Interface Design for Metadata Creation

Abe Crystal
School of Information and Library Science
University of North Carolina at Chapel Hill
acrystal@email.unc.edu

ABSTRACT
The rapid growth of the Web has increased the importance of decentralized metadata creation. Resource authors must create their own metadata to enable enhanced information seeking and retrieval, and they need effective interfaces to support their work. This paper reports a baseline study of author interactions with a metadata system and draws implications for the design of future interfaces.

Keywords
Metadata, interaction design, user interfaces

INTRODUCTION
With the exponential growth of the World Wide Web, the creation of high quality document surrogates has become even more vital [5]. In order to provide surrogates to users, a system must have access to relevant metadata for the documents it manages. Metadata—literally, data about data—may be defined as structured data that facilitates related functions, such as resource discovery [2].

While the value of metadata is well understood, appropriate methods for creating and using web resource metadata are still under development. Initial research suggests that resource authors can create metadata suitable for use in surrogates that can facilitate information seeking and retrieval. Usability problems limit authors' facility with metadata creation and suggest a need for better interfaces.

METHODS
As part of a study [3] conducted at the National Institute for Environmental Health Sciences (NIEHS), we identified scientists who had created the intellectual content of various websites. Screen-capture software was used to record the user interface activities of six scientists as they created Dublin Core metadata [4] using a web application. The resulting logs were then analyzed to extract quantitative (time spent on particular tasks and events) and qualitative (observations of apparent problems) usability data. Pre- and post-task questionnaires were administered to assess users' satisfaction with the process.

CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-630-7/03/0004.

RESULTS
Users spent most of their time focused on only a few subtasks (defined as time spent harvesting and entering information for a specific element).

Subtask	Time (% of total task time)
Relation	23.1
Subject	18.5
URL	9.6
Description	9.3
Date modified	5.8

Table 1: Highest average subtask times. Average task time was 353.5 s.

In general, the most time-consuming subtasks were those associated with problematic interaction design. We observed interaction issues in three primary areas: task flow, metadata quality and the graphical user interface.

Task flow
We observed frequent backtracking and deviations from the expected linear progression through the application. This suggests user uncertainty or confusion, possibly reflecting an anomalous state of knowledge [1] that must be clarified before proceeding. Users frequently spent time editing their inputs to meet the constraints of a particular field, suggesting a need to support looser, more flexible inputs.

Metadata quality
Metadata as generated were generally evaluated positively [3]. However, some authors left useful fields (such as description, which is important for resource discovery and browsing) empty or incomplete. Occasionally, authors misinterpreted field semantics or syntax. One author entered "Microsoft Word" for source, whereas the desired input is the identifier of the original resource (e.g., a URL or ISBN). Another author, apparently confused by a multiline field with the caption "enter one per line," entered the abbreviation ROC (for Report on Carcinogens) as R | O | C—one letter on each line. These misinterpretations lead to meaningless metadata.

Graphical user interface
Users appeared frustrated with some core interface features. Scrolling, particularly horizontal scrolling, greatly slowed users and may have contributed to less

exhaustive metadata creation, as users moved away from metadata elements associated with interface awkwardness. Copying and pasting was often problematic, due to inconsistent behavior. One author, confused by a hidden hard return, selected only one line rather than both lines of a web page's title. Perhaps the most time-consuming interface problem was switching between different contexts—such as the web application and another browser window. This operation evidently challenged users' short-term memory, as we often observed switching contexts multiple times in rapid succession while completing a particular subtask.

IMPLICATIONS FOR DESIGN

For decentralized metadata creation to become a reality, better designs are needed to reduce users' cognitive load and lower barriers to efficient metadata entry. We see many opportunities, which can be organized in three themes: integration, filtering/flagging, and context.

Integration

Currently, there is no apparent integration between metadata harvesting—the collection of useful information for conversion into structured metadata—and entry. Better affordances here could reduce the costs of context switching. For example, agents could automatically harvest simple metadata from web resources, then present their output to authors for editing and annotation. Existing tools (e.g., DC.Dot and Klarity) point the way here, but are fairly crude, and have yet to be integrated with authoring interfaces. Another approach is to combine metadata creation and content creation (e.g., Adobe applications that offer built-in support for Dublin Core metadata records).

Filtering/Flagging

It would be useful to filter or flag "weak" (e.g. incomplete, or syntactically or semantically misconstrued) metadata in order to support later resource discovery. Researchers might consider developing heuristics to identify weakness, such as the number of bytes in a field or record, either absolutely or as a percentage of a "typical" size. Inter-field correlations could be examined to identify overly redundant or disparate metadata. The interface could simply be designed to display newly created metadata in context—for example, as a part of a sample search result, so users could self-evaluate.

Once a metadata record is identified as weak, it could be filtered out, used only for certain purposes, or forwarded to a cataloger for review or annotation.

Context

Metadata creation interfaces would likely benefit from richer contextual information. Additional context could help improve user motivation, if giving users a better understanding of metadata use aids their work. Motivation is critical in decentralized systems because individual authors have little incentive to provide quality metadata for their resources (this is analogous to the "cold-start" problem in recommender systems [6]). Contextual information could also help users enhance input quality, avoiding weak metadata, as well as encouraging the identification of links to related resources.

Information design is crucial here. Users need better examples of good metadata effectively integrated with the interface. They also need better guidelines for determining the granularity of complex elements such as subject descriptors. However, incorporating this additional information into an easily comprehensible and usable interface remains a challenge.

CONCLUSION

Key areas of current research rely heavily on effective metadata interfaces, but very little existing research is relevant to the design and use of these interfaces. In particular, implementation of digital libraries and Semantic Web applications will depend heavily on decentralized metadata creation, which implies the need for highly usable metadata systems.

ACKNOWLEDGMENTS

This work is part of the Metadata Generation Research Project (http://ils.unc.edu/~janeg/mgr/), funded by Microsoft Research and OCLC. I am grateful to my advisors, Dr. Jane Greenberg and Dr. Gary Marchionini, for their advice and encouragement.

REFERENCES

1. Belkin, N.J. Anomalous States of Knowledge as a basis for information retrieval. *Canadian Journal of Information Science 5*: pp. 133-143, 1980.
2. Greenberg, J. Metadata and the World Wide Web. *Encyclopedia of Library and Information Science: 72* (New York: Marcel Dekker, Inc.), pp. 244-261, 2002.
3. Greenberg, J., Pattuelli, M. C., Parsia, B., & W. D. Robertson. Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. *Journal of Digital Information (JoDI), 2(2)*: 2001.
4. Harper, C., Greenberg, J., Robertson, W. D., & Leadem, E. Abstraction Versus Implementation: Issues in Formalizing the NIEHS Application Profile. In *Proceedings of DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence* (Florence, October, 13-17, 2002).
5. Lagoze, C.. From static to dynamic surrogates: resource discovery in the digital age. *D-Lib Magazine*: June 1997.
6. Perguini, S and Goncalves, M. A. Recommendation and personalization: a survey. Technical Report cs.IR/0205059, Computing Research Repository: 2002. Available at http://xxx.lanl.gov/abs/cs.IR/0205059.