The AMeGA Project

Final Report for the AMeGA (Automatic Metadata Generation Applications) Project

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

THE LIBRARY OF CONGRESS

Authors: Jane Greenberg, Kristina Spurgin, and Abe Crystal
Appendix A Authors: Michelle Cronquist and Amanda Wilson

Principal Investigator: Jane Greenberg, Associate Professor
School of Information and Library Science,
University of North Carolina at Chapel Hill

Final Report for the AMeGA (Automatic Metadata Generation Applications) Project

Report Authors: Jane Greenberg (Principal Investigator), Kristina Spurgin, and Abe Crystal

Appendix A Authors: Michelle Cronquist and Amanda Wilson

AMeGA Project Website: http://ils.unc.edu/mrc/amega.htm

Submitted to the Library of Congress February 17, 2005 by:

Jane Greenberg, Associate Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
205 Manning Hall
Chapel Hill, NC 27599
Voice: 919-962-8066
Fax: 919-962-8071
E-mail: janeg@ils.unc.edu

Endorsed by AMeGA Task Force Members

Brad Allen, President, Siderean Software

John D. Byrum, Jr. Chief, Regional & Cooperative Cataloging Division, Library of Congress

Priscilla Caplan, Assistant Director, Digital Library Services, Florida Center for Library Automation

Tim Cole, Mathematics Librarian & Professor of Library Administration, University of Illinois at Urbana-Champaign

Susan J. Neill, Metadata Consultant

Ed O'Neill, Consulting Research Scientist, OCLC, Online Computer Library Center

W. Davenport Robertson, Library Director, National Institute of Environmental Health Sciences

Robin Wendler, Metadata Analyst, Office for Information Systems, Harvard University Library

David Williamson, Cataloging Automation Specialist, Library of Congress

Amanda Wilson, Metadata Librarian, The Ohio State University

Mary Woodley, Collection Development Coordinator, California State University, Northridge

Acknowledgements

We would like to acknowledge the Library of Congress for funding that made this research possible and AMeGA Metadata Task Force Members (http://ils.unc.edu/mrc/amega_task.htm) for their review of metadata expert survey and assistance with participant recruitment. We would also like to thank the survey participants for their time and interest in this project and Michelle Mascaro, Graduate Research Assistant, for her valued assistance.

Cover design by Cynthia McCracken and AMeGA logo design by Michelle Mascaro.

Table of Contents

List of Tables and Figures

Tables

Figures

Executive Summary

Never has there been such a wealth of valuable information accessible to the global public as there is with the World Wide Web (Web). It can also be argued, however, that never has there been such an abundance of easily accessible information that is factually incorrect, misleading, and lacking authentication. This situation has had a tremendous impact on the role of the library—an institution founded on principles of evaluating, collecting, organizing, and providing access to information. Libraries and information centers of every domain look to leading institutions such as the Library of Congress (LC) to demonstrate specifically how to deal with these new and growing information challenges. Libraries want to know how to best leverage new information technologies in order to serve their constituents and support the flow of information that is so vital to society's progress and development.

The Automatic Metadata Generation Applications (AMeGA) project, which was conducted in conjunction with the *Bibliographic Control of Web Resources: A Library of Congress Action Plan* (http://www.loc.gov/catdir/bibcontrol/actionplan.pdf), addresses the challenge of metadata generation for digital resources. The work underlying the AMeGA project was guided by the following three goals:

- To evaluate current automatic metadata generation functionalities supported by content creation software and automatic metadata generation applications; and review automatic metadata generation functionalities supported by integrated library systems (ILSs).
- To survey metadata experts to determine which aspects of metadata generation are most amenable to automation.
- To compile a final report of recommended functionalities for automatic metadata generation applications.

Goal 1

Work Performed:

- A literature review of research in the area of automatic metadata generation was conducted.
- Automatic metadata generation functionalities supported by seven different types of content creation software (software used to create a resource) were reviewed.
- A comparison of Klarity and DC-dot, two state-of-the-art automatic metadata generation applications (tools designed specifically for generating metadata), was conducted (see Greenberg, 2004b, for results).
- A list of automatic metadata generation features advertised by selected ILS vendors and found in current literature was compiled, and an interview was conducted with a cataloger using one of the more advanced ILS systems (see Appendix A for results).

Summary of Findings:

Research in the area of automatic metadata generation falls, primarily, into two areas: *Experimental research*, focusing on information retrieval techniques and digital resource content, and *applications research*, focusing on the development of content creation software and metadata generation tools used in the operational setting. The main finding, presented in this report, is that there is a *disconnect* between experimental research and application development. It seems that metadata generation applications could be vastly improved by integrating experimental research findings.

Metadata generation applications might also improve metadata output if they took advantage of metadata generation functionalities supported by content creation software. For example, Microsoft Word supports the metadata generation of a number of elements that conceptually map to the Dublin Core metadata standard. Some of these elements are generated automatically, while others need to be input by a document author or another person. Content creation software provides a means for generating metadata, which can be harvested by metadata generation applications. More research is needed to understand how metadata creation features in content creation software are used in practice.

Goal 2

Work Performed:

A survey was developed to identify system functionalities desirable for automatic metadata generation applications. The survey gathered data on the following:

- Participants and their metadata/cataloging experience.
- Current organizational metadata practices.
- Participants' knowledge and opinions about automatic metadata generation for Dublin Core metadata standards.
- Participants' knowledge and opinions about automatic metadata generation in general.
- Participants' opinions about desired functionalities for automatic metadata generation applications.

The study of participant and organization metadata practices was primarily restricted to digital document-like objects (DDLOs). A DDLO was defined as "primarily textual resource that is accessible through a Web browser" (Greenberg, 2004a).

Summary of Findings:

- Two-hundred and seventeen (217) survey participants provided responses useful for data analysis (the initial goal was to recruit at least 100 participants).
- Three quarters of participants had three or more years of cataloging and/or indexing experience, verifying their status as metadata experts.
- Organizations are using a variety of different metadata standards (selected examples include: MAchine Readable Cataloging (MARC)—bibliographic format, Dublin Core, Encoded Archival Description, Gateway to Educational Materials, Metadata Object Description Schema, Text Encoding Initiative, and the Government Information Locator Service).
- Most participants (81%) reported using one or two systems for metadata creation in their organization, whereas one participant reported the use of seven different systems.

- Dublin Core element rankings:
    - Participants predicted greater accuracy when using automatic processing for technical metadata (e.g., *ID*, *language*, and *format*) than for metadata requiring intellectual discretion (e.g., *subject* and *description*).
    - Participants thought it was more appropriate to apply automatic processing to technical metadata and machine-readable metadata (e.g., *language*), as opposed to metadata requiring more intellectual discretion.
    - Participants' opinions on resource allocation for automatic generation of Dublin Core elements revealed a *fundamental tension* between metadata *usefulness* and *feasibility*. Participant commentary highlighted the greater need for contextual understanding of metadata and the metadata creation process.
- Participants clearly supported automatic metadata generation, although most participants (96%) were unwilling to recommend fully automatic techniques for metadata generation. Most participants preferred that an application execute automatic metadata generation functionalities first, but then provide a means for human evaluation and manual intervention.
- Participants indicated that it is very important, and in some cases critical, to support automatic metadata generation for nontextual resources.

Goal 3

Work Performed:

A final report of recommended functionalities for automatic metadata generation applications was produced (see Section 8).

Summary of Findings:

The recommendations are organized as follows:

- System Goals
- General System Recommendations
- System Configuration
- Metadata Identification/Gathering

- Support for Human Metadata Generation
- Metadata Enhancement/Refinement and Publishing
- Metadata Evaluation
- Metadata Generation for Nontextual Resources

The recommendations are identified as Version 1.0 because it is likely that they will be enhanced and modified over time, with greater input from the larger bibliographic control/metadata community.

Recommended Next Steps for the Library of Congress

The final portion of this report presents a three-pronged approach to developing an automatic metadata generation application for LC. The plan is comprised of multiple components. The three main tasks are to:

- Build an automatic metadata generation application.
- Foster and facilitate research on automatic metadata generation.
- Implement mechanisms for communicating and negotiating with content creation software vendors.

The recommended next steps also highlight important metadata research questions requiring further examination. Finally, the recommendations stand as a part of LC's initiative to lead many different communities in which metadata plays a vital role.

1. Introduction

The need for metadata supporting World Wide Web (Web) resource description, discovery, and other functions is radically increasing as the Web becomes a major means for communicating and disseminating information. Metadata, drawing from library practices, is a fundamental component of digital libraries and Web initiatives such as the Semantic Web (http://www.w3.org/2001/sw/), Open Archives Initiative (http://www.openarchives.org/), and D-Space (http://www.dspace.org). The sheer mass of resources requiring metadata is a major challenge for all these projects. Library metadata practices, whereby professionals (catalogers and indexers) generate metadata, are rendered prohibitively expensive by the limited availability of qualified persons and financial resources. For libraries to *advance* and *take leadership* in the bibliographic control of Web resources, they must investigate more efficient and less costly metadata creation methods.

Automatic metadata generation can help address this need. Drawing from automatic indexing developments (Anderson & Perez-Carballo, 2001), automatic metadata generation is *more efficient*, *less costly*, and *more consistent* than human-oriented processing. In fact, research indicates that automatic metadata generation can produce acceptable results (Han et al., 2003; Liddy et al., 2002; Takasu, 2003), although results can be problematic at times—particularly for metadata requiring human intellectual discretion (Greenberg, 2004b). Researchers have stated that the most effective results can be achieved by integrating both human and automatic methods (e.g., Schwartz, 2000). This point and the increasing demand for metadata underscore the library community's need to explore how automatic metadata generation can complement or serve as an alternative to traditional library metadata activities.

The *Bibliographic Control of Web Resources: A Library of Congress Action Plan* (LC Action Plan) (http://lcweb.loc.gov/catdir/bibcontrol/actionplan.pdf) recognizes this need and highlights *automatic metadata generation tool development* as a "near-term/high" priority. *LC Action Plan* Section 4.0 targets the development of "automatic tools…to improve bibliographic control of selected Web resources," and Section 4.2 specifically identifies the need for a master specification to guide development of such applications. To address this need, we established the Automatic Metadata Generation Applications (AMeGA) project (http://ils.unc.edu/mrc/amega.htm). The goal of the AMeGA project was to identify and recommend functionalities for applications supporting automatic metadata generation in the library/bibliographic control community. Specific project goals were to:

- Evaluate current automatic metadata generation functionalities supported by *content creation software* and *automatic metadata generation applications*, and review automatic metadata generation functionalities supported by *integrated library systems* (ILSs).
- Survey metadata experts (professional catalogers and indexers, and persons knowledgeable about metadata creation) to determine which aspects of metadata generation are most amenable to automation.
- Compile a final report of recommended functionalities for automatic metadata generation applications.

