Metadata Scope and its Implications for Determining Data Relevance on Relevance for Humans & Machines

The scope of a metadata record has a considerable effect on a user's ability to find relevant resources. Information retrieval experiments performed by Cyril Cleverdon and later confirmed by Alan Seal determined that short entry catalogs were generally easier for users in determining a resource's relevance than long entry catalogs. Applying these findings to metadata for datasets presents questions of scope and granularity.

Enhancing discoverability is a major task affecting the sharing and reuse of data collections. Particularly important is whether or not a dataset, as described by its metadata record, is deemed relevant by the information seeker. As datasets are not self-describing, this places critical importance on the metadata record for aiding discoverability. This poster examines some of the key elements required in creating metadata suitable for relevance judgements by both individuals and machine information retrieval (IR) methods from an analysis of 636 metadata records from a prominent repository

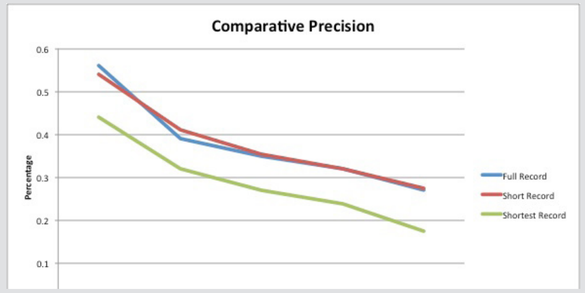
Methods

We collected 636 records from a scientific archive and ran information retrieval tests over the corpus to determine how the number of elements used in a record affected relevance judgements for machines.

Our IR results led us to observe that the "Description" element is particularly important in determining relevance. Effective use of this element required an accurate summary of the nature of the data, potential issues involved with its collection, and the dataset's coverage. Effective descriptions also included:

- Organizations and projects being fully named before being referred to by acronyms
- Specific, measured information on timeframes, locales, and methods
- Avoidance of institutional or discipline specific jargon

Note the very close level of precision between the full and shortened records. Since the subject field in our corpus contained the same value for the majority of the datasets, discoverability critically depended on the description element.



The three versions of each record consist of the following Dublin Core fields:

- **Long** - Title, Creator, Contributor, Subject, Description, Type, Coverage, Format, Date, Identifier, Language, Publisher, Relation, Rights, Source
- **Short** - Title, Subject, Description, Creator
- **Shortest** - Title, Subject, Creator

Conclusions

The finding that shorter catalog records are easier for determining relevance is complemented by our experiment's results in that precision metrics differed little between tests using long and short records. The impact for metadata creators in assessing the level of granularity needed to describe a particular dataset may be guided by the knowledge that a longer record is not necessarily going to enhance discoverability or aid in relevance judgements. Highly detailed provenance data may be more appropriately included in an attached readme file.

There were several issues at play during our analysis, the most pressing of which were issues of quality control and standardization. There was a wide variance in the depth of the values for a given metadata record, especially for dc_description. Further research would look at the length of values in this field and its effects on information retrieval metrics. Future analyses would also repeat this pilot investigation with a substantially larger corpus.

ILLINOIS 1867 UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

More at: metadatascope.wordpress.com

Erik Radio, MS Candidate, radio1@illinois.edu

Brian Balsamo, MS, balsamo1@illinois.edu