-information resource itself with a Web resource that merely represents that resource. Yet how else would fulfilling the Principle of Self-Description for Semantic Web resources be possible? To refuse to allow access to any Web representations would make the Semantic Web completely separate from the Web. Non-information resources need ***associated descriptions**, resources that have as their primary purpose the representation, however incomplete, of some non-information resource*. In other words, associated descriptions are classical examples of metadata. According to the TAG, since the associated description is a separate thing from the non-information resource it represents, the non-information should be given a separate URI. This would fulfill the commonsense requirement that the URI for a thing itself on the Semantic Web should be *separate* from the URI for some information about the thing.

The TAG officially resolved *httpRange-14* by saying that disambiguation between these two types of resource should be done through the `303 See Other` HTTP header. The official resolution to Identity Crisis by the TAG is given below as:

- If an HTTP resource responds to a GET request with a `2xx` response, then the resource identified by that URI is an information resource;
- If an HTTP resource responds to a GET request with a `303 (See Other)` response, then the resource identified by that URI could be any resource;
- If an HTTP resource responds to a GET request with a `4xx (error) response`, then the nature of the resource is unknown.

To give an example, let's say an agent is trying to access a Semantic Web URI that names a non-information resource, the Eiffel Tower itself, as illustrated in Figure 3.5. Upon attempting to access that resource with a HTTP GET request using its Semantic Web URI, since the Eiffel Tower itself is not an information resource, no Web representations are directly available. Instead, the agent gets a `303 See Other` that in turn redirects them to a documentation resource that hosts Web representations about the Eiffel Tower, such as the information resource for the home-

page of the Eiffel Tower. When this URI returns the 200 status code in response to an HTTP GET request, the agent can infer that the homepage is actually an information resource. The Semantic Web URI used to refer to the Eiffel Tower itself, `http://www.example.org/EiffelTower`, could be any kind of resource, and so could be a Semantic Web resource. This `303` redirection then allows the Semantic Web resource given by a Semantic Web URI for the Eiffel Tower itself to comply with the Principle of Self-Description.

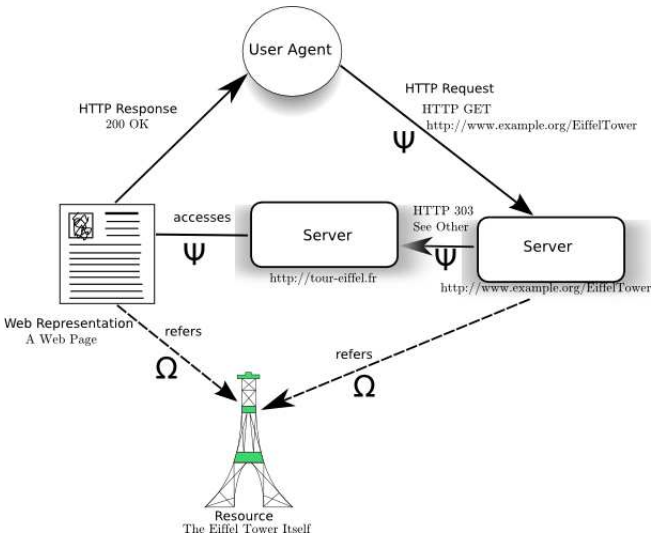

Fig. 3.5 The 303 redirection for URIs

An alternative to the obtuse `303` redirection is the *hash convention*, where one uses the fragment identifier of a URI to get redirection for free. If one wanted a Semantic Web URI that referred to a non-information resource like the Eiffel Tower itself without the hassle of a 303 redirection, one would use the URI `http://www.tour-eiffel.fr/#` to refer to the Eiffel Tower itself. Since browsers, following the follow-your-nose algorithm, either dispose of it or treat the fragment identifier as a fragment of a document or some other Web representation, if an agent tries to access via HTTP GET a Semantic Web URI that uses the hash convention, the server will not return a `404 Not Found` status code, but instead resolve to the URI before the hash, `http://www.tour-eiffel`, which can then be treated as a documentation resource. In this way, Semantic Web inference engines can keep the Semantic Web URI that refers to the Eiffel Tower itself and an associated description about the Eiffel Tower separate by taking advantage of some predefined behaviour in web browsers.

While at first these distinctions between Semantic Web resources and information resources seems ludicrously fine-grained, clarifying them and pronouncing an

official W3C policy on them had an immense impact on the Semantic Web, since once there was definite guidelines on how to publish information on the Semantic Web, users could start creating Semantic Web URIs and connecting them to relevant documentation resources. The TAG's decision on redirection was made part of a tutorial for publishing Semantic Web information called *How to Publish Linked Data on the Web* (Bizer et al, 2007).

3.4 An Ontology of Web architecture

The primary use of a formal ontology in the context of Web architecture is to allow us to formally model the various distinctions used in specifications and debates. Although some other formal logic that deals with actions and events may be more suitable for modelling the temporal transactions of client-server interactions on the Web, an ontology is necessary in order to capture the various distinctions given in specifications first. As even the primary architects of the Web find themselves confused about the distinctions between 'entities' in HTTP and 'representations' in Web architecture (Mogul, 2002), this ontology could be of use as a reference to anyone interested in understanding or even extending existing Web specifications as well as those interested in correctly implementing best practices that are dependent on rather obscure corners of Web architecture, such as Linked Data's 303 redirects. A first attempt to formally model Web concepts was the *Identity, Resources, and Entity* ontology (IRE) (Presutti and Gangemi, 2008), which has evolved in the IRW ontology presented here via several iterations Halpin and Presutti (2009).

IRW is a small ontology at the core of an ontology network. More specifically, IRW defines the core concepts of the Web architecture and can be extended by specialized ontology modules in order to address more specific Web domains such as HTTP transactions and Linked Data. IRW reuses existing ontologies, some of which are ontology design patterns (Gangemi and Presutti, 2009). The following list summarizes the prefixes that are used in the ontology and associates them with their respective ontologies. Terms in IRW ontology will be given in `teletype` font, and if no namespace is given, we will assume the `irw:` namespace. Namespace URIs are given in the footnotes.

Prefix	Ontology name
`irw:`[2]	Identity of Resources on the Web
`ir:`[3]	Information Realization
`comp:`[4]	Composition
`http:`[5]	HTTP concepts based on IRW
`ldow:`[6]	Linked Data concepts based on IRW
`tag:`[7]	'Identity Crisis' concepts based on IRW
`ont:`[8]	Generic Resource
`rdfs:`[9]	RDF Schema
`rdf:`[10]	RDF
`owl:`[11]	OWL

Notice that the stable version of the ontology can also be accessed via its PURL. The latest version of the IRW ontology is at `http://ontologydesignpatterns.org/ont/web/irw.owl#`. While the IRW ontology in full cannot be graphically explicated due to lack of

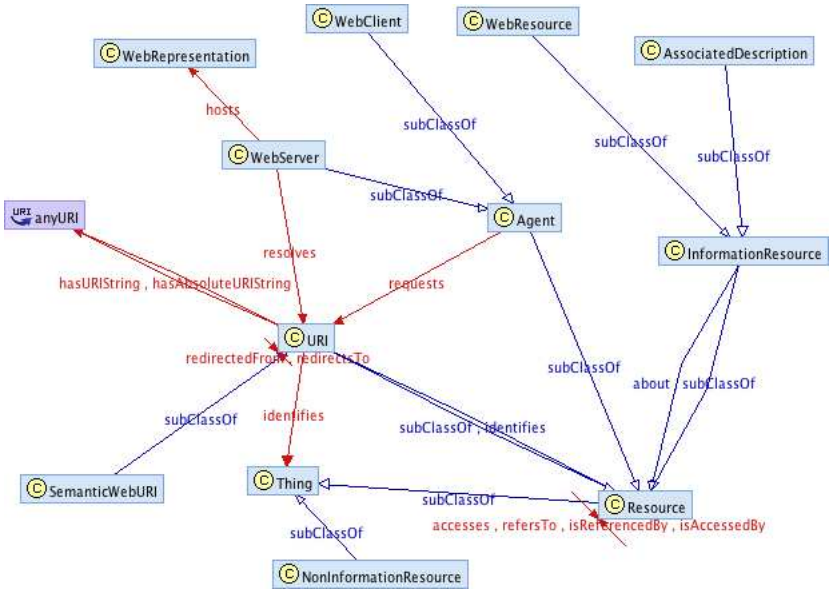

Fig. 3.6 The main elements of the IRW network of ontologies is illustrated as a graph. Boxes with the symbol "C" are classes, while those with a small arrow inside are datatypes. Arcs labelled as "subClassOf" represent `rdfs:subClassOf` relations between classes. The other arcs are either object properties or datatype properties, depending on the range node. The direction of an arc indicates the domain and range of the property. Two arrows that meet on their edges indicate a relation whose domain and range is the given by the same class.

space on a printed page, the primary classes and properties are given in Figure 3.6. The IRW-related elements needed for the example of 303 redirection are given in

Figure 3.5. The IRW ontology defines the class `Resource` to be equivalent to `rdfs:Resource`[12] as it expresses the same intuition.

3.4.1 Resources and URIs

The notion of a URI is modeled as a class, `URI` As XML Schema data-types for URIs are not extensible, modeling URIs as a class allows us to talk about different kinds of URIs, such as IRIs (Internationalized Resource Identifiers) and Semantic Web URIs. A property `identifies` can then connect a URI to a resource. Since we want to associate a URI with character strings (possibly with the XML Schema data-type for URIs) such as 'http://www.example.org,' we also have a property called `hasURIString`. This property has various (functional) sibling children such one relating IRIs to URIs, so that a IRI given in the Japanese character can be a URI. The core properties we include are `hasRelativeURIString` and `hasAbsoluteURIString` for the conversion of relative URIs to absolute URIs.

- **`Resource`**: An OWL Class. "Anything that might be identified by a URI" (Jacobs and Walsh, 2004). This class is meant to express the same intuition of `rdfs:Resource` hence it is defined as equivalent to `rdfs:Resource`.
  - `owl:equivalentTo rdfs:Resource`
- **`URI`**: An OWL Class. An abbreviation for Uniform Resource Identifier. "A global identifier in the context of the World Wide Web" (Jacobs and Walsh, 2004). Any identifier that follows the role given in IETF RFC 3986 can be an instance of this class, even if it is an IRI that has a conversion to a URI or uses a scheme such as URN (Moats, 1997) or URL (Berners-Lee et al, 1994) that has been subsumed by the concept of URIs.[13]
  - `rdfs:subClassOf Resource`
  - `identifies exactly 1 Resource`
- **`identifies`**: An OWL Object Property. The relationship between a URI and a resource. It can be functional as one should "assign distinct URIs to distinct resources" although some users of this ontology may wish to not use this constraint and so use the `refersTo` property (Jacobs and Walsh, 2004).
  - `owl:inverseOf isIdentifiedBy`
  - `rdfs:domain URI`
  - `rdfs:range Resource`
  - `rdfs:subPropertyOf refersTo`
  - `owl:FunctionalProperty`

[12] Notice that the ontology is encoded in OWL2.

[13] Note that this class has itself a URI that is the `irw` class name for URI in the IRW namespace, but concrete individual URIs are instances of this class and could be any URI.

- **accesses**: An OWL Object Property. The relationship between a resource and another resource where the former provides a causal pathway to the latter.
  - owl:inverseOf isAccessedBy
  - rdfs:domain Resource
  - rdfs:range Resource
  - owl:TransitiveProperty
- **refersTo**: An OWL Object Property. The relationship between a resource and another resource where the former may be immediately causally disconnected from the latter but still 'stand in' for it in a syntactic expression. Note that reference in the logicist position is an aspect of an interpretation of the syntax of an ontology, not a property of the use of an ontology itself. So this is actually a meta-property that attempts to make explicit the *intended* interpretation of an agent.
  - owl:inverseOf isReferencedBy
  - rdfs:domain Resource
  - rdfs:range Resource

3.4.2 Information Resources

There is a controversial sub-class of Resource outlined in AWWW known as 'information resources.' The AWWW defines the notion of **information resource** as "a resource which has the property that all of its essential characteristics can be conveyed in a message" (Jacobs and Walsh, 2004), which we model as InformationResource. This definition has widely been thought of as unclear, and defining what set of individuals belong in this class and what do not has been a source of perpetual debate on various list-servs. In order to clarify this notion we decided to reuse a known ontology pattern i.e. the *Information Realization* content ontology pattern, referred to with prefix ir:. Remarkably, this content ontology pattern is extracted from the DOLCE Ultra Light ontology[14] and is implemented also in the Core Ontology for Multimedia (COMM)[15] for addressing a similar modeling issue. The reuse of such a content pattern also supports interoperability with other ontologies that reuse it. This pattern-based approach to ontology design is a strength of IRW.

Notice that the ir: is very small, two classes and two object properties, hence it is convenient to simply directly import all of the *Information Realization* pattern. An InformationResource is viewed to be equivalent to the notion of *information object* from ir:, such as a musical composition, a text, a word, or a picture. An information object is an object defined at a level of abstraction, independently from how it is concretely realized. This means an information resource has, via the ir:realizes property (with inverse ir:isRealizedBy), at least

[14] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

[15] http://comm.semanticweb.org/

one `ir:InformationRealization`, a concrete *realization*. The fact that any information resource's "essential characteristics can be conveyed in a single message" implies that everything from a bound book to the electric voltages that encode a HTTP message can be a realization of an information resource (Jacobs and Walsh, 2004). Furthermore, the property `about` (and inverse property, `isTopicOf`) expresses the relationship between an information resource and other resource (or resources) that an information resource is 'about.'

Examples of realizations are descriptions of a resource using natural language or depictions of a resource using images. Information resources can, but not necessarily, be identified (accessed or referred to) by a URI. In this manner, the text of Moby Dick can be an information resource since it could be conveyed as a single message in English, and can be realized by both a particular book or a web-page containing that text. Thus, the definition of information object and information realization can be thought of as similar to the classic 'type-token' division in philosophy of mind between an object given on a level of abstraction and some concrete thing which realizes that abstraction, where that single abstraction may have multiple realizations. This is similar, but broader than the class-individual distinction as one may want to model the 'token' or 'realization' itself as a class. As such, it's also broader than *TBox* and *ABox* distinction from description logic.

- **`InformationResource`**: An OWL Class. "A resource which has the property that all of its essential characteristics can be conveyed in a message" (Jacobs and Walsh, 2004).
  - `rdfs:subClassOf Resource`
  - `ir:isRealizedBy min 1 ir:InformationRealization`
  - `owl:equivalentTo ont:InformationResource`
  - `owl:equivalentTo ir:InformationObject`, which is defined by `ir:` as "A piece of information, such as a musical composition, a text, a word, a picture, independently from how it is concretely realized" (Gangemi, 2008).
- **`ir:InformationRealization`**: An OWL Class. Imported from `ir:`. "A concrete realization of an expression, e.g. the written document containing the text of a law" (Gangemi, 2008). This is equivalent to the broader notion of **`representation`** as defined in AWWW, "data that encodes information about resource state" (Jacobs and Walsh, 2004).
- **`ir:realizes`**: An OWL Object Property. Imported from `ir:`. "A relation between an information realization and an information object, e.g. the paper copy of the Italian Constitution realizes the text of the Constitution" (Gangemi, 2008).
  - `owl:inverseOf ir:isRealizedBy`
  - `rdfs:domain ir:InformationRealization`
  - `rdfs:range ir:InformationObject`
- **`about`**: An OWL Object Property. An intentional relationship between an information resource and another resource. Note that this property is wider than the inverse functional `foaf:primaryTopic` and `foaf:isPrimaryTopicOf`

properties of the `Friend of a Friend (FOAF)` vocabulary,[16] which could be considered sub-properties of this property, as the `about` property makes no claims about whether a topic is primary or not.

- `owl:inverseOf`: `isTopicOf`
- `rdfs:domain InformationResource`
- `rdfs:range Resource`

3.4.3 Web Resources and Web Representations

Up until this section, the work done by IRW has, outside of mentioning URIs, not been specific to the Web per se, but explicating the more general ideas of information and resources that apply equally as well to books as to web-pages. In this section, we further specialize IRW to the Web domain by considering the notion of 'representations' that can be transferred over a protocol such as HTTP. To avoid confusion with the broader philosophical notion of representation, we call this term from Web architecture **web representations** instead. Also, it is possible our use of the term 'representation' is narrower than the AWWW's use, which could be equivalent to the notion of any information realization in the large, while our use of the term is instead for representations sent over the Web using HTTP. Furthermore, one can distinguish **web resources** (`WebResource`) as a subset of information resources that are *under normal conditions* usually web-accessible, i.e. the server is not down, the browser works normally, etc.

In terms of HTTP, a `WebRepresentation` is an entity (associated with various entity headers and an entity body) that is also subject to content negotiation and so may be transferred as multiple entities. This is because, as given in IETF RFC 2616, a web representation may be defined as "an entity included with a response that is subject to content negotiation" such that "there may exist multiple representations associated with a particular response status" (Fielding et al, 1999). Therefore, we define `WebRepresentation` as a sub-class of a more general `Entity` class as defined by HTTP RFC 2616 (Fielding et al, 1999). The term 'entity' could be confusing as it is often used in many other philosophical and technical contexts. However, in HTTP an **entity** may be the information given by either a HTTP request or response, but a web representation, by virtue of being a 'representation' of a resource, is only for a HTTP response. A web representation is thus a kind of entity that is about the state of a resource as defined in AWWW (Jacobs and Walsh, 2004), but there are entities that only request the state of resources or indicate that requests can or cannot be fulfilled. For example, a HTTP POST request or even a 404 response are entities but they do not necessarily represent the state of a particular web resource. An entity may be transferred as the request or response of many particular actions by a client. For example, different URIs may return the same entity, such as when one URI hosts a copy of a resource given by another

[16] The `foaf:` prefix stands for `http://xmlns.com/foaf/0.1/`

URI. In order to model the complexity of headers and bodies in HTTP entities, we use another popular content ontology pattern, the *Composition* pattern, referred to as `comp:`. This pattern, extracted from the DOLCE Ultra Lite ontology,[17] allows us to model a non-transitive component-whole relationship, which however implies (by subsumption) a transitive part-of relation.

- **`http:Entity`**: An OWL Class. "The information transferred as the payload of a request or response" (Fielding et al, 1999). "An entity consists of metainformation in the form of entity-header fields and content in the form of an entity-body" (Fielding et al, 1999).
  - `rdfs:subClassOf ir:InformationRealization`
  - `comp:hasComponent exactly 1 http:EntityHeader`
  - `comp:hasComponent max 1 http:EntityBody`
- **`http:EntityBody`**: An OWL Class. Whatever information is sent in the request or response is in "a format and encoding defined by the entity-header fields" (Fielding et al, 1999). Also called in HTTP the 'content' of a message (Fielding et al, 1999).
  - `http:hasMediaType some http:MediaType`
- **`http:EntityHeader`**: An OWL Class. "Entity-header fields define metainformation about the entity-body or, if no body is present, about the resource identified by the request" (Fielding et al, 1999). Sometimes called in HTTP "meta-information" (Fielding et al, 1999). Various fields of the entity header can define HTTP status codes (`http:StatusCode`), content encoding (`http:MediaType`), content language (`http:ContentLanguage`), date of creation (`http:CreationDate`), date of modification (`http:ModificationDate`), and so on.
  - `rdfs:subClassOf ir:InformationRealization`
  - `http:hasComponent min 1 http:EntityHeaderField`
- **`http:hasHeaderFieldValue`**: An OWL Object Property. A relation between an entity header field and its field values. It is specialized by several properties, each representing a typical entity header field such as `http:hasStatusCode` and `http:hasContentType`.
  - `rdfs:domain http:HeaderField`
- **`WebRepresentation`**: An OWL Class. A sequence of octets, along with representation metadata describing those octets, that constitutes a record of the state of the resource at the time when the representation is generated (Berners-Lee et al, January 2005). Note that the term 'representation' is used for this class in IETF RFC 3968, but has been changed to 'web representation' to separate it from the more general notion of 'representation' in philosophy (Jacobs and Walsh, 2004)

[17] `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`

  - `rdfs:subClassOf http:Entity`
  - `locatedOn min 1 WebServer`

- **`WebResource`**: An OWL Class. "A network data object or service" (Fielding et al, 1999). As such, this is a resource that is accessible via the Web (Hayes and Halpin, 2008). Therefore, a web resource must have at least one URI and be realized by at least one web representation.

  - `rdfs:subClassOf InformationResource`
  - `isIdentifiedBy min 1 URI`
  - `ir:isRealizedBy min 1 WebRepresentation`

3.4.4 Media Types, Generic, and Fixed Resources

One intriguing problem, central to the notion of web representations and resources, is the connection between media types and resources. Very little work has been done in this area, likely due to the lack of use of content negotiation in general on the hypertext Web. For example, instead of using content negotiation to return versions of the same resource in multiple languages, many sites use explicit links. The only substantial work so far on this issue has been Berners-Lee's note *Generic Resources* where he outlines an ontology of types of resources conditioned by how the resource varies over HTTP requests (Berners-Lee, 1996a). Berners-Lee has informally said that a **generic resource** is equivalent to information resources, since the important part of a generic resource is the information itself, not any particular realization of the information. For example, a resource like 'the weather report for Los Angeles' is a generic resource, as is the text of Moby Dick in any language. However, the 'weather report for Los Angeles today' is not a generic resource as it is indexed to a particular temporal junction nor is Moby Dick in a particular language like English. Resources may also vary over time. For example, the text of Moby Dick will be the same over time and so be **time-invariant**, but the resource for the 'weather report for Los Angeles' will change over time and so be **time-specific** (Berners-Lee, 1996a). Furthermore, resources may vary over media-types. For example, the same information may be given in some custom XML dialect or RDF or the same depiction may be given in different formats like JPG and SVG. These resources are all imported from Berners-Lee's *ont* ontology.[18] There are also **fixed resources** that regardless of aspects like time and natural language always return the same representation. For example, a resource for Moby Dick that always provided the same edition in the same language as plain text would be a fixed resource. The idea of a fixed resource is surprisingly common, as it equates a single web-page with a resource and so matches the folk psychology of most users of the Web.

- **`ont:TimeSpecificResource`**: An OWL Class. A resource of which all representations are in the same version. Representations of the resource will not

[18] `http://www.w3.org/2006/gen/ont`

change as a result of the resource being updated to a version with time. The dates of creation and of last modification of such a resource would be expected to be the same.

  - `rdfs:subClassOf InformationResource`
  - `owl:disjointWith ont:TimeGenericResource`
  - `ir:realizedBy only (WebRepresentation ∧ (comp:hasComponent exaclty 1 CreationDate) ∧ (comp:hasComponent exactly 1 LastModificationDate))`

- **`ont:LanguageSpecificResource`**: An OWL Class. A resource of which all representations are in the same natural language.

  - `rdfs:subClassOf InformationResource`
  - `owl:disjointWith ont:LanguageGenericResource`
  - `ir:realizedBy only (WebRepresentation ∧ (comp:hasComponent exactly 1 ContentLanguage))`

- **`ont:ContentTypeSpecificResource`**: An OWL Class. A resource of which all representations are encoded in the same Internet media-type, also called 'content-type.'

  - `rdfs:subClassOf InformationResource`
  - `owl:disjointWith ont:ContentGenericResource`
  - `realizedBy only (WebRepresentation ∧ (comp:hasComponent only (EntityBody encodedIn exactly 1 MediaType)))`

- **`ont:FixedResource`**: An OWL Class. A resource whose representation type and content will not change under any circumstances.

  - `owl:equivalentTo (ont:ContentTypeSpecificResource ∧ ont:LanguageSpecificResource ∧ ont:TimeSpecificResource)`

3.4.5 Hypertext Web Transactions

The typical web transaction is started by an agent, given in IRW by a class `Agent`, which is some client like a browser in the context of the Web (Jacobs and Walsh, 2004). This agent can issue a **request** (`requests`) through an entity (`http:sendsEntity`) containing a header field with, as value, the URI that the request is acting upon (`hasRequestedURI`). This path is modeled in IRW by means of a property chain axiom, asserted in the module devoted to HTTP, i.e. `http:`. Note that `requests` serves as a hook to the alignment of IRW with *HTTP in RDF*[19] as a URI corresponds a response executed by a server which returns an

[19] `http://www.w3.org/TR/HTTP-in-RDF10/`

entity which includes a status code (`http:StatusCode`). Hence, we also introduce the class `WebServer` for the generic notion of a **web server**, which has a `resolves` property. The property represents the resolution of a URI to a concrete web server, which currently is done by mapping a URI to an IP address or addresses via the Domain Name System (DNS).[20]

Each `WebServer` `resolves` at least one `URI`, and for the resolution to be successful, the web server has also to be the **location of** i.e. it `hosts`, at least one `WebRepresentation`. This indicates that a web server concretely can respond to an HTTP request with a particular web representation. Since `requests` and `resolves` are all sub-properties of the transitive property `accesses`, this part of the ontology models the physical and causal pathway between a given request for a URI and a response with a web representation.

The entity given in the request may have a preferred media-type, and the response should have a media-type as well. The media-type, such as 'application/xml' or 'application/rdf+xml,' tells the agent how to interpret the entity body of the response. Media-types are modeled in IRW through the class `MediaType`. The relationship between a `http:MediaType` and an `http:Entity` is given by the `encodes` relationship. Note that each web representation has a single media-type.

A URI may also have a `redirectsTo` property, a sub-property of `accesses`, that we can use to model HTTP redirection. This can be done via a number of different techniques, ranging from a 'Content-Location' HTTP entity header to a 300-hundred level HTTP status code, and to model these we rely on the *HTTP-in-RDF* ontology.[21] Note that, even in the light of the W3C TAG's *httpRange-14* decision, redirection can be also used between information resources that have nothing to do with the Semantic Web. So, the domain and range say nothing about the type of resource.

- **`Agent`**: An OWL Class. A human or a program that establishes connections for the purpose of sending requests (Fielding et al, 1999). In the W3C AWWW, an agent is "a person or a piece of software acting on the information space on behalf of a person, entity, or process" (Jacobs and Walsh, 2004).
  - `rdfs:subClassOf Resource`
- **`requests`**: An OWL Object Property. "The act of issuing a request message from a client to a server that includes, within the first line of that message, the method to be applied to the resource, the identifier of the resource, and the protocol version in use" (Fielding et al, 1999). A request action is a flow itself characterized by an agent that sends an entity that includes a URI; this is expressed in IRW by a property chain axiom.
  - `rdfs:subPropertyOf accesses`
  - `rdfs:domain Agent`
  - `rdfs:range URI`

[20] Although caching complicates this in actual situations.

[21] `http://www.w3.org/TR/HTTP-in-RDF10/`

  - `http:sendsEntity o comp:hasComponent o http:hasRequestedURI`

- **`WebServer`**: An OWL Class. “An application program that accepts connections in order to service requests by sending back responses”(Fielding et al, 1999). Note that “any server may act as an origin server, proxy, gateway, or tunnel, switching behavior based on the nature of each request” (Fielding et al, 1999). A web server hosts at least one web representation and resolves at least one URI.

  - `rdfs:subClassOf Agent`
  - `hosts min 1 WebRepresentation`
  - `resolves min URI`

- **`resolves`**: An OWL Object Property. The relationship between a web server that hosts a web representation, and the URI of the resource realized by that web representation.

  - `owl:inverseOf resolvesTo`
  - `rdfs:subPropertyOf accesses`
  - `rdfs:domain WebServer`
  - `rdfs:range URI`
  - `hosts o ir:realizes o isIdentifiedBy`

- **`locatedOn`**: An OWL Object Property. A relation between a web representation and a web server, indicating that the entity can be obtained by an HTTP request to the web server.

  - `owl:inverseOf hosts`
  - `rdfs:domain WebRepresentation`
  - `rdfs:range WebServer`

- **`encodedIn`**: An OWL Object Property. The relationship between an information realization and its encoding. In the case of entities its range is the entity’s media type. So given an entity that has a component with a content type header field set to a certain media type, that entity is `encodedIn` that media type.

  - `owl:inverseOf encodes`
  - `rdfs:domain ir:InformationRealization`
  - `comp:hasComponent o comp:hasComponent o irw:hasValueMediaType`

- **`redirectsTo`**: An OWL Object Property. The relationship between two URIs wherein any requested entity is forwarded to the URI given as the object of this property.

  - `owl:inverseOf redirectedFrom`
  - `rdfs:subPropertyOf accesses`
  - `rdfs:domain URI`
  - `rdfs:range URI`

3.4.6 Modeling the Semantic Web and Linked Data

The Semantic Web is supposed to use URIs not only for hypertext documents but also for abstract concepts and things. In order to model explicitly the redirection solution to the 'Identity Crisis' by the W3C TAG, two distinct sub-properties of `redirectsTo` have been added in a specific module of IRW[22] associated with prefix `tag:`. This module contains the `tag:redirects303To` property and the `tag:redirectsHashTo` property. The former models the TAG's 'solution' to *httpRange-14* while the latter represents the hash convention. With these kinds of re-directions in hand, we can now model the typical Semantic Web transaction. A new sub-class of `URI`, `SemanticWebURI` is given. A **Semantic Web URI** refers to a resource that is not accessible on the Web such as the Eiffel Tower, and so the URI must redirect to another URI that can access an information resource containing data encoded in some Semantic Web language like RDF. Therefore, this kind of URI also has a constraint that it must have at least one `redirectsTo` property.

As mentioned earlier, in the 'Linked Data Tutorial' note, the kinds of resources referred to by a Semantic Web URI are called **non-information resources** (Bizer et al, 2007). Although this term is controversial (and explicitly not endorsed by Berners-Lee) and hard to define abstractly, operationally it simply means a resource that is not web-accessible that therefore should, to comply with the Linked Data initiative, use redirection to resolve to an information resource describing the non-information resource. Although the space of non-information resources is relatively large and hard to draw precise boundaries around, we list a few exemplars in order to serve a what Dennett would call "intuition-pumps" in order to help us understand this concept (Dennett, 1981). In particular a new class called `ldow:NonInformationResource` is introduced, which represents things that can not themselves - for whatever reason - be realized as a single digitally encoded message. Naturally, this class is disjoint with `InformationResource`. A number of different kinds of things may be `NonInformationResources`. Since this concept is the cause of much confusion and debate, it can obviously range over physical people, artifacts, places, bodies, chemical substances, biological entities, and the like - or to resources that are created in a social process and can not be completely realized digitally such as legal entities, political entities, social relations, as well as the concept of a horse, and imaginary objects like unicorns or even functions over the integers.

An **associated descriptions** (`ldow:AssociatedDescription`) is an information resource that can be accessed via redirection from a Semantic Web URI (Bizer et al, 2007). In DBpedia[23] the resource `dbpedia:/resource/Eiffel_Tower` redirects to `dbpedia:/data/Eiffel_Tower` in RDF/XML, and to an HTML page at `dbpedia:/page/Eiffel_Tower` depending on the requested media type (Auer et al, 2007). This Linked Data typical scenario can be generalized as follows: a `WebClient requests` a `SemanticWebURI` *x* and the re-

[22] `http://www.ontologydesignpatterns.org/ont/web/tag2irw.owl`

[23] Prefix `dbpedia:` is used for the namespace `http://dpedia.org`

quest is redirected (via hash or 303 redirection) to another URI that identifies a `ldow:AssociatedDescription`[24], which has one `about` property to a non-information resource. The associated description is typically created in order to describe its associated non-information resource. We model `ldow:AssociatedDescription` as a subclass of `WebResource`. For an illustrated example of these classes in action, refer to Figure 3.5.

- **`SemanticWebURI`**: An OWL Class. A URI used to identify any resource that is not accessible on the Web.
  - `rdfs:subClassOf URI`
  - `identifies only NonInformationResource`
  - `redirectsTo min 1 (URI and identifies only ldow:AssociatedDescription)`
- **`NonInformationResource`**: An OWL Class. All resources that are not information resources.
  - `rdfs:subClassOf Resource`
  - `owl:disjointWith InformationResource`
- **`ldow:AssociatedDescription`**: An OWL Class. A resource that exists primarily to describe a non-web accessible resource.
  - `rdfs:subClassOf WebResource`
  - `redirectedFrom some SemanticWebURI`
  - `isAbout exactly 1 ldow:NonInformationResource`
- **`tag:redirects303To`**: An OWL Object Property. A redirection that uses the HTTP 303 status code.
  - `owl:inverseOf redirected303From`
  - `rdfs:domain URI`
  - `rdfs:range URI`
  - `rdf:type owl:FunctionalProperty`
- **`tag:redirectsHashTo`**: An OWL Object Property. A redirection that works via the fragment identifier being removed from the URI.
  - `owl:inverseOf redirectedHashFrom`
  - `rdfs:domain URI`
  - `rdfs:range URI`

[24] Typical Linked Data terminology is represented in a specific module of IRW referred to here by the prefix `ldow:`, which stands for the namespace `http://ontologydesignpatterns.org/ont/web/ldow2irw.owl`

3.5 The Semantic Web: Good Old Fashioned AI Redux?

Despite its apparent utility in crafting formal ontologies, at the present moment the Semantic Web has not taken off as part of the wider Web. To many, it has seemed that the Semantic Web was nothing but a second coming of classical artificial intelligence. As put by Yorick Wilks, "Some have taken the initial presentation of the Semantic Web by Berners-Lee, Hendler and Lassila to be a restatement of the Good Old Fashioned AI agenda in new and fashionable World Wide Web terms" (2008). So why would the Semantic Web succeed where classical knowledge representations failed? The first reason would be a difference in the underlying intellectual project. A second reason would be a difference in technology.

The difference of the project is one both of scope and goal. The Semantic Web is, at first glance at least, a more modest project than artificial intelligence. To review the claims of artificial intelligence in order to clarify their relation to the Semantic Web, we are best served by remembering the goal of AI as stated by John McCarthy at the 1956 Dartmouth Conference, "the study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al, 1955). However, 'intelligence' itself is not even vaguely defined. The proposal put forward by McCarthy gave a central role to "common-sense," so that "a program has common sense if it automatically deduces for itself a sufficient wide class of immediate consequences of anything it is told and what it already knows" (1959).

In contrast, the Semantic Web does not seek to replicate human intelligence and encode all common-sense knowledge in some universal representational scheme. The Semantic Web instead leaves "aside the artificial intelligence problem of training machines to behave like people" but instead tries to develop a representation language that can *complement* human intelligence, for "the Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help" (Berners-Lee, 1998c). Despite appearances, the Semantic Web is in the spirit of Licklider and Engelbart rather than McCarthy, Minsky, and even latter-day proponents of AI like Brooks. Berners-Lee is explicit that the project of encoding human intelligence is not part of the problem, as the Semantic Web "does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing well-defined operations on existing well-defined data" (Berners-Lee, 1998c). Instead, the Semantic Web is an intellectual project whose goal is philosophically the opposite of artificial intelligence, the creation of new forms of collective intelligence. As phrased by Licklider, this would be a "man-machine symbiosis," in which in "the anticipated symbiotic partnership, men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions" (1960).

While the goals of the Semantic Web are different, it does still employ the same fundamental technology as classical artificial intelligence: knowledge representation languages. As put by Berners-Lee, "The Semantic Web is what we will get if we perform the same globalization process to knowledge representation that the Web initially did to hypertext" (Berners-Lee, 1998c). Yet there is a question about whether or not knowledge representation *itself* might be the problem, not just scale. As put by Karen Spärck Jones, one of the founders of information retrieval, "there are serious problems about the core [Semantic Web] idea of combining substantive formal description with world-wide reach, i.e. having your cake and eating it, even if the cake is only envisaged as more like a modest sponge cake than the rich fruit cake that AI would like to have" (2004). So the problem may lie in the very use of *knowledge representation language* itself. So far we have shown that the properties of at least RDF as a a knowledge representation language puts the emphasis on 'Web' as opposed to 'Semantic' in the Semantic Web, as it has a number of properties — a graph structure, the ability to make unconstrained statements, and the like – that have their basis in the tradition of the Web, rather than knowledge representation in AI. As the Web has proved to be extraordinarily successful, the hope of the Semantic Web is that any knowledge representation language which is based on the same principles as the Web may fare better than its ancestors in artificial intelligence. However, these changes in the formalism of RDF due to the influence of the Web are all relatively minor, and while counter-intuitive to traditional knowledge representation, they have yet to be vindicated as the Semantic Web has not yet reached widespread use.

Overlooked by Spärck Jones in her critique of the Semantic Web, the only substantive difference between traditional knowledge representation and the Semantic Web is the central role of URIs. Just as the later principles of Web architecture build upon the Principle of Universality, so the Semantic Web builds on top of the use of URIs as well. The true bet of the Semantic Web is *not* a bet on the return of knowledge representation languages, but a bet on the universality of URIs, namely that agents in a decentralized and global manner can use URIs to share meaning even about non-Web accessible things. As this use of URIs as the basic element of meaning is central to the Semantic Web, and as it is a genuinely *new* technical claim, it is precisely in the understanding of the status of meaning and reference of URIs that any new *theoretical* claim must be made. Furthermore, it is precisely within the realm of URIs that any *technical* claim to advance must be made.

Chapter 4
Theories of Semantics on the Web

Meaning is what essence becomes when it is divorced from the object of reference and wedded to the word. **W.V.O. Quine** (1951).

4.1 The Identity Crisis

How can agents determine what a URI identifies? To use a word more familiar to philosophers, how can anyone determine what a URI refers to or means? On the pre-Semantic Web, a URI trivially identify the hypertext web-pages that the URI accesses. On the Semantic Web, a whole new cluster of questions, dubbed the *Identity Crisis*, emerges. Can a URI for the Eiffel Tower be used to refer to the Eiffel Tower in Paris itself? If one just re-uses a URI for a web-page of the Eiffel Tower, then one risks the URI being ambiguous between the Eiffel Tower itself and a particular representation of the Eiffel Tower. If one gives the Eiffel Tower *qua* Eiffel Tower its own URI, should that URI allow access to any information, such as a hypertext web-page? In the realm of official Web standards, the jury is still out. In the specification of RDF, Hayes notes that "exactly what is considered to be the 'meaning' of an assertion in RDF or RDF(S) in some broad sense may depend on many factors, including social conventions, comments in natural language" so unfortunately "much of this meaning will be inaccessible to machine processing" such that a "a full analysis of meaning" is "a large research topic" (Hayes, 2004).

The comment in the RDF Semantics specification glosses over a huge argument. Unsurprisingly, the reason there is no standardized way to determine the meaning of a URI is because, instead of a single clear answer, there is a conceptual quagmire dominated by two positions in the development of RDF. The first position, the *direct reference position*, is that the meaning of a URI is whatever was intended by the owner. The owner of the URI should be able to unambiguously declare and communicate the meaning of any URI, including a Semantic Web URI. In this position, the referent is generally considered to be some individual unambiguous *single* thing,

like the Eiffel Tower or the concept of a unicorn. This viewpoint is the one generally held by many Web architects, like Berners-Lee, who imagine it holds not just for the Semantic Web, but the entire Web. The second position, which we call the *logicist position* due to its more clear roots in non-modal logic, is that for the Semantic Web, the meaning of a URI is given by whatever things satisfy the model(s) given by the formal semantics of the Semantic Web. Adherents of this position hold that the referent of a URI is ambiguous, as many different things can satisfy whatever model is given by the interpretation of some sets of sentences using the URI. This position is generally held by logicians, who claim that the Semantic Web is entirely distinct from the hypertext Web, with URIs serving as nothing more than particularly funny symbols.

These two antagonistic positions were subterranean in the development of the Semantic Web, until a critical point was reached in an argument between Pat Hayes, the AI researcher primarily responsible for the formal semantics of the Semantic Web, and Berners-Lee. This argument was provoked by an issue called 'Social Meaning and RDF' and was brought about by the following draft statement in the *RDF Concepts and Abstract Syntax Recommendation*, "the meaning of an RDF document includes the social meaning, the formal meaning, and the social meaning of the formal entailments" so that "when an RDF graph is asserted in the Web, its publisher is saying something about their view of the world" and "such an assertion should be understood to carry the same social import and responsibilities as an assertion in any other format" (2004). During the period of comments for the RDF Working Drafts, Bijan Parsia commented that the above-mentioned sentences do not "really specify anything and thus can be ignored" or are "dangerously underthought and underspecified" and so should be removed (Parsia, 2003). While at first these sentences about the meaning of RDF seemed to be rather harmless and in concordance with common-sense, the repercussions on the actual implementation of the Semantic Web are surprisingly large, since "an RDF graph may contain 'defining information' that is opaque to logical reasoners. This information may be used by human interpreters of RDF information, or programmers writing software to perform specialized forms of deduction in the Semantic Web" (Klyne and Carroll, 2004). In other words, a special type of *non-logical* reasoning can therefore be used by the Semantic Web.

An example of this extra-logical reasoning engendered by the fact that URIs identify 'one thing' is as follows. Assume that a human agent has found a URI for the Eiffel Tower from DBpedia, and so by accessing the URI a Semantic Web agent can discover a number of facts about the Eiffel Tower, such that it is in Paris and that its architect is Gustave Eiffel, and these statements are accessed as an RDF graph (Auer et al, 2007). However, a human can have considerable background knowledge about the Eiffel Tower, such as a vague belief that at some point in time it was the tallest building in the world. This information is confirmed by the human agent employing the follow-your-nose algorithm, where by following the subject of any triple, the human would be redirected to the hypertext Wikipedia article about the Eiffel Tower, where the agent discovers via a human-readable description that the Eiffel Tower was in fact the tallest building until 1930, when it was superseded in height by New York City's Chrysler building. This information is *not* explicitly in the RDF graphs

provided. It is furthermore difficult to even phrase this sort of temporal information in RDF. Furthermore, the human agent discovers another URI for the Eiffel Tower, a RDF version of Wordnet in the file `synset-Eiffel_Tower-noun-1.rdf` (van Assem et al, 2006). When the human agent accesses this URI, there is little information in the RDF graph except that this URI is used for a noun. However, the human-readable `gloss` property explains that the referent of this URI is 'a wrought iron tower 300 metres high that was constructed in Paris in 1889; for many years it was the tallest man-made structure.' Therefore, the human agent believes that there is indeed a singular entity called the 'Eiffel Tower' in Paris, and that this entity was in fact at some point the tallest building in the world, and so the two URIs are equivalent in some sense, although the URIs do not formally match. What the 'Social Meaning' clause was trying to state is that the human should be able to *non-logically* infer that both URIs refer to the Eiffel Tower in Paris, and they use this information to merge the RDF graphs, resulting in perhaps some improved inferences in the future.