The following report is arranged as follows: Section 2 provides a brief overview; Section 3 reviews both experimental and application-oriented research; Section 4 presents the study's underlying research questions; Section 5 reviews the research design and procedures; Section 6 presents the study's results; Section 7 provides a contextual discussion of the results; Section 8 outlines recommended functionalities for automatic metadata generation applications; Section 9 highlights the study's key findings, presents conclusions, and points to important research directions; and Section 10 outlines recommended next steps for the Library of Congress.

2. Automatic Metadata Generation

Automatic metadata generation, in its purest form, depends solely on machine processing. It is often defined by distinguishing it from metadata generated by a person. Most automatic metadata generation operations require a human to initiate the process; many operations manipulate metadata previously produced by humans.

Metadata extraction and *metadata harvesting* have been identified as two methods of automatic metadata generation applicable to digital resources (Greenberg, 2004b). Metadata extraction involves the mining of resource content and employs sophisticated automatic indexing techniques to produce structured ("labeled") metadata for object representation. Metadata harvesting relies on machine capabilities to collect tagged metadata previously created by humans, machine processing, or both.

Automatic metadata generation is being explored by researchers because of the important efficiency, cost, and consistency advantages of automatic indexing over human-controlled processes (Andersen, 2002). The use of automatic processing can, in turn, permit human resources to be directed to more intellectually challenging metadata creation and evaluation tasks. These factors underlie automatic metadata generation research efforts and the desire to build superior and robust automatic metadata generation applications, and are central to the AMeGA project.

3. Automatic Metadata Generation Research

Automatic metadata generation research is rooted in automatic indexing, abstracting, and classification research, which began with the availability of electronic text in the early 1950s. Early work in these areas primarily addressed the generation of subject descriptors/keywords and abstracts (e.g., Salton & McGill, 1983). Today, automatic metadata generation has moved beyond subject representation to encompass the production of *author*, *title*, *date*, *format*, and many other types of metadata. In addition, thousands of information databases are now networked via the Internet, and information resources are frequently rendered in open and interoperable standards (e.g., eXtensible Markup Language (XML)). These developments have enabled automatic metadata generation systems to work on far larger corpora. Automatic metadata generation research efforts, stemming from these developments, can be classified into two areas: *Experimental research*, focusing on information retrieval techniques and digital resource content; and *applications research*, focusing on the development of content creation software and metadata generation tools used in the operational setting. These two areas are discussed below.

3.1 Experimental Research and Digital Resource Content

The awesome growth of digital resource repositories provides an abundance of rich collections for studying automatic metadata generation. Researchers manipulating digital resource content for metadata generation have experimented primarily with *document structure* and *knowledge representation systems*.

3.1.1. Document Structure

Researchers have identified relationships between document genre, content, and structure (Toms, Campbell & Blades, 1999). For example, document genre can inform textual density, which can be used to predict metadata extraction algorithm performance for certain types of documents (Greenberg, 2004b). Document genres also exhibit predictable document structures that have proved amenable to algorithmic extraction. For example, research papers include standard information such as document "title," "author," and "author affiliation." Experiments exploiting document structure in this second venue using a Support Vector Machine (SVM) algorithm (e.g., Han et al., 2003) and Variable Hidden Markov Model (DVHMM) (Takasu, 2003) have been fairly successful for metadata generation.

3.1.2 Knowledge Representation Systems

Digital technology has greatly increased the electronic availability of ontologies, thesauri, classificatory systems, authority files, and other knowledge representation tools. This development and the Web's global framework have led to the construction of metadata registries specifically for sharing knowledge representations systems (e.g., Lutes, 1999; Knowledge System Laboratory (KSL) Ontology Server, Stanford University: http://www-ksl-svc.stanford.edu:5915/doc/ontology-server-projects.html) and metadata schemes (SCHEMAS Registry: http://www.schemas-forum.org/registry/; Dublin Core Metadata Registry: http://dublincore.org/dcregistry/). These resources provide another source for automatic metadata generation research.

Patton, Reynolds, Choudhury, and DiLauro (2004) provide an example of automatic name authority control by matching names detected in resource content to names in LC authority file. Liddy et al., (2002) provide another example of research in this area, using natural language processing algorithms, educational domain vocabularies, and the content of educational resources to generate metadata following the Gateway to Education Materials (GEM) metadata standard. Teachers and other users of educational resources, evaluating their work, were nearly as satisfied with the automatically generated metadata as they were with humanly generated metadata.

3.1.3 Summary of Experimental Research

Experimental research focusing on document content has advanced knowledge about automatic metadata generation. One shortcoming is that testing is generally limited to specific subject domains, resource types, resource formats, and metadata elements. Researchers recognize the limitations of algorithms developed for domain vocabulary, however, and have developed prototype tools to employ different ontologies for metadata generation (Hatala & Forth, 2003). More research is needed to determine which approaches would be broadly applicable in metadata applications.

3.2 Applications and Automatic Metadata Generation

Growing recognition of the importance of metadata has led to the development of a variety of applications supporting metadata creation. General *content creation software* and more specialized tools known as *metadata generation applications* both support metadata creation for digital resources. These applications are generally used by resource authors or by other persons who do not have professional training in metadata creation (Greenberg, 2003). ILSs also support metadata creation for digital resources and include automatic functionalities to enhance and maintain metadata quality. ILS cataloging modules are generally used by metadata professionals (catalogers and indexers) and technical assistants who have undergone some training.[1]

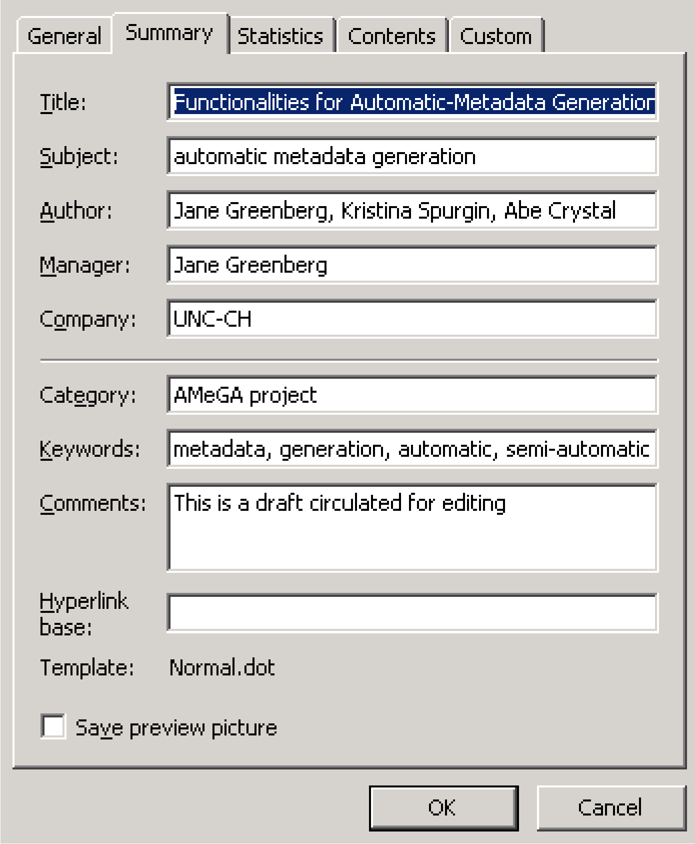
3.2.1 Content Creation Software

Content creation software includes the wide range of software used to create resources. Examples include Microsoft Word, Macromedia Dreamweaver, Adobe Acrobat, and Nullsoft Winamp—essentially any software that can be used to create digital content, whether textual or multimedia. In the context of the Web, content creation software is used to create a digital resource that can be accessed via a standard Web browser and associated software. A bibliographic surrogate can also be considered content. Thompson ResearchSoft's EndNote, which is used to create metadata records, can therefore be considered a form of content creation software.

[1] ILSs are not reviewed in this section due to practical research constraints (see Appendix A).

Content creation software increasingly supports metadata generation via automatic, semi-automatic, and human means. Automatic techniques are frequently employed to produce technical metadata such as *date_created*, *date_modified*, *size* (e.g., bytes), and *format* metadata. Some content creation software extracts metadata from document content in an attempt to provide descriptive representations (e.g., Word automatically assigns a *title* based on the first line of a document). Some content creation software includes a template to facilitate human metadata entry. Automatic techniques may then be employed by the application to convert the entered metadata into a specified encoding language (e.g., XML), embed it in a resource header, or insert it into a metadata database. Word has this functionality. Figure 1 shows Word's Summary dialog box, where users may enter metadata. When a Word document is saved as an HTML file, this metadata is automatically integrated with metadata from other dialog boxes and encoded in the document as an X/HTML header, as represented in Figure 2.

Figure 1: A Properties Dialog Box for a Word File



Title: Functionalities for Automatic-Metadata Generation

Subject: automatic metadata generation

Author: Jane Greenberg, Kristina Spurgin, Abe Crystal

Manager: Jane Greenberg

Company: UNC-CH

Category: AMeGA project

Keywords: metadata, generation, automatic, semi-automatic

Comments: This is a draft circulated for editing

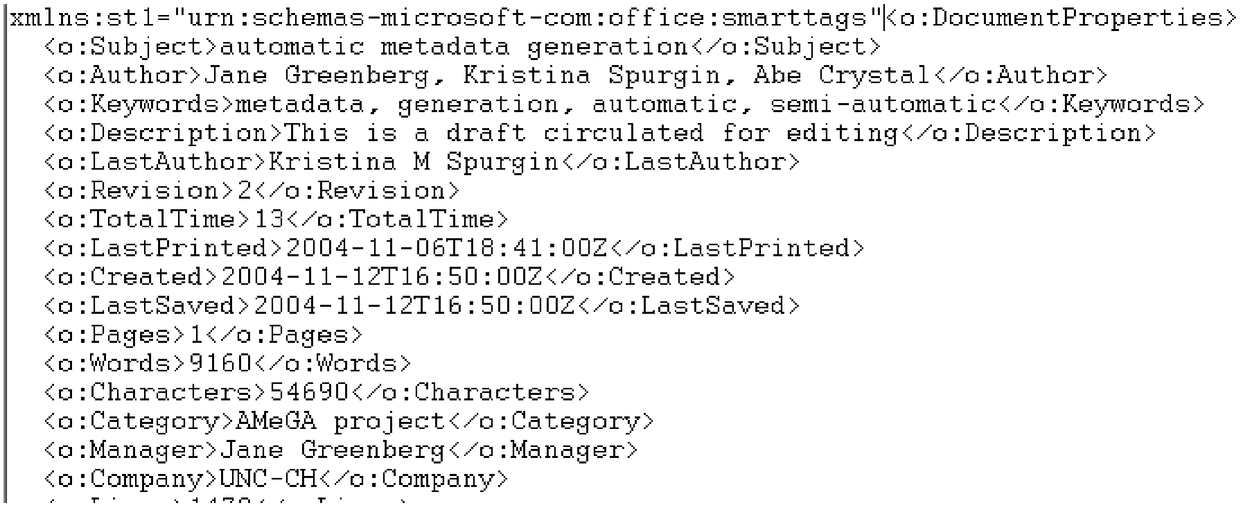
Hyperlink base:

