

A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative

Jewel Ward <jewelw@lanl.gov>

The University of North Carolina

Los Alamos National Laboratory

ACM/IEEE JCDL 2003 Houston, Texas

29 May 2003, updated 10 June 2003

Outline

- Background
- Purpose of the Study
- Methodology
- Results
- Summary and Conclusions
- Further Information
- Questions?

Background

A repository must expose its metadata as unqualified DCMES in addition to or in lieu of a native metadata format because:

- it facilitates cross-domain resource discovery
- it provides for DL interoperability

but critics would like a format that provides more detail about resources.

Purpose of the Study

To examine:

- which elements are used or not used?
- which elements are used the most and least?
- if there are different “types” of DPs, does element usage vary by “type” of DP?

Or to put it another way...is the unqualified DCMES currently used to its fullest extent?

Methodology

- Metadata harvested between May and November, 2002
- Harvested records from 82 of 100 DPs, but only 76 of the 82 could be harvested consistently (OAI-PMH 1.1 was experimental)
- Wrote a Perl program to count the number of metadata records and the number of times a record contained each element
- DP “type” was determined by issuing an `Identify` request or visiting the DPs web site

Results – Data Providers

Three “types” of DPs

- Scientific and Technical (STI): 27
- Humanities: 33
- “Combo” (STI and Humanities): 22

Results – Metadata Records

- Total number of records harvested from 82 DPs: 910,919
- Average number of records per DP: 11,109
- Number of records as a percentage by type of DP (n/910,919):
 - Humanities: 43%
 - STI: 21%
 - Combo: 26%

Results – Metadata Elements

As a ratio against the total number of records:

- Average of 8 elements used per record
- Average of 91,785 elements used in each DP

Results – Metadata Elements

As a percentage of the total number of elements and records (chart p. 316 of JCDL proceedings):

- Top 5 elements account for 71% of all usage:
Most- to least- used: **creator**, **identifier**, **title**, **date**, and **type**
- Bottom 5 elements account for 6% of all usage:
Most- to least- used: **language**, **format**, **relation**, **contributor**, and **source**

Results – Metadata Elements

As a percentage of all DPs (cross tabulated)
(chart p. 316 of JCDL proceedings) :

- Top 5 elements remained the same, but the order changed

Most- to least- used: **title**, **creator**, **date**, **identifier**, and **type**

- Bottom 5 elements changed out two elements compared to the previous order

Most- to least- used: **rights**, **contributor**, **source**, **coverage** and **relation**

Results – Metadata Elements

Most-used elements by type of DP:

- STI: creator, title, identifier, type and date
- Humanities: creator, title, identifier, type and rights
- Combo: creator, title, identifier, type and date

Results – Metadata Elements

Element usage by type of DP:

- STI: **creator** and **identifier** > 1/2 of usage
- Humanities: **title**, **identifier**, **creator** and **type** == 48% of usage
- Combo: **creator**, **identifier**, and **date** == 60% of usage

Results – Metadata Elements

Just 2 elements – **Creator** and **Identifier** –
account for approximately 1/2 the
number of elements used in 54% of
DPs

Summary and Conclusion

- The top five elements are used 71% of the time while the bottom five elements are only used 6% of the time
- 54% of DPs use only two elements (**creator**, **identifier**) for approximately 50% of all usage
- The DCMES is not used to its fullest extent <understatement>
- Due to this underutilization of the DCMES, it will be difficult for the OAI community to build relevant cross-resource services based on it

Further Information

<http://www.foar.net/research/>

Click on the link, “Master’s Paper”.

Questions?