This use-case was put forward primarily by Berners-Lee, and the W3C RDF Working Group decided that deciding on the relationship between the social and formal meaning of RDF was beyond the scope of the RDF Working Group to decide, so the RDF Working Group appealed to the W3C TAG for a decision. As TAG member Connolly noticed, they "didn't see a way to specify how this works for RDF without specifying how it works for the rest of the Web at the same time" (Berners-Lee, 2003b). In particular, Berners-Lee then put forward his own viewpoint that "a single meaning is given to each URI," which is summarized by the slogan that a URI "identifies one thing." (2003c). In response, Hayes said that "it is simply untenable to claim that all names identify one thing" (2003a). Furthermore, he goes on to state that this is one of the basic results of the knowledge representation community and 20th century linguistic semantics, and so that the W3C cannot by fiat render the judgment that a URI identifies one thing. Berners-Lee rejects Hayes's claim that the Semantic Web must somehow build upon the results of logic and natural language, instead claiming that "this system is different from natural language: we designed it such that each URI identifies one and only one concrete thing in the real world or one and only one globally shared concept" (2003a). At this point, in exasperation, Hayes retorted that "I'm not saying that the 'unique identification' condition is an unattainable ideal: I'm saying that it doesn't make sense, that it isn't true, and that it could not possibly be true. I'm saying that it is *crazy*" (2003b). While Hayes did not explain his own position fully, as he was the editor of the formal semantics of RDF and had the support of other logicians in the RDF Working Group, the issue deadlocked and the RDF Working Group was unable to come to a consensus. In order to move RDF from a Working Draft to a Recommendation, the W3C RDF Working Group removed all references to social meaning from the RDF documents.

One should be worried when two prominent researchers such as Berners-Lee and Hayes have such a titanic disagreement, where no sort of consensus agreement seems forthcoming. Yet who is right? Berners-Lee's viewpoint seems intuitive and easy to understand. However, from the standpoint of the formal semantics of logic, the argument would seem to have been won by Hayes. Still, there is reason to pause

to consider the possibility that Berners-Lee is correct. First, while his notion may seem counter to 'common-sense' within formal logic, it should be remembered that as far as practical results are concerned, the project of logic-based modelling of common-sense knowledge in classical artificial intelligence inaugurated by Hayes earlier is commonly viewed to be a failure by current researchers in AI and cognitive science (Wheeler, 2005). In contrast, despite the earlier and eerily similar argument that Berners-Lee had with original hypertext academic researchers about broken links and with the IETF about the impossibility of a single naming scheme for the entire Internet, the Web is without a doubt an unparalleled success. While in general the intuitions of Berners-Lee may seem to be wrong according to academia, history has proven him right in the past. Therefore, one should take his pronouncements seriously.

The Identity Crisis is not just a conflict between merely two differing individual opinions, but a conflict between two entire disciplines: the nascent discipline of 'Web Science' as given by the principles of Web architecture, and that of knowledge representation in AI and logic (Berners-Lee et al, 2006). Berners-Lee's background is in the Internet standardization bodies like the IETF, and it is primarily his intuitions behind Web architecture. Hayes, whose background in logic jumpstarted the field of knowledge representation in artificial intelligence, should be taken equally seriously. If two entire fields, who have joined common-cause in the Semantic Web, are at odds, then trouble at the level of *theory* is afoot.

Troubles at levels of theory invariably cause trouble in practice. So this disagreement would not be nearly as worrisome were not the Semantic Web itself not in such a state of perpetual disrepair, making it practically unusable. In a manner disturbingly similar to classical artificial intelligence, the Semantic Web is always thought of as soon-to-be arriving, the 'next' big thing, but its actual uses are few and far between. The reason given by Semantic Web advocates is that the Semantic Web is suffering from simple engineering problems, such as a lack of some new standard, some easily-accessible list of vocabularies, or a dearth of Semantic Web-enabled programs. Given that the Semantic Web has not yet experienced the dizzying growth of the original hypertext Web, even after an even longer period of gestation, points to the fact that something is fundamentally awry. The root of the problem is the dependence of the Semantic Web on using URIs as names for things non-accessible from the Web.

Far from being a mandarin metaphysical pursuit, this problem is the very first practical issue one encounters as soon as one wants to actually use the Semantic Web. If an agent receives a graph in RDF, then the agent should be able to determine an interpretation. The inference procedure itself may help this problem, but it may instead make it worse, simply producing more uninterpretable RDF statements. The agent could employ the follow-your-nose algorithm, but what information, if any, should be accessible at these Semantic Web-enabled URIs? If a user wants to add some information to the Semantic Web, how many URIs should they create? One for the representation, and another for the referent the representation is *about*? Should the same URI for the Eiffel Tower itself be the one that is used to access a web-page about the Eiffel Tower?

URIs on the Semantic Web can be thought of as analogous to natural language *names*, as names in natural language can be used to refer as well. Therefore, what needs to be done is to distinguish within analytic philosophy the various theories on naming and reference in general, and then see how these various theories either do or do not apply to the Semantic Web. What is remarkable is that the position of Hayes, the logicist position, corresponds to a well-known theory of meaning and reference, the 'descriptivist theory of reference' attributed to early Wittgenstein, Carnap, Russell, and turned into its pure logical form by Tarski (Luntley, 1999). However, it is common currency in philosophical circles that the descriptivist theory of reference was overthrown by the 'causal theory of reference' championed by Kripke and extended by Putnam (Luntley, 1999). It is precisely this causal theory of reference that Berners-Lee justifies in his direct reference position. Thus, the curious coincidence is that both opposing positions on the Semantic Web correspond to *equally* opposing positions in philosophy. Understanding these positions belongs primarily to the domain of philosophy, even if Hayes and especially Berners-Lee do not articulate their positions with the relevant academic citations. In this manner, the precise domain of philosophy that the Identity Crisis falls under is the philosophy of language. The purpose of the rest of this chapter is then the full explication of these two theories of reference in philosophy of language, and then to inspect their practical success (or lack thereof) in the context of the Semantic Web, while at the end offering a critique of both, paving the way for a third theory of meaning.

4.2 Sense and Reference

The original theory of meaning we shall return to is Frege's original controversial theory of sense and reference as given in *Sinn und Bedeutung* (Frege, 1892).[1] This theory is no longer particularly popular, although it has had some revival with an odd dualist variation under the 'two-dimensionalism' of Chalmers Chalmers (2006),[2] and this is likely because Frege himself was quite cryptic with regards to any definition of 'sense.' The key idea lies in Frege's contention that the meaning of any representational term in a language is determined by what Frege calls the "sense" of the sentences that use the term, rather than any direct reference of the term (1892). According to Frege, two sentences could be the same only if they shared the same sense. Take for example the two sentences "Hesperus is the Evening Star" and "Phosphorus is the Morning Star." (Frege, 1892). Since the ancient Greeks did not

[1] The ambiguous translation of this work from original German has been a source of great philosophical confusion. While the word 'Sinn' has almost always been translated into 'sense,' the word 'Bedeutung' has been translated into *either* 'reference' or 'meaning,' depending on the translator. While 'Bedeutung' is most usually translated into the fuzzy English word 'meaning' by most German speakers, the *use* to which Frege puts it is much more in line with how the word 'reference' is used in philosophy. So in the tradition of Michael Dummett, we will translate Frege's 'Bedeutung" into 'reference' Dummett (1973).

[2] Likely Frege himself would not be considered a dualist, but a monist with objective meaning given in the world.

know that 'The Morning Star is the same as the Evening Star,' they did not know that the names 'Hesperus' and 'Phosphorus' share the same referent when they baptized the same star, the planet Venus, with two different names (Frege, 1892). Therefore, Frege says that these two sentences have distinct 'senses' even if they share the same referent. Frege pointed out that, far from being meaningless, statements of identity that would be mere tautologies from the point of view of a theory of reference are actually meaningful if one realizes different terms can have distinct senses. One can understand a statement like 'The Morning Star is the Evening Star' without knowing that both refer to Venus, and one may only know that the 'Morning Star' refers to Venus and by learning the 'Morning Star' and the 'Evening Star' are not distinct senses but a single sense, one can do actual *meaningful cognitive work* by putting these two senses together. While the idea of a notion of 'sense' seems intuitive from the example, it is famously hard to define, even informally. Frege defines 'sense' in terms of the mysterious *mode of presentation*, for "to think of then being connected with a sign (name, combination of words, letters), besides that to which the sign refers, which may be called the reference of the sign, also what I should like to call the sense of the sign, wherein the mode of presentation is contained" (1892). This statement has caused multiple decades of debate by philosophers of language like Russell and Kripke who have attempted to banish the notion of sense and simply build a theory of meaning from the concept of reference.

Regardless of what precisely 'sense' is, Frege believed that the notion of sense is what allows an agent to understand sentences that may not have a referent, for "the words 'the celestial body most distant from Earth' has a sense, but it is very doubtful there is also a thing they refer to...in grasping a sense, one certainly is not assured of referring to anything" (Frege, 1892). So it is the concept of sense that should be given a priority over reference. This is not to deny the role of reference whatsoever, since "to say that reference is not an ingredient in meaning is not to deny that reference is a consequence of meaning ... it is only to say that understanding which speaker of a language has a word in that language ... can never consist merely in his associating a certain thing with it as its referent; there must be some particular *means* by which this association is effected, the knowledge of which constitutes his grasp of its sense" (Dummett, 1973).

Sense is in no way an 'encoded' referent, since the referent is distal from the sense. Instead, the sense of a sentence should naturally lead an agent to correctly guess the referents of the representational sentence. Yet how could this be detected? Again, sense is sense strictly 'in the head' with no effect on behaviour. As put by Wittgenstein, "When I think in language, there aren't 'meanings' going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought" (Wittgenstein, 1953). Sense is the bedrock upon which meaning is constructed, and must be encoded in a language. In fact, according to Frege, sense can only be determined from a sentence in a language, and the sense of a sentence almost always requires an understanding of a whole network of other sentences in a given discourse. Furthermore, without determining which sense of a number of possible senses a sentence *may* have the sentences *does* have, one cannot meaningfully

act, even if the sense used by the agent is incorrect according to the creator of the sentence's purpose.

So, how can sense be determined, or at least detected? After all, almost *anything* counts as meaningful behaviour. While sense determination is a difficult and context-ridden question that seems to require some full or at least 'molecular' language understanding, one account of sense detection so far is given by the earlier notion of assertoric content of Dummett, which is simply that an agent can be thought of as interpreting to a sense if they can answer a number of "yes-no" binary questions about the sense in a way that makes 'sense' to other agents speaking the language (Dummett, 1973). There is a tantalizing connection of Dummett's assertoric content as answers to binary questions to the information-theoretic reduction of uncertainty through binary choices (bits), as the content of information cannot be derived without enough bits in the encoding. Overall, Dummett's notion of sense as grounded in actual language use naturally leads to another question: Is sense objective?

The reason the notion of sense was thought of as so objectionable by many philosophers like Russell and Kripke was that sense was viewed as a private, individual notion, much like the Lockean notion of an *idea*. Frege himself clearly rejects this, strictly separating the notion of a sense from an individual subjective idea of a referent, which he refers to as an 'idea.' Far from a mere subjective idea or impression of a referent, Frege believed that sense was inherently *objective*, "the reference of a proper name is the object itself which we designate by using it; the idea which we have in that case is wholly subjective, in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself" (1892). A sense is objective insofar as it is a shared part of an inherently public language, since a sense is the "common property of many people, and so is not a part of a mode of the individual mind. For one can hardly deny that mankind has a common store of thoughts which is transmitted from one generation to another" (1892). While the exact nature of a sense is still unclear, its main characteristic is that it should be whatever is *objectively shared* between the competent in the use of names in a language.

It is precisely this notion that sense - and therefore meaning as whole - is 'objective' that is crucial for our project of reconstructing meaning on the Web. The Fregean notion of sense is *identical* with our reconstructed notion of informational *content*. These terms should be viewed as identical. The content of information is precisely what is shared between the source and the receiver as a result of the conveyance of a particular message. By definition, this holding of content in common which is the result of the transmission of an information-bearing message *must* by definition involve at least *two* things: a source and a receiver. Furthermore, if the source and receiver are considered to be human agents capable of speaking natural language, then by the act of sharing sentences, which are just encodings shared over written letters or acoustic waves in natural language, the two speakers of language are sharing the content of those sentences. Since the content is possessed by two people, and is by definition of information the *same* content, insofar as *subjective* is defined to be that which is only possessed by a single agent and *objective* is defined

to be that which is possessed by more than one agent (although not necessarily all agents), then *content is objective*.

Most of the productive concepts from Web architecture and philosophy map into the notion of a Fregean sense rather easily. Sentences and terms natural in a language have both a syntactic encoding and a semantic content or sense, that can multiply realized over differing mediums. A sentence is a fully-fledged information-carrying message, that can have multiple realizations in the form of different utterances at different points in space and time. The Gricean notion of a speaker's intentions then maps to the meaningful behavior a sentence is supposed to engender Grice (1957a). The problem of word senses is now revealed to be much larger than previously supposed, as it now stretches across to all sorts of non-natural languages. Everything from messages in computer protocols (formal languages) to paintings (iconic languages) are now just encodings of information, and these too have senses and possible sense ambiguities.

Representations are not just then 'in the head' but also present as an objective component of sentences as the *sense* of *names*. In particular, a name in natural language is no more than some encoding that has as its interpretation the sense of a (possibly and usually distal!) referent. The class of *proper names*, long a source of interest, is just a representation in natural language whose referent is an entity, such that the name 'TimBL' refers to the person Tim Berners-Lee, while the larger class of names such as 'towers' or 'integers' can refer to groups of entities and concepts. There may be some objection that a mere *name* in a sentence is a full-blooded representation. However, unlike some theories of representation such as those put forward by Cummins, we do not require that there be some "isomorphism" or other structured relationship between the representation and its referent (1996), we require the much less-demanding causal relationship with some impact upon the sense (content) and thus the meaningful behaviour of the agent. While it is obvious there is nothing inherent in the term 'Eiffel Tower' that leads the letters or phonemes in the name to correspond in any significant structural way with the Eiffel Tower itself, as long as the sense of the name is dependent on *there being a referent* that the name 'stands-in' for, so a name like the 'Eiffel Tower' is still a representation of the Eiffel Tower itself. The referent itself or some 'image' thereof does not have to be bundled along and carried with the sentence in any meaningful way, as our previous work on the representational cycle shows that is primarily an historical chain with causal efficacy that is the role of the referent.

However, since Frege's time find this notion of meaning as an objective sense has been considered counter-intuitive and controversial, and so with a few exceptions most philosophers of language would throw the notion of sense out the window by grounding theories of meaning in subjective impressions of 'sense-data.' Furthermore, unlike Fregean sense, these theories of semantics have actually been debated in the context of the Web. So before buying into a Fregean notion of sense on the Web, let's see how they fare in their encounter with the Web.

4.3 The Logicist Position and the Descriptivist Theory of Reference

The origin of the logicist semantics is in what is popularly known as 'the descriptivist theory of reference'. In this theory of reference, the referent of a name is given by whatever satisfies the descriptions associated with the name. Usually, the descriptions are thought to be logical statements, so a name is actually a disguised logical description. The referent of the name is then equivalent to the set of possible things, given normally by a mathematical model, such that all statements containing the name are satisfied. To explain a bit further, ***formal languages*** are languages with an explicitly defined syntax at least, and also possibly (although not always) a model-theoretic semantics. The purpose of these formal languages can be interpretation by computers. Many computer languages not considered to be programming languages are languages insofar as they have some normative or even informal interpretation, such as HTML. Furthermore, due to some biases against computer languages being put on the same footing as natural language, sometimes the term *format* is a used as synonym for computer-based language.

As mentioned earlier, an act of interpretation is usually thought of as a mapping from some sentences in a language to the content of some state-of-affairs in a world. This world is often thought to be the everyday world of concrete trees, houses, and landscapes that humans inhabit. Informally an interpretation can be considered to be a mapping from sentences to the physical world itself, a mapping rather ironically and appropriately labelled 'God Forthcoming' (Halpin, 2004). However, often we do not have access to the world itself and it is unclear if a simplistic definition such as "the truth of a sentence consists in its agreement with (or correspondence to) reality" makes any sense, for "all these formulations can lead to various misunderstandings, for none of them is sufficiently precise and clear" Tarski (1944). In an attempt to define a formal notion of truth, Tarksi defined the interpretation of a language, which he terms the "object" language, in terms of a "meta-language" (1944). If both the language and the meta-language are suitably formalized, the interpretation of the language can then be expressed in terms of a satisfaction of a mathematical model, where ***satisfaction*** can be defined as *an interpretation to a mathematical model that defines whether or not every sentence in the language can be interpreted to content*, which in the tradition of Frege is usually thought of as a 'truth' value (i.e. the content is simply the value 'true.'). In this way, formal semantics is distinguished from the jungle of informal semantics by having a precisely defined mathematical model 'stand-in' for the vague and fuzzy world or some portion thereof. While Tarksi originally applied this only to suitably formal languages, others such as Montague have tried to apply this approach, with varying degrees of success and failure, to natural language. To summarize, ***model-theoretic semantics*** is a semantics where *an interpretation of a language's sentences is to a mathematical model*. The ***model*** is *a mathematical structure, possibly a representation of the world or the language itself*. The relationship is summarized below in Figure 4.1, where the relationship between the model and the world is thought to be distal (such

that the model *represents* the world). This is not always the case, as in the model can be thought of as ranging over the world itself.

The adequacy of models is usually judged by whether or not they fulfill the purposes to which the language is designed, or whether or not their behaviour adequately serves as a model of some portion of the world. Given a model-theoretic semantics, an interpretation can be given as "a minimal formal description of those aspects of a world which is just sufficient to establish the truth or falsity of any expression" in the language (Hayes, 2004). While again the history and debate over these terms is outside the scope of this thesis, in general the original notion, as pioneered by Carnap (1947), is that a certain *kind of thing may only be described*, and so given an ***intension***, while the *things that satisfy this description* (which may be more than one thing) are ***extensions***. Sentences are ***consistent*** *if they can be satisfied*, it is ***inconsistent*** if otherwise. Lastly, note that an ***entailment*** is *where an interpretation of one sentence to some content always satisfies the interpretation of another sentence to some content*, i.e. the first statement entails the second. In contrast, an ***inference*** is a *syntactic relationship where one sentence can be used to construct another sentence in a language*. In detail, as shown in Figure 4.1, the syntactic inference mechanisms over time produce more valid inferences, and because these inferences 'line up' with entailments, they also may accurately describe the world outside the formal system. Ideally, this model also 'lines-up' with the world, so the inferences give one more correct statements about the world. Models can be captured in various ways, of which we have primarily described a denotational semantics, but often an axiomatic and operational semantics are equally powerful formalisms. Inference can usually be accomplished by some local inference procedure, like a computer program. The inference procedure of a language is ***sound*** *if every inferred sentence can be satisfied* (i.e. the inference mechanism preserves 'truth'), and it is ***complete*** *if every satisfied sentence can be shown to be entailed* (i.e. all 'true' statements can be proven). This is necessarily a quick overview of the large field of formal semantics, but the general notions are illustrated in Figure 4.1 as the parallel between the causal relationships of the syntactic sentences and their interpretations to a model that *semantically* refers to the world.

4.3.1 Logical Atomism

Obviously, the use of some kind of formal logic to determine what could satisfy a name was appealing, as it appeared that semantics could possibly become a science on the same footing as, say, physics. The roots of the descriptivist theory of reference lay with the confluence of philosophers inspired by this vision who are known as *logical positivists* and *logical atomists*, whose most systematic proponents were Rudolf Carnap and Bertrand Russell respectively. Although logical positivism is a vast school of thought that has proven tremendously influential, even in its current discredited state, for our purposes we will only concern ourselves with one particular doctrine common to both logical positivism and logical atomism, the problem of

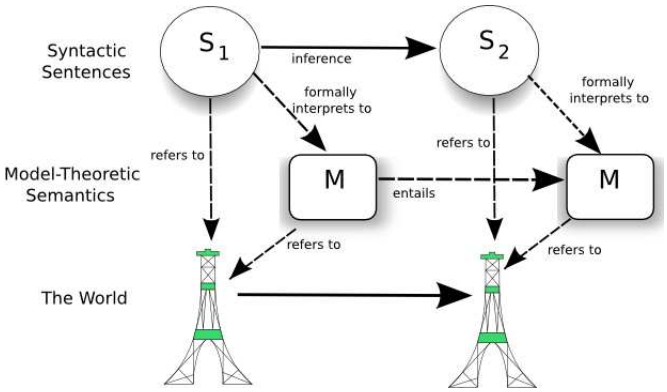

Fig. 4.1 Models, Entailment, and Inference

how natural language terms relate to the logic descriptions, and logical descriptions to the world. The difference between the two schools is mainly one of focus, for the logical positivists hoped to rid the world of metaphysical and epistemological statements through the use of logic and empiricism, while logical atomists thought that the basics metaphysics and even our epistemology should be phrased in terms of logic over elementary sense-data.

The logical positivists and Bertrand Russell were inspired by Wittgenstein's early philosophical work in the *Tractatus Logico-Philosophicus*. In it, Wittgenstein strongly argues for *logical atomism*, that *logic* is the true language of the world; "logic is not a body of doctrine, but a mirror image of the world" for "the facts in logical space" are the world (1921). So logical statements are "laid against reality like a measure" (1921). This is possible because the world is metaphysically determinate at its base, being composed of "simple" and "unalterable" objects that "make up the substance of the world" so that "the configuration of objects produces states of affairs" where "the totality of existing states of affairs is the world" (Wittgenstein, 1921). In other words, there is no – as Brian Cantwell Smith would put it – "flex" or "slop" in this picture, no underlying "metaphysical flux" that somehow resists easily being constrained into these fully determinate "objects" (1996). Although the nature of the world consists of *true* logical facts, humans, since they "picture facts" to themselves, can nonetheless make *false* logical statements, since these pictures merely "model reality" (Wittgenstein, 1921). Contrary to his own logical atomist teacher Russell, Wittgenstein thought that the primary job of the logician is then to state true facts, and "what we cannot speak about" in the form of true logical statements "we must pass over in silence," a phrase he believed was consistently misinterpreted by logical positivism (Wittgenstein, 1921). Note that unlike the more mature standpoint of Hayes, the early logical atomism of Wittgenstein allowed *logical statements* to directly refer to single things in the world, i.e. young Wittgenstein and the logical positivists reified *the formal model* to *be the world* itself.

Carnap's ultimate goal was to use this logical empiricism to render any scientific hypothesis either verifiable by sense experience or not. According to Carnap, in his *The Logical Structure of the World*, all statements (at least, "scientific" statements with "cognitive content") can be reduced to logical statements, where the content of this logical language is given by sensory experiences (1928). These "elementary experiences" (called *eigenpsychische* by Carnap) cannot be directly described, as they are irreducible, but only described by a network of logical predicates that treat these experiences as logical constants (Carnap, 1928). For examples of these kinds of sentences, one would not say "The Eiffel Tower is made of reddish iron." One would say something more elementary like 'hard thing here now' or 'redness here now' when bumping one's toe against the brute fact of the Eiffel Tower. Then these sense-data - which were considered *a priori* true due to their verification by sense experience - could be built up into larger complex sentences and names via logic. Since natural language is part of the world, the structure of language too must be logical, and range over these elementary sense experiences. In this regard, names are given to their referents by concordance with a logical structure ranging over these elementary sensory experiences. Carnap's project was similar in spirit to Chomsky's syntactic theory of language, but focused on semantics rather than syntax: Carnap hoped to develop a semantic and logical definition of meaning that would validate only sentences with 'meaning' and dispose of all metaphysical notions, which would naturally include likely most of Hegel and perhaps even Fregean sense.

Bertrand Russell begins the logical atomist investigation of the connection between logic and names in language is his landmark investigation *On Denoting* with a deceptively simple question: "is the King of France bald?" (Russell, 1905). To what referent does the description "the King of France" refer to? (Russell, 1905) Since in Russell's time there was no King of France, it could not refer to anything like what Carnap later called "elementary sense data" (Carnap, 1928). In this regard, Russell makes a crucial distinction. According to Russell, elementary sensory experiences are known through *acquaintance*, in which we have some sort of direct 'presentation of' the thing (Russell, 1905). According to Russell, these statements of acquaintance with directly present sensory data employ what are known as Russellian demonstratives (such as 'this' or 'that') as exemplified by the statement "That is the Eiffel Tower." Yet knowledge of a thing can be based on *description*, which are those "things we only reach by means of denoting phrases" (Russell, 1905). Russell believed that "all thinking has to start from acquaintance, but it succeeds in thinking *about* many things with which we have no acquaintance" via the use of description (Russell, 1905). Russell was most interested in whether those things with which we have direct acquaintance can be considered true or false, or whether a more mysterious third category such as 'nonsense' is needed. Russell opts to reject creating imaginary but true 'things' as well as any third category, but instead holds that statements such as "the King of France is bald" are false, since "it is false that there is an entity which is now the King of France and is bald" (Russell, 1905). This solution then raises the alarming possibility that "the King of France is not bald" may also come out false, which would seem to violate the Law of the Excluded Middle. So, Russell counters this move by introducing the fact that "the King of France is bald"

is actually a complex logical statement involving scope and quantification, namely $(\exists x.F(x) \wedge G(x)) \wedge (\forall y.F(y) \rightarrow x = y)$, where F is "being the King of France" and G is "being bald" (Russell, 1905). According to the analysis, 'The King of France' is merely a *disguised* complex logical statement. Furthermore, this treatment can be extended to proper names such as 'Sir Walter Scott,' who can be identified with 'the author of Waverly,' so that instead of being a tautology, even a proper name of a person, even if known through acquaintance, is sort of short-hand for a large cluster of logical statements. To use our previous example, the 'Eiffel Tower' can be thought of as a short-hand for not only that 'there exists an entity known as the Eiffel Tower' but also the logical statement was 'the aforementioned entity had Gustave Eiffel as its architect.' If someone did not know that 'the aforementioned entity was also the tallest building in the world up until 1930,' one could then make a statement such as 'The Eiffel Tower is identical to the tallest building in the world up until 1930' without merely stating a tautology, and such a statement would add true and consistent knowledge to a hearer who was not aware of the statement.

As sensible as Russell's programme appeared, there are difficulties in building any theory of reference on, as Quine put it, such a "slender basis" as elementary sense data and logic (1951). One obvious problem for any descriptive theory of names comes for the use of names of any "kind of abstract entities like properties, classes, relations, numbers, propositions," for such entities could not have an interpretation for any content using such a simple sensory epistemology (Carnap, 1950). Carnap's *Empiricism, Semantics, and Ontology* made an argument for basing such entities purely on linguistic form itself. Carnap believed that, despite the difficulty of determining the interpretation of names for abstract entities, "such a language does not imply embracing a Platonic ontology but is perfectly compatible with empiricism" (1950). His position was that while "if someone wishes to speak in his language about a new kind of entity, he has to introduce a system of new ways of speaking, subject to new rules," which Carnap calls the "construction of a linguistic framework for the new entities in question." From *within* a linguistic framework, Carnap believed to commit to any statement about the "existence or reality of the total system of the new entities" was to make a "pseudo-statement without cognitive content" (1950). This particular position of Carnap's was eventually devastated, as Quine showed that even the most unremarkable of sensory expressions such as 'redness here now' were undermined by multiple problems. For example, there is the issue of indeterminacy of translation, in which even the verbal expression of sense experiences assumes a common background, but one can imagine many cases where two creatures would utter "redness here now" in reaction to actually different sensory stimuli (imagine a human with color-blindness). Also, there is the problem where even our sense experiences are not 'raw' but influenced by a complex holistic network of propositions - one does not experience 'hard iron here now' but the Eiffel Tower itself (Quine, 1951).

4.3.2 Tarski's Formal Semantics

Tarski abandoned the quaint epistemology of logical atomism in terms of direct acquaintance with sensory data and defined reference purely in terms of mathematical logic in his *The Concept of Truth in Formal Languages* (Tarski, 1935). Reference was just defined as a consequence of the truth *only* in terms of satisfaction of a formal language (Tarski, 1935). To set up his exposition, Tarski defines two languages, the first being the syntactic *object language L* and the second being the meta-language *M*. The *meta-language* should be more expressive such that it can describe every sentence in the object language, and furthermore, that it contain axioms that allow the truth of every sentence in the object language to be defined. In his first move, Tarski defines the formal conception of truth as 'Convention T,' namely that for a given sentence *s* in *L*, there is a statement *p* in *M* that is a theorem defining the truth of *s*, that is, the truth of *s* is determined via a translation of *s* into *M* (Tarski, 1935). Tarski then later shows that truth can be formally defined as "*s* is true if and only if *p*" (Tarski, 1944). For example, if the object language is exemplified by a sentence uttered by some speaker of English and the meta-language was an English description of the real world; 'The Eiffel Tower is in Paris' is true if and only if the Eiffel Tower is in Paris. The sentence 'The Eiffel Tower is in Paris' must be satisfied by the Eiffel Tower *actually being* in Paris. While this would at first seem circular, its non-circularity is better seen through when the object language is not English, but another language such as German. In this case, "'Der Eiffelturm ist in Paris' is true if and only if the Eiffel Tower is in Paris." However, Tarksi was not interested in informal languages such as English, but in determining the meaning of a new formal language via translations to mathematical models or other formal languages with well-known models. If one was defining a formal semantics for some fragment of a knowledge representation language like RDF, a statement such as `http://www.eiffeltower.example.org ex:location ex:Paris` is true if and only if $\exists ab.R(a,b)$ where *R*, *a*, and *b* are given in first-order predicate logic.

If one is defining a formal Tarski-style semantics for a language, what should one do when one encounters complex statements, such as 'the Eiffel Tower is in Paris and had as an architect Gustave Eiffel'? The answer is at the heart of Tarksi's project, the second component of Tarski's formal semantics is to use the principle of compositionality so that any complex sentence can have its truth conditions derived from the truth conditions of its constituents. To do this, the meta-language has to have finitely many axioms, and each of the truth-defining theorems produced by the meta-language have to be generated from the axioms (Tarski, 1935). So, the aforementioned complex sentence is true if and only if $\exists ab.R(a,b) \wedge Q(a,c)$, where *Q* can be the *architect of* relationship, *c* can be Gustave Eiffel and *a* the Eiffel Tower. Tarksi's theory as explained so far only deals with 'closed' sentences, i.e. sentences containing no variables or quantification. The third, and final component of Tarski's formal semantics is to use the notion of satisfaction via extension to define truth (Tarski, 1935). For a sentence such as 'all monuments have a location,' we can translate the sentence to $\forall a,l.monument(a) \rightarrow hasLocation(a,l)$ which is true if and only if there is an extension *x* from the world that satisfies the logical statements

made about a. In particular, Tarksi has as his preferred extensions infinite ordered pairs, where the ordered set could be anything (Tarski, 1935). For formal languages, a model-theoretic semantics with a model composed by set theory was standard. For example, the ordered pairs in some model of $(EiffelTower, Paris)$ would satisfy, as would $(ScottMonument, Edinburgh)$ but not $(Paris, EiffelTower)$. However, there is no reason why these models could not be "God Forthcoming," things in the the real world itself, albeit given in set-theoretic terms (Smith, 1995). To summarize Tarksi's remarkably successful programme, model-theoretic semantics can produce a theory of truth that defines the semantics of a sentence in terms of the use of a translation of the sentence into some formal language with a finite number of axioms, then using compositionality to define the truth of complex sentences in terms of basic sentences, and finally determining the truth of those basic sentences in terms of what things in a model satisfy the extensions of the basic sentences as given by the axioms. This work marks the high-point of the logical programme, as all questions of meaning are reduced to questions about giving the interpretation of a sentence in terms of a formal notion of truth. This notion of truth is not restricted by the logical atomist's epistemology of elementary sense data, but instead can range over any possible formal language and any possible world. This victory is not without its costs, since while Tarski provides the best account of the relationship between logical descriptions and the world by simply removing all questions that cannot be phrased formally, formal semantics by itself leaves unsolved the fundamental question about how natural language relates to our experience of the world. Ignoring a problem does not make it go away. So when confronted with this vexing problem, champions of formal semantics often revert to the Russellian doctrine of direct acquaintance, thereby returning to the original problems that caused Tarski to abandon epistemology.

4.3.3 Logical Descriptions Unbound on the Web

While the descriptivist theory of reference seems distant from the Identity Crisis of the Web, it is in fact central to the position of Hayes and the Semantic Web as a whole. This is primarily because Hayes's background was in formal logic, with his particular specialty being the creation of Tarski-style semantics for knowledge representation languages. What Hayes calls the "basic results in 20th century linguistic semantics" that Berners-Lee's dictum that "URIs identify one thing" violates is the interpretation of URIs in a Tarski-style formal semantics (Hayes, 2003a). For the logicist position, the *semantics* in the Semantic Web derive from the Tarski-style formal semantics Hayes created for the Semantic Web (Hayes, 2004).

Before delving into the formal semantics of RDF, it should be noticed that these semantics are done by extension, like most other formal languages Hayes (2004). However, the semantics of RDF are purposefully quite weak as not to allow arithmetic or constructs like the negation of a class, and so RDF avoids logical paradoxes like the encoding of Gödel sentences. Yet in order to make RDF triples as

flexible as possible, RDF includes features normally associated with higher-order logic such as "a property may be applied to itself" and classes "may contain themselves" (Hayes, 2004). This is handled semantically by having first an interpretation map the URI to an individual. Then unlike standard first-order logic, this individual then maps to different extensions depending on the role the URI is playing as a property or class in the triple. A simple example should suffice to give a flavour of the formal semantics, where a relation is just another kind of individual. What is the formal semantics of `ex:EiffelTower ex:architect ex:Gustave_Eiffel`? To simplify slightly, Hayes defines the formal semantics in terms of set theory, where there is a set of resources that compose the model of the language, a set of properties, and a set of URIs that can refer to resources. The interpretation of any RDF statement is then given as an extensional mapping from the set of properties to the powerset of resources, to the set of pairs of resources. So, given a set-theoretic model consisting of elements (given by italics) *Gustave Eiffel* and *the Eiffel Tower* and *being the architect of*, then `ex:EiffelTower` $\models$ *the Eiffel Tower*, `ex:Gustave_Eiffel` $\models$ *Gustave Eiffel* and `ex:architect` $\models$ *being the architect of*, so that the entire triple maps to a set of pairs: `ex:EiffelTower ex:architect ex:Gustave_Eiffel` $\models$ *(..., (the Eiffel Tower, Gustave Eiffel), ...)*. Common-sense human intuitions will likely have this interpretation maps to `ex:EiffelTower ex:architect ex:Gustave_EiffelTower`, and using the axioms defined in the RDF formal semantics, a few new triples can be inferred, such as `ex:architect rdf:type rdf:Property`, i.e. *being an architect of* is a property of something.

However, the inherent pluralism of the Tarski approach to models also means that another equally valid interpretation would be the inverse, i.e. the mapping of `ex:EiffelTower` to *Gustave Eiffel* and `ex:Gustave_Eiffel` to *the Eiffel Tower*. In other words, `ex:architect` $\models$ *being the architect of*, so that the entire triple maps to a set of pairs `ex:EiffelTower ex:architect ex:Gustave_Eiffel` $\models$ *..., (Gustave Eiffel, Eiffel Tower), ...)*. Due to the unconstrained nature of RDF, `ex:architect` has no 'natural' relationship to anything in particular, but could easily be assigned either *the Eiffel Tower* or *Gustave Eiffel* just as easily as *being the architect of*. Furthermore, the model could just as easily be given by something as abstract as the integers *1* and *2*, and an equally valid mapping would be for `ex:EiffelTower` $\models$ *1* and `ex:Gustave_Eiffel` $\models$ *2*, so that `ex:architect` $\models$ *being the architect of*, so that the entire triple maps to a set of pairs `ex:EiffelTower ex:architect ex:Gustave_Eiffel` $\models$ *(..., (1,2), ...)*. Indeed, the extreme pluralism of a Tarski-style semantics shows that, at least if all one has is a single lone triple statement, that triple can be satisfied by any model. This is no mere oddity of formal languages, this would also hold for any lone sentence in a language like English – such as "Gustave Eiffel is the architect of the Eiffel Tower" – as long as one subscribed to a Tarski-style semantics for natural language. As the number of triples increased, the amount of possible things that satisfy the model is thought to decrease, but in such a loose language as RDF, Hayes notes that it is "usually impossible to assert enough in any language to completely constrain the interpretations to a single possible world, so there is no such thing as 'the'

unique interpretation" (Hayes, 2004). This descriptivist theory of reference, where descriptions are logical statements in RDF, is illustrated in Figure 4.2.

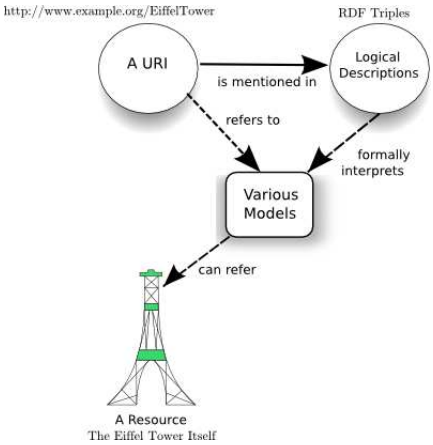

Fig. 4.2 The descriptivist theory of reference for URIs

While Hayes makes no claim that access to some web-pages via URIs is not possible, he claims that such access to Web representations is orthogonal to the question of what a URI could refer to, since "the architecture of the Web determines access, but has no direct influence on reference" (Hayes and Halpin, 2008). Furthermore, Hayes's logical understanding of ambiguity parts path with natural language understandings of ambiguity: Hayes claims that reference to resources is completely *independent* of whatever Web representations can be accessed, even if those contain logical expressions. While much credit should be given to Hayes for creating a logical semantics for RDF, the problem of connecting these descriptions to the world outside of the Web falls outside formal semantics and so opens up a seemingly uncrossable abyss between the logical descriptions and sensory data. One seemingly easy way out of this abyss is to revert to the doctrine of Russellian direct acquaintance, also known as ostentation. In moments, Hayes himself seems to subscribe to the logical atomist epistemology of Russell, as he says that "reference can either be established by either description or ostention" with ostention being defined as the use of a Russellian demonstrative (like 'that' or 'this') identifying a particular "patch of sense data" via a statement such as 'that is the Eiffel Tower' (Hayes, 2006). Since most of the things referred to by names are not accessible, reference can only be determined by description, and these descriptions are inherently ambiguous as regards any sense data (Hayes and Halpin, 2008).

As our example showed, RDF in general says so little inferentially that many different models can satisfy almost any given RDF statement. Therefore, Hayes considers it essential to ditch the vague word 'identify' as used in URIs, and distinguish between the ability of URIs to access and refer. While access is constrained

by Web architecture, according to Hayes, reference is absolutely unconstrained except by formal semantics, and so "the relationship between access and reference is essentially arbitrary" (Hayes and Halpin, 2008). From this philosophical position, the Identity Crisis dissolves into a pseudo-problem, for the same URI can indeed access a web-page and refer to a person unproblematically, as they no longer have to obey the dictum to identify one thing. Hayes compares this situation to that of *overloading*, using a single name to refer to multiple referents, and instead of being a problem, "it is a way of using names efficiently" and not a problem for communication, as "natural language is rife with lexical ambiguity which does not hinder normal communication," as these ambiguities can almost always be resolved by sufficient context (Hayes and Halpin, 2008). Overall, the argument of Hayes against Berners-Lee in the Identity Crisis is the position of keeping the formal semantics of reference separate from the engineering of the Web.

4.4 The Direct Reference Position and the Causal Theory of Reference

The alternative slogan of Berners-Lee, that "URIs identify one thing," may not be completely untenable after all (Berners-Lee, 2003c). It appears to even be intuitive, for when one says 'I went to visit the Eiffel Tower,' one believes one is talking about a very *particular* thing in the *real* world called the 'Eiffel Tower,' not a cluster of descriptions or model of the world. The direct theory of reference of Berners-Lee has a parallel in philosophy, namely Saul Kripke's 'causal theory of reference,' the widely-known argument against the descriptivist theory of reference, and so the reliance upon the purely formal semantics of Hayes (Kripke, 1972). In contrast to the descriptivist theory of reference, where the content of any name is determined by ambiguous interpretation of logical descriptions, in the*causal theory of reference* any name refers via some causal chain directly to a referent (Kripke, 1972).

4.4.1 Kripke's Causal Theory of Proper Names

The causal theory of reference was meant to be an attack on the descriptivist theory of reference attributed to Russell, and its effect in philosophy has been to discredit any neo-Russellian descriptivist semantics for proper names. Unsurprisingly, the causal theory of reference also has its origin in logic, since Kripke as a modal logician felt a theory of reference was needed that could make logical statements about things in different logically possible worlds (Kripke, 1972). However, while Kripke did not directly confront the related position of Tarski, his argument does nonetheless attempt to undermine the ambiguity inherent in Tarski's model-theoretic semantics, although a Tarski-style semantics can merely 'flatten' models of possible

worlds into a singular model. Still, as a response in philosophy of language, it is accepted as a classical refutation of the descriptivist theory of reference.