Template: Normal.dot

☐ Save preview picture

Figure 2: Automatically Encoded X/HTML Document Header

This metadata record includes descriptive metadata humanly input into the Summary dialog box (see **Figure 1**), content creation and descriptive metadata (e.g., *TotalTime*, *LastPrinted*, *Created*) automatically generated in the General dialog box, and document statistical metadata (e.g., *Pages, Word, Characters*) automatically generated in the Statistics dialog box. The "o" prefix preceding each element name is for the "Office" (Microsoft Office) namespace.



```
xmlns:st1="urn:schemas-microsoft-com:office:smarttags"<o:DocumentProperties>
  <o:Subject>automatic metadata generation</o:Subject>
  <o:Author>Jane Greenberg, Kristina Spurgin, Abe Crystal</o:Author>
  <o:Keywords>metadata, generation, automatic, semi-automatic</o:Keywords>
  <o:Description>This is a draft circulated for editing</o:Description>
  <o:LastAuthor>Kristina M Spurgin</o:LastAuthor>
  <o:Revision>2</o:Revision>
  <o:TotalTime>13</o:TotalTime>
  <o:LastPrinted>2004-11-06T18:41:00Z</o:LastPrinted>
  <o:Created>2004-11-12T16:50:00Z</o:Created>
  <o:LastSaved>2004-11-12T16:50:00Z</o:LastSaved>
  <o:Pages>1</o:Pages>
  <o:Words>9160</o:Words>
  <o:Characters>54690</o:Characters>
  <o:Category>AMeGA project</o:Category>
  <o:Manager>Jane Greenberg</o:Manager>
  <o:Company>UNC-CH</o:Company>
```

Content creation software metadata is frequently used for computer desktop file organization and searching. Such metadata can be harvested during the creation of surrogates; in fact many operational metadata applications harvest this type of metadata. However, experimental automatic metadata generation research has not focused on this data source. One reason for this limitation may be that there has been little analysis of the types of metadata that content creation software supports via automatic means. The AMeGA project addresses this limitation by conducting a features analysis for content creation software.

3.2.2 Metadata Generation Applications

Metadata generation applications are increasingly being used to create metadata for Web resources. These applications differ from content creation software in that they are designed specifically, and only, to output metadata records. A list of applications facilitating the creation of records following the Dublin Core metadata standard is found at http://www.dublincore.org. The amount of automatic and human processing required to produce metadata distinguishes *generators*, which are metadata applications relying primarily on automatic techniques, and *editors*, which are applications integrating automatic and human processing (Greenberg, 2003; Meta Matters: http://www.nla.gov.au/meta/).

The increased availability of metadata generation applications is exciting because of the potential to vastly improve the efficiency and effectiveness of metadata production for Web resources. *State-of-the-art* applications are, however, limited by a number of factors:

- Applications rarely support standard bibliographic control functions such as authority control (the standardization of access points) and element qualification (DCMI Metadata Terms, 2004), which can facilitate the production of high-quality standardized metadata. By contrast, ILS cataloging modules generally provide satisfactory support for authority control, using automatic capabilities to link to authority files and insert headings in bibliographic records. ILSs generally support element qualification encoding via the MARC (Machine Readable Cataloging) bibliographic format, although a human needs to record this information.
- Automatic techniques are rarely exploited. It seems that experimental research findings—specifically, the development of sophisticated automatic indexing algorithms focusing on resource structure and knowledge representation systems—have yet to be fully incorporated into the current automatic metadata generation applications. Moreover, a wide-range of automatic indexing algorithms have been developed that could, potentially, support the generation of enhanced metadata by taking advantage of their domain foci.
- Applications are developed in isolation, failing to incorporate previous as well as new advances, partly because of the absence of standards or recommended functionalities guiding the development of metadata generation applications. A standard set of functionalities could inform the development of more robust automatic metadata generation applications. The image metadata community's Automatic Exposure Project (Research Libraries Group, 2003) provides an excellent example of the usefulness of standards for making progress. Sponsored by the Research Libraries Group (RLG), project participants have developed the *Data Dictionary: Technical Metadata for Digital Still Images* standard, National Information Standard Organization (NISO) Z39.87 (2002), which identifies technical metadata (e.g., *shutter speed* or *aperture setting*) that can be automatically recorded by image capture software and harvested by collection management tools for preservation purposes. The standard has been embraced by various industries and cultural heritage institutions, and project members

aim to develop a suite of tools for automatic harvesting and managing of technical metadata supporting the standard.

- Little attention has been directed to examining application usability, let alone effectiveness. Research has shown that the usability of metadata creation applications is an important issue that influences metadata quality as well as the efficiency of metadata creation (Crystal & Greenberg, *in press*; Greenberg et al., 2003). However, there is little evidence of any rigorous review of application usability.

Addressing these limitations could greatly improve the *state-of-the-art* automatic metadata generation applications. The AMeGA project uses these limitations as a basis for surveying metadata experts about desired system functionalities for automatic metadata generation applications.

4. Research Objectives

The primary goal of this research was to identify and recommend functionalities for applications supporting automatic metadata generation for Web resources. Two key objectives were identified to meet this goal. The first was to identify the types of metadata that content creation software supports via automatic and semi-automatic/human means. If metadata application developers have a better understanding of the types of metadata supported by content creation software, they may be able to better employ automatic techniques to take advantage of this rich source of data during metadata generation. The second research objective was to identify functionalities that metadata experts desire in automatic metadata generation applications. Metadata experts are knowledgeable about the range of important bibliographic control functions that facilitate creation of high-quality metadata (e.g., authority control). They are interested in and often aware of research that impacts their work, and they can provide useful insight into the types of system functionalities that require human input and those that can be executed by automatic means. Specific research questions guiding this study included the following:

1. What types of metadata does content creation software facilitate creating? What types of support (automatic and semi-automatic/human) does content creation software provide?

2. What system functionalities should be supported in automatic metadata generation applications?

The AMeGA project's objectives also include an evaluation of automatic metadata generation applications and a review of automatic metadata generation functionalities supported by ILSs. The AMeGA project is an extension of the Metadata Generation Research (MGR) project (http://ils.unc.edu/mrc/mgr_index.htm), which is developing a model for the most efficient and effective means of generating metadata integrating automatic and human means. The MGR project's comparison of two automatic metadata generation applications (Greenberg, 2004b) satisfied the AMeGA requirement in this area. A brief review of automatic metadata generation functionalities supported by ILSs was conducted, satisfying the ILS requirement (results are presented in Appendix A).

5. Research Methods

Two research approaches were used to address the study's research objectives. First, a *features analysis* was conducted to identify the metadata functionalities supported by content creation software frequently used to create digital resources. Second, a *survey of metadata experts* was conducted to identify desired system functionalities for automatic metadata applications.

5.1 Features Analysis

Hundreds of content creation software offerings are available—ranging from packaged software and open-source applications, to proprietary products developed in-house—making it impossible to identify a complete body of applications. For this reason—and because of the exploratory nature of this research—a selective sample was used for the features analysis. Software selection was guided by two main criteria. First, the software analyzed had to be commonly used to create digital resources that can be accessed via a standard Web browser and associated software. Second, the software needed to be accessible via the University of North Carolina at Chapel Hill, or freely accessible via the Web. These criteria were further refined into five document categories: general documents, websites, citations, music, and weblogs (blogs).

Two software applications were examined for the general documents and websites categories, while one software application was examined for each of the three other categories. The sample of software analyzed is presented in Table 1.

Table 1: Content Creation Software Sample

Type of Content Creation Software	Selected Sample
General document software (often heavily textual, although not exclusively)	Word (produced by Microsoft) Acrobat (produced by Adobe)
Website software	Dreamweaver (produced by Macromedia) CityDesk (produced by Fog Creek)
Citation software	EndNote (produced by Thomson ResearchSoft)
Music software (MP3 music)	Winamp (produced by Nullsoft)
Weblog software	Movable Type (produced by Six Apart)

The AMeGA features analysis had three components. First, an element analysis was conducted. Descriptive metadata elements supported by each application were mapped to the Dublin Core metadata standard to the extent possible. Dublin Core was chosen as a base because it is one of the most widely used, well-known, and simple metadata standards developed for Web resources. The mapping activity was guided by conceptual understanding of each element, rather than labeling. For example, EndNote's *URL* element was mapped to the Dublin Core *identifier* element, despite different labels. Elements that did not map to the Dublin Core were also tallied. The mapping activity distinguished elements supported by automatic means and those that required manual input. Crosswalk analyses (e.g., Woodley, 2001) and research on metadata generated by portal system software (Ji & Salendy, 2002) provide frameworks similar to the one underlying the features analysis.

The second component of the features analysis focused on the automatic metadata generation methods supported by each application. The following three methods were identified:

- Derived automatically from system properties. This method included metadata that a system automatically generates (e.g., *date_created*, *date_modified*, or *size*), as well as metadata stored in a user profile (e.g., *institutional_name* or *rights*) and automatically assigned to documents.

- Harvesting humanly generated metadata. For example, using the Z39.50 protocol (http://www.loc.gov/z3950/agency/) to read human-generated bibliographic information from a remote database (e.g., *title*, *subject*, *description*).
- Extraction from document content.

In addition, the features analysis identified user interface features supporting metadata creators during the metadata creation task.

5.2 Metadata Expert Survey

A survey was developed to identify system functionalities desirable for automatic metadata generation applications (see Appendix B). The survey was informed, in part, by the Consortium to Develop an Online Catalog (CONDOC, 1981), an ad hoc consortium formed in 1980, which conducted a survey in order to identify key features for online library catalogs—specifically, for small to medium-sized college and university libraries. Although the scope of AMeGA is much broader than CONDOC, the underlying rationale of *pooling expertise* because "collectively, the knowledge and skills of participants [experts] would be greater than if the project were attempted by a single institution" (Heyman, 1981) was key to the AMeGA project.

AMeGA project participants mainly included metadata experts with extensive experience creating metadata or administering metadata/cataloging activities. The survey gathered data on the participants and their metadata/cataloging experience, current organizational metadata practices, participants' knowledge of and opinions about automatic metadata generation for Dublin Core metadata standards, and participants' opinions about automatic metadata generation and desired functionalities.