In Kripke's *Naming and Necessity*, an agent fixes a name to a referent by a process called *baptism*, in which the referent, known through direct acquaintance is associated with a name via some local and causally effective action by the agent (Kripke, 1972). Afterwards, a historical and causal chain between a current user of the name and past users allows the referent of a name to be transmitted unambiguously through time, even in *other possible worlds*. For example, a certain historical personage was given the name 'Gustave Eiffel' via a rather literal baptism, and the name 'Gustave Eiffel' would still refer to that baptized person, even if he had not been the architect of the Eiffel Tower, and so failed to satisfy that definite description. Later, the causal chain of people talking about 'Gustave Eiffel' would identify that very person, even after Gustave Eiffel was dead and gone. Descriptions aren't entirely out of the picture on Kripke's account; they are necessary for disambiguation when the context of use allows more than one interpretation of a name, and they figure in the process by which things actually get their names, if the thing cannot be directly identified. However, this use of descriptions is a mere afterthought with no causal bearing on determining the referent of the name itself, for as Kripke puts it, "let us suppose that we do fix the reference of a name by a description. Even if we do so, we do not then make the name synonymous with the description, but instead we use the name rigidly to refer to the object so named, even in talking about counterfactual situations where the thing named would not satisfy the description in question" (Kripke, 1972). So what is crucial is not satisfying any description, but the act of baptism and the causal transmission of the name.

4.4.2 Putnam's Theory of Natural Kinds

Kripke's examples of the causal theory of reference used proper names, such as 'Cicero' or 'Aristotle,' and he did not extend his analysis to the whole of language in a principled manner. However, Hilary Putnam, in his *The Meaning of 'Meaning,'* extends Kripke's analysis to all sorts of names outside traditional proper names, and in particular Putnam uses for his examples the names of natural kinds (Putnam, 1975). Putnam was motivated by an attempt to defeat what he believes is the false distinction between intension and extension. The set of logical descriptions, which Putnam identifies with a "psychological state," that something must satisfy to be given a name is the *intension*, while those things in a given interpretation that actually satisfy these descriptions, is the *extension* (Putnam, 1975). Putnam notices that while a single extension can have multiple intensions it satisfies, such as the Eiffel Tower both being "in Paris" and "a monument," a single intension is supposed to have the same extension in a given interpretation. If two people are looking for a "monument in Paris," the Eiffel Tower should satisfy them both, even though the Eiffel Tower can also have many other possible descriptions.

Putnam's analysis can be summarized as follows: Imagine that there is a world "very much like Earth" called 'Twin Earth.' On Twin Earth "the liquid called 'water' is not H_20 but a different liquid" whose chemical formula is abbreviated as XYZ, and that this XYZ is "indistinguishable from water at normal temperatures and pressures" since it "tastes like water and quenches thirst like water" (Putnam, 1975). A person from Earth would *incorrectly* identify XYZ for their normal referent of water, as it would satisfy all their descriptions. In this regard, this shows that meanings "ain't in the head" but are in fact determined, not by individual language use or descriptions, but by some indexical relationship to "stuff that is like water around here" normally. That "stuff" *should* get its name and meaning from *experts*, since "probably every adult speaker even knows the necessary and sufficient condition 'water is H_20,' but only a few adult speakers could distinguish water from liquids which superficially resembled water...in case of doubt, other speakers would rely on the judgment of these 'expert' speakers" who would ideally test XYZ and determine that it was indeed, not water" (Putnam, 1975). Indeed, less outlandish examples, such as the difference between "beech trees" and "elm trees" are trotted out by Putnam to show that a large amount of our names for things, perhaps even extending beyond natural kinds, are actually determined by expert knowledge (Putnam, 1975). In this way, Kripke's baptism can extend to almost all languages, and scientists can be considered a special sort of naming authority capable of baptizing all sorts of things with a greater authority than everyone else. As even Putnam explicitly acknowledges "Kripke's doctrine that natural-kind words are rigid designators and our doctrine that they are indexical are but two ways of making the same point" (Putnam, 1975).

4.4.3 Direct Reference on the Web

This causal theory of reference is naturally close to the direct reference position of Berners-Lee, whose background is in expert-created databases. He naturally assumes the causal theory of reference is uncontroversial, for in database schemas, what a term *refers to* is a matter best left to the expert designer of the database. So Kripke and Putnam's account of unambiguous names can then be transposed to the Web with a few minor variations in order to obey Berners-Lee's "crazy" dictum that "URIs identify one thing" regardless of interpretation or even accessible webpage (Berners-Lee, 2003c). While it may be a surprise to find Berners-Lee to be a closet Kripkean, Berners-Lee says as much, "that the Web is not the final arbiter of meaning, because URI ownership is primary, and the look-up system of HTTP is...secondary" (Berners-Lee, 2003c). There is also an element of Grice in the direct theory of reference, for the *intended* interpretation and perhaps even purpose of the owner is the one that really matters to Berners-Lee, not any publicly accessible particular Web representation (Grice, 1957b). However, ultimately Berners-Lee has far more in common with the causal theory of reference, since although the URI owner's intention determines the referent, after the minting of the new URI for the

resource, the intended interpretation is somehow never supposed to vary (Berners-Lee, 1998a).

To apply the causal theory of reference as to URIs, baptism is given by the registration of the domain names, which gives a domain name and legally binding set of IP addresses, such as *example.org*, a legally binding owner. Of course, the natural question then would be if this Kripkean practice can then be extended to entire URIs such as *http://www.example.org/Eiffel*? For most domain names a specific policy given by the owner could set the allowed referents for the creation of URIs that involve the domain name in question, perhaps as embodied in some software system. One could imagine several variations on this theme, from the URIs being controlled indirectly by systems-programmers or even outsourced to the general public in the form of a user-generated URI registry with a single top-level domain. Regardless of the details, the referent of a URI is established by fiat by the owner(s), and then optionally can be communicated to others in a causal chain in the form of publishing web-page accessible from the URI or by creating Semantic Web statements about the URI. This causal theory of reference for URIs is illustrated in Figure 4.3.

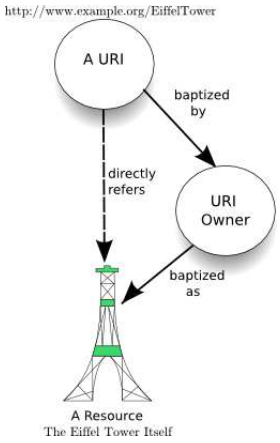

Fig. 4.3 The causal theory of reference for URIs

In this manner, the owner of the URI can thereby determine the referent of the URI and communicate it to others, but ultimately the act of baptism and so the determination of the referent are in the hands of the owner of the URI, the self-professed 'expert' in the new vocabulary term introduced to the Semantic Web by his URI, and the owner has no real responsibility to host any Web representations at the URI. Since the owner can causally establish a name for a non-Web accessible thing via simply minting a new URI without hosting *any* web-page, under the causal theory of reference the Semantic Web can be treated as having a giant translation manual mapping URIs directly to referents, where the URIs refer directly to objects in the world outside of the Web. Realistically, if an agent got a URI like

`http://www.example.org/Gustave_Eiffel` and one wanted to know what the URI referred to, one could use a service such as `whois` to look up the owner of the URI, and then contact the owner of the URI if there was any doubt in the matter. Yet since obviously such URIs cannot access things outside the Web and contacting the owner every time a URI is to be used is absurd, what kinds of web-pages, if any, should this giant Semantic Web dictionary return? If it returns no web-page, how can a user-agent distinguish a URI for a referent outside the Web from that of a URI for a web-page? This question is partially answered by Berners-Lee in a solution called '`303` redirection,' where a distinct URI is given to the thing-in-of-itself, and then when this URI is accessed by an agent such as a web-browser, a particular Web mechanism called the 303 Header redirects to the agent to another URI for a web-page describing the resource, either in RDF or in HTML, or both. However, this mechanism has been considered difficult to use and understand, "analogous to requiring the postman dance a jig when delivering an official letter" (Hayes, 1977b).

4.5 Sense and Reference on the Web

The Semantic Web has still not experienced the tremendous growth of the hypertext Web, and the primary reason appears to be this impasse at the Identity Crisis. For the first few years of its existence (2001-2006), in general the arguments of Hayes prevailed, and the URIs used in RDF graphs did not access any web-pages. However, in this phase of its existence, the Semantic Web did not progress beyond yet another little-used knowledge representation language. In the last few years (2006-2009), the Semantic Web has experienced phenomenal growth under the term 'Linked Data,' as Berners-Lee's position has had more acceptance and users have started deploying RDF using actual URIs. This growth of estimated billions triples, including large-scale projects by biomedical community and in government data in using the Semantic Web, seems to have implicitly validated Berners-Lee's direct reference position. Yet that is far from true; what is apparent from any analysis of the Semantic Web is that there appear to be *too many* URIs for some things, while *no* URIs for other things (Halpin and Lavrenko, 2009). As differing users export their data to the Web in a decentralized manner, new URIs are always minted, and so running the risk of fracturing the Semantic Web into isolated 'semantic' islands instead of becoming a unified web, as the same URIs are not re-used. The critical missing element of the Semantic Web is some mechanism that allows users to come to agreement on URIs and then share and re-use them, a problem ignored both by the logical and direct reference positions on semantics. Given the practical failure of both approaches, one should be suspicious that something is *theoretically* wrong as well.

The philosophical root of the problem may be that both Russell and Kripke - and so both Berners-Lee and Hayes - reject the notion of 'sense.' The Fregean distinction between 'sense' and 'reference' that provoked both Russell and Kripke's intellectual projects to build an entire theory of meaning on top of only reference, where Frege held that the the meaning of any term in a language is determined by the "sense" of

the sentences that use the term, rather than any direct reference of the term (Frege, 1892). It is precisely this notion that sense is ‘objective’ that allows us to construct a new position in the next chapter. Yet how does this notion of sense play out? Dummett provides an insightful hint, “Frege’s thesis that sense is objective is thus implicitly an anticipation of Wittgenstein’s doctrine that meaning is use” (Dummett, 1993). So we must outline a third position, the position of social semantics takes the objective notion of ‘sense’ and Wittgenstein’s analysis of “meaning as use” as its foundation (Wittgenstein, 1953) .

Chapter 5
The Semantics of Tagging

You philosophers ask questions without answers, questions that have to remain unanswered to deserve being called philosophical. According to you, answered questions are only technical matters. That's what they were to begin with. **Jean Lyotard** (1988)

5.1 Making Sense of Tagging

During the last decade the Web has become a space where increasing numbers of users create, share and store content, leading it to be viewed not only as an "information space" Berners-Lee (1996b) but also a "social space" **?**. This new step in the evolution of the Web, often referred to as the "Web 2.0," was shaped by the arrival of the different services that came into existence to support users to easily publish content on the Web, such as photos (Flickr), bookmarks (del.icio.us), movies (YouTube), blogging (Wordpress), and others allow users to *tag* URIs with keywords to facilitate retrieval both for the acting user and for other users. **?**. Almost simultaneously with the growth of user-generated content on the Web came a need create order in this fast growing unstructured data. Tagging refers to the labeling of resources by means of free-form descriptive natural language keywords, and tagging has become the predominant method for organizing, searching and browsing online web-pages, as well as any other resource. Sets of categories derived based on the tags used to characterize some resource are commonly referred to as folksonomies. This approach to organizing online information is usually contrasted with the fomral ontolgoies used by the Semantic Web, as in collaborative tagging systems users themselves annotate resources by tags they freely chose and thus forms a 'flat space of names' without the predefined and hierarchical structure characteristic of the Semantic Web ontologies.

As shown earlier, the Semantic Web has so far been attached to classical theories of semantics that are based on a rejection of the notion of an objective Fregean 'sense' in favor of an approach based purely on reference. The usual critique of

Fregean sense has been that the notion of some objective yet common notion of sense in at least cryptic and even anti-scientific. Yet with the development of collaborative tagging systems, it seems we at long last have an organic notion of a Fregean sense developing that is both objective and common in the form of tagging. In tagging, for each URI a number of users attach tags to a particular URIs, and this common set of tags can be considered the Fregean sense of the URI **?**. While there are some difficulties with this viewpoint, namely that collaborative tagging systems usually conflate a URI with whatever web representations are accessible by that URI (and thus violate the Semantic Web dictum to separate representations from resources and their URIs), such conflation does not at all disqualify tagging as a candidate for a computational theory of sense. First, one can imagine that tagging could be applied to the associated descriptions of Semantic Web URIs, and that these tags would then directly apply to the non-information resource of that URI. To strike deeper, one could also hold that the entire division between Semantic Web URIs and URIs for ordinary hypertext web-pages is fundamentally misbegotten, with 303 redirection being akin to @@DANCING QUOTE. However, it should be also noted that while the Semantic Web has yet to reach widespread usage, collaborative tagging systems are now part and parcel of most major web-sites, and their use seems to be increasing rather than decreasing.

There are concrete benefits to the tagging approach compared to the Semantic Web's traditional focus on formal ontologies. The flexibility of tagging systems is thought to be an asset; tagging is a post-hoc categorization process, in contrast to a pre-optimized classification process such as expert-generated taxonomies. In defining this distinction, Jacob (2004) believes that "categorization divides the world of experience into groups or categories whose members share some perceptible similarity within a given context. That this context may vary and with it the composition of the category is the very basis for both the flexibility and the power of cognitive categorization." Philosophically, tagging is akin to late Wittgenstein's notion of 'family-resemblance.' **?** Classification, on the other hand "involves the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and non-overlapping classes; it mandates consistent application of these principles within the framework of a prescribed ordering of reality" Jacob (2004), a tradition going back to Aristotle **?**. Other authors argue that tagging enables users to order and share data more efficiently than using classification schemes; the free-association process involved in tagging is cognitively much more simple than are decisions about finding and matching existing categories Butterfield (2004). Additionally, proponents of tagging systems show that users of tagging systems only need to agree on the general meaning of a tag in order to provide shared value instead of the more difficult task of agreeing on a specific, detailed taxonomy Mathes (2004).

Yet, what are the *semantics* of a tagging system? A number of problems stem from organizing information through tagging systems, including ambiguity in the meaning of tags and the use of synonyms which creates informational redundancy. Interestingly, Semantic Web ontologies like *NiceTag* have been developed to address the issues of ambiguity in tagging systems by formalizing the tagging process

itself, often by linking a particular tag to a Semantic Web URI **?** . While this may clarify the intended meaning of the tag, this approach does not thereby in some semi-magical manner give semantics to the tag. Also, it seems the most interesting question for our approach is not what the referent of a particular tag act, but whether or not the *collective* sum of individual tagging acts can serve as an objective notion of sense. Since each tag for a given web resource (such as a web page) is repeated a number of times by different users, for any given tagged resource there is a distribution of tags and their associated frequencies. The collection of all tags and their frequencies ordered by rank frequency for a given resource is the *tag distribution* of that resource, which is our candidate for a Fregean sense.

So then, the important open question concerning the use of collaborative tagging to organize metadata is whether the system becomes *stable* over time. By *stable*, we mean that users have collectively developed some implicit consensus about which tags best describe a site, and these tags do not vary much over time. Only this will allow tags to be used as an adequate computational theory of neo-Fregean sense, since otherwise tagging would be subjective rather than objective. We will assume that these tags that best describe a resource will be those that used most often, and new users mostly reinforce already-present tags with similar frequencies. Since users of a tagging system are not acting under a centralized controlling vocabulary, one might imagine that no coherent categorization schemes would emerge at all from collaborative tagging. In this case, tagging systems, especially those with an open-ended number of non-expert users like del.icio.us, would be inherently unstable such that the tags used and their frequency of use would be in a constant state of flux. If this were the case, identifying coherent, stable structures of collective sense produced by users with respect to a site would be difficult or impossible.

The hope among proponents of collaborative tagging systems is that stable tag distributions, and thus, possibly, stable categorization schemes, might arise from these systems. Again, by *stable* we do not mean that users stop tagging the resource, but instead that users collectively settle on a group of tags that describe the resource well and new users mostly reinforce already-present tags with the same frequency as they are represented in the existing distribution. Online tagging systems have a variety of features that are often associated with complex systems such as a large number of users and a lack of central coordination. These types of systems are known to produce a distribution known as a power law over time. A crucial feature of some power laws - and one that we also exploit in this work - is that they can be produced by scale-free networks. So regardless of how large the system grows, the shape of the distribution remains the same and thus *stable*. Researchers have observed, some casually and some more rigorously, that the distribution of tags applied to particular resources in tagging systems follows a power law distribution where there are a relatively small number of tags that are used with great frequency and a great number of tags that are used infrequently Mathes (2004). If this is the case, tag distributions may provide the stability necessary to draw out useful information structures.

This chapter is organized as follows. In the first part of the paper, we examine how to detect the emergence of stable “consensus” distributions of tags assigned to individual resources. In Section 5.2 we demonstrate a method for empirically ex-

amining whether tagging distributions follows a power law distribution. In Section 5.2.4 we show how this convergence to a power law distribution can be detected over time by using the Kullback-Leibler divergence. We further empirically analyze the trajectory of tagging distributions before they have stabilized, as well as the dynamics of the "long tail" of tag distributions. In the second part of the paper, we examine the applications of these stable power law distributions.In Section **??**, we @@ In Section 5.5 we demonstrate how the most frequent tags in a distribution can be used in inter-tag correlation graphs (or folksonomy graphs) to chart their relation to one another. Section 5.6 shows how these folksonomy graphs can be (automatically) partitioned, using community-based methods, in order to extract shared tag vocabularies. Finally, Section 5.7 provides an independent benchmark to compare our empirical results from collaborative tagging, by solving the same problems using a completely different data set: search engine query logs.

5.1.1 Related Work

Existing research on tagging has explored a wide variety of problems, ranging from fundamental to more practical concerns - and much of this research is not relevant to our task at hand, such as discovering the best interfaces for presenting tags to users Halvey and Keane (2007) our using tags to extract data such as event and place locations from tagged photos Rattenbury et al (2007). In a direction of work that bears directly on the larger question of the semantics of collective tagging systems, Mika (2005) addresses the problem of extracting taxonomic information from tagging systems in the form of Semantic Web ontologies, but fails to address the stability of collective tagging. More of interest is studies on the structure of a tagging network for del.icio.us data which examine network characteristics of the tagging system such as the degree distribution (the distribution of the number of other nodes each node is connected to) and the clustering coefficient (based on a ratio of the total number of edges in a subgraph to the number of all possible edges) of this network. Shen and Wu do indeed find that the a snapshot of an entire tagging network is indeed "scale-free" and has the features Watts and Strogatz (1998) found to be characteristic of small world networks: small average path length and relatively high clustering coefficient.

However, we are more interested in the tags applied to individual URIs. An early line of research that has attempted to formalize and quantify the underlying dynamics of a collaborative tagging systems is Golder and Huberman (2006), which also make use of del.ici.ous data. They show the majority of sites reach their peak popularity, the highest frequency of tagging in a given time period, within ten days of being saved on del.icio.us (67% in their data set), though some sites are "rediscovered" by users (about 17% in their data set), suggesting stability in most sites but some degree of "burstiness" in the dynamics that could lead to cyclical patterns of stability characteristic of chaotic systems. Importantly, Golder and Huberman find that the distribution of tags within a given site stabilizes over time, usually around

one hundred tagging events. They do not, however, examine what type of distribution arises from a stabilized tagging process, nor do they present a method for determining the stability of the distribution which we see as central to understanding the possible utility of tagging systems. Thus, the first task should be determine the stability of tagging systems.

5.1.2 The Tripartite Structure of Tagging

To begin, we review the conceptual model of generic collaborative tagging systems theorized by Marlow et al (2006a); Mika (2005) in order to make predictions about collaborative tagging systems based on empirical data and based on generative features of the model.

There are three main types of entities that compose any tagging system:

- The users of the system (people who actually do the tagging)
- The tags themselves
- The resources being tagged (in this case, the websites)

Each of these can be seen as forming separate spaces consisting of sets of nodes, which are linked together by edges (see Fig. 5.1). The first space, the ***user space***, consists of the set of all users of the tagging system, where each node is a user. The second space is the ***tag space***, the set of all tags, where a tag corresponds to a term ("music") or neologism ("toread") in natural language. The third space is the ***resource space***, the set of all resources, where usually each resource is a website denoted by a unique URI.[1] A tagging instance can be seen as the two edges that links a user to a tag and that tag to a given website or resource. Note that a tagging instance can associate a date with its tuple of user, tag(s), and resource.

From Figure 5.1, we observe that tags provide the link between the users of the system and the resources or concepts they search for. This analysis reveals a number of dimensions of tagging that are often under-emphasized. In particular, tagging is often *a methodology for information retrieval*, much like traditional search engines, but with a number of key differences. To simplify drastically, with a traditional search engine a user enters a number of tags and then an automatic algorithm labels the resources with some measure of relevance to the tags *pre-discovery*, displaying relevant resources to the user. In contrast, with collaborative tagging a user finds a resource and then adds one or more tags to the resource manually, with the system storing the resource and the tags *post-discovery*. When faced with a case of retrieval, an automatic algorithm does not have to assign tags to the resource automatically, but can follow the tags used by the user. The difference between this and traditional searching algorithms is two-fold: collaborative tagging relies on human knowledge,

[1] A URI is a "Universal Resource Identifier" such as *http://www.example.com* that can return a webpage when accessed. Some tagging based systems store the entire document, not the URI, but most systems such as del.icio.us store only the URI. The resource space, in this definition, represents whatever is being tagged, which may or may not be websites per se.

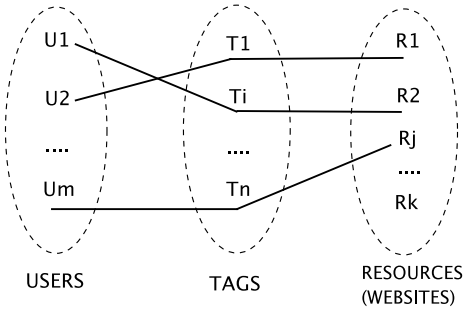

Fig. 5.1 Tripartite graph structure of a tagging system. An edge linking a user, a tag and a resource (website) represents one tagging instance

as opposed to an algorithm, to directly connect terms to documents before a search begins, and thus relies on the collective intelligence of its human users to *pre-filter* the search results for relevance. When a search is complete and a resource of interest is found, collaborative tagging often requires the user to tag the resource in order to store the result in his or her personal collection. This causes a *feedback cycle*. These characteristics motivate many systems like del.icio.us and it is well-known that feedback cycles are one ingredient of complex systems Bar-Yam (2003), giving further indication that a power law in the tagging distribution might emerge.

5.2 Detecting Power Laws in Tags

This section uses data from del.icio.us to empirically examine whether intuitions regarding tagging systems as complex systems exhibiting power law distributions hold.

5.2.1 Power Law Distributions: Definition

A ***power law*** is a relationship between two scalar quantities x and y of the form:

$$y = cx^{\alpha} \tag{5.1}$$

where α and c are constants characterizing the given power law. Eq. 5.1 can also be written as:

$$\log y = \alpha \log x + \log c \tag{5.2}$$

When written in this form, a fundamental property of power laws becomes apparent; when plotted in log-log space, power laws are straight lines. Therefore, the most simple and widely used method to check whether a distribution follows a power law

and to deduce its parameters is to apply a logarithmic transformation, and then perform linear regression in the resulting log-log space. In this paper we used a more powerful regression method to derive α that minimizes the bias in the value of the exponent (see Newman (2005a) for the technical details).

The intuitive explanation of power law parameters in the domain of tagging is as follows: c represents the number of times the most common tag for that website is used, while α gives the power law decay parameter for the frequency of tags at subsequent positions. Thus, the number of times the tag in position p is used (where $p = 1..25$, since we considered the tags in the top 25 positions) can be approximated by a function of the form:

$$Frequency(p) = \frac{c}{p^{-\alpha}} \tag{5.3}$$

where $-\alpha > 0$ and $c = Frequency(p = 1)$ is the frequency of the tag in the first position in the tag distribution (thus, it is a constant that is specific for each site/resource).

5.2.2 Empirical Results for Power Law Regression for Individual Sites

For this analysis, we used two different data sets. The first data set contained a subset of 500 "Popular" sites from del.icio.us that were tagged at least 2000 times (i.e. where we would expect a "converged" power law distribution to appear). The second data set considers a subset of another 500 sites selected randomly from the "Recent" section of del.icio.us. Both sections are prominently displayed on the del.icio.us site, though "Recent" sites are those tagged within the short time period immediately prior to viewing by the user and "Popular" sites are those which are heavily tagged in general.[2] While the exact algorithms used by del.icio.us to determine these categories are unknown, they are currently the best available approximations for random sampling of del.icio.us, both of heavily tagged sites and of a wider set of sites that may not be heavily tagged.

The mean number of users who tagged resources in the "Popular" data set was 2074.8 with a standard deviation of 92.9, while the mean number of users of the "Recent" data set was 286.1 with a standard deviation of 18.2. In all cases, the tags in the top 25 positions in the distributions have been considered and thus all of our claims refer to these tags. Since the tags are rank-ordered by frequency and the top 25 is the subset of tags that are actually available to del.icio.us users to examine for each site, we argue that using the top 25 tags is adequate for this examination.

[2] All data used in the convergence analysis was collected in the week immediately prior to 19 Nov 2006.

Results are presented in Figure 5.2. In all cases, logarithm of base 2 was used in the log-log transformation. [3]

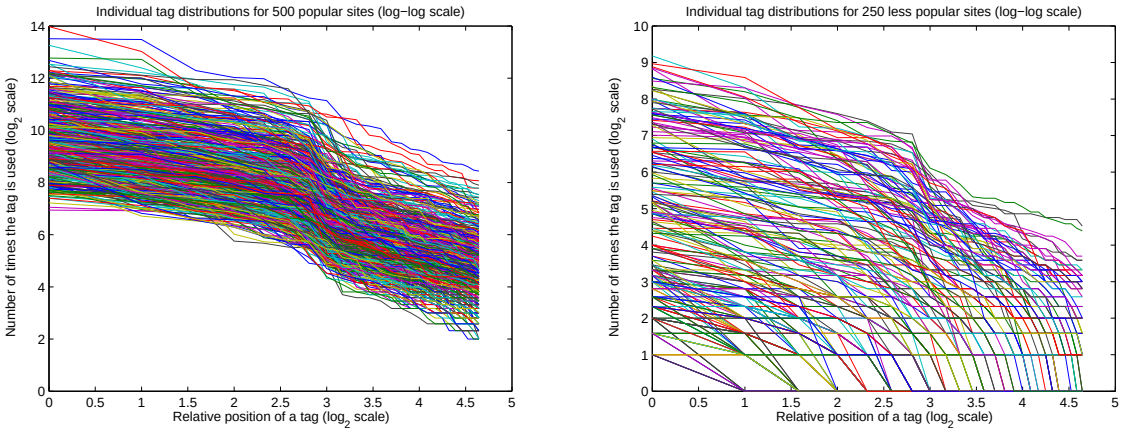

Fig. 5.2 Frequency of tag usage relative to tag position. For each site, the 25 most frequently used tags were considered. The plot uses a double logarithmic (log-log) scale. The data is shown for a set of 500 randomly-selected, heavily tagged sites (left) and for a set of 500 randomly-selected, less-heavily tagged sites (right).

As shown by Newman (2005a) and others, the main characteristic of a power law is its slope parameter α. On a log-log scale, the constant parameter c only gives the "vertical shift" of the distribution with respect to the y-axis. For each of the sites in the data set, the corresponding power law function was derived and the slopes of each (α parameters) were compared. The slopes indicate the fundamental characteristic of the power laws, as vertical shifts can and do vary significantly between different sites.

Our analysis shows that for the subset of heavily tagged sites, the slope parameters are very similar to one another, with an average of $\alpha = -1.22$ and a standard deviation ± 0.03. Thus, it appears that the power law decay slope is relatively consistent across all sites. This is quite remarkable, given that these sites were chosen randomly with the only criteria being that they were heavily tagged. This pattern where the top tags are considerably more popular than the rest of the tags seems to indicate a fundamental effect of the way tags are distributed in individual websites which is independent of the content of individual websites. The specific content of the tags themselves can be very different from one website to the other and this obviously depends on the content of the tagged site.

[3] Note that the base of the logarithm does not actually appear in the power law equation (c.f. Eq. 5.1), but because we use empirical and thus possibly noisy data, this choice might introduce errors in the fitting of the regression phase. However, we did not find significant differences from changing the base of the logarithm to *e* or 10.

For the set of less-heavily tagged sites, we found the slopes differed from each other to a much greater extent than with the heavily tagged data, with an average $\alpha = -5.06$ and standard deviation ± 6.10. Clearly, the power law effect is much less pronounced for the less-heavily tagged sites as opposed to the heavily tagged sites, as the standard deviation reveals a much poorer fit of the regression line to the log-log plotted aggregate data. For sites with relatively few instances of tagging, the results reveal mostly noise.

5.2.3 Empirical Results for Power Law Regression Using Relative Frequencies

In the previous section, we applied power law regression techniques to individual sites, using the number of hits for a tag in a given position in the distribution. In this section, we examine the aggregate case where we no longer use the raw number of tags (because these are not directly comparable across sites), and instead use the relative frequencies of tags. The relative frequency is defined as the ratio between the number of times a tag in a particular position is used for a resource and the total number of times that resource is tagged [4]. Thus, relative frequencies for a given site always sum to one. These relative frequencies based on data from all 500 sites of the "Popular" data set were then averaged. Results are presented in Figure 5.3.

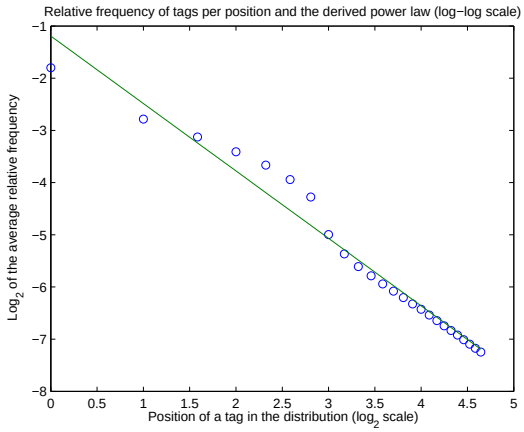

Fig. 5.3 Average relative frequency of tag usage, for the set of 500 "Popular" sites from above. On the y-axis, the logarithm of the relative frequency (probability) is given. (The plot uses a double logarithmic (log-log) scale, thus on the y-axis values are negative since relative frequencies are less than one.)

[4] To be more precise, the denominator is taken as the total number of times the resource is tagged with a tag from the top 25 positions, given available data.

As before, a power law was derived in the log-log space using least-means squares (LMS) regression. This power law was found to have the slope $\alpha = -1.25$. The regression error, computed through the LMS method in the normal, not logarithmic space, was found to be 0.038. Note that the LMS regression error computation only makes sense when converted back in the normal space, since in the log-log space exponents are negative and, furthermore, deviations on the y-axis only denote actual error only after the exp_2 function is applied. This corresponds to a LMS error rate in the power law regression of 3.8% over the total number of tags in the distribution, which is low enough to allow us to conclude that tag distributions do follow power laws.

We note, however, that there is a deviation from a perfect power law in the del.icio.us data in the sense that there is a change of slope after the top seven or eight positions in the distribution. This effect is also relatively consistent across the sites in the data set. This may be due to the cognitive constraints of the users themselves or an artifact of the way the del.icio.us interface is constructed, since that number of tags are offered to the users as a suggestion to guide their search process. Nevertheless, given that the LMS regression error is rather low, we argue the effect is not strong enough to change the overall conclusion that tag distributions follow power laws.

5.2.4 The Dynamics of Tag Distributions

In Section 5.2, we provide a method for detecting a power law distribution in the tags of a site or collection of sites. In this section, we study another aspect of the problem, namely how the shape of these distributions develops in time from the tagging actions of the individual users. First, we examine the how power law distributions form at the top (the first 25 positions) of tag distributions for each site. For this, we employ a method from information theory, namely the Kullback-Leibler divergence. Second, we study the dynamics of the entire tag distributions, including all tags used for a site, and we show that the relative weights of the top and tail of tag distributions converge to stable ratios in the data sets.

5.2.4.1 Kullback-Leibler Divergence: Definition

In probability and information theory, the Kullback-Leibler divergence (also known "relative entropy" or "information divergence") represents a natural distance measure between two probability distributions P and Q (in our case, P and Q are two vectors representing discrete probability distributions). Formally, the Kullback-Leibler divergence between P and Q is defined as:

$$D_{KL}(P||Q) = \sum_x P(x) log(\frac{P(x)}{Q(x)}) \tag{5.4}$$

The Kullback-Leibler distance is a non-negative, convex function, i.e. $D_{KL}(P,Q) \geq 0, \forall P,Q$ (note that $D_{KL}(P,Q) = 0$ iff. P and Q coincide). Also, unlike other distance measures it is not symmetric, i.e. in general $D_{KL}(P,Q) \neq D_{KL}(Q,P)$.

5.2.4.2 Application to Tag Dynamics

We use two complementary ways to detect whether a distribution has converged to a steady state using the Kullback-Leibler divergence:

- The first is to take the relative entropy between every two consecutive points in time of the distribution, where each point in time represents some change in the distribution. Again, in our data, tag distributions are based on the rank-ordered tag frequencies for the top 25 highest-ranked unique tags for any one website. Each point in time was a given month where the tag distribution had changed; months where there was no tagging change were not counted as time points. Using this methodology, a tag distribution that was "stable" would show the relative entropy converging to and remaining at zero over time. If the Kullback-Leibler divergence between two consecutive time points becomes zero (or close to zero), it suggests that the shape of the distribution has stopped evolving. This technique may be most useful when it is completely unknown whether or not the tagging of a particular site has stabilized at all.
- The second method involves taking the relative entropy of the tag distribution for each time step with respect to the final tag distribution, the distribution at the time the measurement was taken or the last observation in the data, for that site. This method is most useful for heavily tagged sites where it is already known or suspected that the final distribution has already converged to a power law.

The two methods are complementary; the first methodology would converge to zero if the two consecutive distributions are the same, and thus one could detect whether distributions converged if even temporarily. Cyclical patterns of stabilization and destabilization may be detected using this first method. The second method assumes that the final time point is the stable distribution so this method detects convergence only towards the final distribution. If both of these methods produce relative entropies that approach zero, then one can claim that the distributions have converged over time to a single distribution, the distribution at the final time point. Given our interest in distributions that have converged to power laws, we are actually examining the dynamics of convergence to a power law.

5.2.4.3 Empirical Results for Tag Dynamics

The analysis of the intermediate dynamics of tagging is considerably more involved than the analysis of final tag distributions. Because the length of the histories varies widely, there is no meaningful way to compute a cumulative measure across all sites as in Section 5.2, so our analysis has to consider each resource individually. In

Figure 5.4 (A and B), we plot the results for the convergence of the 500 “Popular” sites, on the basis that their final distribution must have converged to a power law, that their complete tagging history was available from the first tagging instances, and that this history was of substantial length. In the data set considered, up to 35 time points are available for some sites (which roughly corresponds to three years of data, since one time point represents one month).

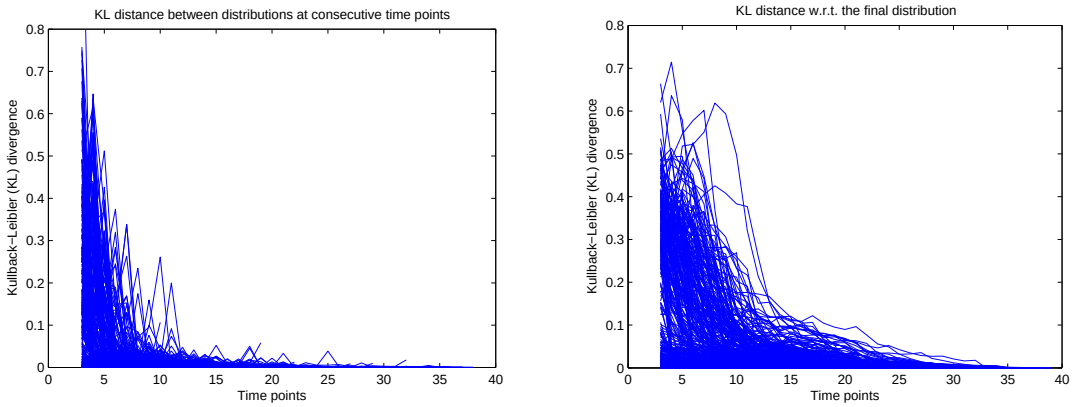

Fig. 5.4 A (left). Kullback-Leibler divergence between tag frequency distributions at consecutive time steps for 500 ”Popular” sites. B (right). Kullback-Leibler divergence of tag frequency distribution at each time step with respect to the final distribution.

There is a clear effect in the dynamics of the above distributions.[5] At the beginning of the process when the distributions contain only a few tags, there is a high degree of randomness, indicated by early data points. However, in most cases this converges relatively quickly to a very small value, and then in the final ten steps, to a Kullback-Leibler distance which is graphically indistinguishable from zero (with only a few outliers). If the Kullback-Leibler divergence between two consecutive time points (in Figure 5.4A) or between each step and the final one (Figure 5.4B) becomes zero or close to zero, it indicates that the shape of the distribution has stopped changing. The results here suggest that the power law may form relatively early on in the process for most sites and persist throughout. Even if the number of tags added by the users increases many-fold, the new tags reinforce the already-formed power law. Interestingly, there is a substantial amount of variation in the initial values of the Kullback-Leibler distance prior to the convergence. Future work might explore the factors underlying this variation and whether it is a function of the content of the sites or of the mechanism behind the tagging of the site. Additionally, convergence to zero occurs at approximately the same time period (often within a few months) for these sites.

[5] Note that in Figure 5.4, the first two time points were omitted because their distribution involved few tags and were thus very highly random.

The results of the Kullback-Leibler analysis provide a powerful tool for analyzing the dynamics of tagging distributions. This very well might be the result of the "scale-free" property of tagging networks, so that once the tagging of users have reached a certain threshold, regardless of how many tags are added, the distribution remains stable Shen and Wu (2005). This method can be immensely useful in analyzing real-world tagging systems where the stability of the categorization scheme produced by the tagging needs to be confirmed.

5.2.4.4 Examining the dynamics of the entire tag distribution

In the previous sections, we focused on the distributions of the tags in the top 25 positions. However, heavily tagged or popular resources, such as those considered in our analysis, can be tagged several tens of thousands of times each, producing hundreds or even thousands of distinct tags. It is true that many of these distinct tags are simply personal bookmarks which have no meaning for the other users in the system. However, it is still crucial to understand their dynamics and the role they play in tagging, especially with respect to the top of the tag distribution. Some sources (e.g. Anderson Anderson (2006)), have argued that the dynamics of long tails are a fundamental feature of Internet-scale systems. Here we were particularly interested in two questions. First, how does the number of times a site is tagged (including the long tail) evolve in time? Second, how does the relative importance of the head (top 25 tags) to the long tail change as tags are added to a resource?

Results for the same set of 500 "Popular" sites described above are shown in Figure 5.5. Note that the tag distributions were reconstructed through viewing the tagging history of the individual site as available through del.icio.us and collecting the growth of this tagging distribution over time, thus allowing us to record the growth of tags outside the 25 most popular.

As seen in Figure 5.5, the total number of times a site is tagged grows continuously at a rate that is specific to each site and this probably depends on its domain and particular context. Though the results are not shown here due to space constraints, a similar conclusion can be formulated for the number of distinct tags, given that the number of distinct tags varies considerably per site and does not seem to stabilize in time. However for virtually all of the sites in the data set considered, the proportion of times a tag from the top 25 positions is used relative to the total number of times that a resource is tagged did stabilize over time. So, while the total number of tags per resource grows continuously, the relative weight of the tags in the head of the tag distribution compared to the those in the long tail does stabilize to a constant ratio. This is an important effect and it represents a significant addition to our analysis of the stability analysis of the top 25 positions, since it shows the relative importance of the long tail with respect to the head of the distribution does eventually stabilize regardless of the growth of tags in the long tail.

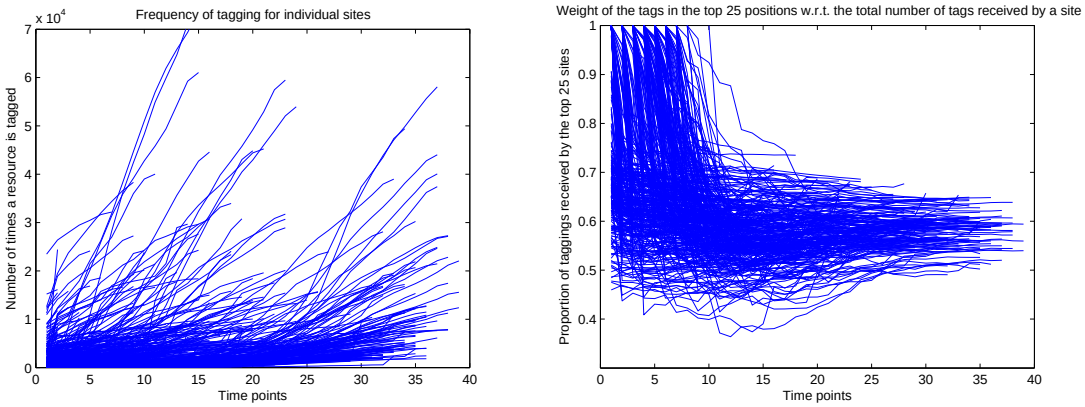

Fig. 5.5 A (left). Cumulative number of times a resource is tagged for each time point. B (right). Proportion of times a tag in the top 25 spots of the distribution has been used to tag a resource to the total number of times the resource has been tagged with any tag.

5.3 The Effect of Suggestions on Tagging

So far, we have explored the important question of whether a coherent, stable way of characterizing sense can emerge from collaborative tagging systems and has presented several novel methods for analyzing data from such systems. We have shown that tagging distributions of heavily tagged resources tend to stabilize into power law distributions and present a method for detecting power law distributions in tagging data, and see the emergence of stable power law distributions as an aspect of what may be seen as collective consensus around the categorization of information driven by tagging behaviour. Thus groups of tags are indeed an adequate candidate for a notion of Fregean sense.

However, one could argue that the stabilization is just a mere artifact of tag suggestions. Tag suggestions are when a tagging system, instead of letting the user tag the resource, automatically (as the product of some algorithm) presents a list of 'suggested' tags for the user. The user can then easily accept these tags or choose through them, rather than choose their own. This could lead to the stabilization of the tagging system not as a product of the actual collaborative sense-making of users, but as an artificial and predictable result of the tag suggestion algorithm. However, the reasons behind the emergence of a power-law distribution in tagging systems are yet unknown, although explanations fall into two general categories. The first of these explanations is relatively simple: the tags stabilize into a power-law because users are imitating each other via tag suggestions put forward by the tagging system Golder and Huberman (2006). The second and more recent explanation is that in addition to imitation, the users share through a similar tag generation procedure based on the informationon the webpage, most likely because the users have the same background knowledge **?**. However, drawing these two influences apart has

not yet been tested scientifically, which we will do. However, first let's inspect these the existing explanations for tagging stabilization more deeply.

5.3.1 Models of collaborative tag behavior

5.3.1.1 A simple model: The Polya Urn

The most elementary model of how a user selects tags when annotating a resource is simple imitation of other users. Note that 'imitation' in tagging systems means that the tags are being reinforced via a 'tag suggestion' mechanism, and so the terms "imitation", "reinforcement", "feedback", and 'tag suggestion' can be considered to be synonymous in the context of tagging systems. The user can imitate other users precisely because the tagging systems tries to support the user in the tag selection process by providing tag suggestions based on tags other people used when tagging the same resource. There are minor variants of this theme, such as the possibility of using a combination of tags of other users in combination with a user's own previously used tags. In most tagging systems like del.icio.us these tag suggestions are presented as a list of tags that the user can select in order to add them to their tagging instance. The selections of tags from the tag recommendation forms a positive feedback loop in which more frequent tags are being reinforced, thus causing an increase in their popularity, which in turn causes them to be reinforced further and exposed to ever greater numbers of users. This simple type of explanation is easily amendable to preferential attachment models, also known as 'rich get richer' explanations, which are well-known to produce power-law distributions. Intuitively, the earliest studies of tagging observed that users imitate other pre-existing tags Golder and Huberman (2006). Golder and Huberman proposed that the simplest model that results in a "power-law" would be the classical Polya urn model Golder and Huberman (2006). Imagine that there is urn containing balls, each of some finite number of colors. At every time-step, a ball is chosen at random. Once a ball is chosen, it is put back in the urn along with another ball of the same color, which formalizes the process of feedback given by tag suggestions. As put by Golder and Huberman, "replacement of a ball with another ball of the same color can be seen as a kind of imitation" where each color of a ball is made equal to a natural language tag and since "the interface through which users add bookmarks shows users the tags most commonly used by others who bookmarked that URL already; users can easily select those tags for use in their own bookmarks, thus imitating the choices of previous users" Golder and Huberman (2006). Yet, this model is too limited to describe tagging, as it features only reinforcement of existing tags, not the addition of *new* tags.

5.3.1.2 Imitation and The Yule-Simon Model

The first model that formalized the notion of new tags was proposed by Cattuto et al. **?**. In order for new tags to be added, a single parameter p must be added to the model, which represents the probability of a new tag being added, with the probability $\bar{p} = (1-p)$ that an already-existing tag is reinforced by random uniform choice over all already-existing tags. This results in a Yule-Simon model, a model first employed by Yule Yule (1925) to model biological genera and later Simon to model the construction of a text as a stream of words Simon (1955). This model has been shown to be equivalent to the famous Barabasi and Albert algorithm for growing networks Bornholdt and Ebel (2001). Yet the standard Yule-Simon process does not model vocabulary growth in tagging systems very well, as noticed by Cattuto et al. as it produces exponents "lower than the exponents we observe in actual data" **?**.

Cattuto et al. hypothesize that this is because the Yule-Simon model assumes users are choosing to reinforce ($\bar{p}$) tags uniformly from a distribution of *all* tags that have been used previously, so Cattuto concludes that "it seems more realistic to assume that users tend to apply recently added tags more frequently than old ones" **?**. This behavior could be caused by the exposure of a user to a feedback mechanism, such as del.icio.us tag suggestion system. This suggestions exposes the user only to a subset of previously existing tags, such as those most recently added. Since the tag suggestion mechanism only encourages more recently-added tags to be re-enforced with a higher probability, Cattuto et al. added a memory kernel with a power-law exponent to standard Yule-Simon model. This means that the weight of a previously existing tag being reinforced is weighted according to a power-law itself, so that a tag that has been applied x steps in the past is chosen with a probability $\bar{p}(x) = a(t)/(x+\tau)$, where $a(t)$ is a normalization factor and τ "is a characteristic time scale over which recently added words have comparable probabilities" **?**. While the parameter p controls the probability of reinforcing an existing tag, this second parameter τ, controls how fast the memory kernel decays and so over what time-scale a tag may likely count as 'new' and so be more likely to be reinforced. As Cattuto et al. notes, "the average user is exposed to a few roughly equivalent top-ranked tags and this is translated mathematically into a low-rank cutoff of the power-law" **?**. This model produces an "excellent agreement" with the results of tag-correlation graphs **?**. It should be clear that the original Yule-Simon model simply parametrizes the probability of the imitation of existing tags. The modified Yule-Simon model with a power-law memory kernel also depends on the imitation of existing tags, where the probability of a previously-used tag is decaying according to a power-law function.

5.3.1.3 Adding Parameters and Background Knowledge

Although Cattuto et al.'s model is without a doubt an elegant minimal model that captures tag-correlation distributions well, it was not tested against tag-resource distributions **?**. Furthermore, as noticed by Dellschaft and Staab, Cattuto et al.'s model

also does not explain the sub-linear tag vocabulary growth of a tagging system **?**. Dellschaft and Staab propose an alternative model, which adds a number of new parameters that fit the data produced by tag-growth distributions and tag-resource distributions better than Cattuto et al.'s model **?**. The main points of interest in their model is that instead of a new tag being chosen uniformly, the new tag is chosen from a power-law distribution that is meant to approximate "background knowledge." So besides "background knowledge" ($\bar{p}$), their model also features the inverse of "background knowledge," i.e. the "probability that a user imitates a previous tag assignment" (p) **?**. In essence, Dellschaft and Staab have added (at least) two new parameters to a Yule-Simon process, and these additional parameters allows the reinforcement of existing tags to be more finely tuned. Instead of a single power-law memory kernel with a single parameter τ, these additional parameters allow the modeling of "an effect that is comparable to the fat-tailed access of the Yule-Simon model with memory" while keeping tag-growth sub-linear **?**. The model proposed by Cattuto et al. kept the tag-growth parameter equal to 1 and so makes tag growth linear to p **?**. Yet for us, most important advantage of Dellschaft and Staab over Cattuto et al.'s model is that their added parameters lets their model match the previously unmatched observation by Halpin et al. of the frequency rank distribution of resources being a power-law Halpin et al (2007). The match is not as close as the match with vocabulary growth and tag correlations, as resource-tag frequency distributions vary highly per resource, with the exception of the drop in slope around rank 7-10 Halpin et al (2007).

5.3.2 Research Questions

What unifies all of these models is that they assume that imitation, usually assumed to be tag suggestions from the tagging system, has a major impact on the emergence of a power-law distribution. With concern to the modified Yule-Simon model and the more highly parametrized model that takes into account 'background knowledge,' different claims are made of where the imitated tags come from. Cattuto et al. proposes that they come from a random uniform distribution of tags while Dellschaft and Staab propose a more topic-related distribution that itself has a power-law distribution **?**. However, just because a simple model based on imitation of tag suggestions can lead to a power-law distribution does not necessarily mean that tag suggestions are actually the causal mechanism that causes the power-law distribution to arise in tagging systems. The research question posed then is: Is the tag suggestion mechanism, the main force behind the observed power-law distributions in tagging systems?

In order to measure the effects of tag suggestions on the tag behavior of users we developed a Web-based experiment in which users were asked to tag 11 websites, with two varying conditions: the 'tag suggestion' condition (Condition A) in which 7 tag suggestions were presented to the user, and a 'no tag suggestion' condition (Condition B) in which no tag suggestions were presented to the user.

In this experiment we focus on del.icio.us which is the one of the earliest and well-known tagging systems. Del.icio.us was the first to introduce a tag based collaborative bookmark system. Del.icio.us has more than five million users and 150 million tagged URIs and so provides a vast data-set. The user interface used in our experiment presented the tag suggestions in a similar way to del.icio.us to avoid confusion.

The 11 websites used in the experiment were selected according to two criteria. First, the topics of the websites needed to appeal to the general public. Second, the website needed to have over 200 tagging instances. The appeal to the general public was operationalized by randomly choosing sites that were tagged with the tag "lifestyle" on del.icio.us. The tag "lifestyle" is a popular tag with 72,889 tagged web-pages as of October 2008. This was chosen in order to not bias our study to one particular specialized subject matter, and so exclude web-pages on del.icio.us that have a highly technical content. Specialized content may not lead to normal tagging behavior from users in the experiment who might not be familiar with the specialist subject matter. The second criteria of using only web-pages with over 200 tagging instances was chosen since it has been shown that stable power-law tag distributions emerge around the 100-150th tagging Golder and Huberman (2006). We did not want the tag suggestions to be from non-stable tag distributions, as it has been shown that the variance between the top popular tag could vary widely before 100-150th tag. The 11 websites selected for this experiment, with the popular tags provided from del.icio.us and the number tags. Note that while the number of URIs 11 may appear to be small, it is larger than previous experiments over tag suggestions Suchanek et al (2008) and was enough to give the experiment enough power to be statistically significant. It was far more critical for this experiment to get enough subjects in order for power-law distributions to be given the chance to arise without tag suggestion, and this would require at least 100 experimental subjects tagging each URI.

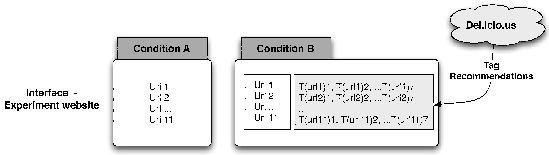

Fig. 5.6 Experimental Design

Figure 5.6 shows the experimental design. In the 'no tag suggestion' condition (Condition A), as shown in Figure 5.6, a user is presented the 11 websites he needs to tag without any form of tag suggestions. In the 'tag suggestion' condition (Condition B), also shown in Figure 5.6, a user is presented the 11 websites with 7 suggested tags. While the details of the tag suggestion algorithm applied by del.icio.us is unknown, for our experiment the suggested tags in condition B were aggregated from del.icio.us and the 7 suggested tags given by del.icio.us for each of the 11

websites. For the experiment the 7 popular tags were aggregated and presented to the participants in manner similar to how tags are suggested to users of del.icio.us, being shown to the user before they commence their tagging. Each of the 300 participants was randomly assigned to either the 'tag suggestion' or 'no tag suggestion' condition. Of these 300 users, 78 did not tag any website (37 in the 'tag suggestion' condition, 41 in the 'tag suggestion' condition) and are therefore excluded from further analysis. The users were randomized over age, gender, computer, Internet and their past tagging usage.

5.4 Results

In total the 222 participants applied 7,250 tags over all websites in both conditions, with 3,694 tags applied in the 'tag suggestion' condition and 3,556 in the 'no tag suggestion' condition. On average every user in the 'tag suggestion' condition applied 32.69 ($S.D. = 9.77$) tags over all 11 URIs and for the no tag suggestion conditions 32.61 ($S.D. = 6.80$) tags over 11 URIs.

5.4.1 Detecting Power-Law Distributions

The power-law distribution is defined by the function:

$$y = cx^{-\alpha} + b \tag{5.5}$$

in which c and α are the constants that characterize the power-law and b being some constant or variable dependent on x that becomes constant asymptotically. The α exponent is the scaling exponent that determines the slope of the distribution before the long tail behavior begins. A power-law function can be transformed to a log-log scale as in the following equation:

$$log(y) = -\alpha log(x) + log(c) \tag{5.6}$$

This equation shows the characteristic property of power-law function is that when transformed to a log-log scale the power-law distribution takes the shape of a linear function with slope α. So transforming a function to a log-log scale and determining the slope α is one of the first steps in examining whether a distribution has a power-law. We averaged the tag-resource distributions for all 11 web-pages, and this distribution in log-log space is given in Figure 5.7. In a log-log scale, *both* conditions appear visually to exhibit power-law behavior.

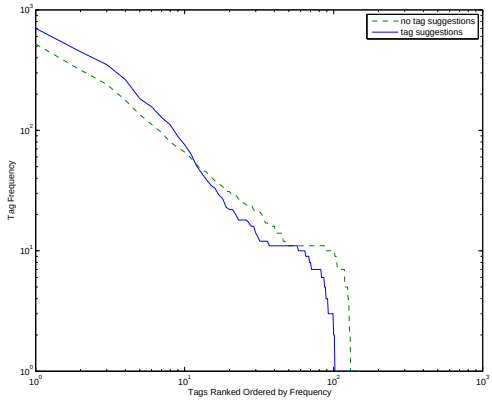

Fig. 5.7 Averaged tag-resource distributions for both experimental conditions on a log-log scale. The solid line depicts the 'tag suggestion' condition, the dotted line the 'no tag suggestion' condition.

5.4.1.1 Parameter Estimation via Maximum-Likelihood

The most widely used method to check whether a distribution follows a power-law is to apply a logarithmic transformation, and then perform linear regression, estimating the slope of the function in logarithmic space to be α. However, this least-square regression method has been shown to produce systematic bias, in particular due to fluctuations of the long tail Clauset et al (2007). To determine a power-law accurately requires minimizing the bias in the value of the scaling exponent and the beginning of the long tail via maximum likelihood estimation. See Newman Newman (2005b) for the technical details. To determine the α of the observed distributions, we fitted the data using the maximum likelihood method recommended by Newman Newman (2005b). Figure 5.8 shows the different α parameters for the 'tag suggestion' and 'no tag suggestion' conditions, as well as the α determined via aggregation of tagging data from del.icio.us for the 11 URIs. Overall, for the 'no tag suggestion' condition, the average α was 2.1827 (S.D. 0.0799) while for the 'tag suggestion' condition the average α was 2.0682 (S.D. 0.0941). The α values for both conditions and the aggregated data from del.icio.us are situated in the interval $[1.732391 < \alpha < 2.249359]$. Figure 5.8 shows that both experimental conditions and the aggregated data from del.icio.us have similar exponents. Our results shows that a similar α holds for both the 'tag suggestion' and 'no tag suggestion' condition. Further updates to determine if there is an actual difference between the two conditions as regards if a power-law distribution actually holds.

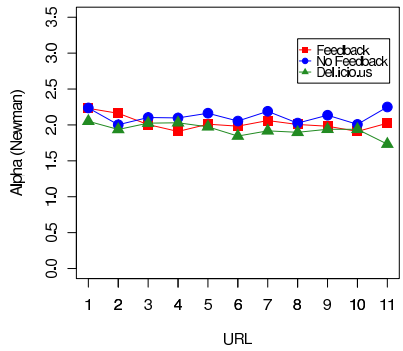

Fig. 5.8 X axis depicts the URI used in the experiment, Y axis depicts the different α values

5.4.1.2 Kolmogorov-Smirnov Complexity

Determining whether a particular distribution is a 'good fit' for a power-law is difficult, as most goodness-of-fit tests employ some sort of normal Gaussian assumption that is inappropriate for non-normal power-law distributions. However, the Kolmogorov-Smirnov Test (abbreviated as the 'KS Test') can be employed as a 'goodness-of-fit' test for any distribution without implicit parametric assumptions and is thus ideal for use measuring goodness-of-fit of a given finite distribution to a power-law function. Intuitively, given a reference distribution P (perhaps produced by some well-known function like a power-law) and a sample distribution Q of size n, where one is testing the null hypothesis that Q is drawn from P, then one simply compares the cumulative frequency of both P and Q and then the greatest discrepancy (the D-statistic) between the two distributions is tested against the critical value for n, which varies per function.

For a power-law distribution generating function, we can get a critical p-value by generating artificial data using the scaling exponent α and lower-bound equal to those found in the supposed fitted power-law distribution. A power-law is fit to this artificial data, and then the KS test is then done for each distribution that was artificially generated comparing it to its *own* fitted power-law. The p-value is then just the fraction of the amount of times the D-statistic is larger for the artificially-generated distribution than the D-statistic of the empirically-found distribution. Therefore, the larger the p-value, the more likely a genuine power-law has been found in the empirical data. According to Clauset, "once we have calculated our p-value, we need to make a decision about whether it is *small enough to rule out* the power-law hypothesis" (emphasis added) Clauset et al (2007). The power-law hypothesis is simply that the distribution was generated by a power-law generating function. The null hypothesis is that by chance a function would generate the power-law distribution observed in the empirical data. We shall also use $p \leq 0.1$.

The KS test for all 11 tagged web-pages, testing both the 'tag suggestion' and 'no tag suggestion' condition, is given in Figure 5.9. The average D statistic for the 'no tag suggestion' condition is 0.0313 (S.D. 0.0118) with $p = .48(p > .1$, power-law

found). For the 'tag suggestion' condition the average *D*-statistic is 0.0724 (S.D. 0.0256) with $p = .08 (p \leq .1$, no power-law found). These results show that the power-law function exhibited *only* in the 'no tag suggestion' conditions is significant, the fit is closer for the 'no tag suggestion' condition than the 'tag suggestion' condition. The *D*-statistic showed a range from 0.0170 to 0.0552 for 'no tag suggestion' condition yet a range of 0.0428 to 0.1318 for 'tag suggestion.' Thus, the power-law only significantly appears without tag suggestions, and with tag suggestions a power-law cannot be reliably found. This is surprising, as tag suggestions do not only *not* cause the power-law to form, but they seems that they somehow prevent it from being formed. On the other hand, the 'no tag suggestion' condition results in a significantly good fit to a power-law. Therefore, the result is somewhat counter-intuitive, as according to our experimental data a simple tag-based suggestion mechanism is unlikely the main cause of the power-law formation.

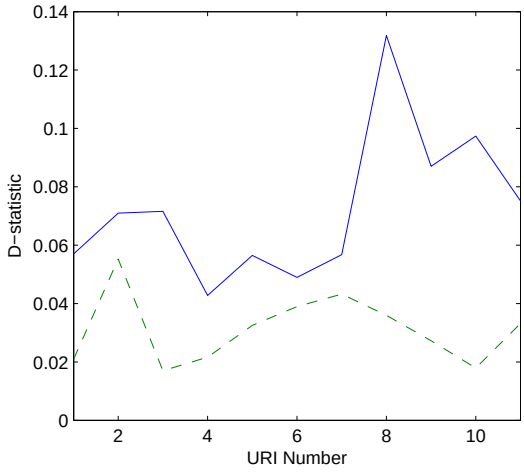

Fig. 5.9 X axis depicts the URI used in the experiment, Y axis depicts the different D Statistics from the KS Test. The dotted line is the 'no tag suggestion' condition, while the solid line is the 'tag suggestion' condition.

5.4.2 Influence of tag suggestion on the tag distribution

Given that the KS test shows that there is a significant and perhaps counter-intuitive difference in the emergence of the power-law distributions between the conditions, we need a more fine-grained way to tell what the differences are in the distributions for the two conditions. A number of differing techniques will be deployed to answer this question.

5.4.2.1 Kullback Leibler Divergence

The Kullback-Leibler divergence (also known as *relative entropy*), which we abbreviate as 'KL divergence,' can be used an intuitive information-theoretic measure of the distance between two distributions P and Q. Unlike many other methods, it takes the entire distribution (in our case, the long tail is of particular interest) into account. Note that it is not a true metric as it is an asymmetric, however, it is a useful measure of the difference between two distributions as it is a non-negative, convex function with well-known properties. The KL divergence is zero if and only if the two distributions are the same, otherwise a positive distance results that is larger the greater the divergence between the distributions. Intuitively in information theory, the KL divergence is the expected difference in bits required to encode to distribution Q when using a code based on distribution P. The KL divergence between P and Q is given as:

$$D_{KL}(P||Q) = \sum_{x} P(x) log(\frac{P(x)}{Q(x)}) \tag{5.7}$$

The KL divergence (using the 'tag suggestion' condition for P and the 'no tag suggestion' condition for Q) for each URI in the experiment are given in Figure 5.10. While some URIs (like number 6 and 7) have almost no difference between the 'tag suggestion' and 'no tag suggestion' conditions, other URIs like number 11 have large differences. This average KL divergence between the 'tag suggestion' condition and 'no tag suggestion' condition is 0.1617 (S.D. 0.0820). This is small but not insubstantial. As shown in the observation of Figure 5.7, the long tail of the 'tag suggestion' condition is often shorter than the 'no tag suggestion' condition, while the top of the 'tag suggestion' distribution has a higher frequency than the top of the 'no tag suggestion' distribution. The KL divergence takes this into account, while merely finding the α does not. The effect on the top of the distribution should be investigated further.

5.4.2.2 Ranked frequency distribution

In order to observe the micro-behavior of the 'tag suggestion' and 'no tag suggestion' distributions, we investigate whether or not the tag suggestion tags are 'forced' higher in the distribution, so leading to a more sparse long tail and an exaggerated top of the distribution in the 'tag suggestion' condition. In order to provide a measurement of the number of suggested tags in the top of the distribution, the percentage of suggested tags that were found in the top 7 and top 10 tags were calculated. We compared the percentage of suggested tags in the top 7 and top 10 ranks for both conditions with del.icio.us. For this we assume that the 7 suggested tags provided by del.icio.us represent the top 7 tags in the ranked frequency distribution so that the percentage of suggested tags in the top 7 and top 10 ranks for del.icio.us is equal to 100%. We averaged the percentages for all URIs per experimental condition.

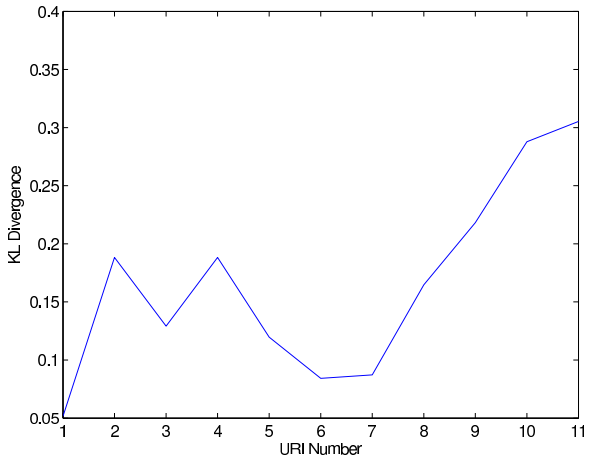

Fig. 5.10 X axis depicts the URI used in the experiment, Y axis depicts the different KL Divergence values

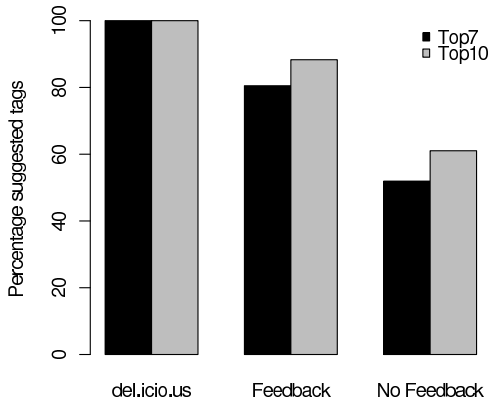

Fig. 5.11 Ranked Frequency Distribution Repeating Suggested Tags

Figure 5.11 shows that for the percentage of suggested tags available in the top 7 rank for the ‘tag suggestion’ condition is 80.51% and for the ‘no tag’ suggestion condition 51.93%. This means that only half of the suggested tags can be found in the top 7 of the ranked frequency distribution in the ‘no tag suggestion’ condition. So unsurprisingly, in the ‘tag suggestion’ condition, we observed more of the suggested tags than in the ‘no tag suggestion’ condition. There is an influence of tag suggestions on the ranked position and the frequency of the suggested tags. Tag suggestions do influence the tag-resource distribution, as tag suggestion causes a net

gain of nearly one in three tags being imitated that would otherwise not be. However, when users are not guided by tag suggestions and tag freely they still choose for themselves half of the tags that would have been otherwise suggested had they had a 'tag suggestion' mechanism available. Further we look at the availability of suggested tags in the top 10 as an indication how dispersed the suggested tags are in the ranked frequency distribution for both conditions. For the top 10 rank figure 5.11 shows that the percentage of suggested tags in the 'tag suggestion' condition is 88.30% and for the "no tag suggestion" condition is 61.03%.

5.4.2.3 Imitation Rates

Another metric that measures the influence of tag suggestion on the tag distribution is the matching and imitation rate as proposed by Suchanek et al. Suchanek et al (2008). The matching rate measure the proportion of applied tags that are available in the suggested tags. This metric provides insight in how the user is influenced by the tag suggestion provided by the tagging system. For our experiment the *matching rate* (*mr*) is being defined as:

$$mr(X) = \frac{\sum_{i=1}^{n} \mid T(X,i) \cap S(X) \mid}{\sum_{i=1}^{n} \mid T(X,i) \mid} \tag{5.8}$$

X denotes the tag suggestion method that is being used in both our conditions. The 'tag suggestion' condition provides 7 suggested tags while the 'no tag suggestion' condition provided no suggested tags. For a given URI, $T(X,i)$ denotes the set of tags at the ith tag entry and $S(X)$ denotes the suggested tags for that URI. For a tagging instance in which all tags are given by the suggested tags the matching rate will be 1.

The matching rate for the 11 URIs in the experiment and over the both conditions was calculated. The resulting matching rates can be found in Table 5.1. The 'no tag suggestion' condition serves as a reference point. The results in Table 5.1 show that users in the 'tag suggestion' condition are being influenced by the appearance of tag suggestions. The average matching rate for the 'tag suggestion' condition is 0.57 (S.D. 0.086) and for the 'no tag suggestion' condition 0.35 (S.D. 0.068). The main drawback of the matching rate is that it can't account for the application of suggested tags when tag suggestion is absent.

This ability to account for tag repetition even when the tag is missing is given by the *imitation rate* (*ir*), defined as Suchanek et al (2008):

$$\alpha_n(S) = \frac{prec_n(X,S) - prec_n(NONE,S)}{1 - prec_n(NONE,S)} \tag{5.9}$$

With $prec_n(X,S)$ defined as:

Table 5.1 Matching rate

URI No.	Tag Suggestion	No Tag Suggestion
1	0.47	0.31
2	0.57	0.34
3	0.53	0.32
4	0.65	0.48
5	0.45	0.29
6	0.52	0.29
7	0.58	0.38
8	0.65	0.38
9	0.74	0.46
10	0.63	0.30
11	0.59	0.31

$$prec_n(X,S) = \frac{\sum_{i=1}^{n} \mid T(X,i) \cap S \mid [S(X,i) = S]}{\sum_{i=1}^{n} \mid T(X,i) \mid [S(X,i) = S]} \tag{5.10}$$

The term $prec_n(X,S)$ defines the proportion of applied tags that are available in the single tag suggestion set S. Since the tags S in our experiment is always static, $prec_n(X,S)$ is equal to the calculation of the matching rate for the tag suggestion condition in Equation 5.8. $prec_n(NONE,S)$ defines the proportion of suggested tags that are available in the tags applied by the user when no tag suggestion is given. This is similar to the calculation of the matching rate for the 'no tag suggestion' condition. Therefore we can rewrite the imitation rate as:

$$ir = \frac{mr(ConditionA) - mr(ConditionB)}{1 - mr(ConditionB)} \tag{5.11}$$

Table 5.2 shows the imitation rates for the different experimental URIs. An imitation rate of 1 will denote full imitation. The results show that users tend to select suggested tags when the are available with a chance of 1 out of 3 with a mean imitation rate of 0.36 (S.D. 0.097).

Table 5.2 Imitation rate

URI No.	Imitation Rate
1	0.22
2	0.35
3	0.29
4	0.35
5	0.20
6	0.34
7	0.31
8	0.42
9	0.50
10	0.48
11	0.43

Combining this insight with our previous work in KL divergence and looking at Figure 5.7, it appears that 'tag suggestion' condition 'compresses' the distribution that naturally arises without tag suggestions. This 'compression' of the distribution that the 'no tag suggestion' generates can be defined as highly frequent tags being reinforced more and less frequent tags reinforced less or not used at all, leading to more imitation in the top of the distribution and a 'shorter' long tail. It is because of this 'compression' caused by tag suggestions that the averaged 'tag suggestion' distributions does not significantly fit power-law distributions while the averaged 'tag suggestion' distribution does fit a power-law distribution. Taking a 'scale-free' power-law as an ideal stable tag distribution, rather counter-intuitively a simple tag suggestion scheme based on frequency may actually hurt rather than help the stabilization of tagging as a power-law distribution.

5.4.2.4 Tag Suggestions Do Not Cause Tag Stabilization

This experiment provides a first step that leads to a new interpretation of the accepted theories and models that explain the emergence of power-laws in tagging systems. Common wisdom in tagging suggested that the power-law was unlikely to form without tag suggestions. As put by Marlow, Boyd, and others, "a convergent folksonomy is likely to be generated when tagging is not blind," blind tagging being tagging without tag suggestions Marlow et al (2006b). The results show that the tags of users *without* tag suggestions converge into a power-law distribution. Moreover, a power-law function fits *more closely* the behavior of users when the users are *not* given tag suggestions than when the users are given tag suggestions. This means that tag suggestions distorts the power-law function that would already naturally occur when users tag blindly without tag suggestions. These results are not unexpected. After all, *words in natural language naturally follow a power-law*, and there exists purely information-theoretic arguments why this is the case Mandelbrot (1953).

This helps clarify a number of experimental results from previous experiments in tagging. First, this result clarifies how the power-law distribution was observed by Cattuto et al. even before del.icio.us began using tag suggestion via the tag interface **?**. Second, it also helps explain how the majority of users in Suchanek et al.'s experiment had a high matching rate, even when in their report-back most of them said they didn't use or even notice tag suggestions Suchanek et al (2008). Our experiment does have a number of limitations, in particular our experiment should be extended to deal with more web-pages as well as expert and non-expert users dealing with different kinds of expert subject matters. In this situation, tag suggestions may have more of an influence on tagging behavior. Although the presented results indicate that some of the previous assumptions underlying the emergence of power-laws do not hold, a power-law distribution alone does not provide the necessary information needed to determine the role of tag suggestion on tag behavior. One line of research that seems promising is to understand how human categorize in general, which could easily influence how they decide which tags to use to annotate web-pages. While the large amount of tagging data on the web made it easy to de-

velop simple mathematical models of human behavior, it seems that a more detailed understanding of what users are *actually* doing is needed, the role of language in the use of the Web by human agents. Therefore, we need to inspect the collective use of language in tags more thoroughly to get a grasp of what is occuring with tagging systems as a kind of sense.

5.5 Constructing Tag Correlation Graphs

While earlier we have discovered the kinds of of tag frequency distributions that emerge from the collective tagging actions of individual users, as well as the dynamics of this process of sense-making, we have come into a key problem. If the tag stabilization simply reflects the large-scale dynamics of English language usage, then the result is not very surprising. However, tags are often domain specific terms, and thus may not actually reflect English language use. Therefore, it would be uesful to see if ay latent structure could be extracted from the stabilized tag distributions, and if those latent structures reflected the domain-specific organization of information. We look at one of the most simple latent structures that can be derived through collaborative tagging: inter-tag correlation graphs (or, perhaps more simply, "folksonomy graphs") . We discuss the methodology used for obtaining such graphs and then illustrate our approach through an example domain study.

5.5.1 Methodology

The act of tagging resources by different users induces, at the tag level, a simple distance measure between any pair of tags. This distance measure captures a degree of co-occurrence which we interpret as a similarity metric, between the content represented by the two tags. The collaborative filtering Sarwar et al (2001); Robu and Poutré (2006) and natural language processing Manning and Schutze (2002) literature proposes several distance or similarity measures that can be employed for such problems. The metric we found most useful for this problem is *cosine distance*. Note that this should not be interpreted as a conclusion on our part that cosine distance is always an optimal choice for this problem. This issue probably requires further research on larger data sets.

Formally, let T_i, T_j represent two random tags. We denote by $N(T_i)$ and $N(T_j)$ respectively the number of times each of the tags was used individually to tag all resources, and by $N(T_i, T_j)$ the number of times two tags are used to tag the same resource. Then the similarity between any pair of tags i and j is defined as:

$$similarity(T_i, T_j) = \frac{N(T_i, T_j)}{\sqrt{N(T_i) * N(T_j)}} \tag{5.12}$$

In the rest of the paper, we use the shorthand: sim_{ij} to denote $similarity(T_i, T_j)$. From these similarities we can construct a tag-tag correlation graph or network, where the nodes represent the tags themselves weighed by their absolute frequencies, while the edges are weighed with the cosine distance measure. We build a visualization of this weighed tag-tag correlation, by using a "spring-embedder" or "spring relaxation" type of algorithm. We tested two such algorithms: Kawada-Kawai and Fruchterman-Reingold Batagelj and Mrvar (1998); the two graphs included in this paper are based on the latter. An analysis of the structural properties of such tag graphs may provide important insights into both how people tag and how structure emerges in collaborative tagging.

5.5.2 Constructing the tag correlation (folksonomy) graphs

In order to exemplify our approach, we collected the data and constructed visualizations for a restricted class of 50 tags, all related to the tag "complexity." Our goal in this example was to examine which sciences the user community of del.icio.us sees as most related to "complexity" science, a problem which has traditionally elicited some discussion. The visualizations were made on Pajek Batagelj and Mrvar (1998). The purpose of the visualization was to study whether the proposed method retrieves connection between a central tag "complexity" and related disciplines. We considered two cases:

- Only the dependencies between the tag "complexity" and all other tags in the subset are taken into account when building the graph (Fig. 5.12).
- The weights of all the 1175 possible edges between the 50 tags are considered (Fig. 5.13).

In both figures, the size of the nodes is proportional to the absolute frequencies of each tag, while the distances are, roughly speaking, inversely related to the distance measure as returned by the "spring-embedder" algorithm.[6] We tested two energy measures for the "springs" attached to the edges in the visualization: Kamada-Kawai and Fruchterman-Reingold Batagelj and Mrvar (1998). For lack of space, only the visualization returned by Kamada-Kawai is presented here, since we found it more faithful to the proportions in the data.

The results from the visualization algorithm match relatively well with the intuitions of an expert in the organization of content in this field. Some nodes are much larger than others which again shows that taggers prefer to use to general, heavily used tags (e.g. the tag "art" was used 25 times more than "chaos"). Tags such as "chaos", "alife", "evolution" or "networks" which correspond to topics generally seen as close to complexity science are close to it. At the other end, the tag "art" is a large, distant node from "complexity." This is not so much due to the absence of

[6] For two of the tags, namely "algorithms" and "networks," morphological stemming was employed. So both absolute frequencies and co-dependencies were summed over the singular form tag, i.e. "network" and the plural "networks," since both forms occur with relatively high frequency.

sites discussing aspects of complexity in art as there are quite a few of such sites, but instead due to the fact that they represent only a small proportion of the total sites tagged with “art,” leading to a large distance measure.

In Figure 5.13, the distances to “complexity” change significantly, due to the addition of the correlations to all other tags. However, one can observe several clusters emerging which match reasonably well with intuitions regarding the way these disciplines should be clustered. Thus, in the upper-left corner one can find tags such as “mathematics”, “algorithmics”, “optimization”, “computation”, while immediately below are the disciplines related to AI (“neural” [networks], “evolutionary” [algorithms] and the like). The bottom left is occupied by tags with biology-related subjects, such as “biology”, “life”, “genetics”, “ecology” etc, while the right-hand side consists of tags with more “social” disciplines (“markets”, “economics”, “organization”, “society” etc.). Finally, some tags are both large and central, pertaining to all topics (“research”, “science”, “information”).

We also observed some tags that are non-standard English words, although we filtered most out as not relevant to this analysis. One example is “complexsystems” (spelled as one word), which was kept as such, although the tags ”“complex” and “system” taken individually are also present in the set. Perhaps unsurprisingly, the similarity computed between the tags “complexsystems” and “complex” is one of the strongest between any tag pair in this set. One implication of this finding is that tag distances could be used to find tags that have minor syntactic variance with more well-known tags, such as “complesystems,” but which cannot simply detected by morphological stemming.

5.6 Identifying tag vocabularies in folksonomies using community detection algorithms

The previous sections analyzed the temporal dynamics of distribution convergence and stabilization in collaborative tagging as well as some latent information structures, like tag correlation (or folksonomy) graphs, that can be created from these tag distributions. In this section, we look at how these folksonomy graphs could be used to identifying shared tag vocabularies.

The problem considered in this section can be summarized as: given a heterogeneous set of tags (which can be represented as a folksonomy graph), how can we partition this set into subsets of related tags? We call this problem a “vocabulary identification” problem. It is important to note that we use the term “vocabulary” only in a restricted sense, i.e. as a collection of related terms, relevant to a specific domain. For instance, a list of tropical diseases is a “vocabulary”, a list of electronic components in a given electronic device is a vocabulary, and a list of specialized terms connected to a given scientific subfield would all be “vocabularies” in our definition. We acknowledge that this is a restricted definition the type of structural information from formal ontologies is difficult to extract only from tags, given the simple structure of folksonomies. Nevertheless, our approach could still prove use-

ful in such applications: for example, one could construct the set of related terms as a first rough step and then a human expert (or, perhaps, another [semi]-automated method) could be used to add more more detail to the extracted vocabulary set.

Note that the complexity-related disciplines data set (already introduced in Sect. 4) is a useful tool to examine this question, since the initial set of tags are heterogeneous (complexity science is, by its very nature, an interdisciplinary field), but there are natural divisions into sub-fields, based on different criteria. This allows easier intuitive interpretation of the obtained results (besides the mathematical modularity criteria described below). The technique we will use in our approach is based on the so-called "community detection" algorithms, developed in the context of complex systems and network analysis theory Newman and Girvan (2004); Newman (2004). Such techniques have been well studied at a formal level and have been used to study large-scale networks in a variety of fields from social analysis (e.g. analysis of co-citation networks), analysis of biological nets (e.g. food chains) to gene interaction networks. Newman and Girvan (2004) provide an overview of existing applications of this theory, while Newman (2004) presents a formal analysis of the algorithm class used.

5.6.1 Using community detection algorithms to partition tag graphs

In network analysis theory, a community is defined as a subset of nodes that are connected more strongly to each other than to the rest of the network. In this interpretation, a community is related to clusters in the network. If the network analyzed is a social network (i.e. vertexes represent people), then "community" has an intuitive interpretation. For example, in a social network where people who know each other are connected by edges, a group of friends are likely to be identified as a community, or people attending the same school may form a community. We should stress, however, that the network-theoretic notion of community is much broader, and is not exclusively applied to people. Some examples Newman and Girvan (2004); Jin et al (2007) are networks of items on Ebay, physics publications on arXiv, or even food webs in biology. We will use a community detection algorithm to identify "vocabularies" within a folksonomy graph, identifying "communities" as "vocabularies."

5.6.1.1 Community detection: a formal discussion

Let the network considered be represented a graph $G = (V,E)$, when $|V| = n$ and $|E| = m$. The community detection problem can be formalized as a partitioning problem, subject to a constraint. The partitioning algorithm will result in a finite number of explicit partitions, based on clusters in the network, that will considered "communities." Each $v \in V$ must be assigned to exactly one cluster $C_1, C_2, ...C_{n_C}$, where all clusters are disjoint, i.e. $\forall v \in V, v \in C_i, v \in C_j \Rightarrow i = j$.

Generally speaking, determining the optimal partition with respect to a given metric is intractable, as the number of possible ways to partition a graph G is very large. Newman (2004) shows there are more than 2^{n-1} ways to form a partition, thus the problem is at least exponential in n. Furthermore, in many real life applications (including tagging), the optimal number of disjoint clusters n_C is generally not known in advance.

In order to compare which partition is "optimal", the global metric used is *modularity*, henceforth denoted by Q. Intuitively, any edge that in a given partition has both ends in the same cluster contributes to increasing modularity, while any edge that "cuts across" clusters has a negative effect on modularity. Formally, let $e_{ij}, i, j = 1..n_C$ be the fraction of all edges in the graph that connect clusters i and j and let $a_i = \frac{1}{2}\sum_j e_{ij}$ be the fraction of the ends of edges in the graph that fall within cluster i (thus, we have $\sum_i a_i = \sum_{i,j} e_{ij} = m$).

The modularity Q of a graph $|G|$ with respect to a partition C is defined as:

$$Q(G,C) = \sum_i (e_{i,i} - a_i^2) \tag{5.13}$$

Informally, so Q is defined as the fraction of edges in the network that fall within a partition, minus the expected value of the fraction of edges that would fall within the same partition if all edges would be assigned using a uniform, random distribution. These partitions are identified as communities by Newman and Girvan (2004). In tagging, each of these partitions is identified as a vocabulary.

As shown in Newman (2004), if $Q = 0$, then the chosen partition c shows the same modularity as a random division.[7] A value of Q closer to 1 is an indicator of stronger community structure - in real networks, however, the highest reported value is $Q = 0.75$. In practice, Newman (2004) found (based on a wide range of empirical studies) that values of Q above around 0.3 indicate a strong community structure for the given network. We will return shortly to define the algorithm by which this optimal partition can actually be computed, but first some additional steps are needed to link this formal definition to our tagging domain.

5.6.2 Edge filtering step

As shown in tag graph construction step above, for our data set the initial inter-tag graph contains $\binom{50}{2} = 1225$ pairwise similarities (edges), one for each potential tag pair.