The study of participant and organization metadata practices was restricted to *digital document-like objects* (DDLOs), defined as a "primarily textual resource that is accessible through a Web browser. DDLOs may contain images, sound, and non-textual formatting, but they must contain textual content (e.g., HTML/XHTML resources, Microsoft Word documents, Acrobat PDF documents)" (Greenberg, 2004a). The restriction was implemented because of research resource constraints. The survey was designed using SurveyMonkey.com and included both structured and open-ended questions (see Appendix A). The survey was extensively reviewed by AMeGA Metadata Generation Task Force (MGTF) members (http://ils.unc.edu/mrc/amega_task.htm), a group of 11 metadata experts, and was pilot-tested before being officially launched.

Participants were recruited via the following four methods: MGTF members recruited participants via personal and e-mail communication from their respective institutions; flyers were distributed at selected metadata/cataloging sessions at the annual American Library Association conference in Orlando, Florida, in June 2004; recruitment messages were distributed via electronic mailing lists of interest to communities of metadata experts working with digital resources (Table 2, column 1); and recruitment messages were distributed to three blogs of interest in the cataloging/metadata community (Table 2, column 2).

Table 2: Electronic Distribution for AMeGA Recruitment Message

Electronic Mailing Lists	**Weblogs**
▪ AutoCat (AUTOCAT@LISTSERV.ACSU.BUFFALO.EDU), ▪ METS (mets@loc.gov), ▪ Dublin Core General (DC-GENERAL@JISCMAIL.AC.UK) ▪ Open Archives Initiative General Interest List (oai-general-request@openarchives.org) ▪ CIC (Big Ten) Academic Libraries OAI List (OAI-CIC-L@LISTSERV.UIUC.EDU), ▪ CIC Library Metadata (cic-lib-metadata@cic.net), ▪ Serialst (SERIALST@LIST.UVM.EDU), ▪ OLAC (olac@listserv.acsu.buffalo.edu).	▪ Catalogablog (http://catalogablog.blogspot.com/2004_06_27_catalogablog_archive.html#108861938970165296) ▪ Infomusings (http://www.infomuse.net/blog/archives/2004_06.html#000794) ▪ Bibliolatry (http://www.bibliolatry.net/2004/07/meta.html).

A note at the bottom of the electronic mail recruitment encouraged forwarding the participant call to other electronic mailing lists of interest; the recruitment message was probably forwarded to other forums in addition to those listed in Table 2.

6. Results

6.1 Features Analysis Results

The features analysis identified the metadata elements supported by each application, the types of automatic metadata generation methods used, and the interface features supporting metadata generation.

The results of the element analysis are given in Table 3. The top section of the table identifies metadata elements that map to the Dublin Core metadata standard, while the lower part identifies additional elements that did not map to the Dublin Core. A code system of "☼" and "■" designates manual and automatic/semi-automatic support for element generation, respectively. Winamp supported the specialized ID3 format (http://www.id3.org), a metadata standard used with audio files, primarily MP3 files. None of the other applications directly supported any standard metadata scheme such as the Dublin Core, although a couple of applications (e.g., Acrobat) allowed for metadata scheme customizations.

Results of the element analysis show that EndNote had the most comprehensive metadata system, as well as the highest number of metadata elements that could be mapped to the Dublin Core. *Title* was the only metadata element supported by all seven of the examined software applications. *Creator*, *subject*, and *date* metadata followed second, with each of these elements supported by six of the seven applications reviewed.

Winamp was the only software examined that supports automatic techniques for all of its descriptive metadata elements. Word, Acrobat, EndNote, and Movable Type all support automatic techniques to process at least half of their descriptive metadata elements. Specifically, 7 of 10 of Word elements, 3 of 6 of Acrobat elements, 12 of 15 EndNote elements, and 3 of 4 of Movable Type elements were generated from some automatic procedures. Dreamweaver and CityDesk, the two applications examined for website creation, demonstrated poor support of automatic techniques, with CityDesk using automatic techniques only for the *date* element and Dreamweaver not demonstrating the use of any automatic techniques.

Table 4 presents the automatic metadata generation methods underlying the examined software applications. Winamp and EndNote, which had the most elements supported by automatic means, were the only two applications to use all three automatic processing methods (deriving metadata automatically from system properties, harvesting human-generated metadata, and extracting metadata from resource content). Word, Acrobat, and Movable Type all derive metadata automatically from system properties and extract metadata from resource content, although they do not seem to harvest human-generated metadata. CityDesk only derives metadata from system properties.

The lower part of Table 4 identifies user interface features supporting metadata creators during the metadata creation task. All of the applications analyzed provide a form to facilitate metadata creation. Word, EndNote, Winamp, and Movable Type standardize metadata input with menus of values (commonly identified as pick-lists or dropdown menus). EndNote facilitates metadata creation with both a type-ahead feature (automatically completes a word being typed by a user) and an abbreviation expansion feature (automatically spells out a full name). In order to support this functionality, the terms and full names of abbreviations must be stored in a user profile first.

Table 3: Content Creation Software Support of Metadata Elements

Key: ☼ = Element requires manual input.
■ = Element creation supported by some automatic processing.

Type of Software→	**General Document**		**Website**		**Citation**	**Music**	**Weblog**
Software →	**Word**	**Acrobat**	**Dream-weaver**	**CityDesk**	**EndNote**	**Winamp**	**Movable Type**
Dublin Core metadata elements							
Title	■	■	☼	☼	■	■	■
Creator	■	■		☼	■	■	■
Subject	☼	☼	☼	☼	■		☼
Description		☼	☼		■		
Publisher					■	■[2]	
Contributor					■		
Date	■[3]	■		■	■	■	■
Type					■		
Format					■		
Identifier			☼		■		
Source	☼		☼		■	■	
Language					■		
Relation			☼		☼		
Coverage							
Rights		☼				■	
Additional metadata elements							
Manager	■						
Company	■						
Comments	☼				☼		
Document statistics	■					■	
Editing time	■						
Audience				☼			
Genre				☼		■	
Custom/user-definable			☼		☼		

[2] Including: Composer, original artist, encoder.

[3] Including: Created, modified, accessed, printed, last saved by, revision number.

Table 4: Automation Techniques and Interface Features Supporting Metadata Generation

Type of Software→	**General Document**		**Website**		**Citation**	**Music**	**Weblog**
Software →	**Word**	**Acrobat**	**Dream-weaver**	**CityDesk**	**EndNote**	**Winamp**	**Movable Type**
Automation techniques							
Deriving from stored/system properties	■	■		■	■	■	■
Harvesting from networked resource					■	■	
Extracting from existing document	■	■			■	■	■
Interface features							
Fill-in form	■	■	■	■	■	■	■
Listboxes or other selection control	■				■	■	■
Type-ahead					■		
Abbreviation/acronym expansion					■		

6.2 Metadata Expert Survey Results

The metadata expert survey data analysis focused on participants and their metadata/cataloging experience, current organizational metadata practices specific to DDLOs, participants' knowledge and opinions about automatic generation of Dublin Core metadata and metadata generation in general, as well as the features participants would like to see incorporated into automatic metadata generation applications.

6.2.1 Participant Profile

Two-hundred and seventeen (217) survey participants provided responses useful for data analysis (the initial goal was to recruit at least 100 participants). A total of 320 people actually started the survey; approximately one-third of survey participants did not complete it, mainly because they found it was beyond the scope of their work experience. Research has confirmed that a large percentage of people who start invited Web surveys often fail to complete them (Hayslett & Wildemuth, 2004). Even so, researchers have found that online surveys yield a "higher response quality" than do self-completion postal surveys and other offline methods (Gunter, Huntington, & Williams, 2002). All survey questions were optional, and the reporting that follows includes valid percentages (percentages based on the response rate per question).

Participant categories are presented in Table 5. (Percentages for tables hereafter do not all add up to exactly 100% because of rounding to the one-point decimal.) The largest portion of participants providing information on their professional role was identified either as administrators/executives (51 participants, 29.5%) or catalogers/metadata librarians (49 participants, 28.3%). Among the five persons identified as *other* were a Freedom of Information Officer, a consultant, an artistic-scientific assistant, a person holding a masters degree, and a bioinformatician. The largest portion of participants (70 participants, 40.7%) providing institutional affiliation information were active in an academic library environment—although, as shown in Table 6, representatives from other metadata generation environments participated in the study. Figure 3 shows the distribution of participants by organization.

Table 5: Participants by Professional Role

Professional Role	# of Participants
Administrator/executive	51 (29.5%)
Catalogers/metadata librarian	49 (28.3%)
Information/Web architect	15 (8.7%)
Professor/researcher	11 (6.3%)
Information technologist/systems analyst	10 (5.8%)
Librarian (general)	10 (5.8%)
Digital librarian	9 (5.2%)
Archivist	7 (4.1%)
Technical services librarian	6 (3.5%)
Other	5 (2.9%)

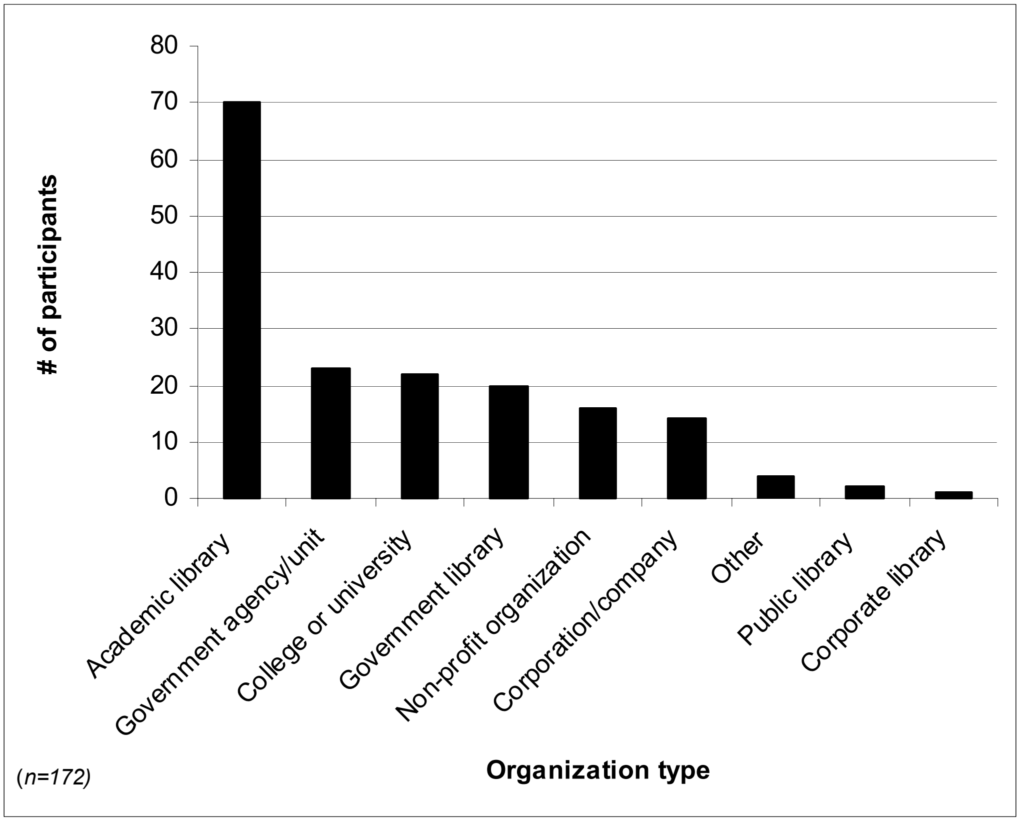
(*n=173*)

Table 6: Participants by Organization

Organization	# of Participants
Academic library	70 (40.7%)
Government agency/department	23 (13.4%)
Academic community (not in the library)	22 (12.8%)
Government library	20 (11.6%)
Nonprofit organization	16 (9.3%)
Corporation/company	14 (8.1%)
Public library	2 (1.2%)
Corporate library	1 (0.1%)
Other	4 (2.3%)

(*n=172*)

Figure 3: Distribution of Participants by Organization



The geographic distribution of participants was assessed by examining their institutional affiliations. Participants were identified from nine different countries: Australia, Canada, Czech Republic, Germany, Italy, The Netherlands, Spain, United Kingdom, and the United States. Other countries may have been represented but could not be identified.