In this paper, we make the choice to filter and use in further analysis only the top $m = k_d * n$ edges, corresponding to the strongest pairwise similarities. Here, k_d is a parameter that controls the density of the given graph (i.e. how many edges are

[7] Note that Q can also take values smaller than 0, which would indicate that the chosen partition is worse than expected at random.

there to be considered vs. the number of vertexes in the graph). In practice, we take values of $k_d = 1..10$, which for the tag graph we consider means a number of edges from 500 down to 50.

5.6.3 Normalized vs. non-normalized edge weights

The graph community identification literature Newman and Girvan (2004) generally considers considers graphs consisting of discrete edges (for example, in a social network graph, people either know or do not know each other, edges do not usually encode a "degree" of friendship). In our graph, however, edges represent similarities between pairs of tags (c.f. Eq. 5.12). There are two ways to specify edge weights. The non-normalized case assigns each edge that is retained in the graph, after filtering, a weight of 1. Edges filtered out are implicitly assigned a weight of zero. The normalized case assigns each edge a weight proportional to the similarity between the tags corresponding to the ends. Formally, using the notations from Eq. 5.12 and 5.13 from above, we initialize the values e_{ij} as:

$$e_{ij} = \frac{m}{\sum_{ij} sim_{ij}} sim_{ij} \tag{5.14}$$

Where $\frac{m}{\sum_{ij} sim_{ij}}$ is simply a normalization factor, which assures that $\sum_{ij} ei_{ij} = m$.

5.6.4 The graph partitioning algorithm

Since we have established our framework, we can now formally define the graph partitioning algorithm. As already shown, the number of possible partitions for this problem is at least 2^{n-1} (e.g. for our 50 tag setting $2^{50} > 10^{15}$). Therefore, to explore all these partitions exhaustively would be clearly unfeasible. The algorithm we use to determine the optimal partition (Alg. 1) is based on Newman (2004), and it falls into the category of "greedy" clustering heuristics.

Informally described, the algorithm runs as follows. Initially, each of the vertexes (in our case, the tags) are assigned to their own individual cluster. Then, at each iteration of the algorithm, two clusters are selected which, if merged, lead to the highest increase in the modularity Q of the partition. As can be seen from lines 5-6 of Alg. 1, because exactly two clusters are merged at each step, it is easy to compute this increase in Q as: $\Delta Q = (e_{ij} + e_{ji} - 2a_i a_j)$ or $\Delta Q = 2 * (e_{ij} - a_i a_j)$ (the value of e_{ij} being symmetric). The algorithm stops when no further increase in Q is possible by further merging.

Note that it is possible to specify another stopping criteria in Alg. 1, line 9, e.g. it is possible to ask the algorithm to return a minimum number of clusters (subsets), by letting the algorithm run until n_C reaches this minimum value.

Algorithm 1 *GreedyQ Determination*: Given a graph $G = (V,E), |V| = n, |E| = m$ returns partition $< C_1, ...C_{n_C} >$

1. $C_i = \{v_i\}, \forall i = \overline{1,n}$
2. $n_C = n$
3. $\forall i, j, e_{ij}$ initialized as in Eq. 5.14
4. repeat
5. $\quad < C_i, C_j >= \text{argmax}_{c_i,c_j}(e_{ij} + e_{ji} - 2a_i a_j)$
6. $\quad \Delta Q = \max_{c_i,c_j}(e_{ij} + e_{ji} - 2a_i a_j)$
7. $\quad C_i = C_i \bigcup C_j, C_j = \emptyset$ *//merge* C_i *and* C_j
8. $\quad n_C = n_C - 1$
9. until $\Delta Q \leq 0$
10. $maxQ = Q(C_1, ..C_{n_C})$

5.6.5 Experimental results

The experimental results from applying Alg. 1 to our data set are shown in Fig. 5.15. In Fig. 5.14 we present a detailed "snapshot" of the partition obtained for one of the experimental configurations. There are several interesting results. First, it becomes clear that using normalized edge weights produces partitions with higher modularity than assigning all the top edges the same weight of 1. This was intuitively hypothesized by us, since edge weights represent additional information we can use, but it was confirmed experimentally. Second, we are clearly able to identify partitions with a modularity higher than around 0.3, which exhibit a strong community structure according to Newman and Girvan (2004). Yet perhaps the most noteworthy feature of the partitions is the rapid increase both in the modularity factor Q and in the number of partitions, as the number of edges filtered decreases (from left to right, in our figure). The filtering decision represents, in fact, a trade-off. Having too many edges in the graph may stop us from finding a partition with a reasonable modularity, due to the high volume of "noise" represented by weaker edges. However, keeping only a small proportion of the strongest edges (e.g. 100 or 50 for a 50-tag graph, in our example), may also have disadvantages, since we risk throwing away useful information. While a high modularity partition can be obtained this way, the graph may become too "fragmented": arguably, dividing 50 tags into 10 or 15 vocabularies may not be a very useful.

Note that it is difficult to establish a general rule for what a "good" or universally "correct" partition should be in this setting. For example, even the trivial partition that assigns each tag to its own individual cluster cannot be rejected as "wrong" but such a trivial partition would not be considered a useful result for most purposes. In this paper we generally report the partitions found to have the highest modularity for the setting. However, for many applications, having a partition with a certain number of clusters, or some average cluster size, may be more desirable. The clustering algorithm propose here (Alg. 1) can be easily modified to account for such desiderata, by changing the stop criteria in line 9.

Fig. 5.14 shows the solution with the highest modularity Q for a graph with 200 edges, in which 7 clusters are identified. This partition assigns tags related to mathematics and computer science to Cluster 1, tags related to social science and phenomena to Cluster 2, complexity-related topics to Cluster 4 etc., while "art" is assigned to its own individual cluster. This matches quite well our intuition, and its modularity $Q = 0.34$ is above (albeit close) to the theoretical relevance threshold of 0.3.

5.6.5.1 Eliminating tags from resulting partitions to improve modularity

The analysis in the previous section shows that community detection algorithms were able to produce useful partitions, with above-relevance modularity. Still, there are a few general-meaning tags that would fit well into any of the subsets resulting after the partition. These tags generally reduce the Q modularity measure significantly, since they increase the inter-cluster edges. Therefore, we hypothesized that the modularity of the resulting partitions could be greatly improved by removing just a few tags from the set under consideration. In order to test this hypothesis, we tested another greedy tag elimination algorithm, formally defined as Alg. 2. Result graphs are shown in Fig 5.16, while in Fig. 5.14 we show the top 5 tags that, if eliminated, would increase modularity Q from 0.34 to 0.43.

Algorithm 2 *GreedyQ Elimination*: Given a partition $C_1,...C_{n_C}$ of graph $G = (V,E)$ removes all vertexes $v_i \in V$ that increase Q

1. repeat
2. $v_i = \text{argmax}_{v_i}[Q(..,C_k \setminus \{v_i\},..) - Q(..,C_k,..)]$
3. $\Delta Q = \max_{v_i}[Q(..,C_k \setminus \{v_i\},..) - Q(..,C_k,..)]$
   where $v_i \in C_k$ *//C_k is the partition of vertex i*
4. until $\Delta Q \leq 0$

As seen in Fig. 2, for this data set only 5-6 tags need to be eliminated as eliminating more does not lead to a further increases in Q. In the example in Fig. 5.14, we see which these are, in order of elimination: theory, science, research, simulation, networks. In fact, these tags, that are marked for elimination automatically by Alg. 2, are exactly those that are the most general in meaning and would fit well into any of the subsets.

Regarding scalability, it is relatively straightforward to show that both Alg. 1 and 2 have linear running time the number of vertexes n, i.e. in this case, number of tags considered in the initial set. In the case of Alg.1, exactly two clusters of tags are merged at each step, so one cluster increases in size by a minimum of one, until the algorithm terminates. In case of Alg. 2, one tag is eliminated per step, until termination. In practice, this scalability property means they are easily applicable to analyze much larger folksonomy systems.

We leave some aspects open to further work. For instance, in the current approach, similarity distances between pairs of tags are computed using all the tagging instances in the data set. In some applications, it might be useful to first partition the set of users that do the tagging, and then consider only the tags assigned by a certain class of users. For example, for tags related to a given scientific field, expert taggers may come up with a different vocabulary partition than novice users. This may require a two-fold application of this algorithm: first to partition and select the set of users, and then the set of tags based on the most promising category of users.

5.7 Comparing Tags to Search Keywords

While these applications of tagging distributions have shown promise, one question that can be reasonably asked is how well these applications of tagging compare to some benchmark that does not use tagging distributions? In other words, is the notion of a Fregean sense inherently limited to only tags explicitly created in tagging systems? The most compelling other in which natural language terms are attached to URIs is that of search engines. One can consider the query terms of a user in a search engine as the implicit tagging of a resource, as is done in what has been termed 'query flow graphs' **?**. Thus, the main difference between search engine terms and tags is that in search engines natural language terms are used to discover a resource *a priori*, while tagging are terms attached to a resource *post-hoc*. Regardless, this also means that the Fregean notion of a sense does not have to be confined to the collective tags attached to a resource, but can include search terms as well. However, as the data for the stabilization of search terms is not publically available like tagging systems, for the time being we will have to compare tagging to search terms using the more limited correlation graph techniques.

The idea of approximating semantics by using search engine data has, in fact, been proposed before, and is usually found in existing literature under the name of "Google distance." Cilibrasi and Vitanyi (2007) were the first to introduce the concept of "Google distance" from an information-theoretic standpoint, while other researchers Gligorov et al (2008) have recently proposed using it for tasks such as approximate ontology matching. It is fair to assume (although we have no way of knowing this with certainty), that current search engines and related applications, such as Google Sets http://labs.google.com/sets (2008), also use text or query log mining techniques (as opposed to collaborative tagging) to solve similar problems.

There are two ways of comparing terms (in this case, keywords) using a search engine. One method would be to compare the number of resources that are retrieved using each of the keywords and their combinations. Another method is to use the query log data itself, where the co-occurrence of the terms in the same queries vs. their individual frequency is the indicator of semantic distance. We employ this latter method as it is more amendable to comparison with our work on tagging. In the latter method, the query terms are comparable to tags, where instead of basing our folksonomy graphs and vocabulary extraction on tags, we used query terms. In gen-

eral, query log data is considered proprietary and much more difficult to obtain than tagging data. We were fortunate to have access to a large-scale data set of query log data, from two separate proposals awarded through Microsoft's "Beyond Search" awards. In the following we describe our methodology and empirical results.

5.7.1 Data set and methodology employed

The data set we used consists of 101,000,000 organic search queries, produced from Microsoft search engine Live.com, during a 3-month interval in 2006. Based on this set of queries, we computed the bilateral correlation between all pairs from the set of of complexity related terms considered in Sect. 5.5 and 5.6 above. The set of terms are, however, no longer treated as tags, but as search keywords.[8] The correlation between any two keywords T_i and T_j is computed using the cosine distance formula in Equation 5.12 from Section 5.5 above. However, here $N(T_i, T_j)$ represents the number of queries in which the keywords T_i and T_j appear in together, while $N(T_i)$ and $N(T_j)$ are the numbers of queries in which T_i, respectively T_j appear in total (irrespective of other terms in the query), from the 100 million queries in the data set.

The rest of the analysis mirrors closely the steps described in Sections 5.5 and 5.6, but optimizing the learning parameters which best fit this data set, in order to give both methods a fair chance in the comparison. More specifically, the Pajek visualization of the keyword graphs in Figs.5.17 and 5.18 were also built by using a spring-embedder algorithm based on the Kamada-Kawai distance, while Fig. 5.19 shows the keyword vocabulary partition that maximizes the modularity coefficient Q in the new setting, considering the top 200 edges. For clarity, the graph pictures are depicted in a different color scheme, to clearly show they result from entirely different data sets: Figures 5.12 and 5.13 from del.icio.us collaborative tagging data, and Figures 5.17 and 5.18 from Microsoft's Live.com query logs.

5.7.2 Discussion of the results from the query log data and comparison

When comparing the graphs in Figures 5.12 and 5.17 (i.e. the ones which only depict the relations to the central term "complexity") an important difference can be observed. While the graph in Fig. 5.12, based on collaborative tagging data, shows 48 terms related to complexity, the one is Fig. 5.17, based on query log data, shows just 6. The basic reason is that no relationship between the term "complexity" and

[8] We acknowledge this method has some drawbacks, as a few terms in the complexity-related set, such as "powerlaw" and "complexsystems" (spelled as one word) or "alife" (for "artificial life") are natural to use as tags, but not very natural as search keywords. However, since there are only 3 such non-word tags, they do not significantly affect our analysis.

the other 40+ terms can be inferred from the query log data. These relationships either do not appear in the query logs or are statistically too weak (only based on a few instances).

It is important to emphasize here that this result is not an artifact of the cosine similarity measure we use. Even if we use another, more complex distance measure between keywords, such as some suggested in the previous literature Cilibrasi and Vitanyi (2007), we get very similar results. The fundamental reason for the sparseness of the resulting graph is that the query log data itself does not contain enough relevant information about complexity-related disciplines. For example, among the 101,000,000 queries, the term complexity appears exactly 138 times, a term such as "networks" 1074 times. Important terms such as "cognition" or "semantics" are even less common, featuring only 47 and 26 times respectively among more than 100 million queries. Therefore, it is fair to conclude that the query log data, while very large in size, is quite poor in useful information about the complexity-related sciences domain. As a caveat, we do note that more common terms, such as "community" (78,862 times), "information" (36,520 times), "art" (over 52.000), or even "agent" (about 7,000) do appear more frequently, but these words have a more general language usage and are not restricted to the scientific domain. Therefore, these higher frequencies do not actually prove very useful for identifying the relationship of these terms to complexity science, which was our initial target question.

Turning our attention to the second graph in Fig. 5.18 and the partition in Fig. 5.19, we can see that query logs can also produce good results in comparison with tagging, although they are somewhat different from the ones obtained from tagging. For example, if we compare the partitions obtained in Fig. 5.14 (resulting from tagging data) and the one in Fig. 5.19 (from query log data), we see that tagging produces a more precise partition of the disciplines into scientific sub-fields. For instance, it is clear from Fig. 5.14 that cluster 1 corresponds to mathematics, optimization and computation, cluster 2 to markets and economics, cluster 5 to biology and genetics, cluster 4 to disciplines very related to complexity science and so forth. The partition obtained from query log data in Fig. 5.19, while is still very reasonable, reflects perhaps how a general user would classify the disciplines, rather than a specialist: organization is related to both information, systems and community (cluster 2), research is either qualitative or quantitative (cluster 6), and the like. There are also some counter-intuitive associations, such as putting biology and markets in the same cluster (number 1). Note that the clustering (or modularity) coefficient Q is higher in Fig. 5.19 than 5.14, but this is only because there are less inter-connections between terms in general in the query log data, thus there are less edges to "cut" in the clustering algorithm.

5.8 Conclusions

To conclude, user-generated collaborative tags can serve as a Fregean *sense*. Using KL divergence, we can show that tagging distributions per resource do indeed stabi-

lize the a scale-free power law distribution, so that the 'tag cloud' of a resource after a certain point stabilizes into what is widely-accepted in a particular community to be a good description of the resource. Furthermore, this behavior of stabilization is a function of time and number of users, and does not simply reflect an artifact of the tag suggestion mechanism. Tagging can indeed be the foundation for a sense-based semantics on the Web.

Also, it seems tagging produces a richer notion of sense than search terms. This can probably be explained by the fact the del.icio.us users have more expertise and interest in complexity-related topics than general web searchers. Furthermore, they are probably more careful in selecting resources to tag and in selecting labels for them that would be useful to other users as well (general web searchers are known to be "lazy" in typing queries). As a caveat, we note that this target domain (i.e. complexity-related disciplines) is scientific and very specialized. If the target would be more general (for example, if we selected a set of terms related to pop-culture), the comparison might lead to different results. Also, people who sign up to use a collaborative tagging system are implicitly more willing to share their knowledge and expertise with a community of other users. By contrast, web search is implicitly a private activity, where not only may tracing users' actual identity may be undesirable to the user, but also the user is not even aware their activity is being tracked and the keywords they use can then be used by search engines or other programs to change the results for other users.

The question remains: while one can operationalize some notion of Fregean sense-based semantics on the Web in the form of collaborative tags, is this enough? After all, many URI are not tagged at all! Superficially, the preliminary results from search engine keyword analysis seem to show that keywords are a much sparser source of sense than tags. However, these results only were shown on a tiny group of keywords gathered from a search engine on a particular topic. To think more broadly, perhaps *all* associated keywords with a particular resource could serve as a better sense-based semantics for a URI. This may include not only the keywords from tags explicitly given to that URI and from keywords used to reach a URI, but also from the terms accessible from the web representations hosted at the URI, ranging from Semantic Web documents to hypertext web-pages. It is to this more comprehensive notion of computational sense that we turn to next.

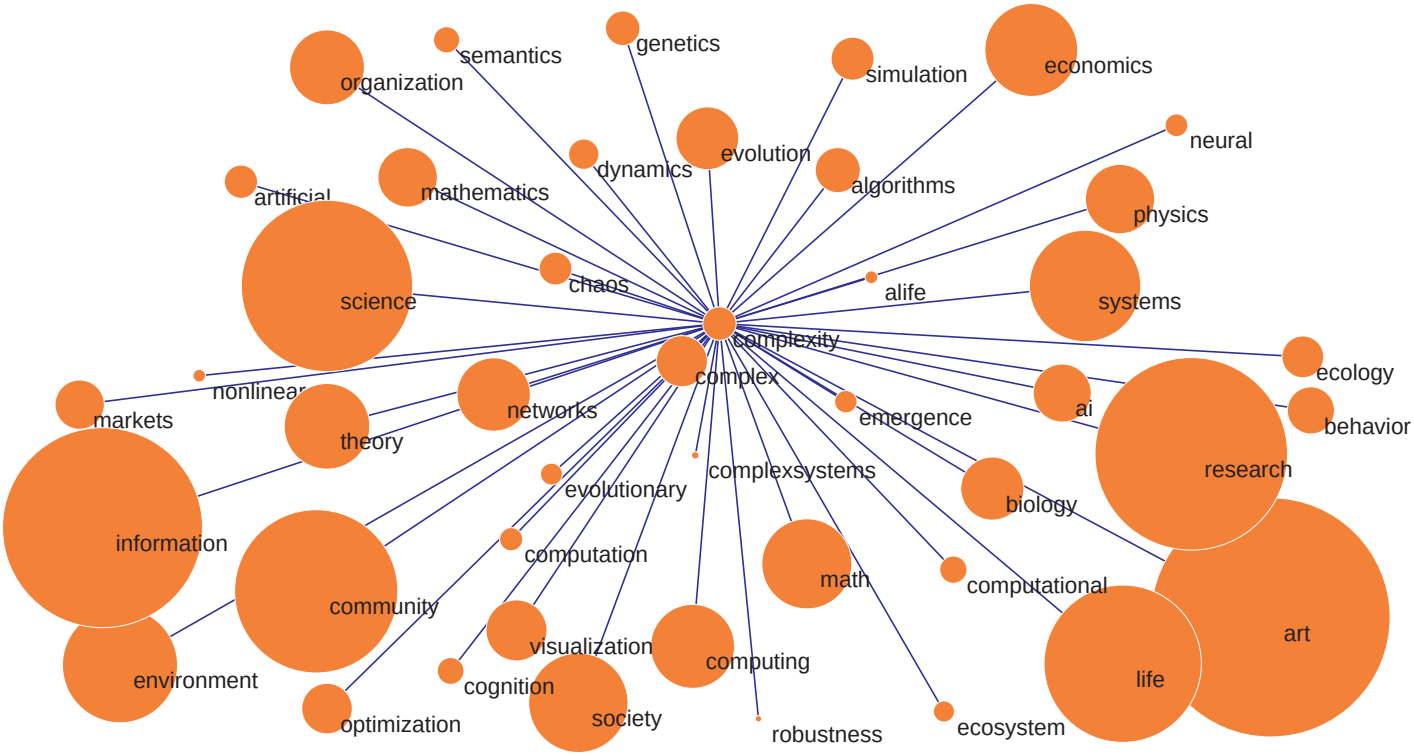

Fig. 5.12 Folksonomy graph, considering only correlations corresponding to central tag “complexity”

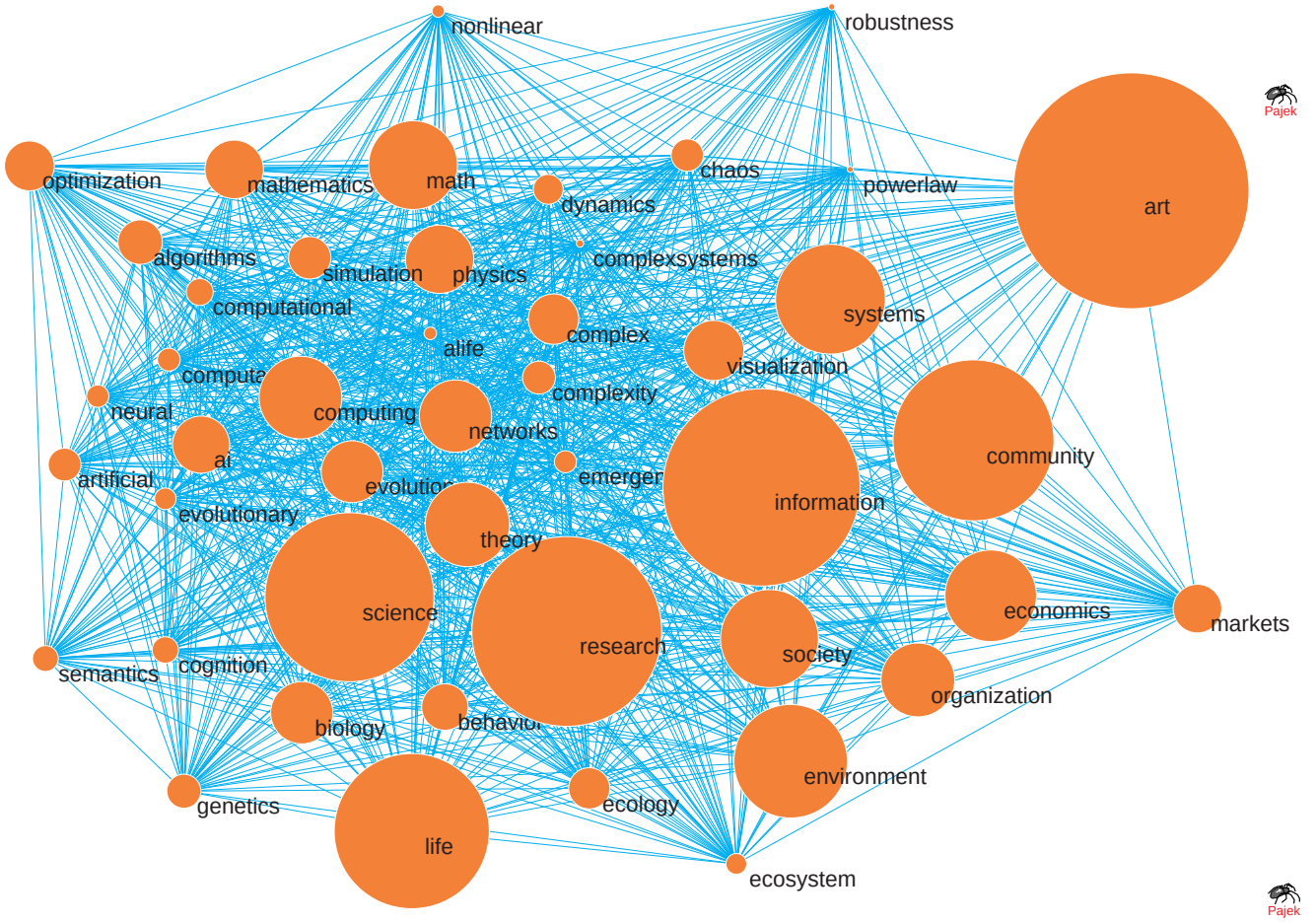

Fig. 5.13 Folksonomy graph, considering all relevant inter-tag correlations

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	
computation	markets	semantics	powerlaw	genetics	robustness	art	
optimization	economics	cognition	nonlinear	biology			
visualization	society	neural	complexsystems	evolution			
physics	community	ai	dynamics	evolutionary			
mathematics	organization	alife	chaos	science			
math	ecology	artificial	emergence				
computational	ecosystem	life	networks				
algorithms	environment	behavior	systems				
information		simulation	complex				
computing		research	complexity				
theory							
Tags that increase modularity the most, if eliminated: theory, science, research, simulation, networks.							

Fig. 5.14 Optimal partition in tag clusters (i.e. "communities") of the folksonomy graph, when the top 200 edges are considered. This partition has a Q=0.34. After eliminating the 5 tags mentioned at the bottom, Q can increase to 0.43.

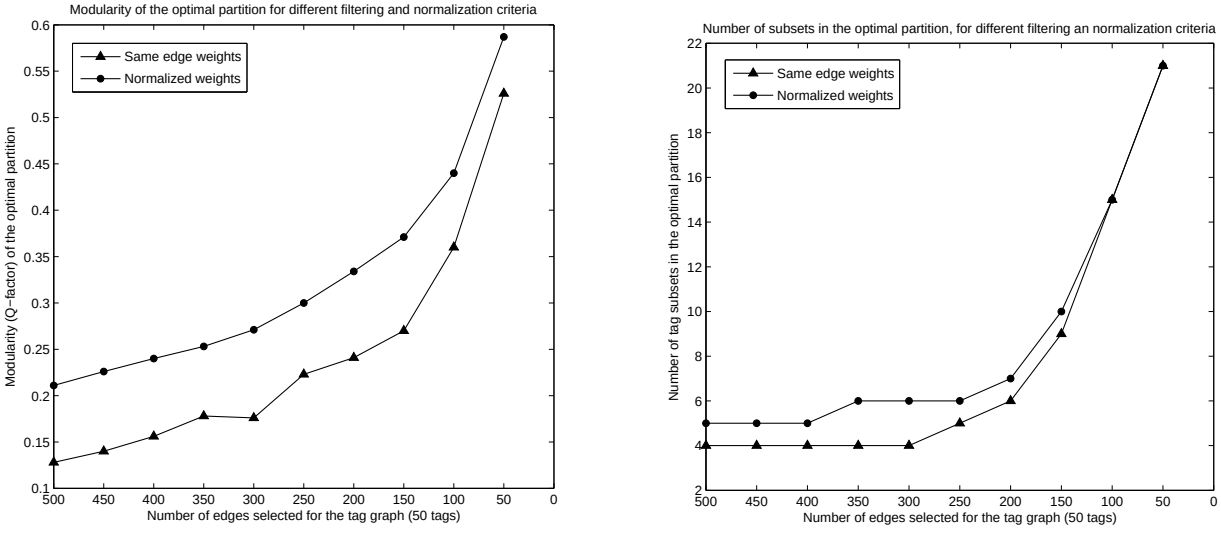

Fig. 5.15 Modularity (Q-factor) and number of partitions obtained from applying community detection algorithms to the scientific disciplines data set

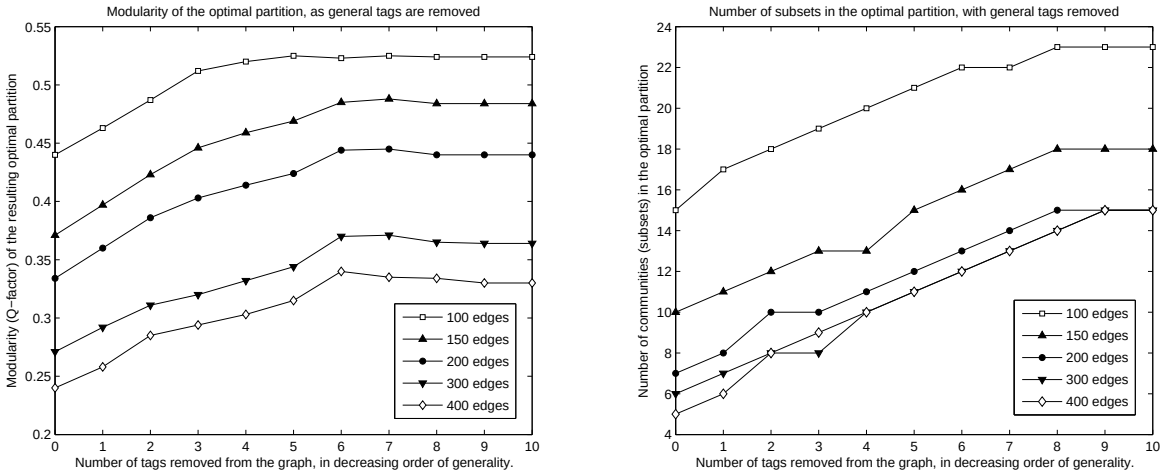

Fig. 5.16 Modularity (Q-factors) and number of partitions obtained after gradually eliminating tags from the data set, such as to increase the modularity. At each step, the tag that produced the highest increase in modularity between the initial and resulting partition was selected. In these results, all edge weights are normalized.

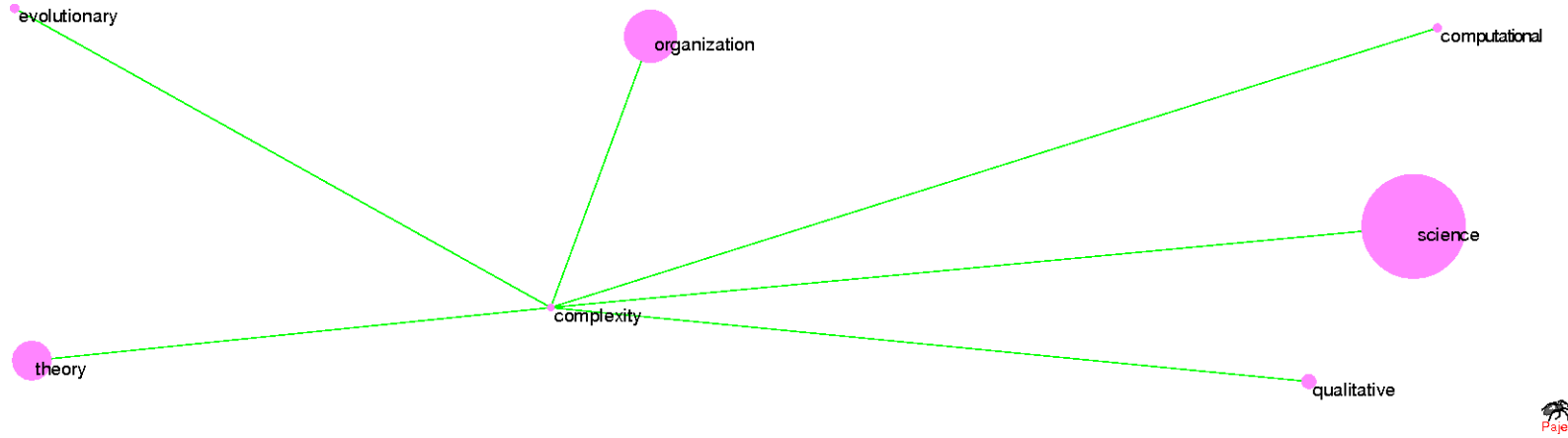

Fig. 5.17 Correlation graph from Microsoft queries, showing only correlations to the term “complexity”.

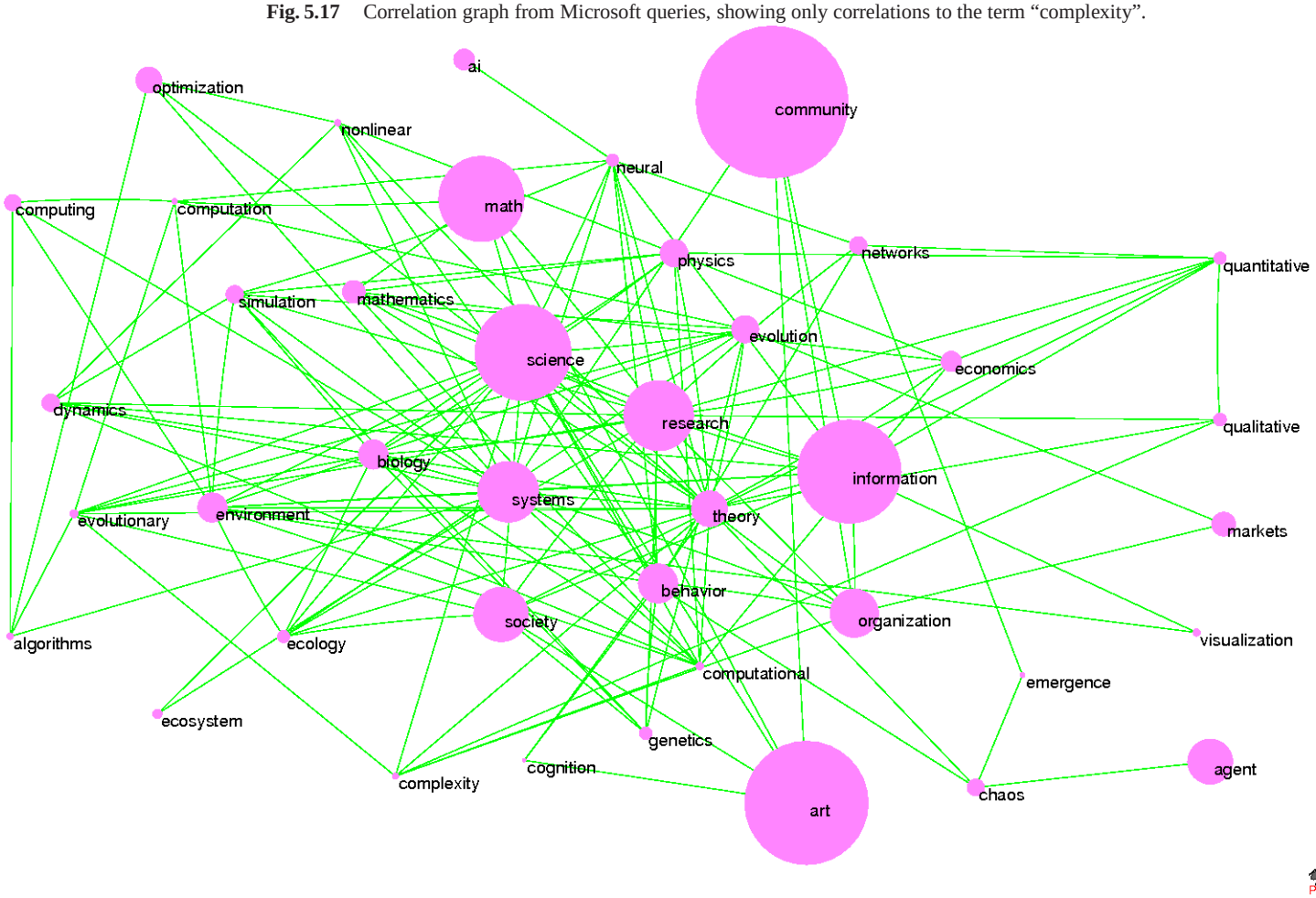

Fig. 5.18 Correlation graph obtained from Microsoft query logs, considering all relevant search terms.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	
complexity	systems	networks	algorithms	mathematics	research	
evolution	visualization	ai	ecology	physics	quantitative	
evolutionary	organization	emergence	math	economics	qualitative	
chaos	information	neural	computing	art	society	
cognition	community		optimization	science		
biology			computation	simulation		
theory			environment	dynamics		
behavior				nonlinear		
markets				computational		
genetics				ecosystem		
agent						
Terms left unclassified (i.e. one word clusters): complex, complexsystems, robustness, multi-agent, life, artificial, semantics, powerlaw, alife.						

Fig. 5.19 Optimal partition into clusters, obtained from the Microsoft query data, when the top 200 edges are considered. The resulting partition has a Q=0.536. However, 9 terms were assigned to their own cluster, thus basically left unclassified.

Chapter 6
The Semantics of Search

The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of implementations are rather sketchy **R.M. Duck-Lewis** (Hirst, 2000)

6.1 Introduction

What kinds of information should be used in the construction of the sense of a resource? Given our previous work, there appears to be a priori reason why we should confine ourselves to tags when constructing the sense of a resource. Up till now, we have been considering the sense-based semantics of a particular URI in terms of a term frequency distribution. However, this seems limited. There is always the case of co-referential URIs, where a single resource is identified by multiple URIs. Should the semantics somehow combined the distributions of the various Web representations? If so, precisely how - and in particular if the web representations are in multiple encodings? If one wanted the most thorough description of a resource, would it not make sense to define the semantics of these representations in terms of as many representations as possible, as it is well-known in statistical machine-learning that there's 'no data like more data,' such that simply adding more data under the right conditions can increase the likelihood of a stable and rich distributional semantics **?**.

Yet the intuition that simply adding more representations to the sense will increase its effectivness needs to be operationalized and tested. A number of difficult questions immediately appear, such as how to identify possibly co-referential URIs for the same reosurce? Or to make matters worse, how to limit the kinds of encodings that the sense will be constructed with? These questions can be answered by attempting to fit the intuition within a well-understood experimental paradigm, which we believe can be the well-studied paradigm of information retrieval. To extend further, ***relevance feedback*** is the *use of explicit relevance judgments from users of a query in order to expand the query*. By 'expand the query,' we mean that

the usually rather short query is expanded into a much larger query by adding words from known relevant documents. For example, a query on the hypertext Web for the Eiffel Tower given as 'eiffel' might be expanded into 'paris france eiffel tour.' If the relevant pages instead were about an Eiffel Tower replica in Texas, the same results query could be expanded into 'paris texas eiffel replica.' The same principle applies to the Semantic Web, except that the natural language terms may include Semantic Web URIs and terms resulting from inference or URI processing. The hypothesis of relevance feedback, as pioneered by Rocchio in the SMART retrieval system, is that the relevant documents will disambiguate and in general give a better description of the information need of the query than the query itself Rocchio (1971). Relevance feedback has been shown in certain cases to improve retrieval performance significantly. Extending this classical work, *textbfrelevance models*, as formalized by Lavrenko et al. Lavrenko (2008)) create relevance models directly from the indexed documents rather than explicitly waiting for the user to make a relevance judgment. Relevance models are especially well-suited to our hypothesis that multiple kinds of encodings should be part of the same sense, as relevance models consider each source of data (query, documents, perhaps even tags and Semantic Web data) as 'snapshots' from some underlying generative model.

Since we will use representations from different sources of data, we cannot simply contain the notion of resource to a single URI, as currently - as content negotiation amongst various encodings is currently barely deployed on the Web - hypertext web-pages and Semantic Web documents encoded in RDF without exception almost always have different URIs. However, a web-page for the Eiffel Tower encoded in HTML and a Semantic Web document encoded in RDF can still share the same content of the Eiffel Tower, despite having differing URIs. So, the information pertaining to a resource will be spread amongst multiple co-referntial URIs. Therefore, the best way to determine the set of URIs relevant to a particular resource is to attach the resource to the *information need* of a ordinary web user as expressed by a query in a search engine. Then the next step is to have humans judge a set of web representations - either Semantic Web documents, hypertext web document, or both - and consider the set of these web representations and attendant URIs to be a partial snapshot of the relevant information pertaining to a sense.

This technique can be transformed into a testable hypothesis; the hypothesis put forward by Baeza-Yates that search on the Semantic Web can be used to improve traditional ad-hoc information retrieval for hypertext Web search engines and vice-versa Baeza-Yates (2008). Currently, there exist several nascent Semantic Web search engines that specifically index and return ranked Linked Data in RDF in response to keyword queries. Yet their rankings are much less well-studied than hypertext Web rankings, and so are thought likely to be sub-optimal. While we realize the amount and sources of structured data on the Web are huge, to restrict and test the hypothesis of Baeza-Yates, from hereon we will assume that 'semantic search' refers to indexing and retrieving of Linked Data by search engines like Sindice and FALCON-S Cheng et al (2008), and hypertext search refers to the indexing and retrieval of hypertext documents on the World Wide Web by search engines like Google and Yahoo! Search. Our experimental hypothesis is that the statistical se-

mantics of sense created from Semantic Web documents can help hypertext search and vice versa, and this can be empirically shown via the use of relevance feedback.

On an aside, we realize that our reduction of 'semantic search' to keyword-based information retrieval over the Semantic Web is very restrictive, as many people use 'semantic search' to mean simply search that relies on anything beyond surface syntax, including the categorization of complex queries Baeza-Yates and Tiberi (2007) and entity-recognition using Semantic Web ontologies Guha et al (2003). We will not delve into an extended explanation of the diverse kinds of semantic search, as surveys of this kind already exist Mangold (2007). Yet given the relative paucity of publicly accessible data-sets about the wider notion of semantics and the need to start with a simple rather than complex paradigm, we will restrict ourselves to the Semantic Web and assume a traditional, keyword-based ad-hoc information retrieval paradigm for both kinds of search, leaving issues like complex queries and natural language semantics for future research. Keyword search consisting of 1-2 terms should also be explored as it is the most common kind of query in today's Web search regardless of whether any results from this experiment can generalize to other kinds of semantic search Silverstein et al (1999). In order to thoroughly test our system, Until recently semantic search suffered from a lack of a thorough and neutral Cranfield-style evaluation, and so we carefully explain and employ the traditional information retrieval evaluation frameworks in our experiment to evaluate semantic search. At the time of the experiment, our evaluation was the first Cranfield-style evaluation for searching on the Semantic Web. This evaluation later generalized into the annual 'Semantic Search' competition,[1] which has since become a standard evaluation for search over RDF data Blanco et al (2011). However, our particular evaluation presented here is still the only evaluation to determine relevance judgments over both hypertext and RDF using the same set of queries.

In Section 6.2 we first elucidate the general nature of search from hypertext documents to *semantic search* over Semantic Web documents. A general open-domain collection of user queries from a real hypertext query-log against the Semantic Web and then have human judges construct a 'gold-standard' collection of queries and results judged for relevance, from both the Semantic and hypertext Web. Then in Section 6.3 we give a brief overview of information retrieval frameworks and ranking algorithms. While this section may be of interest to Semantic Web researchers unfamiliar with such techniques, information retrieval researchers may wish to proceed immediately past this section. Our system is described in Section 6.4. In Section 6.5, these techniques are applied to the 'gold standard' collection created in Section 6.2 so that the best parameters and algorithms for relevance feedback for both hypertext and semantic search can be determined. In Section 6.6 and Section 6.7 the effects of using pseudo-feedback and Semantic Web inference are evaluated. The system is evaluated against 'real-world' deployed systems in Section 6.8. Finally, in Section 6.9 future work on this particular system is detailed, and conclusion on the veracity of our method of sense-making are given in Section 6.10.

[1] Sponsored by Yahoo! Research for both 2010 and 2011.

6.2 Is There Anything Worth Finding on the Semantic Web?

In this section we demonstrate that the Semantic Web does indeed contain information relevant to ordinary users by sampling the Semantic Web according to a real-world queries referring to entities and concepts from the query log of a major search engine. The main problem confronting of any study of the Semantic Web is one of *sampling*. As almost any large-data database can easily be exported to RDF, statistics demonstrating the actual deployment of the Semantic Web can be biased by the automated release of large, if useless, data-sets, the equivalent of 'Semantic Web' spam. Also, large specialized databases like Bio2RDF can easily dwarf the rest of the Semantic Web in size. A more appropriate strategy would be to try to answer the question: What information is available on the Semantic Web that users are actually interested in? The first large-scale analysis of the Semantic Web was done via an inspection of the index of Swoogle by Ding and Finin Ding and Finin (2006). The primary limitation of that study was that the large majority of the Semantic Web resources sampled did not contain rich information that many people would find interesting. For example, the vast majority of data on the Semantic Web in 2006 was Livejournal exporting every user's profile as FOAF and RSS 1.0 data that used Semantic Web techniques to structure the syntax of news feeds. Yet with information-rich and interlinked databases like Wikipedia being exported to the Semantic Web, today the Semantic Web may contain information needed by actual users. As there is no agreed-upon fashion to sample the Semantic Web (and the entire Web) in a fair manner, we will for our evaluation create a sample driven by queries from real-users using easily-accessible search engines that claim to have a Web-scale index, although independent verification of this is difficult if not impossible.

6.2.1 Inspecting the Semantic Web

In order to select real queries from users for our experiment, we used the query log of a popular hypertext search engine, the Web search query log of approximately 15 million distinct queries from Microsoft Live Search. This query log contained 6,623,635 unique queries corrected for capitalization. The main issue in using a query log is to get rid of navigational and transactional queries. A straightforward gazetteer-based and rule-based named entity recognizer was employed to discover the names of people and places Mikheev et al (1998), based off a list of names maintained by the Social Security Administration and a place name database provided by the Alexandria Digital Library Project. From the query log a total of 509,659 queries were identified as either (fundamentally analog) people or places by the named-entity recognizer, and we call these queries *entity queries*. Employing WordNet to represent abstract concepts, we chose queries recognized by WordNet that have *both* a hyponym and hypernym in WordNet. This resulted in a more restricted

16,698 queries that are supposed to be about abstract concepts realized by multiple entities, which we call *concept queries*.

A sample entity query from our list would be 'charles darwin,' while a sample concept query would be 'violin.' In our data-set using hypertext search, both queries return almost all relevant results. The query 'charles darwin' gives results that are entirely encyclopedia pages (Wikipedia, eHow, `darwin-online.org.uk`) and other factual sources of information, while 'violin' returns 8 out of 10 factual pages, with 2 results just being advertisements for violin makers. On the contrary for the Semantic Web, the query 'charles darwin' had 6 relevant results, with the rest being for places such as the city of Darwin and books or products mentioning Darwin. For 'violin,' only 3 contain relevant factual data, with the rest being the names of albums called 'Violin' and movies such as 'The Violin Maker.' From inspection of entities with relevant results, it appears the usual case for semantic search is that DBpedia and WordNet have a substantial amount of overlap in the concepts to which they give URIs. For example, they have distinct URIs for such concepts as 'violin' (`http://dbpedia.org/resource/Violin` vs. W3C WordNet's `synset-violin-noun-1`). Likewise, most repetition of entity URIs comes from WordNet and DBpedia, both of which have distinct URIs for famous people like Charles Darwin. In many cases, these URIs do not always appear at the top, but in the second or third position, with often an irrelevant URI at top. Lastly, much of the RDF that is retrieved seems to have little information in it, with DBPedia and WordNet being the most rich sources of information.

The results of running the selected queries against a Semantic Web search engine, FALCON-S's Object Search Cheng et al (2008), were surprisingly fruitful. For entity queries, there was an average of 1,339 URIs (S.D. 8,000) returned for each query. On the other hand, for concept queries, there were an average of 26,294 URIs (S.D. 14,1580) returned per query, with no queries returning zero documents. Such a high standard deviation in comparison to the average is a sure sign of a non-normal distribution such as a power-law distribution, and normal statistics such as average and standard deviation are not good characteristic measures of such distributions. As shown in Figure 6.1, when plotted in logarithmic space, both entity queries and concept queries show a distribution that is heavily skewed towards a very large number of high-frequency results, with a steep drop-off to almost zero results instead of the characteristic long tail of a power law. For the vast majority of queries, far from having no information, the Semantic Web of Linked Data appears to have *too much data*, but for a minority of queries there is just *no data*. This is likely the result of the releasing of Linked Data in large 'chunks' from data-silos about specific topics rather than the more organic development of the hypertext Web that typically results in power-law distributions. Also, note that hypertext web-pages are updated as regards trends and current events much more quickly than the relatively slow-moving world of Linked Data.

Another question is whether or not there is any correlation between the amount of URIs returned from the Semantic Web and the popularity of the query. As shown by Figure 6.2, there is *no* correlation between the amount of URIs returned from the Semantic Web and the query popularity. For entity queries, the correlation co-

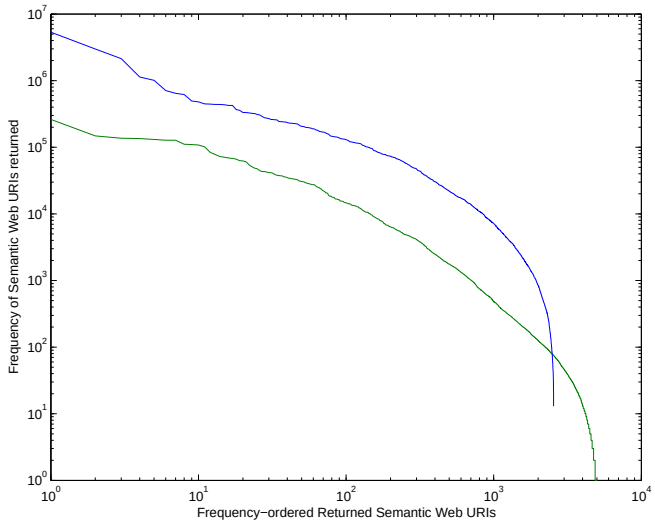

Fig. 6.1 The rank-ordered frequency distribution of the number of URIs returned from entity and concept queries, with the entity queries given by green and the concept queries by blue.

efficient was 0.0077, while for concept queries, the correlation coefficient was still insignificant, at 0.0125. The popularity of query is not related to how much information the Semantic Web possesses on the information need expressed by the query: Popular queries may have little data, while infrequent queries may have a lot. This is likely due to the rapidly changing and event-dependent nature of hypertext Web queries versus the Semantic Web's preference for more permanent and less temporally-dependent data. For a more full exploration of the data-set used in this experiment, including types of URIs, see the paper on 'A Query-Driven Characterization of Linked Data' Halpin (2009a). Since this data was collected in spring of 2009 it may not be currently accurate as a characterization of either FALCON-S or the state of Linked Data currently, but for evaluation purposes this sample should suffice, and using random selections from a real human query log is a definite advance, as randomly sampling all of Linked Data would result in an easily biased evaluation, away from what human users are interested in and towards what happens to be available as Linked Data.

Surprisingly, there is a large amount of information that may be of interest to ordinary hypertext users on the Semantic Web, although there is no correlation between the popularity of queries and the availability of that information on the Semantic Web. The Semantic Web is not irrelevant to ordinary users as there is data on the Semantic Web ordinary users are interested in, even if it is distributed unevenly and does not correlate with the popularity of their queries.

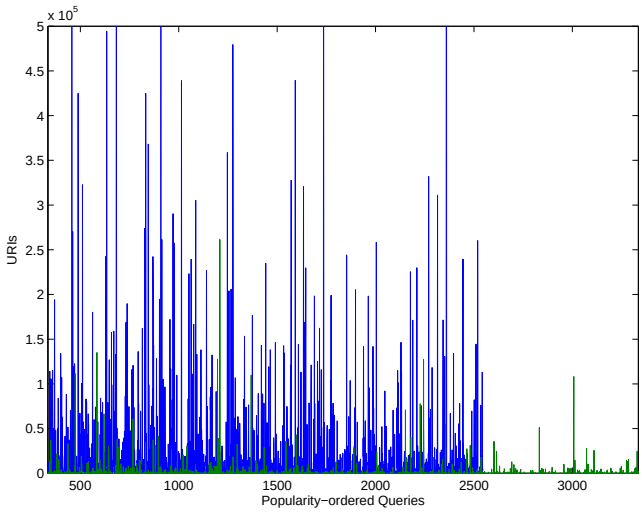

Fig. 6.2 The rank-ordered popularity of the queries is on the *x*-axis, with the *y* axis displaying the number of Semantic Web URIs returned, with the entity queries given by green and the concept queries by blue.

6.2.2 Selecting Queries for Evaluation

In order to select a subset of informational queries for evaluation, we randomly selected 100 queries identified as abstract concepts by WordNet and then 100 queries identified as either people or places by the named entity recognizer, for a total of 200 queries to be used in evaluation. Constraints were placed on the URIs resulting from semantic search, such that at least 10 Semantic Web documents (a file containing a valid RDF graph) had to be retrieved from the URI returned by the Semantic Web search engine. This was necessary as some queries returned zero or less than 10 URIs, as explained in Section 6.2.1. For each query, hypertext search always returned more than 10 URIs. So for each query, 10 Semantic Web documents were retrieved using the FALCON-S Object Search engine Cheng et al (2008), leading to a total of 1,000 Semantic Web documents about entities and 1,000 Semantic Web documents about concepts, for a total of 2,000 Semantic Web documents for relevance judgments. Then, the same experimental query log was used to retrieve pages from the hypertext Web using Yahoo! Web search, resulting in the same number of web-pages about concepts and entities (2,000 total) for relevance judgments. The total number of all Semantic Web documents and hypertext web-pages gathered from the queries is 4,000.

The queries about entities and concepts are spread across quite diverse domains, ranging from entities about both locations (El Salvador) and people (both fictional such as Harry Potter and non-fictional such as Earl May) to concepts ranging over a

large degree of abstraction, from sociology to ale. A random selection of ten queries from the entity and concept queries is given in Table 1. This set of 4,000 hypertext web-pages and Semantic Web documents are then used to evaluate our results in Section 6.5.

Entity	Concept
ashville north carolina	sociology
harry potter	clutch
orlando florida	telephone
ellis college	ale
university of phoenix	pillar
keith urban	sequoia
carolina	aster
el salvador	bedroom
san antonio	tent
earl may	cinch

Table 6.1 10 Selected Entity and Concept Queries

6.2.3 Relevance Judgments

For each of the 200 queries selected in Section 6.2.2, 10 hypertext web-pages and 10 Semantic Web documents need to be judged for relevance by three human judges, leading to a total of 12,000 judgments for relevance for our entire experiment, with the correct relevance determined by 'voting' amongst the three judges per document. Human judges each judged 25 queries presented in a randomized order, and were given a total of 3 hours to judge the entire sample for relevancy. No researchers were part of the rating. The judges were each presented first with ten hypertext web-pages and then with ten Semantic documents that could be about the same query. Before starting judging, the judges were given instructions and trained on 10 sample results (5 web-pages and 5 Semantic Web documents). The human judges were forced to make binary judgments of relevance, so each result must be either relevant or irrelevant to the query. They were given the web-page selected by the human user from the query log as a 'gold standard' to determine the meaning of the keyword.

The standard TREC definition for relevance is "If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant" Hawking et al (2000). As semantic search is supposed to be about entities and concepts rather than documents, semantic search needs an definition of relevance based around information about entities or concepts independent of documents. In one sense, this entity-centric relevance should have both a wider remit than document-centric relevance definition, as any information about the entity that could be relevant should be included. Yet in another sense, this

definition is more restrictive, as if one considers the world (perhaps fuzzily) partitioned into distinct entities and concepts, then merely related information would not count. In the instructions, relevance was defined *as whether or not a result is about the same thing as the query, which can be determined by whether or not accurate information about the information need is expressed by the result.* The following example was given to the judges: "Given a query for 'Eiffel Tower,' a result entitled 'Monuments in Paris' would likely be relevant if there was information about the Eiffel Tower in the page, but a result entitled 'The Restaurant in the Eiffel Tower' containing only the address and menus of the restaurant would not be relevant."

Kinds of Web results that would ordinarily be considered relevant are therefore excluded. In particular, there is a restriction that the relevant information must be present in the result itself. This excludes possibly relevant information that is accessible via outbound links, even a single link. All manner of results that are collections of links are thus excluded from relevancy, including both 'link farms' purposely designed to be highly ranked by page-rank based search engines, as well as legitimate directories of high-quality links to relevant information. These hubs are excluded precisely because the information, even if it is only a link transversal away, is still not directly present in the retrieved result. By this same principle, results that merely redirect to another resource via some method besides the standardized HTTP methods are excluded, since a redirection can be considered a kind of link. They would be considered relevant only if additional information was included in the result besides the redirection itself.

In order to aid the judges, a Web-based interface was created to present the queries and results to the judges. Although an interface that presented the queries and the search interface in a manner similar to search engines was created, human judges preferred an interface that presented them the results for judgments one-at-a-time, forcing them to view a rendering of the web-page associated with each URI originally offered by the search engine. For each hypertext web-page, the web-page was rendered using the Firefox Web Browser and PageSaver Pro 2.0. For each Semantic Web document, the result was rendered (i.e. the triples and any associated text in the subject) by using the open-source Disco Hyperdata Browser with Firefox.[2] In both cases, the resulting rendering of the Web representation was saved at 469×631 pixel resolution. The reason that the web-page was rendered instead of a link given directly to the URI is because of the unstable state of the Web, especially the hypertext Web. Even caching the HTML would have risked losing much of the graphic element of the hypertext Web. By creating 'snapshot' renderings, each judge at any given time was guaranteed to be presented with the result in the same visual form. One side-effect of this is that web-pages that heavily depend on non-standardized technologies or plug-ins would not render and were thus presented as blank screen shots to the user, but this formed a small minority of the data. The user-interface divided the evaluation into two steps:

[2] The Disco Hyperdata Browser, a browser that renders Semantic Web data to HTML, is available at `http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/`.

- *Judging relevant results from a hypertext Web search:* The judge was given the search terms created by an actual human user for a query and an example relevant web-page whose full snapshot could be viewed by clicking on it. A full rendering of the retrieved web-page was presented to the user with its title and summary (as produced by Yahoo! Search) easily viewed by the judge as in Figure 6.3. The judge clicked on the check-box if the result is considered relevant. Otherwise, the web-page was by default recorded as not relevant. The web-page results were presented to the judge one at a time, ten times for each query.
- *Judging relevant results from a Semantic Web search:* Next, the judge assessed all the Semantic Web results for relevancy. These results were retrieved from the Semantic Web using the same interface displayed to the judge in the first step as shown in Figure 6.4, and a title was displayed by retrieving any literal values from `rdfs:label` properties and a summary by retrieving any literal values from `rdfs:comment` values. Using the same interface as in the first step, the judge had to determine whether or not the Semantic Web results were relevant.

Fig. 6.3 The interface used to judge web-page results for relevancy.

After the ratings were completed, Fleiss' κ statistic was taken in order to test the reliability of inter-judge agreement on relevancy judgments Fleiss (1971). Simple percentage agreement is not sufficient, as it does not take into account the likelihood of purely coincidental agreement by the judges. Fleiss' κ both corrects for chance agreement and can be used for more than two judges Fleiss (1971). The null hypothesis is that the judges cannot distinguish relevant from irrelevant results, and so are judging results randomly. Overall, for both relevance judgments over Semantic Web results and web-page results, $\kappa = 0.5724$ ($p < .05$, 95% Confidence interval $[0.5678, 0.5771]$), indicating the rejection of the null hypothesis and 'moderate' agreement. For web-page results only, $\kappa = 0.5216$ ($p < .05$, 95% Confidence interval $[.5150, 0.5282]$), also indicating the rejection of the null hypothesis and 'moderate' agreement. Lastly, for only Semantic Web results, $\kappa = 0.5925$ ($p < .05$, 95%

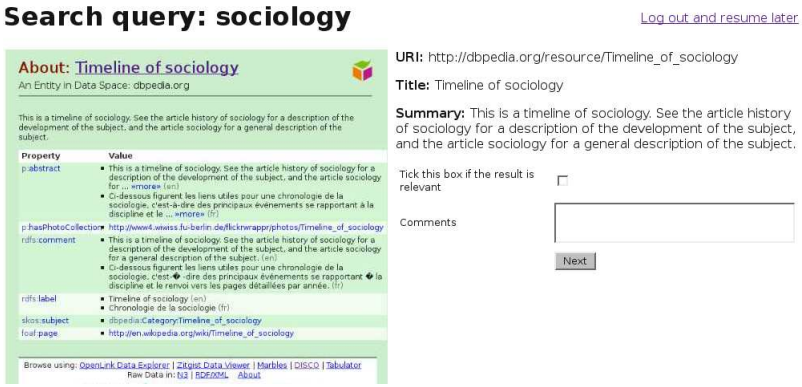

Fig. 6.4 The interface used to judge Semantic Web results for relevancy

Confidence interval $[0.5859, 0.5991]$), also indicating the null hypothesis is to be rejected and 'moderate' agreement. So, in all cases there is 'moderate' agreement, which is sufficient given the general difficulty of producing perfectly reliable relevancy judgments. Interestingly enough, the difference in κ between the web-page results and Semantic Web results show that the judges were actually *slightly* more reliable in their relevancy judgments of information from the Semantic Web rather than the hypertext Web. This is likely due to the more widely varying nature of the hypertext results, as compared to the more consistent informational nature of Semantic Web results.

Were judges more reliable with entities or concepts? Recalculating the κ for all results based on entity queries, $\kappa = 0.5989$ ($p < .05$, 95% Confidence interval $[0.5923, 0.6055]$), while for all results based on concept queries was $\kappa = 0.5447$ ($p < .05$, 95% Confidence interval $[0.5381, 0.5512]$). So it appears that judges are slightly more reliable discovering information about entities rather than concepts, backing the claim made by Hayes and Halpin that there is more agreement in general about 'less' abstract things like people and places rather than abstract concepts Hayes and Halpin (2008). However, agreement is still very similar and 'moderate' for both information about entities and concepts. It is perhaps due to the entity-centric and concept-centric definition of relevance that the agreement was not higher.

For the queries, much of the data is summarized in Table 3. ***Resolved*** queries are *queries that return at least one relevant result* in the top 10 results, while ***unresolved*** queries are *queries that return no relevant queries in the top 10 results*. 'Hypertext' means that the result was taken only over the hypertext Web results and 'Semantic Web' indicates the same for the Semantic Web results. The percentages for resolved and unresolved for 'hypertext' and 'Semantic Web' were taken over *all* the hypertext and Semantic Web relevancy corpora in order to allow direct comparison. On the contrary, the percentages for 'Top Relevant' and 'Top Non-Relevant' were com-

puted as percentages over only resolved queries, and so excludes unresolved queries. For ease of reference, a pie-chart for the hypertext relevancy is given in Figure 6.5 and for the Semantic Web relevancy in Figure 6.6.

Results:	Hypertext	Semantic Web
Resolved:	197 (98%)	132 (66%)
Unresolved:	3 (2%)	68 (34%)
Top Relevant:	121 (61%)	76 (58%)
Top Non-Relevant:	76 (39%)	56 (42%)

Table 6.2 Results of Hypertext and Semantic Web Search Relevance Judgments: Raw numbers followed by percentages. The top two row percentages are with respect to all queries, while the latter two columns are with respect to the total of resolved queries.

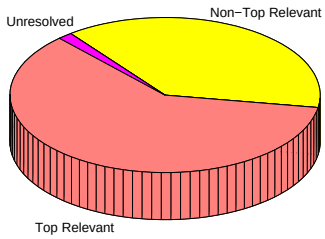

Fig. 6.5 Results of Querying the Hypertext Web.

For both hypertext and Semantic search, there were 71 (18%) unresolved queries that did not have any results. For the hypertext Web search, only 3 (2%) queries were unresolved, while 68 (34%) of the queries were unresolved for the Semantic Web. This simply means that the hypertext search engines almost always returned at least one relevant result in the top 10, but that for the Semantic Web almost a third of all queries did not return any relevant result in the top 10. This only means there is much information that does not yet have a relevant form on the Semantic Web, unless it is hidden by the perhaps sub-optimal ranking by FALCON-S.

Another question is how many queries had a relevant result as their top result? In general, 197 queries (50%) had top-ranked relevant results over both Semantic Web and hypertext search. While the hypertext Web search had 121 (61%) top-ranked relevant results, the Semantic Web only had 76 (58%) top-ranked results. What is more compelling for relevance feedback is the number of relevant results that were *not* the top-ranked result. Again for both kinds of searches, there were 132 (33.0%)

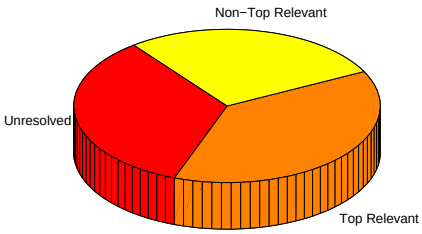

Fig. 6.6 Results of Querying the Semantic Web.

queries where a relevant result was *not* in the top position of the returned results. For the hypertext Web, there were 76 (39%) queries with a top non-relevant result. Yet for the Semantic Web there were 56 (42%) queries that had a top non-relevant result. So queries on the Semantic Web are more likely to turn up no relevant results in the top 10. When a relevant query is returned in the top 10 results it is quite likely that a non-relevant result will be in the top position for both the hypertext Web and the Semantic Web.

6.3 Information Retrieval for Web Search

In our evaluation we tested two general kinds of information retrieval frameworks: vector-space models and language models. In the *vector-space model*, document models are considered to be vectors of terms (usually called 'words' as they are usually, although not exclusively, from natural language, as we transform URIs into 'pseudo-words') where the weighing function and query expansion has no principled basis besides empirical results. Ranking is usually done via a comparison using the cosine distance, a natural comparison metric between vectors. The key to success with vector-space models tends to be the tuning of the parameters of their weighing function. While fine-turning these parameters has led to much practical success in information retrieval, the parameters have little formally-proven basis but are instead based on common-sense heuristics like document length and average document length.

Another approach, the *language model* approach, takes a formally principled and probabilistic approach to determining the ranking and weighting function. Instead of each document being considered some parametrized word-frequency vector, the documents are each considered to be samples from an underlying probabilistic language model M_D, of which D itself is only a single observation. In this manner,

the query Q can itself also be considered a sample from a language model. In early language modeling efforts the probability that the language model of a document would generate the query is given by the ranking function of the document. A more sophisticated approach to language models considers that the query was a sample from an underlying *relevance model* of unknown relevant documents, but that the model could be estimated by computing the co-occurrence of the query terms with every term in the vocabulary. In this way, the query itself was just considered a limited sample that is automatically expanded before the search has even begun by re-sampling the underlying relevance model.

In detail, we will now inspect the various weighting and ranking functions of the two frameworks. A number of different options for the parameters of each weighting function and the appropriate ranking function will be considered.

6.3.1 Vector Space Models

6.3.1.1 Representation

Each vector-space model has as a parameter the factor m, the maximum *window size*, which is the number of words, ranked in descending order of frequency, that are used in the document models. In other words, the size of the vectors in the vector-space model is m. Words with a zero frequency are excluded from the document model.

6.3.1.2 Weighting Function: BM25

The current state of the art weighting function for vector-space models is *BM*25, one of a family of weighting functions explored by Roberson Robertson et al (1994) and a descendant of the *tf.idf* weighting scheme pioneered by Spärck Jones and Robertson Robertson and Spärck Jones (1976). In particular, we will use a version of *BM25* with the slight performance-enhancing modifications used in the InQuery system Allan et al (2000). This weighting scheme has been carefully optimized and routinely shows excellent performance in TREC competitions Craswell et al (2005). The InQuery BM25 function assigns the following weight to a word q occurring in a document D:

$$D_q = \frac{n(q,D)}{n(q,D)+0.5+1.5\frac{dl}{avg(dl)}} \frac{\log(0.5+N/df(q))}{\log(1.0+\log N)} \tag{6.1}$$

The *BM*25 weighting function is summed for every term $q \in Q$. For every q, *BM*25 calculates the number of occurrences of a term q from the query in the document D, $n(q,D)$, and then weighs this by the length of document dl of document D in comparison to the average document length $avg(dl)$. This is in essence the equivalent of term frequency in $tf.idf$. The *BM*25 weighting function then takes

into account the total number of documents N and the document frequencies $df(q)$ of the query term. This second component is the idf component of classical $tf.idf$.

6.3.1.3 Ranking Function: Cosine and InQuery

The vector-space models have an intuitive ranking function in the form of cosine measurements. In particular, the cosine ranking function is given by Equation 6.2, for a document D with query Q, where both D and Q contain q words, iterating over all words.

$$cos(D,Q) = \frac{D \cdot Q}{|D||Q|} = \frac{\sum_q Q_q D_q}{\sqrt{\sum_q Q_q^2}\sqrt{\sum_q D_q^2}} \tag{6.2}$$

The only question is whether or not the vectors should be normalized to have a Euclidean weight of 1, and whether or not the query terms themselves should be weighted. We investigate both options. The classical cosine is given as *cosine*, which normalizes the vector lengths and then proceeds to weight both the query terms and the vector terms by $BM25$. The version without normalization is called *inquery* after the *InQuery* system Allan et al (2000). The *inquery* ranking function is the same as *cosine* except without normalization each word in the query can be considered to have uniform weighing.

6.3.1.4 Relevance Feedback Algorithms: Okapi, LCA, and Ponte

There are quite a few options on how to expand queries in a vector-space model. One popular and straightforward method, first proposed by *Rocchio* Rocchio (1971) and at one point used by the *Okapi* system Robertson et al (1994), is to expand the query by taking the average of the j total relevant document models R, with a document $D \in R$, and then simply replacing the query Q with the top m words from averaged relevant document models. This process is given by Equation 6.3 and is referred to as $okapi$:

$$okapi(Q) = \frac{1}{j}\sum_{D \in R} D \tag{6.3}$$

Another state of the art query expansion technique is known as *Local Content Analysis* (*lca*) Xu and Croft (1996). Given a query Q with query terms $q_1 ... q_k$ and a set of results D and a set of relevant documents R, then lca ranks every $w \in V$ by Equation 6.4, where n is the size of the relevant documents R, idf_w is the inverse document frequency of word w, and D_q and D_w are the frequencies of the words w and $q \in Q$ in relevant document $D \in R$.

$$lca(w;Q) = \prod_{q \in Q} \left(0.1 + \frac{1/\log n}{1/id\,f_w} \log \sum_{r \in R} D_q D_w \right)^{id\,f_q} \tag{6.4}$$

After each word $w \in V$ has been ranked by *lca*, then the query expanded by LCA is just the top m words given by *lca*. Local Content Analysis attempts to select words from relevant documents to expand the query that have limited ambiguity, and so it does extra processing compared to the *okapi* method that simply averages the most frequent words in the relevant documents. In comparison, Local Content Analysis performs an operation similar in effect to $tf.idf$ on the possibly relevant terms, and so attempting by virtue of weighing to select only words w that both appear frequently with terms in query q but have a low overall frequency ($id\,f_w$) in all the results.

The final method we will use is the heuristic method developed by Ponte Ponte (1998), which we call *ponte*. Like *lca*, *ponte* ranks each word $w \in V$, but it does so differently. Instead of taking a heuristic-approach like *Okapi* or *LCA*, it takes a probabilistic approach. Given a set of relevant documents $R \in D$, Ponte's approach estimates the probability of each word $w \in V$ being in the relevant document, $P(w|D)$, divided by its overall probability of the word to occur in the results $P(w)$. Then the *Ponte* approach gives each $w \in V$ a score as given in Equation 6.5 and then expands the query by using the m most relevant words as ranked by their scores.

$$Ponte(w;R) = \sum_{D \in R} log \left(\frac{P(w|D)}{P(w)} \right) \tag{6.5}$$

6.3.2 Language Models

6.3.2.1 Representation

Language modeling frameworks in information retrieval represent each document as a language model given by an underlying multinomial probability distribution of word occurrences. Thus, for each word $w \in V$ there is a value that gives how likely an observation of word w is given D, i.e. $P(w|u_D(v))$. The document model distribution $u_D(v)$ is then estimated using the parameter ε_D, which allows a linear interpolation that takes into account the background probability of observing w in the entire collection C. This is given in Equation 6.6.

$$u_D(w) = \varepsilon_D \frac{n(w,D)}{|D|} + (1 - \varepsilon_D) \frac{n(w,C)}{\sum_{v \in V} n(v,C)} \tag{6.6}$$

The parameter ε_D just takes into account the relative likelihood of the word as observed in the given document D compared to the word given the entire collection of documents C. $|D|$ is the total number of words in document D, while $n(w,D)$ is the frequency of word d in document D. Further, $n(w,C)$ is the frequency of occurrence

of the word w in the entire collection C divided by the occurrence of all words v in collection C.

6.3.2.2 Language Modeling Baseline

When no relevance judgments are available, the language modeling approach ranks documents D by the probability that the query Q could be observed during repeated random sampling from the distribution $u_D(\cdot)$. The typical sampling process assumes that words are drawn independently, with replacement, leading to the following retrieval score being assigned to document D:

$$P(Q|D) = \prod_{q \in Q} u_D(q) \tag{6.7}$$

The ranking function in Equation 6.7 is called *query-likelihood* ranking and is used as a baseline for our language-modeling experiments.

6.3.2.3 Language Models and Relevance Feedback

The classical language-modeling approach to IR does not provide a natural mechanism to perform relevance feedback. However, a popular extension of the approach involves estimating a relevance-based model u_R in addition to the document-based model u_D, and comparing the resulting language models using information-theoretic measures. Estimation of u_D has been described above, so this section will describe two ways of estimating the relevance model u_R, and a way of measuring distance between u_Q and u_D for the purposes of document ranking.

Let $R = r_1 \ldots r_k$ be the set of k relevant documents, identified during the feedback process. One way of constructing a language model of R is to average the document models of each document in the set:

$$u_{R,avg}(w) = \frac{1}{k} \sum_{i=1}^{k} u_{r_i}(w) = \frac{1}{k} \sum_{i=1}^{k} \frac{n(w, r_i)}{|r_i|} \tag{6.8}$$

Here $n(w, r_i)$ is the number of times the word w occurs in the $i'th$ relevant document, and $|r_i|$ is the length of that document. Another way to estimate the same distribution would be to *concatenate* all relevant documents into one long string of text, and count word frequencies in that string:

$$u_{R,con}(w) = \frac{\sum_{i=1}^{k} n(w, r_i)}{\sum_{i=1}^{k} |r_i|} \tag{6.9}$$

Here the numerator $\sum_{i=1}^{k} n(w, r_i)$ represents the total number of times the word w occurs in the concatenated string, and the denominator is the length of the concatenated string. The difference between Equations 6.8 and 6.9 is that the former treats

every document equally, regardless of its length, whereas the latter favors longer documents (they are not individually penalized by dividing their contributing frequencies $n(w,r_i)$ by their length $|r_i|$).

6.3.2.4 Ranking Function: Cross Entropy

We now want to re-compute the retrieval score of document D based on the estimated language model of the relevant class u_R. What is needed is a principled way of comparing a relevance model u_R against a document language model u_D. One way of comparing probability that has shown the best performance in empirical information retrieval research Lavrenko (2008) is cross entropy. Intuitively, cross entropy is an information-theoretic measure that measures the average number of bits needed to identify the probability of distribution p being generated if p was encoded using given probability distribution p rather than q itself. For the discrete case this is defined as:

$$H(p,q) = -\sum_{x} p(x) log(q(x)) \tag{6.10}$$

If one considers that the $u_R = p$ and that document model distribution $u_D = q$, then the two models can be compared directly using cross-entropy, as shown in Equation 6.11. This use of cross entropy also fulfills the Probability Ranking Principle and so is directly comparable to vector-space ranking via cosine Lavrenko (2008).

$$-H(u_R||u_D) = \sum_{w \in V} u_R(w) \log u_D(w) \tag{6.11}$$

Note that either the *averaged* relevance model $u_{R,avg}$ or the *concatenated* relevance model $u_{R,con}$ can be used in Equation 6.11. We refer to the former as *rm* and to the latter as tf in the following experiments.

6.4 System Description

We present a novel system that uses the same underlying information retrieval system on both hypertext and Semantic Web data so that relevance feedback can be done in a principled manner from both sources of data with language models. In our system, the query is run first against the hypertext Web and relevant hypertext results can then be used to expand a Semantic Web search query with terms from resulting hypertext web-pages. The expanded query is then ran against the Semantic Web, resulting in a different ranking of results than the non-expanded query. We can also then run the process backwards, using relevant Semantic Web data as relevance feedback to improve hypertext Web search.

This process is described using pseudo-code in Figure 6.7 where the set of all queries to be ran on the system is given by the *QuerySet* parameter. The two different kinds of relevance feedback are given by the *SearchType* parameter, with *SearchType=RDF* for searching over RDF data using HTML documents as data for relevance feedback-based query expansion, and *HTML* for searching over HTML documents with RDF as the data for relevance-feedback query expansion. *Representation* is the internal data model used to represent the documents, either vector-space models or language models. The feedback used to expand the query is given by *Feedback* with the kind of relevance feedback algorithm used to expand the query is given by *Algorithm*, which for relevance models are directly built into the representation. The ranking function (cross-entropy for language models, or some variation of cosine for vector-space models) is given by *Ranking*. The final results for each query are presented to the user in *PresentResults*.

We can compare both Semantic Web data and hypertext documents by considering both to be 'bags of words' and using relevance modelling techniques to expand the queries Lavrenko and Croft (2001). We consider both to be 'bags of words.' Semantic Web data can be flattened, and URIs can be reduced to 'words' by the following steps:

- Reduce to the rightmost hierarchical component.
- If the rightmost component contains a fragment identifier (#), consider all characters right of the fragment identifier the rightmost hierarchical component.
- Tokenize the rightmost component on space, capitalization, and underscore.

So, `http://www.example.org/hasArchitect` would be reduced to two tokens, 'has' and 'architect.' Using this system, we evaluated both the vector-space and language models described in Section 6.3 on queries selected in Section 6.2.2 with relevance judgments on these queries selected in Section 6.2.3.

Algorithm 6.4.1: SEARCH(*QuerySet*, *SearchType*)

if $SearchType = RDF$
$\begin{cases} Data1 \in Representation(HTMLdata) \\ Data2 \in Representation(RDFdata) \end{cases}$
else $SearchType = HTML$
$\begin{cases} FeedbackData \in Representation(RDFdata) \\ ResultData \in Representation(HTMLdata) \end{cases}$
for each $Query \in QuerySet$
$\begin{cases} FeedbackResults \leftarrow Feedback(Query, Data1) \\ ExpandedQuery \leftarrow Algorithm(FeedbackResults) \\ FinalResults \leftarrow Ranking(ExpandedQuery, Data2) \\ PresentResults(FinalResults) \end{cases}$

Fig. 6.7 Feedback-Driven Semantic Search

6.5 Feedback Evaluation

In this section we evaluate algorithms and parameters using relevance feedback against the same system without relevance feedback. In Section 6.8 we evaluate against deployed systems such as FALCON-S and Yahoo! Web Search. To preview our final results in Section 6.8, relevance feedback from the Semantic Web shows an impressive 25% gain in average precision over Yahoo! Web Search with a 16% gain in precision over FALCON-S without relevance feedback.

6.5.1 Hypertext to Semantic Web Feedback

6.5.1.1 Results

A number of parameters for our system were evaluated to determine which parameters provide the best results. For each of the parameter combinations, we compared the use of relevance feedback to a baseline system which did not use relevance feedback, yet used the same parameters with the exception of any relevance feedback-related parameters. The baseline system without feedback can also be considered an unsupervised algorithm, while a relevance feedback system can be thought of as a supervised algorithm. For example, the relevant hypertext web-pages R can be considered to be training data, while the Semantic Web documents D we wish to re-rank can be considered to be test data. The hypertext web-pages and Semantic Web documents are disjoint sets ($D \cap R = \emptyset$). For evaluation we used mean average precision (MAP) with the standard Wilcoxon sign-test, which we will often just call 'average precision.'

For vector-space models, the *okapi*, *lca*, and *ponte* relevance weighting functions were all run, each trying both the *inquery* and *cosine* ranking functions. The primary parameter to be varied was the *window size* (m), the number of top frequency words to be used in the vectors for both the query model and the document models. Baselines for both *cosine* and *inquery* were run with no relevance feedback. The parameter m was varied over $5, 10, 20, 50, 100, 300, 1000, 3000$. Mean average precision results are given in Figure 6.8.

Interestingly enough, *okapi* relevance feedback weighting with a window size of 100 and an *inquery* comparison was the best, with a mean average precision of 0.8914 ($p < .05$). It outperformed the baseline of *inquery*, which has an average precision of 0.5595 ($p < .05$). Overall, *lca* did not perform as well, often performing below the baseline, although its performance increased as the window size increased, reaching an average precision of 0.6262 with $m = 3000$ ($p < .05$). However, given that a window size of 10,000 covered most documents, increasing the window size will not likely result in better performance from *lca*. The *ponte* relevance feedback performed very well, reaching a maximum MAP 0.8756 with a window size of 300 using *inquery* weighing, and so was insignificantly different from *inquery* ($p > .05$). Lastly, both *ponte* and *okapi* experienced a significant decrease in per-

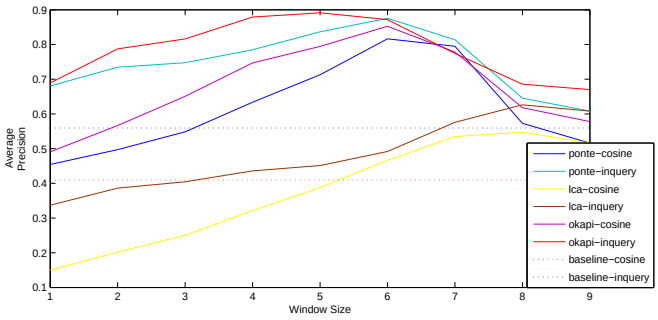

Fig. 6.8 Average Precision Scores for Vector-space Model Parameters: Relevance Feedback From Hypertext to Semantic Web

formance as m was increased, so it appears that the window sizes of 300 and 100 are indeed optimal. Also, as regards comparing baselines, *inquery* outperformed *cosine* ($p < .05$).

For language models, both averaged relevance models *rm* and concatenated relevance models tf were investigated, with the primary parameter being m, the number of non-zero probability words used in the relevance model. The parameter m was varied between 100, 300, 1000, 3000,and 10000. Remember that the query model *is* the relevance model for the language model-based frameworks. As is best practice in relevance modeling, the relevance models were not smoothed, but a number of different smoothing parameters for ε were investigated for the cross entropy ranking function, ranging from ε between .01, .1, .2, .5, .8, .9, and 0.99. The results are given in Figure 6.9.

The highest performing language model was tf with a cross-entropy ε of .2 and a m of 10,000, which produced an average precision of 0.8611, which was significantly higher than the language model baseline of 0.5043 ($p < .05$) using again an m of 10,000 for document models and with a cross entropy ε of .99). Rather interestingly, tf always outperformed *rm*, and *rm*'s best performance had a MAP of 0.7223 using an ε of .1 and a m of 10,000.

6.5.1.2 Discussion

Of all parameter combinations, the *okapi* relevance feedback works best in combination with a moderate sized word-window ($m = 100$) and with the *inquery* weighting scheme. It should be noted its performance is identical from a statistical standpoint with *ponte*, but as both relevance feedback components are similar and both use *inquery* comparison and *BM*25 weighing, and not surprisingly the algorithms are very similar. Why would *inquery* and *BM*25 be the best performing? The area of optimizing information retrieval is infamously a black art. In fact, *BM*25 and

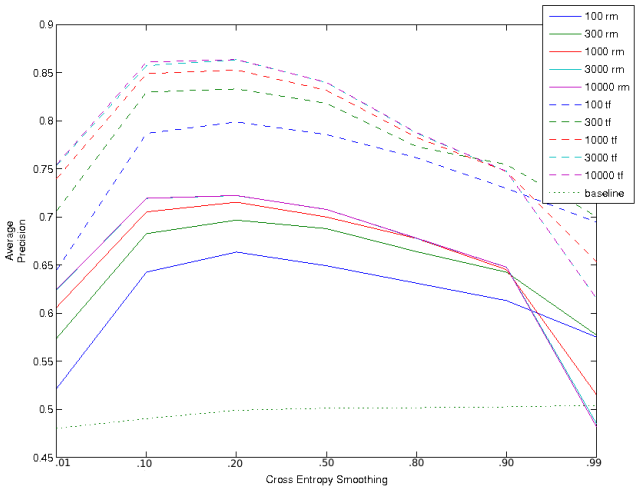

Fig. 6.9 Average Precision Scores for Language Model Parameters: Relevance Feedback From Hypertext to Semantic Web

inquery combined present the height of heuristic-driven information retrieval algorithms as explored in Robertson and Spärck Jones Robertson and Spärck Jones (1976). While its performance increase over *lca* is well-known and not surprising, it is interesting that *BM*25 and *inquery* perform significantly better than the language model approach.