6.2.2 Participants' Professional Cataloging/Metadata Experience

The survey gathered data on participants' cataloging/metadata experience and the activities in which they were involved in order to verify their status as experts. Three quarters of participants (161 participants, 75.2%) had three or more years of cataloging and/or indexing experience. Approximately 10% of the participants had one year or less than one year of experience in this area. Table 7 summarizes the years of experience of the participants.

Participants were also asked specifically about how many years of experience they had working with metadata (cataloging, indexing, other metadata activities), specifically for DDLOs. The question was posited to gather information about activities beyond cataloging, such as administrative activities, specifically in relation to DDLOs, a parameter of the study. The results show that half of the participants (110 participants, 50.9%) had been involved in metadata activities related to DDLOs for more than three years. Table 8 summarizes these results.

Information was also gathered on the variety of metadata activities in which participants were involved. Participants were presented with a list of tasks and asked to check all that applied. These results are presented in Table 9. Most participants (192 participants, 90.1%) were involved in metadata creation and metadata maintenance/quality control activities (150 participants, 70.4%). Well over half of the participants (172 participants, 80.8%) were also involved in some sort of administrative or supervision activity related to metadata creation. Among the types of activities participants noted in the "other" category were facilitating metadata interoperability, aggregating metadata, programming, designing metadata schemas, and consulting.

Table 7: Participants' Cataloging/Indexing Experience

# of Years	# of Participants
<1	19 (8.9%)
1	2 (0.9%)
2	17 (7.9%)
3	15 (7.0%)
>3	161 (75.3%)

(*n=214*)

Table 8: Participants' Experience Cataloging/Indexing, or Other Metadata Activities with DDLOs

# of Years	#of Participants
<1	29 (13.4%)
1	18 (8.3%)
2	34 (15.7%)
3	25 (11.6%)
>3	110 (50.9%)

(*n=216*)

Table 9: Participants' Metadata Activities

Metadata Activity	# of Participants
Metadata creation	192 (90.1%)
Administration/ supervision	150 (70.4%)
Maintenance/ quality control	172 (80.8%)
Other	97 (45.5%)

(*n=213*)
Participants checked all that applied.

6.2.3 Organizational Metadata Practices

Data were gathered on the different metadata practices taking place in the various organizations, consortia, or initiatives with which the participants were affiliated (the term organization will be used hereafter). Table 10 identifies the prevalence of various metadata creators. Although most people involved in metadata creation appear to be metadata professionals (152 participants, 91.6%), the results show that a variety of people also undertake this task. Table 11 shows the prevalence of metadata creators in the library environment, and Table 12 shows the prevalence of metadata creators in the nonlibrary environment. The number of mentions given for these two environments (library and nonlibrary) do not add up to the total number of mentions given in Table 10 because organizational affiliation was not provided by every participant.

Table 10: Metadata Creators in Various Organizations

Metadata Creators	# of Mentions
Metadata professionals	152 (91.6%)
Other professionals (e.g., reference librarians)	104 (62.7%)
Information/Web architects	93 (56.0%)
Nonprofessionals/ technicians	108 (65.1%)
Resource authors	75 (41.8%)
Volunteers	14 (8.4%)
Others	28 (16.9%)

(*n=166*)
Participants checked all that applied.

Table 11: Metadata Creators in Libraries

Metadata Creators	# of Mentions
Metadata professionals	76 (100%)
Other professionals (e.g., reference librarians)	48 (63.2%)
Information/Web architects	37 (48.7%)
Nonprofessionals/ technicians	55 (72.4%)
Resource authors	26 (34.2%)
Volunteers	7 (9.2%)
Others	15 (19.7%)

(*n=76*)
Participants checked all that applied.

Table 12: Metadata Creators in Nonlibrary Environments

Metadata creators	# of Mentions
Metadata professionals	48 (53.3%)
Other professionals (e.g., reference librarians)	37 (41.1%)
Information/Web architects	36 (40.0%)
Nonprofessionals/ technicians	32 (35.5%)
Resource authors	36 (40.0%)
Volunteers	4 (4.4%)
Others	8 (8.8%)

(*n=90*)
Participants checked all that applied.

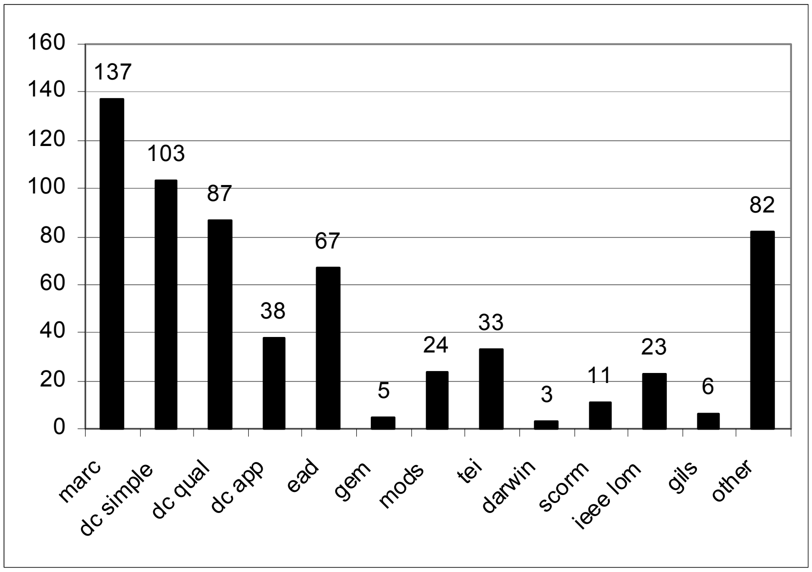
Survey results show that organizations are using a variety of metadata standards software systems for metadata generation. Figure 4 shows the range of different metadata standards being used. Among several of the metadata standards in the "other" category were the DDI (Data Documentation Initiative), ONIX (Online Information Exchange) International, AGLS (Australian Government Locator Service) metadata standard, METS (Metadata Encoding and Transmission Standard,) RLG (Research Libraries Group) preservation metadata standard, NBII (National Biological Information Infrastructure), RSS (RDF Site Summary, previously Rich Site Summary), and SMEF (Standard Media Exchange Framework) metadata model; several respondents also noted the use of "homegrown" metadata schemes. Table 13 shows the prevalence of metadata standards being used in the library environment, and Table 14 shows the prevalence of metadata standards being used in the nonlibrary environment. Similar to Tables 11 and 12 in comparison to Table 10, the number of mentions for these two environments do not add up to the total number of mentions given in Figure 4 because organizational affiliation was not provided by every participant.

Finally, in relation to metadata standards, data were gathered on the metadata function being supported in participants' organizations. Results are presented in Table 14. The results show that resource discovery is the most important function, although other functions are also supported. Among functions noted in the "other" category were geo-referencing, temporal referencing, content management, and workflow. Although 20 participants checked "other," few specified those other functions.

Key:

- marc = Machine Readable Cataloging
- dc simple = Dublin Core (unqualified)
- dc qual = Dublin Core (qualified)
- dc app = Dublin Core application profile
- ead = Encoded Archival Description
- gem = Gateway to Educational Materials
- mods = Metadata Object Description Schema
- tei = Text Encoding Initiative (header)
- darwin = Darwin Core
- scorm = Sharable Courseware Object Reference Model
- ieee lom = IEEE Learning Objects Metadata
- gils = Government Information Locator Service

Figure 4: Metadata Standards Being Used in Organizations



(n=169)

Table 13: Metadata Standards Used in Libraries

Metadata Standard	# of Mentions
MARC	80 (87.0%)
DC simple	44 (47.8%)
DC qual	39 (42.4%)
DC app	16 (17.4%)
EAD	39 (42.4%)
GEM	N/A
MODS	18 (19.6%)
TEI	22 (23.9%)
Darwin	2 (2.2%)
SCORM	2 (2.2%)
IEEE LOM	6 (6.5%)
GILS	3 (3.3%)
Other	29 (31.5%)

(*n=92*)
Participants checked all that applied.

Table 14: Metadata Standards Used in Nonlibrary Environments

Metadata Standard	# of Mentions
MARC	28 (35.9%)
DC simple	33 (42.3%)
DC qual	32 (41.0%)
DC app	17 (21.8%)
EAD	12 (15.4%)
GEM	3 (3.8%)
MODS	4 (5.1%)
TEI	4 (5.1%)
Darwin	1 (1.3%)
SCORM	4 (5.1%)
IEEE LOM	13 (16.7%)
GILS	N/A
Other	N/A

(*n=78*)
Participants checked all that applied.

Table 15: Metadata Functions

Metadata Functions	# of Mentions
res disc	208 (96.7%)
pres	104 (48.4%)
admin	107 (49.8%)
rights	79 (36.7%)
ext harv	120 (55.8%)
other	20 (9.3%)

(*n=215*)
Participants checked all that applied.

key:

- res disc = resource discovery
- pres = preservation
- admin = administration
- ext harv = external harvesting/resource sharing

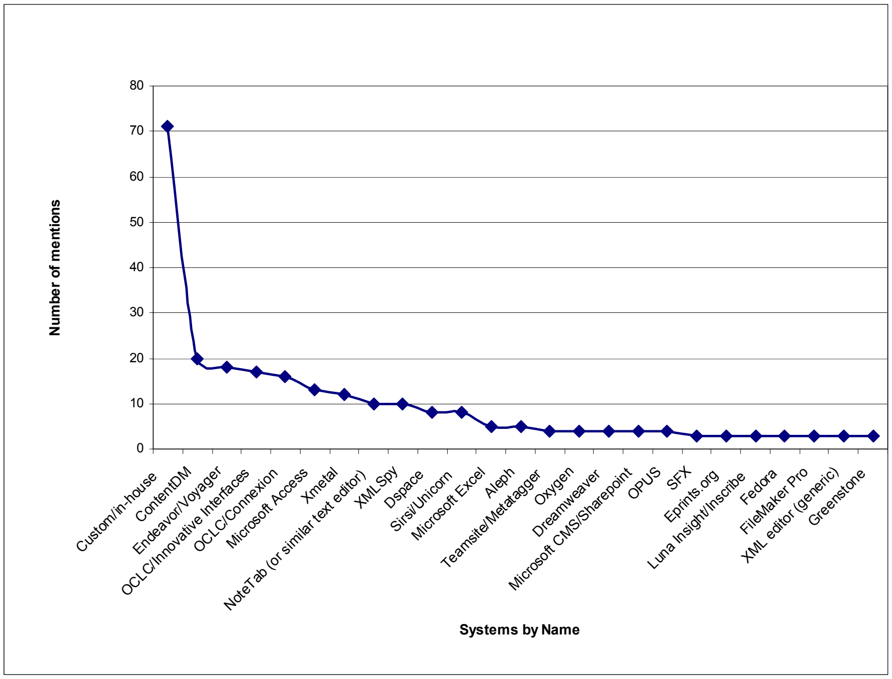
Participants reported using from one to seven different software applications or systems for metadata generation in their organizations (Table 16). Most participants (150 participants, 81.5%) reported using one or two systems, whereas one participant reported the use of seven different applications. Identification of systems being used was diverse and ranged from a simple mention of a software or system name, to detailed analysis of system functionality and usability. Each mention of a specific system was recorded to obtain a sense of the most prevalent systems in the community of metadata members participating in the study. Ninety-one distinct systems were recorded (see Figure 5). The systems noted in Figure 5 show that participants' interpretation of "system" ranged from metadata generation software to word processing tools and editors.

Table 16: Number of Metadata Systems Being Used

# of Systems	# of Organizations
1	94 (51.1%)
2	55 (29.9%)
3	22 (12.0%)
4	6 (3.3%)
5	4 (2.2%)
6	2 (1.1%)
7	1 (0.5%)

(*n=184*)

Figure 5: Distribution of Systems Used for Metadata Generation Per Organization



(*n=184*)

The examination of organizational practices also examined participants' use of metadata creation tools with automatic capabilities. Participants were presented with the definitions shown in Table 17.

Table 17: Automatic Metadata Generation Definitions (from Greenberg, 2004a)

Automatic Metadata Generation Concept	Example(s)
Metadata extraction. The process of automatically pulling (extracting) metadata from a resource's content. Resource content is mined to produce structured ("labeled") metadata for object representation.	*Metadata extraction for a Web page involves extracting metadata from the resource's content that is displayed via a Web browser.*
Metadata harvesting. The process of automatically collecting resource metadata already *embedded in* or *associated with* a resource. The harvested metadata is originally produced by humans or by fully or semi-automatic processes supported by software.	*Metadata harvested from a Web page is found in the "header" source code of an HTML (or XHTML) resource (e.g., "Keywords" META tags). Metadata for a Microsoft WORD document is found under file properties (e.g., "Type of file," which is automatically generated, and "Keywords," which can be added by a resource author).*
Fully-automatic metadata generation. Complete (or total) reliance on automatic processes to create metadata.	*Web editing software (e.g., Macromedia's Dreamweaver and Microsoft's FrontPage) and selected document software (e.g., Microsoft WORD and Acrobat) automatically produce metadata at the time a resource is created or updated (e.g., "Date of creation" or "Date modified") without human intervention.*
Semi-automatic metadata generation. Partial reliance on software to create metadata; a combination of fully-automatic and human processes to create metadata.	*(1) Fully-automatic techniques are used to generate metadata (e.g., "Keywords") as a first pass, and software then presents the metadata to a person, who may manually edit the metadata. (2) Software may present a person (e.g., resource author or Web architect) with a "template" that guides the manual input of metadata, and then automatically converts the metadata to appropriate encoding (e.g., XML tags). The software may even automatically embed metadata in a resource.*

*Results from the content creation software analysis found that Dreamweaver does not automatically produce META tags for date metadata.

More than half of the participants responding to this question (125 participants, 58.1%) indicated they had worked with tools that included some automatic metadata generation capabilities, while fewer than half (90 participants, 41.9%) indicated they had not worked with tools using automatic capabilities.

A final open-ended question specific to organizational metadata practices asked participants if they were performing any evaluation/quality control (QC) in relation to digital resources, and to explain the process and evaluation criteria used. QC was defined as a separate task from the metadata creation (post-metadata creation); it is generally carried out by a person other than the record creator, who may be too close to the record to discover errors. The QC analysis focused only on record content, not processes to make the QC activity possible (e.g., automatic routing of metadata records). Close to three-quarters of the participants (160 participants, 73.7%) answered this question, although answers were too vague to assess for 26 of these participants. Results for the 134 participants (61.8% of the total survey population) whose answers could be assessed are presented in Table 18. Eighty-five of these participants (63.4%) noted formal QC activities. Twenty-three participants (17.2%) described their process of maintaining standards while creating metadata records. Processes described included selecting high-quality metadata records for use in copy cataloging, using authority files or controlled vocabularies, proofreading before saving or exporting metadata records, and working with systems that will not produce metadata records missing required fields. It is common practice for catalogers/metadata creators to follow these processes, despite the fact that most participants did not indicate this. Participants who passed over this question may have been members of organizations without a formal evaluation/QC program, although more data are needed for verification.

Table 18: Evaluation/QC

Process	Organizational Practice
Formal QC activities	85 (63.4%)
Evaluation at time of metadata creation	23 (17.2%)
No evaluation/QC	22 (16.4%)
Planning QC	4 (3.0%)

(*n=134*)

Table 19: QC Processes

Process	Organizational Practice
Manual	47 (69.1%)
Multiple processes (manual, automatic, & semi-automatic)	8 (11.8%)
Semi-automatic	8 (11.8%)
Automatic	5 (7.4%)

(*n=68*)

Results from the 85 participants noting formal QC activities were further analyzed to determine the extent to which automatic capabilities are being used specifically to assess metadata content. Results from the 68 participants whose QC methods could be assessed are presented in Table 19 (answers for 17 of the 85 participants noting formal QC were too vague to assess). The QC analysis focused specifically on activities directed at metadata content. For example, a system scanning and flagging non-uniform subject headings or names in metadata records, for a human to analyze and update at a latter date, was identified as automatic; a system using automatic routing processes, but depending on a human for content evaluation, was identified as manual. Most QC activities involved manual processes. A small percentage of participants (7.4%) relied fully on automatic means for QC activities; QC in these cases mostly involved scripts or software to automatically check links in metadata records, and one participant mentioned automatic validation/checking of record encoding.

6.2.4 Automatic Metadata Generation of Dublin Core

Participants' opinions about the feasibility and usefulness of automatic generation of Dublin Core metadata for DDLOs were recorded. To help assess these results, background data were first gathered on participants' knowledge and experience with Dublin Core (Table 20). With the exception of one participant who skipped this question, all of the participants had at least heard of the Dublin Core. More than three quarters of the participants (174 participants, 80.6%) had worked with the Dublin Core (Table 20, last four rows), with approximately a third of the participants (32.9%) indicating extensive work with the Dublin Core and 13 participants indicating involvement in the development of the Dublin Core.

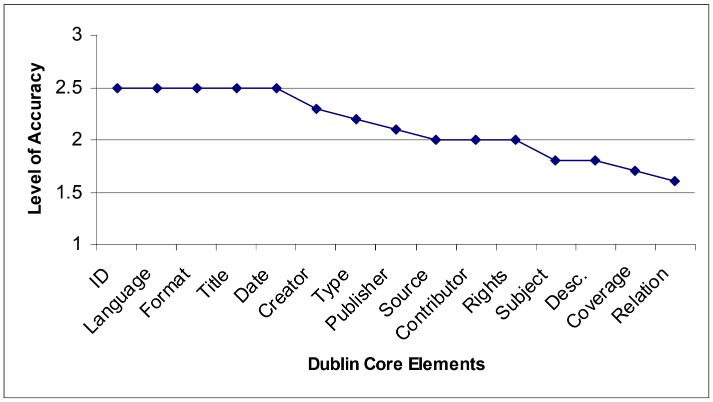
Table 20: Dublin Core Knowledge/Experience

Knowledge/ Experience	# of Participants
Heard of DC, but not familiar with it	17 (7.9%)
Read DC standard and/or have had DC training , but not worked with it	25 (11.6%)
Worked with DC a little	90 (41.7%)
Worked with DC extensively	71 (32.9%)
Involved in DC development	6 (2.8%)
Worked with DC extensively and involved in DC development	7 (3.2%)

(*n=216*)

The feasibility/usefulness analysis focused on *expected accuracy* and *appropriate application levels* for automatic Dublin Core generation. A semantic differential scale, with "3" meaning "very accurate," "2" meaning "moderately accurate," and "1" meaning "not very accurate" was used to record expected accuracy levels for automatic generation of Dublin Core metadata. Averages for all 15 Dublin Core elements are graphed in Figure 6. In general, greater accuracy was predicted for technical metadata such as *ID*, *language*, and *format*—all of which resulted in an average score of 2.5. Less accuracy was expected for metadata requiring intellectual discretion, such as *subject* and *description*, which resulted in an average score of 1.8. *Coverage* metadata, which is used for *temporal* or *spatial* subject-like metadata, had a similar ranking, with an average score of 1.7. Participants expected the least degree of accuracy for *relation* metadata, with an average score of 1.6. This element deals with intellectual bibliographic-like relationships defined as Dublin Core qualifiers (DCMI Metadata Terms, 2004).