The answer is rather subtle. Another observation is in order; note that for vector models, *inquery* always outperformed *cosine*, and that for language models tf always outperformed *rm*. Despite the differing frameworks of vector-space models and language models, both *cosine* and *rm* share the common characteristic of normalization. In essence, both *cosine* and *rm* normalize by documents: *cosine* normalizes term frequencies per vector before comparing vectors, while *rm* constructs a relevance model on a per-relevant document basis before creating the average relevance model. In contrast, *inquery* and tf do not normalize: *inquery* compares weighted term frequencies, and tf constructs a relevance model by combining all the relevance documents and then creating the relevance model from the *raw pool* of all relevant document models.

Thus it appears the answer is that any kind of normalization by length of the document hurts performance. The reason for this is likely because the text automatically extracted from hypertext documents is 'messy,' being of low quality and bursty, with highly varying document lengths. As observed informally earlier Ding and Finin (2006) and more formally later Halpin (2009a), the amount of triples in Semantic Web documents follow a power-law, so there are wildly varying document lengths of both the relevance model and the document models. Due to these

factors, it is unwise to normalize the models, as that will almost certainly dampen the effect of valuable features like crucial keywords (such as 'Paris' and 'tourist' in disambiguating various 'eiffel'-related queries).

Then the reason *BM*25-based vector models in particular perform so well is that, due to its heuristics, it is able to effectively keep track of a term's both document frequency and inverse document frequency accurately. Also, unlike most other algorithms, *BM*25 provides a slight amount of rather unprincipled non-linearity in the importance of the various variables Robertson et al (2004). This is important, as it provides a way of extenuating the effect of one particular parameter (in our case, likely term frequency and inverse term frequency) and then massively lowering the power of another parameter (in our case, likely the document length). While *BM*25 can be outperformed normally by language models Lavrenko (2008) in TREC competitions featuring high-quality samples of English, in the non-normal conditions of comparing natural language and pseudo-natural language terms extracted from structured data in RDF, it is not surprising that *okapi*, whose non-linearity allows certain highly relevant terms to have their frequency 'non-linearly' heightened, provides better results than more principled methods that derive their parameters by regarding the messy RDF and HTML-based corpus as a sample from a general underlying language model.

6.5.2 Semantic Web to Hypertext Feedback

In this section, we assume that the user or agent program has accessed or otherwise examined the Semantic Web documents from the URIs resulting from a Semantic Web search, and these Semantic Web documents then be used as relevance feedback to expand a query for the hypertext Web so that the feedback cycle has been reversed.

6.5.2.1 Results

The results for using Semantic Web documents as relevance feedback for hypertext Web search are surprisingly promising. The same parameters as explored in Section 6.5.1.1 were again explored. The average precision results for vector-space models are given in Figure 6.10. The general trends from Section 6.5.1.1 were similar in this new data-set. In particular, *okapi* with a window size of 100 and the *inquery* comparison function again performed best with an average precision of 0.6423 ($p < .05$). Also *ponte* performed almost the same, again an insignificant difference from *okapi*, producing with the same window size of 100 an average precision of 0.6131 ($p > .05$). Utilizing again a large window of 3,000, *lca* had an average precision of 0.5359 ($p < .05$). Similarly, *inquery* consistently outperformed *cosine* in comparison, with *inquery* having a baseline average precision of 0.4643 ($p < .05$) in comparison with the average precision of *cosine* being 0.3470 ($p < .05$).

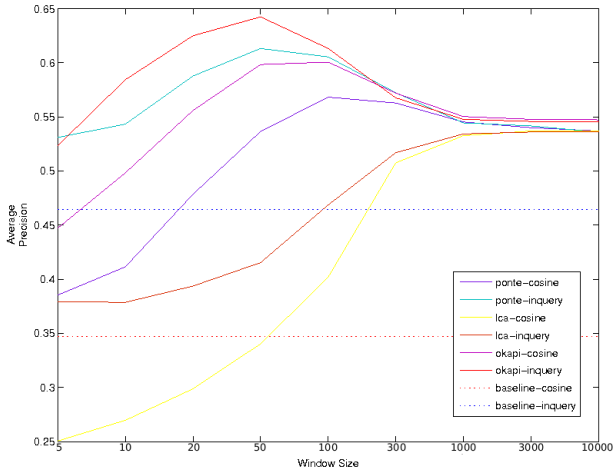

Fig. 6.10 Average Precision Scores for Vector-space Model Parameters: Relevance Feedback From Semantic Web to Hypertext

The results for language modeling were similar to the results in Section 6.5.1.1 and are given in Figure 6.11, although a few differences are worth comment. The best performing language model was tf with a m of 10,000 and a cross entropy smoothing factor ε to .5, which produced an average precision of .6549 ($p < .05$). In contrast, the best-performing *rm*, with a m of 3,000 and ε=.5, only had an average precision of 0.4858 ($p < .05$). The tf relevance models consistently performed better than *rm* relevance models ($p < .05$). The baseline for language modeling was also fairly poor with an average performance of 0.4284 ($p < .05$). This was the 'best' baseline using again an m of 10,000 for document models and cross entropy smoothing ε of .99. The general trends from the previous experiment then held, except the smoothing factor was more moderate and the difference between tf and *rm* was even more pronounced. However, the primary difference worth noting was that best performing tf language model outperformed, if barely, the *okapi* (*BM*25 and *inquery*) vector model by a relatively small but still significant margin of .0126. Statistically, the difference was significant ($p < .05$).

6.5.2.2 Discussion

Why is tf relevance modeling better than *BM*25 and *inquery* vector-space models in using relevance feedback from the Semantic Web to hypertext? The high performance of *BM*25 and *inquery* has already been explained, and that explanation about why document-based normalization leads to worse performance still holds. Yet the

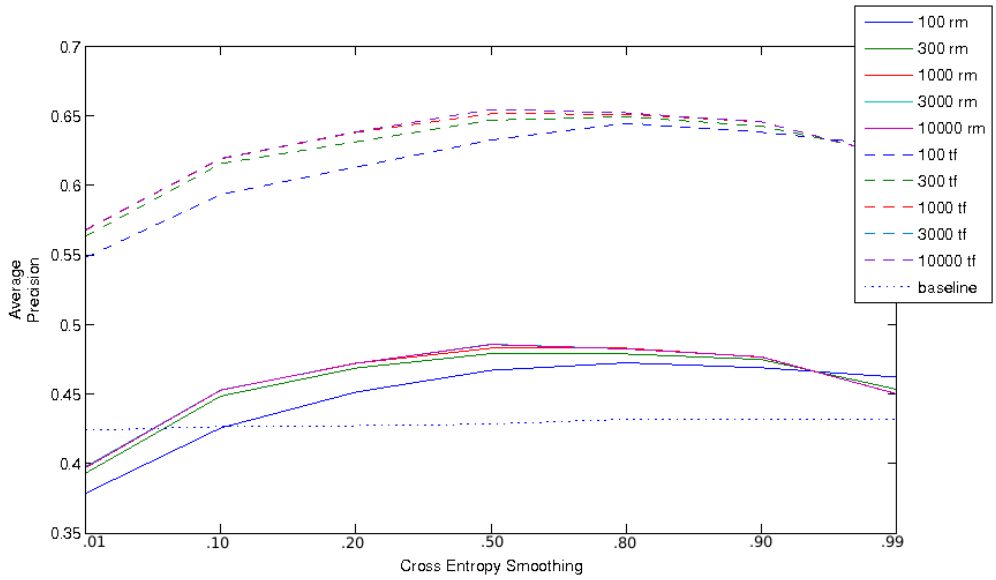

Fig. 6.11 Average Precision Scores for Language Model Parameters: Relevance Feedback From Hypertext to Semantic Web

rise in performance of tf language models seems odd. However, it makes sense if one considers the nature of the data involved. Recalling previous work Halpin (2009a), there are two distinct conditions that separated this data-set from the more typical natural language samples as encountered in TREC Hawking et al (2000). In the case of using relevant hypertext results as feedback for the Semantic Web, the relevant document model was constructed from a very limited amount of messy hypertext data, which had many text fragments, with a large percentage coming from irrelevant textual data to deal with issues like web-page navigation. However, in using the Semantic Web for relevance feedback, these issues are reversed: the relevant document model is constructed out of relatively pristine Semantic Web documents and compared against noisy hypertext documents.

Rather shockingly, as the Semantic Web is mostly manually high-quality curated data from sources like DBpedia, the actual natural language fragments found on the Semantic Web, such as Wikipedia abstracts, are much better samples of natural language than the natural language samples found in hypertext. Furthermore, the distribution of 'natural' language terms extracted from RDF terms (such as 'sub class of' from `rdfs:subClassOf`), while often irregular, will either be repeated very heavily or fall into the sparse long tail. These two conditions can then be dealt with by the generative tf relevance models, since the long tail of automatically generated words from RDF will blend into the long tail of natural language terms, and the probabilistic model can properly 'dampen' without resorting to heuristic-driven

non-linearities. Therefore, it is on some level not surprising that even hypertext Web search results can be improved by Semantic Web search results, because used in combination with the right relevance feedback parameters, in essence the hypertext search engine is being 'seeded' with high-quality structured and accurate descriptions of the information need of the query to be used for query expansion.

6.6 Pseudo-feedback

In this section we explore a very easy-to-implement and feasible way to take advantage of relevance feedback without manual selection of relevant results by human users. One of the major problems of relevance feedback-based approaches is their dependence on manual selection of relevant results by human users. For example, in our experiments we used judges manually determining if web-pages were relevant using an experimental set-up that forced them to judge every result as relevant or not, which is not feasible for actual search engine use.

A well-known technique within relevance feedback is *pseudo-feedback*, namely simply assuming that the top x documents returned are relevant. Then, one can use this as a corpus of relevance documents to expand the queries in the same manner using language models as described in Section 6.3. However, in general pseudo-relevance feedback is a more feasible method, as human intervention is not required.

Using the same optimal parameters as discovered in Section 6.5.1.1, tf with a $m = 10,000$ and $\varepsilon = .2$ was again deployed, but this time using pseudo-feedback. Can pseudo-feedback from hypertext Web search help improve the rankings of Semantic Web data? The answer is clearly positive. Employing all ten results as pseudo-relevance feedback and the same previously optimized parameters, the best pseudo-relevance feedback result had an average precision of 0.6240. This was considerably better than the baseline of just using relevance pseudo-feedback from the Semantic Web to itself, which only had an average precision of 0.5251 ($p < .05$), and also clearly above the 'best' baseline of 0.5043 ($p < .05$). However, as shown by Figure 6.12, the results are still not nearly as good as using hypertext pages judged relevant by humans, which had an average precision of 0.8611 ($p < .05$). This is likely because, not surprisingly, the hypertext Web results contain many irrelevant text fragments that serve as noise, preventing the relevant feedback from boosting the results.

Can pseudo-feedback from the Semantic Web improve hypertext search? The answer is yes, but barely. The best result for average precision is 0.4321 ($p < .05$), which is better than the baseline of just using pseudo-feedback from hypertext Web results to to themselves, which has an average precision of 0.3945 ($p < .05$) and the baseline without feedback at all of 0.4284 ($p < .05$). However, the pseudo-feedback results are both still significantly worse performance by a large margin than using Semantic Web documents judged relevant by humans, which had an average precision of 0.6549 ($p < .05$). These results can be explained because, given the usual ambiguous and short one or two word queries, the Semantic Web tends to return

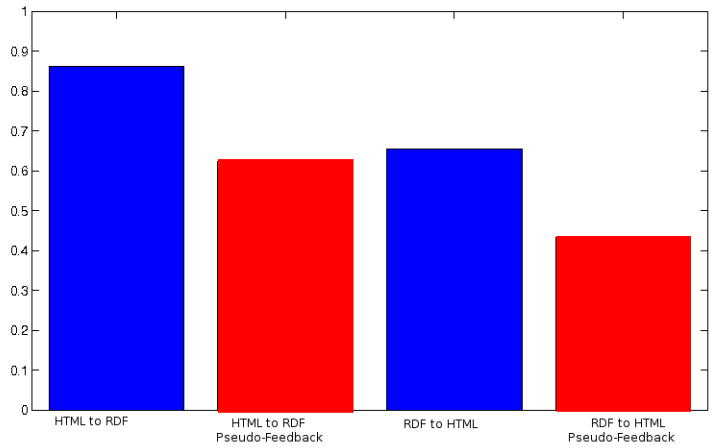

Fig. 6.12 Comparing Relevance Feedback (red) to Pseudo-Relevance Feedback (blue) on the Semantic Web (RDF) and Hypertext Web (HTML)

structured data spread out of over multiple subjects even moreso than the hypertext Web. Therefore, adding pseudo-relevance feedback increases the amount of noise in the language model as opposed to using actual relevance feedback, hurting performance while still keeping it above baseline.

6.7 Inference

In this section the effect of inference on relevance feedback is evaluated by considering inference to be document expansion. One of the characteristics of the Semantic Web is that the structure should allow one 'in theory' to discover more relevant data. The Semantic Web formalizes this in terms of type and sub-class hierarchies in RDF using RDF Schema Brickley and Guha (2004). While inference routines are quite complicated as regards the various Semantic Web specifications, in practice the vast majority of inference that can be used is on the Semantic Web is of two types (as shown by our survey of Linked Data Halpin (2009a)), *rdf:subClassOf* that indicates a simple sub-class inheritance hierarchy and *rdf:type* that indicates a simple type. For our experiment, we followed all explicit *rdf:subClassOf* statements up one level in the sub-class hierarchy and explicit *rdf:type* links. The resulting retrieved Semantic Web data was all concatenated together, and then concatenated yet again with their source document from the Semantic Web. In this way, Semantic Web inference is considered as *document expansion*.

Inference was first tried using normal relevant feedback, again with the same best-performing parameters (tf with $m = 10,000$ and $\varepsilon = .2$). In the first case, the in-

ference is used to expand Semantic Web documents in semantic search, and then the hypertext results are used as relevance feedback to improve the ranking. However, as shown in Figure 6.13, deploying inference only causes a drop in performance. In particular, using hypertext Web results as relevance feedback to the Semantic Web, the system drops from a performance of 0.8611 to a performance of 0.4991 ($p < .05$). With pseudo-feedback over the top 10 documents, the performance drops even more, from 0.6240 to 0.4557 ($p < .05$). The use of inference actually makes the results worse than the baseline performance of language models of 0.5043 ($p < .05$) without either relevance feedback or inference.

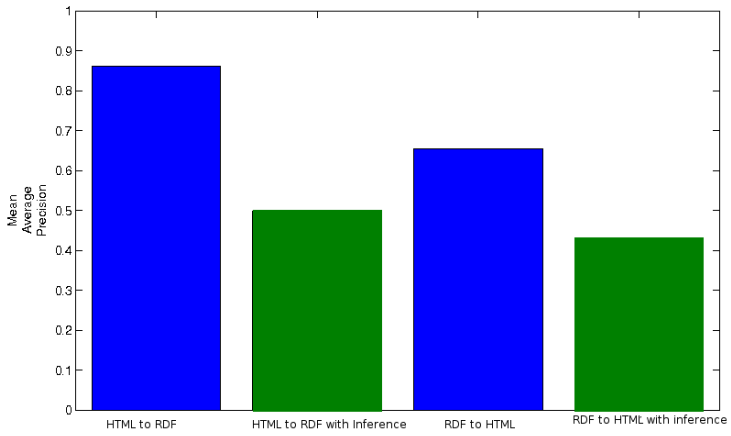

Fig. 6.13 Comparing the Relevance Feedback on the Semantic Web (RDF) and Hypertext Web (HTML) both without (blue) and with (green) Semantic Web inference

The results of using inference to boost hypertext Web results using Semantic Web equally fail to materialize any performance gains. In this case, inference is used to expand Semantic Web documents, which are then used via relevance feedback to improve the ranking of hypertext search. Using the same parameters as above, the feedback from the expanded Semantic Web data to the hypertext Web results in an average precision of 0.4273, which is insignificantly different from the baseline of not using relevance feedback at all of 0.4284 ($p < .05$) and considerably worse than not using inference at all, which has a MAP of 0.6549 ($p < .05$). When pseudo-feedback is used, the results fall to the rather low score of 0.3861, which is clearly below the baseline of 0.4284 ($p < .05$). So, at least one obvious way of use of simple type and sub-class based Semantic Web inference seems to only lead to a decline in performance.

Why does inference hurt rather than help performance? One would naively assume that adding more knowledge in the form of Semantic Web would help the results. However, this assumes the knowledge gained through inference would some-

how lead to the discovery of new relevant terms. However, in the case of much inference with the Semantic Web, this is not the case. For example, simply consider the Semantic Web data about the query for the singer 'Britney Spears.' While the first Semantic Web document about Britney Spears gives a number of useful facts about her, such as the fact that she is a singer, determining that Britney Spears is a person via inference is of vastly less utility. For example, the Cyc ontology Lenat (1990) declares that Britney Spears is a person, namely that "Something is an instance of Person if it is an individual Intelligent Agent with perceptual sensibility, capable of complex social relationships, and possessing a certain moral sophistication and an intrinsic moral value." In this regard, inference only serves as noise, adding irrelevant terms to the language models. For example, adding 'sophistication' to a query about 'Britney Spears' will likely not help discover relevant documents. Inference would be useful if it produced surprising information or reduced ambiguity. However, it appears that at least for simple RDF Schema vocabularies, information higher in the class hierarchy is usually knowledge that the user of the search engine already possesses (like Britney Spears is a person) and that the reduction of ambiguity is already done by the user in their selection of keywords. However, it is possible that more sophisticated inference techniques are needed, and that inference may help in specialized domains rather than open-ended Web search. Further experiments in parametrization of inference would be useful given that our exploration in this direction showed no performance increase, only performance decrease.

6.8 Deployed Systems

In this section we evaluate our system against 'real-world' deployed systems. One area we have not explored is how systems based on relevance feedback perform relative to systems that are actually deployed, as our previous work has always been evaluated against systems and parameters we created specifically for experimental evaluation. Our performance in Section 6.5.1.1 and Section 6.5.2.1 was only compared to baselines that were versions of our weighting function without a relevance feedback component. While that particular baseline is principled, the obvious needed comparison is against actual deployed commercial or academic systems where the precise parameters deployed may not be publicly available and so not easily simulated experimentally.

6.8.1 Results

The obvious baseline to choose to test against is the Semantic Web search engine, FALCON-S, from which we derived our original Semantic Web data in the experiment. The decision to use FALCON-S as opposed to any other Semantic Web search engine was based on the fact that FALCON-S returned more relevant results in the

top 10 than other existing semantic search engines at the time using a random sample of 20 queries from the set of queries described in Section 6.2.2. Combined with the explosive growth of Linked Data over the last year and the changes in ranking algorithms of various semantic search engines, it is difficult to judge whether a given Semantic Web search engine is representative of semantic search. However, we would find it reasonable that if our proposed hypothesis works well on FALCON-S, it can be generalized to other Semantic Web search engines.

We used the original ranking of the top 10 results given by FALCON-S to calculate its average precision, 0.6985. We then compared both the best baseline, *rm*, as well as the best system with feedback in Figure 6.14. As shown, our system with feedback had significantly ($p < .05$) better average precision (0.8611) than FALCON-S (0.6985), as well better ($p < .05$) than the 'best' language model baseline without feedback (0.5043) as reported earlier as given in Section 6.5.1.1.

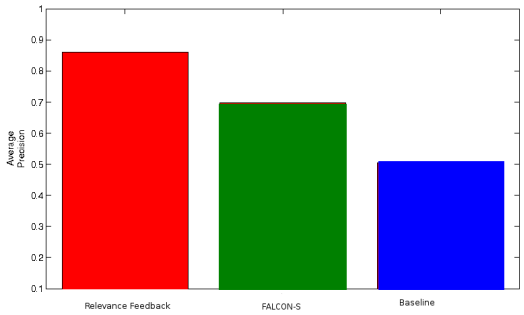

Fig. 6.14 Summary of Best Average Precision Scores: Relevance Feedback From Hypertext to Semantic Web

Average precision does not have an intuitive interpretation, besides the simple fact that a system with better average precision will in general deliver more accurate results closer to the top. In particular, one scenario we are interested in is having *only* the most relevant RDF data accessible from a single URI returned as the top result, so that this result is easily consumed by some program. For example, given the search 'amnesia nightclub,' a program should be able to consume RDF returned from the Semantic Web to produce with high reliability a single map and opening times for a particular nightclub in Ibiza in the limited screen space of the browser, instead of trying to display structured data for every nightclub called 'amnesia' in the entire world. In Table 3, we show that for a significant minority of URIs (42%), FALCON-S returned a non-relevant Semantic Web URI as the top result. Our feedback system achieves an average precision gain of 16% over FALCON-S. While a 16% gain in average precision may not seem huge, in reality the effect is quite dramatic, in particular as regards boosting relevant URIs to the top rank. So in Table 3, we present results of how our best parameters tf with $m = 10,000$ lead to

the most relevant Semantic data in the top result. In particular, notice that now 89% of resolved queries now have relevant data at the top position, as opposed to 58% without feedback. This would result in a noticeable gain in performance for users, which we would argue allows Semantic Web data to be retrieved with high-enough accuracy for actual deployment.

While performance is boosted for both entities and concepts, the main improvement comes from concept queries. Indeed, as concept queries are often one word and ambiguous, not to mention the case where the name of a concept has been taken over by some company, music band, or product, it should not be surprising that results for concept queries are considerably boosted by relevance feedback. Results for entity queries are also boosted. A quick inspection of the results reveals that the entity queries were the most troublesome, and that these entity queries gave both FALCON-S and our feedback system problems. These problematic queries were mainly very difficult queries where a number of Semantic Web documents all share similar natural language content. An example would be a query for 'sonny and cher,' which results in a number of distinct Semantic Web URIs: one for *Cher*, another one for *Sonny and Cher* the band, and another for 'The Sonny Side of Cher,' an album by Cher. For concepts, one difficult concept was the query 'rock.' Although the system was able to disambiguate the musical sense from the geological sense, there was a large cluster of Semantic Web URIs for rock music, ranging from *Hard Rock* to *Rock Music* to *Alternative Rock*. These types of queries seem to present the most difficulties for Semantic Web search engines.

Results:	Feedback	FALCON-S
Top Relevant:	118 (89%)	76 (58%)
Non-Relevant Top:	14 (11%)	56 (42%)
Non-Relevant Top Entity:	9 (64%)	23 (41%)
Non-Relevant Concept:	5 (36%)	33 (59%)

Table 6.3 Table Comparing Hypertext-based Relevance Feedback and FALCON-S

Although less impressive than the results for using hypertext web-pages for relevance feedback for the Semantic Web, the feedback cycle from the Semantic Web to hypertext does improve significantly the results of even commercial hypertext web-engines, at least for our set of queries about concepts and entities. Given the unlimited API-based access offered by Yahoo! Web Search in comparison to Google and Microsoft web search, we used Yahoo! Web Search for hypertext searching in this experiment, and we expect that the results in a coarse-grained manner should generalize to other Web search engines. The hypertext results for our experiment were given by Yahoo! Web Search, and we calculated a mean average precision for Yahoo! Web Search to be 0.4039. This is slightly less than our baseline language model ranking, which had an average precision of of 0.4284. As shown in Figure 6.15, given that our feedback based had an average precision of 0.6549, our relevance feedback system performs significantly ($p < .05$) better than Yahoo! Web Search and ($p < .05$) the baseline *rm* system.

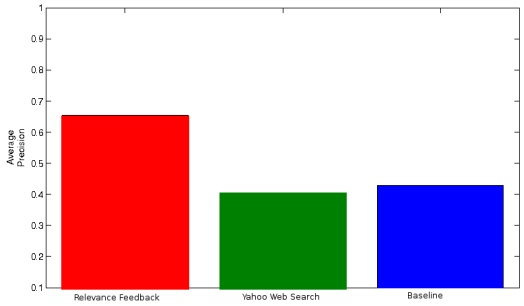

Fig. 6.15 Summary of Best Average Precision Scores: Relevance Feedback From Semantic Web to Hypertext

6.8.2 Discussion

These results show our relevance feedback method works significantly better than various baselines, both internal baselines and state of the art commercial hypertext search engines and Semantic Web search engines. The parametrization of the precise information retrieval components used in our system is not entirely arbitrary, as argued above in Section 6.5.1.2 and Section 6.5.2.2. The gain of our relevance feedback system, a respectable 16% in average precision over the engine FALCON-S, intuitively makes the system's ability to place a relevant structured Semantic Web data in the top rank acceptable for most users.

More surprisingly, by incorporating human relevance judgments of Semantic Web documents, we make substantial gains over state of the art systems for hypertext Web search, a 25% gain in average precision over Yahoo! search. One important factor is the constant assault of hypertext search engines by spammers and others. Given the prevalence of a search engine optimization and spamming industry, it is not surprising that the average precision of even a commercial hypertext engine is not the best, and that it performs less well than Semantic Web search engines. Semantic Web search engines have a much smaller and cleaner world of data to deal with than the unruly hypertext Web, and hypertext Web search must be very fast and efficient. Even without feedback from the Semantic Web, an average precision of 40% is impressive, although far from the 65% precision using relevance feedback from the Semantic Web.

Interestingly enough, it seemed that pseudo-feedback only helps marginally in improving hypertext Web search using Semantic Web data. Therefore, it is somewhat unrealistic to expect the Semantic Web to instantly improve hypertext Web search. Even with the help of the Semantic Web, hypertext search is unlikely to achieve near perfect results anytime soon. This should not be a surprise, as pseudo-feedback in general performs worse than relevance feedback. However, the loss of performance given by pseudo-feedback in comparison with traditional relevance

feedback show that for the Semantic Web using pseudo-feedback for concepts and entities is difficult, as many results that are about highly different things and subject matters may be returned. However, both pseudo-feedback and traditional relevance feedback help a fair amount in improving Semantic Web search using hypertext results, and as relevance judgments can be approximated by click-through logs of hypertext Web search engines, it is realistic and feasible to try to improve semantic search using relevance feedback from hypertext search. In fact, it is simple to implement pseudo-feedback from hypertext Web search using hypertext search engine APIs, as no manual relevance judgments must be made at all and the API simply can produce the top 10 results of any query quickly.

6.9 Future Work on Relevance Feedback

There are a number of areas where our project needs to be more thoroughly integrated with other approaches and improved. The expected criticism of this work is likely the choice of FALCON-S and Yahoo! Web search as a baseline, and that we should try this methodology over other Semantic Web search engines and hypertext Web search engines. Lastly, currently it is unknown how to combine traditional word-based techniques from information retrieval with structural techniques from the Semantic Web, and while our experiment with using inference as document expansion did not succeed, a more subtle approach may prove fruitful. At this point, we are currently pursuing this in context of creating a standardized evaluation framework for all Semantic search engines. The evaluation framework presented here has led to the first systematic evaluation of Semantic Web search at the Semantic Search 2010 workshop over Structured Web Data (2011). Yet in our opinion the most exciting work is to be done as regards scaling our approach to work with live large-scale hypertext Web search engines.

While language models, particularly generative models like relevance models Lavrenko (2008), should have theoretically higher performance than vector-space models, the reason why large-scale search engines do not in general implement language models for information retrieval is that the computational complexity of calculating distributions over billions of documents does not scale. However, there is reason to believe that relevance models could be scaled to work with Web search if they built their language sample from suitably large ‘clean’ sample of natural language and also compressed the models by various means.

One of the looming deficits of our system is that for a substantial amount of our queries there are *no* relevant Semantic Web URIs with accessible RDF data. This amount is estimated to be 34% of all queries. However, these queries with no Semantic Web URIs in general *do* have relevant information on the hypertext Web, if not the Semantic Web. The automatic generation of Semantic Web triples from natural language text could be used in combination with our system to create automatically generated Semantic Web data, in response to user queries.

Another issue is how to determine judgments for relevance in a manner that scales to actual search engine use. Manual feedback, while providing the more accurate experimental set-up for testing relevance feedback, does not work in real search scenarios because users do not exhaustively select results based on relevance, but select on a small subset. However, pseudo-feedback does not take advantage of users selecting web-pages, but just assumes the top x are relevant. A better approach would be to consider click-through logs of search engines incomplete approximations of manual relevance feedback Cui et al (2002). As we only had a small sample of the Microsoft Live Query log, this was unfeasible for our experiments, but would be compelling future work. There is a massive amount of human user click-through data available to commercial hypertext search engines although Semantic Web data has little relevance feedback data itself. While it is easy enough to use query logs to determine relevant hypertext Web data, no such option exists for the Semantic Web. However, there are possible methodologies for determining the 'relevance' of Semantic Web data, even if machines rather than humans are consuming the data. For example, Semantic Web data that is consumed by applications like maps and calendar programs can be ascertained to be actually relevant.

Finally, while generic Semantic Web inference may not help in answering simple keyword-based queries for entities and concepts, further research needs to be done to determine if inference can help answer complex queries. While in most keyword-based searches the name of the information need is mentioned directly in the query, which in our experiment results from choosing the queries via a named entity recognizer, in complex queries only the type or attributes of the information need are mentioned directly. The name of particular answers is usually unknown. Therefore, some kind of inference may be crucial in determining what entities or concepts match the attributes or type mentioned in the query terms. For example, the SemSearch 2011 competition's 'complex query' task was very difficult for systems that did well on keyword search, and the winning system used a customized crawling of the Wikipedia type hierarchy over Structured Web Data (2011).

6.10 The Representational Nexus

This study features a number of results that impact the larger field of semantic search. First, it shows a rigorous information retrieval evaluation, the 'Cranfield paradigm', can be applied to semantic search despite the differences between the Semantic Web and hypertext. These differences are well-recorded in our sample of the Semantic Web as taken via FALCON-S using a query log, and reveals a number of large differences between the Semantic Web data and hypertext data, in particular that while relevant data for ordinary open-domain queries does appear on the Semantic Web, Semantic Web data is in general more sparse than hypertext data when given a keyword query from an ordinary user's hypertext Web search. However, when the Semantic Web does contain data relevant to a given query, that data is likely to be accurate information, a fact we exploit in our techniques.

Unlike previous work in semantic search that focuses usually on some form of PageRank or other link-based ranking, we concentrate on using techniques from information retrieval, including language models and vector-space models, over Semantic Web data. Relevance feedback from hypertext Web data can improve Semantic Web search, and even *vice versa*, as we have rigorously and empirically shown. While relevance feedback is known to in general improve results, our use of wildly disparate sources of data such as the structured Semantic Web and the unstructured hypertext Web to serve as relevance feedback for each other is novel. Furthermore as regards relevance feedback, we show using vector-space models over hypertext data is optimal while language models are optimal when operating over Semantic Web. These techniques (as evidenced by the failure of relevance feedback to beat baseline results with incorrect parametrizations) must be parametrized correctly and use the correct weighting and ranking algorithm to be successful. It is shown by our results to be simply false to state that relevance feedback always improves performance over hypertext and Semantic Web search, but only under certain (although easily obtainable) parameters. We do this by treating both data sources as 'bags of words' and links in order to make them compatible and find from the Semantic Web high quality terms for use in language models. Also, untraditionally, we turn the URIs themselves into words. Our results of demonstrate that our approach of using feedback from hypertext Web search helps users discover relevant Semantic Web data. The gain is significant over both baseline systems without feedback and the state of the art page-rank based mechanism used by FALCON-S and Yahoo! Web search. Furthermore, the finding of relevant structured Semantic Web data can even be improved by pseudo-feedback from hypertext search.

More exciting to the majority of users of the Web is the fact that apparently relevance feedback from the Semantic Web can improve hypertext Web. However, pseudo-feedback also improves the quality of results of hypertext Web search engines, albeit to a lesser degree. Interestingly enough, using inference only hurt performance, due to the rather obscure terms from higher-level ontologies serving functionally as 'noise' in the feedback. Lastly, pseudo-feedback from the hypertext Web can help Semantic Web search today and can be easily implemented. Indeed, the key to high performance for search engines is the use of high quality data of any kind for query expansion, whether it is stored in a structured Semantic Web format or the hypertext Web. However, the Semantic Web, by its nature as a source of curated and formalized data, seems to be a better source of high quality data than the hypertext Web itself, albeit with less coverage. While it is trivial to observe that as the Semantic Web grows, semantic search will have more importance, it is even more interesting to demonstrate that as the Semantic Web grows, the Semantic Web can actually improve hypertext search.

The operative philosophical question is: Why does does relevance feedback work between such diverse encodings? Although there appears to be a huge gulf between the Semantic Web and the hypertext Web, it is precisely because the same *content* is encoded in the unstructured hypertext and the structured Semantic Web representations that these two disparate sets of data can be used as relevance feedback for each other. This leads to an exciting conclusion, and one that complexifies the ear-

lier picture of semantics considerably. If the Semantic Web is fundamentally about extending the Web to those things outside the Web, then we have to acknowledge that *most of the current hypertext Web is already representational*.

We call *the multitude of representations that share the same content and so can be used to compose its sense* the ***representational nexus*** of the referent, a potentially large collection of representations in a variety of formal, natural, and even iconic languages that all share the same referent. For example, if one uses a search engine to look for the 'Eiffel Tower,' one gets a large number of web-pages that are to some extent all *about* the Eiffel Tower by virtue of having some meaningful relationship with it, ranging from pictures of the Eiffel Tower, maps to the Eiffel Tower, and even possibly even videos of the Eiffel Tower. These would all count as representations of the Eiffel Tower, and so would be part of the representational nexus of the Eiffel Tower. Therefore, the aggregate 'bag-of-words' of all these representations would be an even more adequate notion of sense than just the tags explicitly given to a resource. Yet imagine how large of a landscape this opens for sense, for it allows us to apply search terms, documents, queries, Semantic Web representations - almost anything! - as part of the creation of sense in aggregate. This large aggregation has been phrased as the "database of intentions" by John Batelle, "the aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result." (2003). This should remind us that behind all of these representations are the concrete needs of ordinary users of the Web. What our task is to now attempt to phrase a philosophical theory of meaning adequate to this enlarged position of sense on the Web: the position of social semantics.

Glossary

This glossary presents the technical terminology used in this book, both from Web architecture and work in philosophy. Some of this terminology is presented as a formal Semantic Web ontology in Chapter 3, which my clarify the relationships of various sundry terms to each other.

absolute URI A URI in which there must a single scheme and the scheme must identify a name of a resource.

access The use of a identifier to create immediately a causal connection to the thing identified.

agent A thing capable of having an interpretation.

analog Every thing that is not digital.

arc role The URI of a link that provides information about what kind of link the link itself belongs to.

authority In a URI, a name that is usually a domain name, naming authority, or a raw IP address, and so is often the name of the server

AWWW The ***A***rchitecture of the ***W***orld ***W***ide ***Web***, a W3C Recommendation produced by the W3C to describe the defining characteristics of the Web, available at `http://www.w3.org/TR/webarch/`.

cache When a user-agent has a local copy of a Web representation that it accesses in response to a request rather than getting a Web representation from the server itself.

causal If one thing is connected with another thing and a change in the former thing is follows a change in the latter process in an interpretation.

causal theory of reference Any name refers via some causal chain directly to a referent

channel The physical substrate that determines whether or not the information is preserved over time or space.

client-server architecture Protocols that take the form of a request for information and a response with information.

client The agent that is requesting information. In the context of the Web, called a *user-agent*, which may be a Web browser or Web spider.

complete The inference procedure of a language if every satisified sentence can be shown to be entailed.

compositionality The content of a sentence is related systematically to terms in the which it is composed.

concept The regularities of the thing or set of things at a level of abstraction that are different than a realization. Often formalized as *classes* in formal ontologies and languages such as OWL and RDF Schema.

connected Those things that are not separated by time and space. Also called *proximal* and *local*.

content Whatever is held in common between the source and the receiver as a result of the conveyance of a particular information-bearing message.

consistent A sentence or sentence that can not be satisified.

content negotiation A mechanism defined in a protocol that makes it possible to respond to a request with different web representations of the same resource depending on the preference of the user-agent

content types The types of formal languages that can be explicitly given in a response or request in HTTP.

convention The use of a thing based purely on previous history, without regard to imitation or natural selection.

depictions A sentence or sentences in a natural or formal language whose primary purpose is to be a visual representation.

descriptions A sentence or sentences in an iconic language whose primary purpose is to be a linguistic or formal representation.

disconnected Things that are separated by time and space. Also called *distal*.

descriptivist theory of reference The referent of a name is given by whatever satisfies the descriptions associated with the name.

dialect A language created with or as a subset of another language.

digital When the boundaries in a particular encoding converge with a regularity in a physical realization. So there must be some finitely differentiable physical regularities that serves as a boundary.

direct reference position : A theory of semantics on the Web where the meaning of a URI is whatever was intended by the owner.

domain names A specification for a tree-structured namespace, where each component of the domain name (part of the name separated by a period) could direct the user-agent to more specific "domain name server" until the translation from an identifier to the name to IP address was complete.

encoding A set of regularities that can then be used to realize content-bearing messages.

ending resource The resource a link is directed to. @@OBJECT

endpoint Any thing that either requests or responds to a protocol.

entailment Where an interpretation of one sentence to some content always satisfies the interpretation of another sentence.

entity A thing where the regularities of the thing can only be realized by the thing itself, not in another realization. For the use of the term in HTTP, see *HTTP entity*.

entity body See *HTTP entity body*.

expression A particular message in a language.

extension Things that satisfy a description. @@RIGHT?

generic resource Web resources that vary over time, media type, and natural language.

graph merge When two formerly separate RDF graphs combine with each other when they use any of the same URI.

finitely differentiable When it is possible to determine for any given mark whether it is identical to another mark or marks. From @@GOODMAN

fixed resource A Web resource equivalent to a particular realization, a Web representation that should not change.

follow-your-nose algorithm An agent can follow the following steps in to help intepret a resource identified by a URI: dispose of any fragment identifier, inspect the media type of the retrieved Web representation, follow any namespace declarations, and follow any links. Available in full in Section 2.3.3.

formal languages A language with an explicitly defined syntax and possibly model-theoretic semantics, so suitable for interpretation by computers.

format A synonym for *formal language*, particularly for on computer-based digital formal language.

fragment identifier In a URI, either identifies fragment of a hypertext document in the case of media-type `text/html` being returned, or identifies some other resource that has has some relationship to the URI without the fragement identifier.

headers In HTTP, the part of the method that specify some information that may be of used by the server to determine the response or that specifies to the client information about the response.

hierarchical component The left to right dominant component of the URI that syntactically identifies the resource.

holism A sentence has meaning only in the context of a whole language. @@MOLECULARISM?

HTTP **H**yper**T**ext **T**ransfer **P**rotocol, a protocol originally purposed for the transfer of hypertext documents, although its now ubiquitous nature often lets it be used for the transfer of almost any encoding over the Web.

HTTP Entity The information transferred as the payload of a request or response excluding any optional headers. Confusingly, also sometimes also called the 'content,' although we use that term in a different sense, see *content* for our use.

iconic language A language based on visual images.

identifier A term that can be used to either access, refer to, or both access and refer to a thing.

inbound links Where the ending resource is a local Web representation and the distal starting resource is given by an identifier.

inconsistent A statement or statements that can not be satisified.

intension Kind of thing may only be described.@@NOT right EXTENSION?

inference A syntactic relationship where one sentence can be used to construct another sentence in a language

information Whatever in common between two things, where one thing is called the *sender* and the other is called the *receiver*. To have something in common means to share the same regularities, e.g. parcels of time and space that cannot be distinguished at a given level of abstraction. Information has at least one *encoding* that has some *content* in relationship to an agent capable of *interpretation*. When the term 'information' by itself is used, we are referring to both *abstract information* and any of its particular *realizations*. as well as both the *content* of any information as well as any *encoding* that transmits the content.

information resource A resource that is information with the possibility of a digital encoding.

interpretation The relationship between an encoding and its content. In formal semantics this is deployed in two distinct but related ways, an *interpretation mapping* that denotes the relationship between a language and a model, and the *interpretation structure* is a model that satisfies a particular interpretation mapping.

knowledge representation language A language whose primary purpose is the representation of non-digital content in a digital formal language.

level of abstraction The set of certain physical differences and regularities that have a causal effect on an agent and so may have a causal effect on the agent's meaningful behavior and may be captured in an interpretation.

link A connection between resources. @@predicate

linkbase Where the links can be represented outside of any Web representation of the starting or ending resource. @@RDF

location An identifier that can be used to access a thing.

logicist position For the Semantic Web, the meaning of a URI is given by whatever model(s) satisfy the formal semantics of the Semantic Web.

mark A physical characteristic.

meaning The causal effect of information on agents, often demonstrated by the behavior of agents.

media type A generalization of content types to any Internet protocol. It consists of a two-part scheme (separated by the '́) that separates the type and a subtype of an encoding.

message In HTTP, messages are also things that have headers and optional HTTP entity bodies. For its wider use in information theory, see *information* and *realization*, although HTTP messages also realize information, and so are inline with the broader user of the term.

method A request for a certain type of response from a user-agent to the server.

model A mathematical representation of the world or the language itself.

model-theoretic semantics When an interpretation of a language's sentences is to a mathematical model

monotonic In a system capable of inference, when the inference relationshiop $\vdash$ is monotonic if and only if for all sets of statements s_1 and s_2, and all inferred statements s_3, if $s_1 \vdash s_3$ and $s_2 \supset s_3$ then $s_2 \vdash s_3$.

language A system in which information is related to other information systematically. In a language, this is a relationship is between how the encoding of some information can change the interpretation of other encodings.

name An identifier that can be used to refer to a thing.

namespace declaration within a given Web representation in a particular dialect, the information that specifies the namespace URI of the dialect.

namespace document A Web representation that provides more information about the dialect.

namespace URI A URI that identifies that particular language or dialect thereof.

natural language A language based on human linguistic expressions.

non-monotonic When montonocity does not hold for a system capable of inference.

payload The information transmitted by a protocol.

path component A number of text strings delimited by special reserved characters that identify a resource.

Principle of Least Power A Web representation given by a resource can be described in the least powerful but adequate language.

Principle of Linking Any URI or Web representation can be linked to another resource identified by a URI.

Principle of the Open World The number of resources on the Web can always increase.

Principle of Self-Description The information an agent needs to have an interpretation of a resource should be accessible from its URI. This is often informally called the "follow-your-nose" algorithm.

Principle of Universality Any resource can be identified by a URI.

proper function Whatever characteristics which a a thing has in lieu of those characteristics promoting the reproduction or imitation of the thing. From Millikan @@.

protocol A convention for transmitting information between two or mobile agents.

proxy A cache that is not stored on the user-agent itself, but are shared among multiple user-agents by a server or group of servers.

public language position The Web is a form of language, and language exists as a public mechanism among multiple agents, then the meaning of a URI is the use of the URI, which must be a public mechanism that easily fits in the form of life of agents on the Web, which lets them in turn establish, find, and re-use URIs.
@@Social Semantics

purpose The intended meaning of information, often given by the behavior of the receiver intended by the sender of a message.

Open World Assumption Statements that cannot be proven to be true can not be assumed to be false.

Open World Principle See *Principle of the Open World*.

owner The agent that have the ability to create and alter the Web representation accessible from the URI.

outbound links Links that are inserted into Web representations directly and go beyond the local Web representation to an distal ending resources @@PREDICATE?

realization The physical thing that realizes the regularities of the information due to its local characteristics.

receiver See *information*.

regularity A lack of difference in time and space at a given level of abstraction.

relative URI A URI in a scheme where the path component itself is enough to identify a resource within certain contexts.

reference The relationship of an thing to another thing to which one is immediately causually disconnected.

referent The distal thing referred to by a representation. Also called *denotation*.

representation Any encoding of information that has distal content in some respect. Also called *symbol*. Note that this word "representation" has a distinct meaning in terms of its usage in Web standards, which we disambiguate by using the term *Web representation*. See *Web representation* for details.

resource Any thing capable of having identity. A resource is typically not a particular encoding of the information but the content of the information that can be given by many encodings.

request In HTTP, the method used by the agent and the headers, along with a blank line and an optional message body.

response In HTTP, the combination of the status code and the entities.

REST (**Re**presentational **S**tate **T**ransfer, an architectural style in which all state where the information state of the interaction between the between the server and client is stored on the client.

role A URI that can be attached to a link to provide information about the ending resource. @@predicate

satisfaction An interpretation to a mathematical that defines whether or not every sentence in the language can be interpreted to content.

scheme The name of a protocol or other naming convention, used as the first part of a URI.

sentence any combination of terms that is valid according the language's syntax. @@FORMAL? @@DIGITAL

semantics A system in which the content of information is related to each other systematically.

Semantic Web The use of the Web as a formal language to represent things, including things not accessible from the Web.

Semantic Web resource A resource that is analog.

Semantic Web URI A URI for a Semantic Web resource.

server the agent That is responding to the request.

specific resource Web resources that does not vary over one or more of the dimensions of time, media type, or natural language. These are called *time invariant*, *media-type invariant*, and *natural language invariant* respectively.

standard A convention for the encoding and possibly interpretation of information, often created by the explicit consensus of multiple parties via a standards body like the IETF or W3C.

statement Any combination of terms that has an interpretation to content according to the language's semantic

status code One of a finite number of codes gives the user-agent information about the server's HTTP response itself.

sound The inference procedure of a language if every inferred sentence can be satisfied.

source See *information*. Also called *sender*.

starting resource The resource that the link is directed from, also called the *subject* in RDF.

state Information about a resource that is not given as part of its identity, so it is information that may change over time.

syntax A system in which the encoding of information is related to each other systematically.

term regularities in marks @@.

thing Events, processes, objects, and proto-objects where the thing can be defined by having some regularity in time and space that can distinguish it from other possible thing.

user-agent A client in the context of the Web.

URI **U**niform **R**esource **I**dentifier) A unique identifier whose syntax is given in Berners-Lee et al (January 2005) that may be used to either or both refer to or access a resource.

URI Collision When the same resource has multiple URIs.

URI Opacity A URI should never itself have an interpretation, only the information referred to or accessed by that URI should have an interpretation.

URL **U**niform **R**esource **L**ocations) A scheme for locations that allows user-agents to via an Internet protocol access an realization of information.

URN **U**niform **R**esource **N**ame) A scheme whose names that could refer to things outside of the causal reach of the Internet.

@@web-page
@@web-resource

Web representation The content given by a resource given in response to a request whose encoding is capable of being determined by content negotiation.

WWW The ***W***orld ***W***ide ***Web***, an information space in which resources are identified by URIs.

References

Allan J, Connell M, Croft WB, Feng FF, Fisher D, Li X (2000) INQUERY and TREC-9. In: Proceedings of the Ninth Text REtrieval Conference (TREC-9), pp 551–562

Althusser L (1963) Marxism and Humanism. In: For Marx, Verso, republished in 2005 by Verso. Trans. Ben Brewster.

Anderson C (2006) The Long Tail. Random House Business Books

Andrews K, Kappe F, Maurer H (1995) The Hyper-G network information system. Journal of Universal Computer Science 1(4):206–220

Anklesaria F, McCahill M, Linder P, Johnson D, Torrey D, Alberti B (1993) IETF RFC 1436 the Internet Gopher protocol. Category: Informational. http://www.ietf.org/rfc/rfc1436.txt (Last accessed on Oct. 5th 2008)

van Assem M, Gangemi A, Brickley D (2006) RDF/OWL Representation of WordNet. Editor's draft, W3C, http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion (Last accessed Nov. 20th 2008)

Auer S, Bizer C, Lehmann J, Kobilarov G, Cyganiak R, Ives Z (2007) DBpedia: A nucleus for a web of open data. In: Proceedings of the International and Asian Semantic Web Conference (ISWC/ASWC2007), Busan, Korea, pp 718–728

Baeza-Yates R (2008) From capturing semantics to semantic search: A virtuous cycle. In: Proceedings of the 5th European Semantic Web Conference, Tenerife, Spain, pp 1–2

Baeza-Yates RA, Tiberi A (2007) Extracting semantic relations from query logs. In: Proceedings of the Conference on Knowledge Discovery and Data-mining (KDD), pp 76–85

Bar-Yam Y (2003) Dynamics of Complex Systems (Studies in Nonlinearity). Westview Press

Batagelj V, Mrvar A (1998) Pajek - A program for large network analysis. Connections 21:47–57

Batelle J (2003) The database of intentions. Http://battellemedia.com/archives/000063.php (Last accessed Dec. 11th 2008)

Bateson G (2001) Steps to an Ecology of Mind. University of Chicago Press, Chicago, Illinois, USA

Berners-Lee T (1989) Information management: A proposal. Tech. rep., CERN, http://www.w3.org/History/1989/proposal.html (Last accessed on July 12th 2008)

Berners-Lee T (1991) Document naming. Informal Draft. http://www.w3.org/DesignIssues/Naming (Last accessed on July 28th 2008)

Berners-Lee T (1994a) IETF RFC 1630 Universal Resource Identifier (URI). Http://www.ietf.org/rfc/rfc1630.txt (Last accessed on May 3rd 2008)

Berners-Lee T (1994b) World wide web future directions. Plenary Talk. http://www.w3.org/Talks/WWW94Tim/ (Last accessed on Oct. 5th 2008)

Berners-Lee T (1996a) Generic resources. Informal Draft. http://www.w3.org/DesignIssues/Generic.html (Last accessed on Dec. 4th 2008)

Berners-Lee T (1996b) Universal Resource Identifiers: Axioms of Web Architecture. Informal Draft. http://www.w3.org/DesignIssues/Axioms.html (Last accessed Sept. 5th 2008

Berners-Lee T (1998a) Cool URIs don't Change. Http://www.w3.org/Provider/Style/URI (Last accessed on Nov 19th 2008)

Berners-Lee T (1998b) Semantic web road map. Informal Draft. http://www.w3.org/DesignIssues/Semantic.html (Last accessed on April 12th 2008

Berners-Lee T (1998c) What the Semantic Web can represent. Informal Draft. http://www.w3.org/DesignIssues/rdfnot.html (Last accessed on Sept. 12th 2008)

Berners-Lee T (2000) Weaving the Web. Texere Publishing, London

Berners-Lee T (2003a) Message on www-tag@w3.org list. Http://lists.w3.org/Archives/Public/www-tag/2003Jul/0158.html

Berners-Lee T (2003b) Message to www-tag@w3.org. Http://lists.w3.org/Archives/Public/www-tag/2003Jul/0127.html

Berners-Lee T (2003c) Message to www-tag@w3.org. Http://lists.w3.org/Archives/Public/www-tag/2003Jul/0022.html

Berners-Lee T, Connolly D (June 1993) IETF Working Draft HyperText Markup Language (HTML): A Representation of Textual Information and MetaInformation for Retrieval and Interchange. Http://www.w3.org/MarkUp/draft-ietf-iiir-html-01.txt

Berners-Lee T, Cailliau R, Groff JF, Pollermann B (1992) World-Wide Web: The Information Universe. In: Electronic Networking: Research, Applications and Policy, Meckler, Westport, Connecticut, USA, pp 74–82

Berners-Lee T, Fielding R, McCahill M (1994) IETF RFC 1738 Uniform Resource Locators (URL). Proposed Standard. http://www.ietf.org/rfc/rfc1738.txt (Last accessed on Sept. 3th 2008)

Berners-Lee T, Fielding R, Frystyk H (1996) IETF RFC 1945 Hypertext Transfer Protocol (HTTP/1.0). Http://www.ietf.org/rfc/rfc1945.txt (Last accessed on Oct. 5th 2008)

Berners-Lee T, Fielding R, Masinter L (1998) IETF RFC 2396 Uniform Resource Identifier (URI): Generic Syntax. Http://www.ietf.org/rfc/rfc2396.txt (Last accessed on Sept. 15th 2008)

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284(5):35–43, URL http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21catID=2

Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner DJ (2006) Creating a science of the web. Science 313(5788):769–771

Berners-Lee T, Fielding R, Masinter L (January 2005) IETF RFC 3986 Uniform Resource Identifier (URI): Generic Syntax. Http://www.ietf.org/rfc/rfc3986.txt(Last accessed on April 2th 2008)

Bizer C, Cygniak R, Heath T (2007) How to publish Linked Data on the Web. Http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (Last accessed on May 28th 2008)

Bizer C, Heath T, Idehen K, Berners-Lee T (2008) Linked data on the web. In: Proceedings of the WWW2008 Workshop on Linked Data on the Web, URL http://CEUR-WS.org/Vol-369/paper00.pdf

Blanco R, Halpin H, Herzig D, Mika P, Pound J, Thompson H, Duc TT (2011) Repeatable and Reliable Search System Evaluation using Crowd-Sourcing. In: Proceedings of the 34th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Beijing, China

Boley H, Kifer M (2008) RIF Basic Logic Dialect. Working draft, W3C, http://www.w3.org/TR/rif-bld/ (Last accessed Aug. 8th 2008

Bollen D, Halpin H (2009) An experimental analysis of suggestions in collaborative tagging. In: Web Intelligence, Milan, Italy, pp 108–115

Borden J, Bray T (2002) Resource Directory Description Language (RDDL). Http://www.rddl.org/

Bornholdt S, Ebel H (2001) World Wide Web scaling exponent from Simon's 1955 model. Physical Review E 64(3):(R)–1 035,104–4

Bouquet P, Stoermer H, Tummarello G, Halpin H (eds) (2007) Proceedings of the WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007, CEUR Workshop Proceedings, CEUR-WS.org

Bouquet P, Stoermer H, Tummarello G, Halpin H (eds) (2008) Proceedings of the ESWC2008 Workshop on Identity, Reference, and the Web, Tenerife, Spain, June 1st, 2008, CEUR Workshop Proceedings

Box D, Ehnebuske D, Kakivaya G, Layman A, Mendelsohn N, Nielsen H, Thatte S, Winer D (2000) Simple Object Access Protocol (SOAP) 1.1. Http://www.w3.org/TR/2000/NOTE-SOAP-20000508/

Brachman R (1983) What IS-A is and isn't: An analysis of taxonomic links in semantic networks. IEEE Computer 16(10):30–36

Brachman R, Schmolze J (171-216) An overview of the KL-ONE knowledge representation system. Cognitive Science 9(2):151–160

Brachman R, Smith B (1980) Special issue on knowledge representation. SIGART Newsletter 70:1–38

Bray T, Paoli J, Sperberg-McQueen C (1998) Extensible Markup Language (XML). Recommendation, W3C, http://www.w3.org/TR/1998/REC-xml-19980210 (Last accessed on March 10th 2008)

Brickley D, Guha RV (2004) RDF Vocabulary Description Language 1.0: RDF Schema. Recommendation, W3C, http://www.w3.org/TR/rdf-schema/ (Last accessed on Nov. 15th 2008)

Brooks R (1991) Intelligence without representation. Artificial Intelligence 47(1-3):139–159

Bush V (1945) As we may think. Atlantic Monthly 1(176):101–108

Butterfield S (2004) Folksonomy. Http://www.sylloge.com/personal/2004/08/folksonomy-social-classification-great.html

Carnap R (1928) The Logical Structure of the World. University of California Press, Berkeley, California, USA, republished in 1967

Carnap R (1947) Meaning and Necessity: a Study in Semantics and Modal Logic. University of Chicago Press, Chicago, Illinois, USA

Carnap R (1950) Empiricism, semantics, and ontology. Revue Internationale de Philosophie 4:20–40

Carnap R, Bar-Hillel Y (1952) An outline of a theory of semantic information. Tech. Rep. RLE-TR-247-03150899, Research Laboratory of Electronics, Massachusetts Institute of Technology

Carpenter B (June 1996) IETF RFC 1958 Architectural Principles of the Internet. Http://www.ietf.org/rfc/rfc1958.txt (Last accessed on March 12th 2008)

Cerf V, Kahn R (1974) A protocol for packet network intercommunication. IEEE Transactions on Communications 22(4):637–648

Chalmers D (1995) Facing up to the problem of consciousness. Journal of Consciousness Studies 2(3):200–219

Chalmers DJ (2006) Two-dimensional semantics. In: Oxford Handbook of the Philosophy of Language, Oxford University Press

Cheng G, Ge W, Qu Y (2008) FALCONS: Searching and browsing entities on the Semantic Web. In: Proceedings of the the World Wide Web Conference

Chomsky N (1957) Syntactic Structures. Mouton, Paris, France

Cilibrasi R, Vitanyi P (2007) The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3):370–382

Clark A (1997) Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge, MA

Clark K (1978) Negation as failure. In: Gallaire H, Minker J, Nicolas J (eds) Logic and Databases, Plenum, New York City, New York, United States

Clauset A, Shalizi C, Newman M (2007) Power-law distributions in empirical data. Http://arxiv.org/abs/0706.1062v1 (Last accessed October 13th 2008)

Connolly D (1998) The XML revolution. Nature Http://www.nature.com/nature/webmatters/xml/xml.html (Last accessed on April 3rd 2008)

Connolly D (2002) An evaluation of the World Wide Web with respect to Engelbart's requirements. Informal Draft. http://www.w3.org/Architecture/NOTE-ioharch (Last accessed on Dec. 4th 2008)

Connolly D (2006) A pragmatic theory of reference for the web. In: Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference, http://www.ibiblio.org/hhalpin/irw2006/dconnolly2006.pdf (Last accessed November 22nd 2008)

Connolly D (2007) Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Tech. rep., W3C, URL http://www.w3.org/TR/grddl/, recommendation

Craswell N, Zaragoza H, Robertson S (2005) Microsoft Cambridge at trec-14: Enterprise track. In: Proceedings of the Seventh Text REtrieval Conference (TREC-7), p http://research.microsoft.com/apps/pubs/default.aspx?id=65241 (Last accessed January 10th 2009

Cui H, Wen JR, Nie JY, Ma WY (2002) Probabilistic query expansion using query logs. In: Proceedings of the 11th International Conference on World Wide Web (WWW 2002), ACM, New York, NY, USA, pp 325–332

Cummins R (1996) Representations, Targets, and Attitudes. MIT Press, Cambridge, Massachusetts, USA

Delugach H (2007) ISO common logic. Standard, ISO, http://cl.tamu.edu/ (Last accessed on March 8th 2008)

Dennett D (1981) Brainstorms: Philosophical Essays on Mind and Psychology. Cambridge, MA USA

DeRose S, Maler E, Orchard D (2001) XML Linking Language (Xlink) Version 1.0. Tech. rep., W3C Recommendation, http://www.w3.org/TR/xlink/ (Last accessed on Nov. 12th 2008)

Detlefsen M (1990) Brouwerian intuitionism. Mind 99(396):501–34

Ding L, Finin T (2006) Characterizing the Semantic Web on the Web. In: Proceedings of the International Semantic Web Conference (ISWC), pp 242–257

Dowty D (2007) Compositionality as an Empirical Problem. In: Barker C, Jacobson P (eds) Direct Compositionality, Oxford University Press, Oxford, United Kingdom, pp 23–101

Dretske F (1981) Knowledge and the Flow of Information. MIT Press, Cambridge, Massachusetts, USA

Dreyfus H (1979) What Computers Still Can't Do: A critique of artificial reason. MIT Press, Cambridge, Massachusetts, USA

Dummett M (1973) Frege: Philosophy of Language. Duckworth, London, United Kingdom

Dummett M (1993) What is a Theory of Meaning. In: The Seas of Language, Oxford University Press, Oxford, United Kingdom, pp 1–33, originally published in *Truth and Meaning: Essays in Semantics* in 1976.

Engelbart D (1962) Augmenting Human Intellect: A Conceptual Framework. Tech. rep., Stanford Research Institute, aFOSR-3233 Summary Report

Engelbart D (1990) Knowledge-domain interoperability and an open hyperdocument system. In: Proceedings of the Conference on Computer-Supported Collaborative Work, pp 143–156

Engelbart D, Ruilifson J (1999) Bootstrapping our collective intelligence. ACM Computer Survey 31(4):38, URL http://portal.acm.org/citation.cfm?id=346040

Ferraiolo J (2002) Scalable vector graphics (svg) 1.0 specification. Recommendation, W3C, http://www.w3.org/TR/2001/REC-SVG-20010904/ (Last accessed April 22nd 2008

Fielding R (2010) Architectural styles and the design of network-based software architectures. PhD thesis, University of California, Irvine

Fielding R, Gettys J, Mogul J, Frystyk H, Berners-Lee T (1999) IETF RFC 2616 Hypertext Transfer Protocol - HTTP 1.1. Http://www.ietf.org/rfc/rfc2616.txt (Last accessed on April 2nd 2008)

Fleiss J (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76:378–382

Floridi L (2004) Open problems in the philosophy of information. Metaphilosophy 35(4):554–582

Foucault M (1970) The Order of Things: An Archaeology of the Human Sciences. Pantheon Books, New York City, New York, USA

Fountain A, Hall W, Heath I, Davis H (1990) Microcosm: An open model for hypermedia with dynamic linking. In: Proceedings of Hypertext: Concepts, Systems and Applications (ECHT), Paris, France, pp 298–311

Frege G (1892) Uber sinn und bedeutung. Zeitshrift fur Philosophie and philosophie Kritic 100:25–50, reprinted in The Philosophical Writings of Gottlieb Frege (1956), Blackwell, Oxford, United Kingdom (1956), Max Black (trans.)

Galloway A (2004) Protocol: How Control Exists After Decentralization. MIT Press, Boston, Massachusetts, USA

Gangemi A (2008) Norms and plans as unification criteria for social collectives. Journal of Autonomous Agents and Multi-Agent Systems 16(3)

Gangemi A, Presutti V (2009) Handbook on Ontologies, 2nd edn, Springer, chap Ontology Design Patterns

Gerber A, van der Merwe A, Barnard A (2008) A Functional Semantic Web Architecture. In: Proceedings of the 5th European Semantic Web Conference, Tenerife, Spain, pp 273–287

Gligorov R, Aleksovski Z, ten Cate W, van Harmelen F (2008) Using google distance to weight approximate ontology matches. In: Proc. of 16th Int. World Wide Web Conference (WWW'07), ACM Press, pp 767–775

Golder S, Huberman B (2006) Usage patterns of collaborative tagging systems. Journal of Information Science 32(2):198–208

Goodman N (1968) Languages of Art: An Approach to a Theory of Symbols. Bobbs-Merrill, Indianapolis, Indiana, USA

Grice P (1957a) Meaning. Philosophical Review 66:377–388

Grice P (1957b) Meaning. The Philosophical Review 66:377–388

Guha R, McCool R, Miller E (2003) Semantic search. In: Proceedings of the International Conference on World Wide Web (WWW), ACM, New York, NY, USA, pp 700–709

Guha RV (1996) Meta Content Framework : A Whitepaper. Http://www.guha.com/mcf/wp.html (Last accessed Aug. 11th 2008)

Hafner K, Lyons M (1996) Where Wizards Stay Up Late: The Origins of the Internet. Simon and Schuster, New York City, New York, USA

Halasz F, Schwartz M (1994) The Dexter hypertext reference model. Communications of the ACM 37(2):30–39

Halpin H (2004) The semantic web: The origins of artificial intelligence redux. In: Proceedings of Third International Workshop on the History and Philosophy of Logic, Mathematics, and Computation (HPLMC-04 2005), Donostia San Sebastian, Spain, republished in 2007 by Icfai University Press in The Semantic Web. http://www.ibiblio.org/hhalpin/homepage/publications/airedux.pdf (Last accessed April 2nd 2008)

Halpin H (2006) Representationalism: The hard problem for artificial life. In: Proceedings of Artificial Life X, Bloomington, Indiana, pp 527–534

Halpin H (2008a) Foundations of a philosophy of collective intelligence. In: Proceedings of Convention for the Society for the Study of Artificial Intelligence and Simulation of Behavior

Halpin H (2008b) Philosophical engineering: Towards a philosophy of the web. APA Newsletter on Philosophy and Computers 7(2):5–11

Halpin H (2009a) A query-driven characterization of linked data. In: Proceedings of the Linked Data Workshop at the World Wide Web Conference, Madrid, Spain

Halpin H (2009b) Sense and Reference on the Web. Phd thesis, University of Edinburgh, School of Informatics, Institute for Communicating and Collaborative Systems, Edinburgh, UK

Halpin H (2011) Sense and reference on the web. Minds and Machines 21(2):153–178

Halpin H, Lavrenko V (2009) Relevance feedback between hypertext search and semantic search. In: Proceedings of the Semantic Search Workshop at the World Wide Web Conference, Madrid, Spain

Halpin H, Lavrenko V (2011) Relevance feedback between web search and the semantic web. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Catalonia, Spain, pp 2250–2255

Halpin H, Presutti V (2009) An ontology of resources: Solving the identity crisis. In: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications, Springer-Verlag, Berlin, Heidelberg, ESWC 2009 Heraklion, pp 521–534

Halpin H, Thompson HS (2009) Social meaning on the web: From wittgenstein to search engines. IEEE Intelligent Systems 24(6):27–31

Halpin H, Hayes P, Thompson HS (eds) (2006) Proceedings of the WWW2006 Workshop on Identity, Reference, and the Web, Edinburgh, United Kingdom, May 23, 2008, CEUR Workshop Proceedings

Halpin H, Robu V, Shepherd H (2007) The complex dynamics of collaborative tagging. In: Proc. of the 16th International World Wide Web Conference (WWW'07), pp 211–220

Halpin H, Clark A, Wheeler M (2010) Towards a philosophy of the web : Representation , enaction , collective intelligence. In: Proceedings of the Web Science Conference: Extending the Frontiers of Society On-Line (WebSci 2010, Raleigh, North Carolina, USA, pp 1–5, URL http://journal.webscience.org/324/2/websci10$_{s}ubmission_{1}$20.pdf

Halvey M, Keane MT (2007) An assesment of tag presentation techniques. In: Proceedings of the 16th Int. World Wide Web Conference (WWW 2007), ACM Press, pp 1313–1314

Haugeland J (1981) Analog and analog. In: Mind, Brain, and Function, Harvester Press, New York City, New York, USA, pp 213–226

Haugeland J (1991) Representational genera. In: Philosphy and Connectionist Theory, Erlbaum, Mahwah, New Jersey, USA, pp 61–89

Hawking D, Voorhees E, Craswell N, Bailey P (2000) Overview of the trec-8 web track. In: Proceedings of the Text REtrieval Conference (TREC), ACM, pp 131–150

Hayes P (1977a) In defence of logic. In: Proceedings of International Joint Conference on Artificial Intelligence, pp 559–565

Hayes P (1977b) In defense of logic. In: Proceedings of International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, USA, pp 559–565

Hayes P (1979) The Naive Physics Manifesto. In: Expert Systems in the Micro-Electronic Age, Edinburgh University Press, Edinburgh, Scotland, pp 242–270

Hayes P (2002) Catching the dream. Http://www.aifb.uni-karlsruhe.de/ sst/is/WebOntologyLanguage/hayes.htm (Last accessed Oct. 17th 2008)

Hayes P (2003a) Message to www-rdf-comments@w3.org. Http://lists.w3.org/Archives/Public/www-tag/2003Jul/0147.html

Hayes P (2003b) Message to www-rdf-comments@w3.org. Http://lists.w3.org/Archives/Public/www-tag/2003Jul/0198.html

Hayes P (2004) RDF Semantics. Recommendation, W3C, http://www.w3.org/TR/rdf-mt/ (Last accessed Sept. 21st 2008)

Hayes P (2006) In defense of ambiguity. In: Invited talk at the Identity, Reference, and the Web Workshop at the WWW Conference, http://www.ibiblio.org/hhalpin/irw2006/hayes.pdf

Hayes P, Halpin H (2008) In defense of ambiguity. International Journal of Semantic Web and Information Systems 4(3)

Hayles NK (2005) My Mother was a Computer: Digital Subjects and Literary Texts. University of Chicago Press, Chicago, Illinois

Hegel G (1959) Säammtliche Werke. Fromann, Stuttgart, Germany

Hirst G (2000) Context as a spurious concept. In: Proceedings of Context in Knowledge Representation and Natural Language, AAAI Fall Symposium, pp 273–287

http://labsgooglecom/sets (2008) Google sets. Retrieved: 1st Sept.

Israel D, Perry J (1990) What is information? In: Hanson P (ed) Information, Language, and Cognition, University of British Columbia Press, Vancouver, Canada, pp 1–19

Jacob E (2004) Classification and categorization: A difference that makes a difference. Library Trends 52(3):515–540

Jacobs I (1999) W3C Mission Statement. Tech. rep., W3C, http://www.w3.org/Consortium/

Jacobs I, Walsh N (2004) Architecture of the World Wide Web. Tech. rep., W3C, http://www.w3.org/TR/webarch/ (Last accessed Oct 12th 2008)

Jameson F (1981) The Political Unconscious. Cornell University Press, Ithaca, New York, USA

Jin RKX, Parkes DC, Wolfe PJ (2007) Analysis of bidding networks in eBay: Aggregate preference identification through community detection. In: Proc. AAAI Workshop on Plan, Activity and Intent Recognition (PAIR)

Jones KS (2004) What's new about the Semantic Web?: Some questions. SIGIR Forum 38(2):18–23

Klyne G, Carroll J (2004) Resource description framework (rdf): Concepts and abstract syntax. Recommendation, W3C, http://www.w3.org/TR/rdf-concepts/

Koller D, Pfeffer A (1998) Probabilistic frame-based systems. In: Proceedings of the 15th National Conference on Artificial Intelligence, Madison, Wisconsin, pp 580–587

Kripke S (1972) Naming and Necessity. Harvard University Press, Cambridge, Massachusetts, USA

Lavrenko V (2008) A Generative Theory of Relevance. Springer-Verlag, Berlin, Germany

Lavrenko V, Croft WB (2001) Relevance-based language models. In: Proceedings of the Twenty-Fourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New Orleans, Louisiana, USA, pp 120–127

Leiner B, Cerf V, Clark D, Kahn R, Kleinrock L, Lynch D, Postel J, Roberts L, Wolff S (2003) A brief history of the internet. Http://www.isoc.org/internet/history/brief.shtml (Last accessed March 20th 2008)

Lenat D (1990) Cyc: Towards programs with common sense. Communications of the ACM 8(33):30–49

Levensque H, Brachman R (1987) Expressiveness and tractability in knowledge representation and reasoni ng. Computational Intelligence 3(1):78–103

Lewis D (1971) Analog and digital. Nous 1(5):321–327

Licklider J (1960) Man-computer symbiosis. IRE Transactions on Human Factors in Electronics 1:4–11

Luntley M (1999) Contemporary Philosophy of Thought. Blackwell, London, United Kingdom

Lyotard JF (1988) The Inhuman: Reflections on Time. Editions Galilee, Paris, France, republished 1998 by Blackwell. Trans. Bennington and Rachel Bowlby

Mandelbrot B (1953) An informational theory of the statistical structure of languages. In: Jackson W (ed) Communication Theory, Academic Press, New York, USA

Mangold C (2007) A survey and classification of semantic search approaches. International Journal of Metadata, Semantics, and Ontologies 2(1):23–34

Manning C, Schutze H (2002) Foundations of statistical natural language processing. MIT Press, London

Marlow C, Naaman M, Boyd D, Davis M (2006a) Position paper, tagging, taxonomy, flickr, article, toread. In: Collaborative Web Tagging Workshop at WWW'06, Edinburgh, UK

Marlow C, Naaman M, Boyd D, Davis M (2006b) Position paper, tagging, taxonomy, flickr, article, toread. In: Collaborative Web Tagging Workshop at WWW'06

Masterman M (1961) Semantic message detection for machine translation, using an interlingua. In: Proceedings of International Conference on Machine Translation of Languages and Applied Language Analysis, London, United Kingdom

Mathes A (2004) Folksonomies: Cooperative classification and communication through shared metadata. Http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

McCarthy J (1959) Programs with common-sense. Nature 188:77–91, http://www-formal.stanford.edu/jmc/mcc59.html

McCarthy J (1980) Circumspection – a form of nonmonotonic reasoning. Artificial Intelligence 1(13):27–39
McCarthy J (1992) 1959 memorandum. IEEE Annals of the History of Computing 14(1):20–23, reprint of original memo made in 1952
McCarthy J, Hayes P (1969) Some philosophical problems from the standpoint of Artificial Intelligence. In: Meltzer B, Michie D (eds) Machine Intelligence, vol 4, Edinburgh University Press, pp 463–502
McCarthy J, Minksy M, Rochester N, Shannon C (1955) A Proposal for the Dartmouth Summer Research Project on Artificial intelligence. Tech. rep., Dartmouth College, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (Last accessed March 12th 2008)
McDermott D (1987) A critique of pure reason. Computational Intelligence 33(3):151–160
McKay D (1955) The place of meaning in the theory of information. In: Cherry E (ed) Information Theory, Basic Books, New York City, New York, USA, pp 215–225
Mealling M, Daniel R (1999) IETFRFC 2483 URI resolution services necessary for URN resolution. Experimental. http://www.ietf.org/rfc/rfc2483.txt (Last accessed April 13th 2008)
Mendelsohn N (2006) The self-describing web. Draft TAG finding, W3C, http://www.w3.org/2001/tag/doc/namespaceState-2006-01-09.html (Last accessed March 7th 2008)
Mika P (2005) Ontologies are us: A unified model of social networks and semantics. In: Proc. of the 4th Int. Semantic Web Conference (ISWC'05), Springer LNCS vol. 3729
Mikheev A, Grover C, Moens M (1998) Description of the LTG system used for MUC. In: Seventh Message Understanding Conference: Proceedings of a Conference
Millikan R (1984) Language, Thought and Other Biological Categories: New Foundations for Realism. MIT Press, Cambridge, Massachusetts, United States
Millikan R (2004) Varieties of Meaning. MIT Press, Cambridge, Massachusetts, United States
Minsky M (1975) A framework for representing knowledge. In: Winston P (ed) The Psychology of Computer Vision, McGraw Hill, Columbus, Ohio, USA, pp 211–277
Moats R (1997) IETF RFC 2141 URN Syntax. Http://www.ietf.org/rfc/rfc2141.txt
Mockapetris P (Novemeber 1983) IETF RFC 882 Domain Names - Concpets and Facilities. Http://www.ietf.org/rfc/rfc882.txt (Last accessed on March 12th 2008)
Mogul J (2002) Clarifying the fundamentals of HTTP. In: Proceedings of the 11th International World Wide Web Conference(WWW), Honolulu, Hawaii, USA, pp 444–457
Mueller V (2007) Representation in digital systems. In: Proceedings of Adapation and Representation, http://www.interdisciplines.org/adaptation/papers/7 (Last accessed March 8th 2008)
Nelson T (1965) Complex information processing: a file structure for the complex, the changing and the indeterminate. In: Proceedings of 20th National Conference of the Association for Computing Machinery, pp 84–100
Newell A (1980) Physical symbol systems. Cognitive Science 1(4):135–183

Newman M (2005a) Power laws, pareto distributions and zipf's law. Contemporary Physics 46:323–351

Newman M (2005b) Power laws, pareto distributions and zipf's law. Contemporary Physics 46:323–351

Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69, 066133

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69, 026113

Parsia B (2003) Message to www-rdf-comments@w3.org. Http://lists.w3.org/Archives/Public/www-rdf-comments/2003JanMar/0366.html

Pennebaker W, Mitchell J (1992) Joint photographic still image data compression standard. Standard, ISO

Ponte JM (1998) A language modeling approach to information retrieval. Phd dissertation, University of Massachusets

Postel J (August 1982) IETF RFC 821 Simple Mail Transfer Protocol. Http://www.ietf.org/rfc/rfc821.txt

Postel J (March 1994) IETF RFC 1590 Media Type Registration Procedure. Category: Informational. http://www.ietf.org/rfc/rfc1590.txt

Postel J, Reynolds J (October 1985) IETF RFC 959 File Transfer Protocol: FTP. Http://www.ietf.org/rfc/rfc959txt

Presutti V, Gangemi A (2008) Identity of resources and entities on the web. International Journal of Semantic Web and Information Systems 4(2):49–72

Prud'hommeaux E, Seaborne A (2008) Sparql query language for rdf. Recommendation, W3C, http://www.w3.org/TR/rdf-sparql-query/

Putnam H (1975) The meaning of meaning. In: Gunderson K (ed) Language, Mind, and Knowledge, University of Minnesota Press, Minneapolis, Minnesota, USA

Quillian MR (1968) Semantic memory. In: Minsky M (ed) Semantic Information Processing, MIT Press, Cambridge, Massachusetts, USA, pp 216–270

Quine WVO (1951) Two dogmas of empiricism. The Philosophical Review 60:20–43

Raggett D, LeHors A, Jacobs I (1999) HTML 4.01 Specification. Recommendation, W3C, http://www.w3.org/TR/REC-html40/

Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from flickr tags. In: Press A (ed) Proceedings of SIGIR'07, Amsterdam, The Netherlands, pp 103–110

Reiter R (1978) On closed world data bases. In: Logic and Data bases, Plenum Publishing, New York City, New York

Robertson S, Zaragoza H, Taylor M (2004) Simple bm25 extension to multiple weighted fields. In: Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), ACM, Washington, D.C., USA, pp 42–49

Robertson SE, Spärck Jones K (1976) Relevance weighting of search terms. Journal of the American Society for Information Science 27:129–146

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M (1994) Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference (TREC-3), pp 109–126

Robu V, Poutré JAL (2006) Retrieving utility graphs used in multi-item negotiation through collaborative filtering. In: Proc. of RRS'06, Hakodate, Japan (Springer LNCI, to appear)

Robu V, Halpin H, Shepherd H (2009) Emergence of consensus and shared vocabularies in collaborative tagging systems. ACM Transactions on the Web 3:14:1–14:34

Rocchio J (1971) Relevance feedback in information retrieval. In: Salton G (ed) The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Inc., Uppder Saddle River, New Jersey, USA, pp 313–32

Russell B (1905) On denoting. Mind 14:479–493

RVGuha, DLenat (1993) Language, representation and contexts. Journal of Information Processing 15(3)

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Tenth International WWW Conference (WWW10), Hong Kong

Schank R (1972) Conceptual dependency: A theory of natural language understanding. Cognitive Psychology 3(4):532–631

Schmidt-Schauss M (1989) Subsumption in kl-one is undecidable. In: Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning, pp 421–431

Shannon C, Weaver W (1963) The Mathematical Theory of Communication. University of Illinois Press, republished 1963

Shen K, Wu L (2005) Folksonomy as a complex network. Http://arxiv.org/abs/cs.IR/0509072

Silverstein C, Marais H, Henzinger M, Moricz M (1999) Analysis of a very large web search engine query log. SIGIR Forum 33(1):6–12

Simon H (1955) On a class of skew distribution functions. Biometrika 42(3/4)

Simondon G (1958) Du mode d'existence des objets techniques. Aubier, Paris, France, english Translation accessed on the Web at http://accursedshare.blogspot.com/2007/11/gilbert-simondon-on-mode-of-existence.html (Last accessed September 7th 2008)

Smith BC (1984) Reflection and semantics in LISP. Proceedings of 11th ACM SIGACT-SIGPLAN symposium on Principles of programming languages pp 23–35

Smith BC (1986) The correspondence continuum. In: Proceedings of the Sixth Canadian Conference on Artificial Intelligence, Montreal, Canada

Smith BC (1991) The Owl and the Electric Encyclopedia. Artificial Intelligence 47:251–288

Smith BC (1995) The Origin of Objects. MIT Press, Cambridge, MA

Smith BC (1996) On the Origin of Objects. MIT Press, Cambridge, Massachusetts

Smith BC (1997) One hundred billion lines of C++. LeHigh Cog Sci News 1(10), http://www.ageofsignificance.org/people/bcsmith/papers/smith-billion.html

Smith BC (2002) The Foundations of Computing. In: Scheutze M (ed) Computationalism: New Directions, MIT Press, Cambridge, Massachusetts, USA

Sollins K, Masinter L (1994) IETF RFC 1737 Functional Requirements for Uniform Resource Names. Http://www.ietf.org/rfc/rfc1737.txt (Last accessed April 20th 2008

Sowa J (1976) Conceptual graphs for a data base interface. IBM Journal of Research and Development 20(4):336–357

Sowa J (1987) Semantic Networks. In: Shapiro S (ed) Encyclopedia of Artificial Intelligence, Wiley and Sons, New York City, New York, USA, pp 1011–1024

over Structured Web Data ESE (2011) Roi blanco and harry halpin and daniel herzig and peter mika and jeffrey pound and henry thompson and thanh tran duc. In: Proceedings of the 1st International Workshop on Entity-Oriented Sarch workshop on Entity-Oriented Search (SIGIR 2011), ACM, New York, NY, USA

Suchanek FM, Vojnovic M, Gunawardena D (2008) Social Tags: Meaning and Suggestions. In: 17th ACM Conference on Information and Knowledge Management (CIKM 2008)

Tarski A (1935) The concept of truth in formalized languages. Studia Philosophia 1:261–405, reprinted in Logic, Semantics and Metamathematics (1956), Oxford University Press, Oxford United Kingdom, (1956), JH Woodger (trans.)

Tarski A (1944) The semantic conception of truth and the foundations of semantics. Philosophy and Phenomenological Research 4:341–375

Thompson H, Beech D, Maloney M, Mendelsohn N (2004) XML Schema Part 1: Structures. Recommendation, W3C, http://www.w3.org/TR/xmlschema-1/

Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460

Wadler P (2001) The Girard-Reynolds Isomorphism. In: International Symposium of Theoretical Aspects of Computer Software

Waldrop MM (2001) The Dream Machine: J.C.R. Licklider and the Revolution That Made Computing Personal. Penguin, New York City, New York, USA

Walsh N, Thompson H (2007) Associating resources with namespaces. TAG Finding. http://www.w3.org/2001/tag/doc/nsDocuments/

Watts D, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

Wheeler M (2005) Reconstructing the Cognitive World: The Next Step. MIT Press, Cambridge, Massachusetts, USA

Wheeler M (2008) The Fourth Way: A comment on Halpin's 'Philosophical Engineering'. APA Newsletter on Philosophy and Computers 8(1):9–12

Wiener N (1948) Cybernetics or Control and Communication in the Animal and the Machine. MIT Press, Cambridge, Massachusetts, United States

Wilks Y (1975) A preferential, pattern-seeking, semantics for natural language inference. Artificial Intelligence 6(1):53–74

Wilks Y (2008) The semantic web: Apotheosis of annotation, but what are its semantics? IEEE Intelligent Systems 23(3):41–49

Winograd T (1972) Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. Cognitive Psychology 3(1)

Winograd T (1976) Towards a procedural understanding of semantics. Stanford Artificial Intelligence Laboratory Memo AIM-292

Wittgenstein L (1921) Tractatus Logico-Philosophicus. Routledge, New York City, New York, USA, republished 2001

Wittgenstein L (1953) Philosophical Investigations. Blackwell Publishers, London, United Kingdom, republished 2001

Woods W (1975) What's in a link: Foundations for semantic networks. In: Representation and Understanding: Studies in Cognitive Science, Academic Press, Inc., Orlando, Florida, USA, pp 35–82

Xu J, Croft WB (1996) Query expansion using local and global document analysis. In: Proceedings of the Nineteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp 4–11

Yule G (1925) A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. Philosophical Transactions of the Royal Society of London, Ser B 213:21–87

Zimmerman H (1980) The ISO model of architecture for Open Systems Interconnection. IEEE Transactions on Communications 28(4):425–432