Figure 6: Expected Accuracy Level for Automatic Generation of Dublin Core

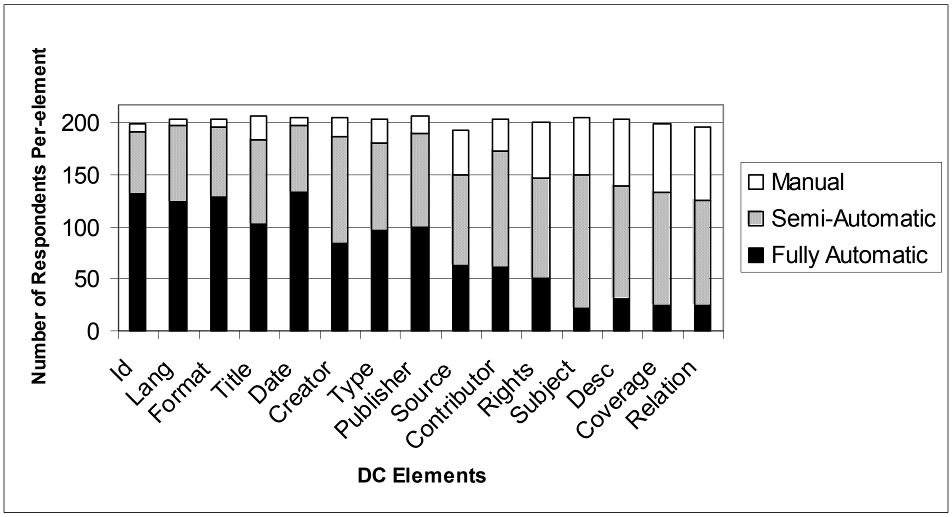


(*n varied slightly per metadata element*)

Open-ended comments on accuracy ratings were analyzed and revealed a number of themes, the most prevalent of which was a perceived skepticism about the accuracy of automatic techniques for the generation of metadata requiring intellectual discretion (primarily *subject metadata*). A number of participants emphasized the value of controlled vocabulary and were skeptical about controlled vocabulary assignment via automatic techniques. A few participants also voiced concerns about automatic metadata generation for element definitions that they perceived as too vague. One participant said, “How can we automate even elements w/o agreement on semantics?” Finally, a few participants advocated taking a more holistic approach to metadata creation, highlighting the need for information systems to consider context and incorporating metadata extraction into workflow. For example, one participant suggested that systems “import…context-sensitive information from the authoring environment,” such as metadata creator profiles and intended users. Another participant said “We have taken a systems approach to this [metadata generation]” and described how they integrated the various stages of workflow.

In examining appropriate metadata generation levels, participants were asked to check one of three options (manual, semi-automatic, and fully-automatic) for all 15 Dublin Core elements. The results are shown in Figure 7. In general, greater support for automatic processing was found for technical metadata such as *ID* and *format*, which can be extracted with little difficulty, and other types of metadata such as *language*, which are easily machine readable. Manual processes were considered more appropriate for metadata requiring greater intellectual discretion, such as *subject*, *description*, *coverage*, and *relation* metadata. The results depicted in Figure 7 parallel, to some degree, the accuracy expectancy results shown in Figure 6.

Figure 7: Appropriate Metadata Generation Levels for Dublin Core



(*n varied slightly per metadata element*)

A final question on automatic metadata generation of Dublin Core metadata records asked participants about application design and funding allocation per Dublin Core element—assuming limited resources. Participants were asked to select "High" if they would devote extensive resources, "Medium" if they would devote a moderate amount of resources, and "Low" if they would devote few resources to developing and implementing automatic metadata generation techniques for each element. Participants were united in their assessment of automatic metadata generation as potentially valuable. As one participant noted, metadata creators must "reallocate budget from the traditional processing by hand to high-tech solutions." Participants, however, were divided as to how research and development efforts in this area should be focused. This division centered on a *fundamental tension* in thinking about how to allocate funding. The tension was between *usefulness*, focusing on the elements "most important for resource discovery," and *feasibility*, focusing on those elements that are easiest or "most clear-cut" to generate automatically.

Participants appeared split into two camps—optimists and skeptics—reflecting their assessments of this difficulty. The optimists were forward-looking, anticipating advances that would make automatic generation of intellectual metadata realistic. They argued for funding these more research-intensive areas: "I'd spend my money on areas that require the most amount of AI [artificial intelligence] or lexical analysis and comparison to develop sound output." This was in direct contrast to the skeptics, who argued for focusing resources on areas where full automation is feasible, particularly "physical" fields such as *identifier* or *format*. Skeptics asserted that attempting automatic generation of "intellectual" fields such as *subject* or *description* is pointless or impossible. "I am not convinced the tool would work," wrote one skeptic; "a total waste," said another. The skeptics often referred to unsuccessful experiences with automatic tools: "I haven't yet seen software that can really identify subject and keywords automatically," one participant wrote. Many also noted that the elements most important for resource discovery are also the most difficult to generate automatically. In summary, the comments indicate that metadata experts view automatic generation as an unsolved problem and are divided as to how future efforts should be focused.

6.2.5 Automatic Metadata Generation Challenges and Preferences

The last section of the survey briefly addressed automatic metadata generation for nontextual and foreign language resources, and then focused on additional desired functionalities for automatic metadata generation applications.

6.2.5.1 Automatic metadata generation for nontextual and foreign language resources

Although the survey emphasized DDLOs, several questions in the last section were posed to gather baseline data on automatic metadata generation of nontextual and foreign language material. Participants were asked about the importance of developing applications to support automatic metadata generation for nontextual digital resources (e.g., multimedia). Results presented in Table 21 indicate that participants thought it was very important to develop automatic or semi-automatic methods of generating metadata for nontextual content, although many emphasized this was a difficult task. Several participants indicated that it may be even more important to develop some automatic methods for nontextual resources because of the absence of text for indexing. One participant said, "There is only the metadata to rely on for resource discovery rather than full-text indexing." Another participant added that automatic metadata generation for non-textual resources "will be more important in the long run than for textual records since multimedia records cannot be easily searched by their contents." Several participants stressed the availability of technical metadata, stating that "technical metadata for nontextual resources (such as digital still images) is a prime candidate for automated metadata creation and metadata extraction."[4]

Table 21: Automatic Metadata Generation for Nontextual Resources

Importance Value	Response Rating
Very important	121 (57.3%)
Somewhat important	82 (38.9%)
Not important	8 (3.8%)

(*n=211*)

Participants also called for developing applications that would support linking and cross-referencing between metadata records in general because nontextual objects are frequently to be associated with or related to other objects (e.g., a video news clip may be linked to its transcript). Responses highlighted both the importance and difficulty of automating linking mechanisms. One reply clearly articulated the difficulty of this task by stating that "only a person can really grasp how the items inter-relate and whether a single part is the dominant part with accompanying material or if all the parts have equal value and make a whole resource in themselves."

Similar to the *usefulness/feasibility* responses for automatic metadata generation for Dublin Core elements requiring intellectual discretion, a small group of pessimists responded that it is not possible to automatically or semi-automatically generate metadata for nontextual resources. In fact, one participant recommended that "efforts might be better put toward making textual metadata generation as automatic as possible. That way human intervention and expertise could be spent on the more subjective description of nontextual materials."

Participants' support of automatic metadata generation for foreign language resources is presented in Table 22. Most participants indicated that this function was "somewhat important," followed by 95 participants (44.8%) indicating it was "very important," and a few participants indicating it was "not important" (15 participants, 7.1%).

[4] The use of the word "extraction" in this quote is more synonymous with the word "harvesting" given in the discussion of automatic metadata generation research in this report.

Table 22: Automatic Metadata Generation for Foreign Language Resources

Importance Value	Response Rating
Very important	95 (44.8%)
Somewhat important	102 (48.1%)
Not important	15 (7.1%)

(*n=212*)

Several participant comments indicate problems with existing tools that do not support the diacritics and special characters used in some languages. Participants highlighted working with collections containing items in multiple languages and serving communities with diverse language needs as reasons why this functionality is important. As one participant said, "I answered very important…because we have several projects that are exchanges with foreign institutions and lots of materials that are foreign-language. These projects have faced challenges with diacritics and character sets in existing tools, so built-in foreign language functionality would be extremely useful."

As shown in Table 23, little more than half of the participants (112 participants, 53.1%) indicated that it is "somewhat important" for an automatic metadata generation tool to provide machine translation of metadata records into multiple languages. Slightly more participants indicated that this function was "not important" (51 participants, 24.2%), compared to those participants who indicated it was "very important" (48 participants, 22.7%).

Table 23: Automatic Metadata Generation Machine Translation

Importance Value	Response Rating
Very important	48 (22.7%)
Somewhat important	112 (53.1%)
Not important	51 (24.2%)

(*n=211*)

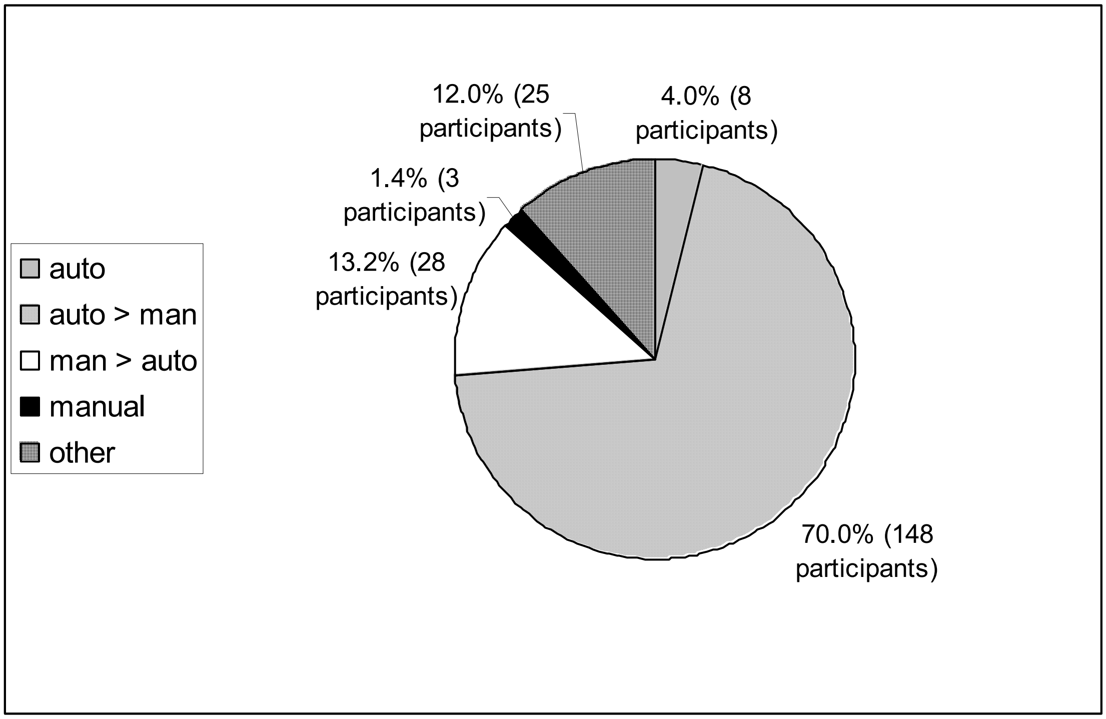
Participant comments seemed to relate to practical work scenarios. For example, one participant commented that "there should be no difference between metadata creation for different languages. As long as we use standard formats a title is just a title regardless of the language." Another participant responded that "in officially bilingual environments like Canada…we prefer to see English and French as parallel and not as translations in order to protect the integrity of the original text and all its linguistic nuance." Many respondents point out that multilingual mapping of subject terminology would be more useful than machine translation of records: "Where schemas or taxonomies used are bilingual we want the values from the alternate language resource to be autopopulated."

6.2.5.2 Additional functionalities for automatic metadata generation applications

The final portion of the survey examined workflow and automatic metadata sequence preferences, the integration of selected cataloging/metadata creation aids, and additional desired functionalities for automatic metadata generation applications.

Participants were asked to indicate the metadata generation workflow they would like, with several options for integrating automatic processing during the metadata creation process. Results to this question are shown in Figure 8. Most participants (148 participants, 70.0%) indicated that they would prefer an application to first execute automatic algorithms, and afterwards allow a human to evaluate and edit the results. Only 3 participants (1.4%) exclusively supported manual processes. Workflow options described in the "other" category were almost unanimous and steadfast about the use of automatic processes, with flexible manual review options based on need and the metadata creator. "As fully automatic as possible, but I am afraid some editing by a person will be needed every now and then," one participant responded. Two others responded, "Automatically created as much as possible then edit," and "Fully automatic with the capability of editing." The latter participant added, "The creation could occur anytime then notify persons to view. Then if inaccuracies then [*sic*] we would want to be able to edit." In general, participants wanted a flexible workflow, where a "person can choose to start it [an automatic process] or not."

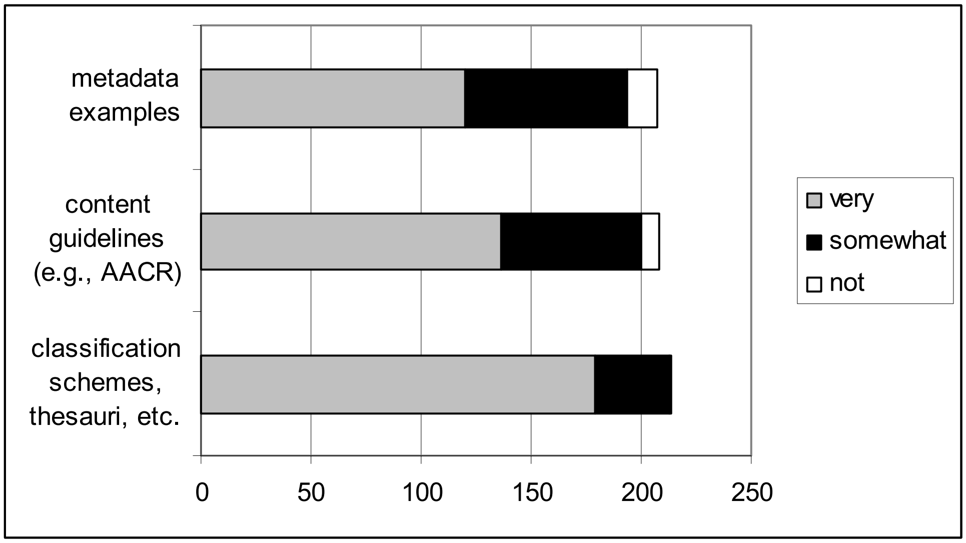
Figure 8: Preferred Metadata Generation Workflow



(*n=212*)

Participants were asked about the desirability of integrating metadata/cataloging examples, content creation guidelines, and subject-oriented schemes and vocabulary into automatic metadata generation applications. These results are shown in Figure 9. Participants indicated it was generally "very desirable" or "somewhat desirable" to integrate any of these aids, with the greatest support for integrating subject classification schemes and vocabulary tools.

Figure 9: Integration Desirability of Examples, Content Guidelines, and Schemes



(*n varied slightly per feature*)

The examination of functionalities also included an open-ended question asking participants to comment on “other features” they thought would be desirable in automatic metadata generation applications. Themes that emerged when analyzing the results include the following:

- System should integrate name authority files for personal and organizational names.

- System should have the ability to import and export metadata in standard formats. Platform independence in formats is desired.

- System should support automatic and semi-automatic quality control routines, error checking, and validation of encoding against schemas.

- System should support the creation or administration of rights management metadata and the embedding of digital signatures into metadata records to support privacy and use restrictions.

- System should support automatic linking of metadata records, including referencing and cross-referencing between related items.

- System should support user/organizational customizability and flexibility and should include intelligent defaults.

- System should support the extraction and creation of technical and preservation metadata.

The themes listed here and the results of all the analyses underlying the AMeGA project have been incorporated into the recommendations for Automatic Metadata Generation Applications presented in Section 8 of this report.

7. Discussion

The discussion below addresses the features analysis and the metadata expert survey. Both research undertakings helped to identify progress and limitations with metadata generation, and provide data for identifying recommended functionalities for automatic metadata generation applications.

7.1 Features Analysis

The features analysis found that content creation software supports descriptive metadata, and many of the metadata elements supported can be conceptually mapped to the Dublin Core metadata standard (Table 3, top portion). This finding is not surprising, given the underlying purposes of metadata in these two venues: Content creation software metadata is produced to support computer desktop file organization and searching; the Dublin Core was developed to facilitate resource discovery primarily, although not exclusively, for digital resources.

Elements across the software reveal traditional resource discovery practices. For example, the *title* element was supported by all of the applications, and *author* (*creator*) and *subject* metadata were found in six out of seven of the applications. These findings pertain to Charles Amni Cutter's first two objectives outlined in his *Rules for a Dictionary Catalogue* (1904):

1. To enable a person to find a book of which the author, title, or subject is known.
2. To show what the library has by a given author, on a given subject, and in a given kind of literature.

These metadata elements (*title*, *author*, and *subject*) are fundamental to resource discovery on the Web, and it is sensible that content creation software provides means for their creation—whether manual, semi-automatic, or fully automatic. A current limitation is that these elements are not always populated with data values by resource creators or other persons making resources Web accessible. Moreover, state-of-the-art metadata creation applications do not always exploit metadata created with content creation software. Research is needed to address these shortcomings.

The features analysis also examined the employment of different automation techniques (Table 4, top portion). It was discouraging to find that Dreamweaver and CityDesk, the two applications examined for website creation, demonstrated poor support of automatic techniques (CityDesk uses automatic techniques only for the *date* element, and Dreamweaver does not demonstrate any automatic metadata functionalities). These two tools are well-known Web editors in many venues, and it is difficult to determine why their support for metadata creation is limited. One possible reason is that these applications focus on the creation of HTML documents and emphasize resource appearance (e.g., color and font size) over structure and content. Even so, it seems development of X/HTML, which supports structured metadata, would have had an impact on such applications, improving metadata functionalities. Regardless of the underlying reasons, improving Web editor metadata creation functions could lead to an increase in metadata production, and ultimately improve resource discovery and other metadata-supported functions. Perhaps better communication between selected metadata communities and Web editor vendors would improve current metadata functionalities for these tools.

Despite the poor metadata support found with Web editors, it was encouraging to find that the other software examined employs automatic techniques to generate at least half of their elements. EndNote generated the greatest number of metadata elements automatically by harvesting from online bibliographic databases (e.g. an OPAC, ISI Web of Knowledge). This approach is appropriate, given that EndNote is a bibliographic tool.

The features analysis also found that the most frequent use of automatic techniques was deriving metadata from system properties (e.g., *date created* and *format*). Content creation software likely derives and encodes this metadata because it is available, easy to derive, and immune to human error. This does not mean that automatic processing is error free. A case in point is that automatic techniques cannot always detect the actual publication date of a printed resource that has been digitized. Even so, deriving system-generated metadata is a worthwhile development and permits human resources to be directed to other metadata generation and evaluation tasks. The fact that most of the applications automatically encoded humanly generated metadata, which was input through a system template, is also encouraging.

Use of metadata extraction techniques was found to be uncommon, and the methods employed were quite primitive. Metadata generation functionalities employing extraction algorithms could be improved by incorporating more sophisticated algorithms developed through experimental research (e.g., Han, et. al., 2003; Takasu, 2003). However, it will be a major challenge to incorporate such algorithms without restricting content creation software to a particular subject or discipline domain. For example, incorporating a high-performance automatic indexing algorithm for "blood disease" research would not be very useful for indexing and assigning subject metadata to resources on "terrorism." Automatic matching and/or human computer interaction techniques might allow a metadata creator to pick an algorithm labeled by a subject domain. This step may be too detailed for content creation software developers to consider, and more applicable to tools specifically identified as automatic metadata generators. Regardless, some improvement in general domain extraction algorithms should be incorporated into content creation software, and perhaps automatic indexing techniques could be used to identify an appropriate domain-specific algorithm to assist with assigning *subject* metadata. More research is needed in this area, both for content creation software and, more importantly, metadata generation applications.

Winamp's metadata generation practice was among one of the most exciting findings. This software can effectively leverage large, existing repositories of human-created metadata and automatically assign metadata to users' resources. Winamp automatically creates all of its associated metadata elements following the ID3 standard (http://www.id3.org) using open repositories such as CDDB (http://www.gracenote.com/music/). This type of automation could be used with other content creation software to produce better quality metadata.

Final discoveries regarding content creation include several interface features such as input forms and value lists to aid metadata creation, and the use of word processing features (e.g., a type-ahead function and automatic generation of full words for abbreviations and acronyms stored in a profile) (Table 4, bottom portion). It would be useful to incorporate these functionalities in all content creation software and metadata generation applications, as they could aid in quality and speed of metadata generation and also encourage resource authors to become more involved in the metadata generation process, without overwhelming them.

7.2 Metadata Expert Survey

The metadata expert survey identified a number of areas in which more robust automatic metadata generation applications may aid metadata creation. The following discussion highlights these areas while covering participants and organizational practices, Dublin Core element rankings, and desired system functionalities.

7.2.1 Participant and Organizational Practices

The study confirmed that participants were metadata experts, with approximately three quarters of them (161 participants, 75.3%, Table 7) having three or more years of cataloging/indexing experience, and 90.1% (Table 9) involved in metadata creation and/or other types of metadata activities (e.g., administration/supervision, maintenance/quality control, etc.). Additionally, more than three quarters of the participants (174 participants, 80.6%, Table 20, summation of the last four rows) had worked with the Dublin Core. Participant experiences help validate their answers and the conclusions drawn.

Most participants were working in academic libraries. This is not surprising, given both the leadership role played by academic libraries in metadata developments and the participant recruitment methodologies underlying the AMeGA project. Many metadata-related electronic mailing list subscribers and blog authors and readers work in academic settings, and the survey recruitment message was disseminated through these venues. Additionally, most AMeGA Task Force members who recruited participants from their organizations work in academic libraries. Another influential factor was that the AMeGA study was conducted in conjunction with the LC Action Plan, which aims to provide leadership to libraries in the new millennium. Despite all these factors, participants' organizational affiliations extended well beyond academic libraries to include government agencies, college and university departments outside of the library, government libraries, nonprofit organizations, corporations/companies, public libraries, and corporate libraries, demonstrating the importance of metadata in many environments (Table 6 and Figure 3).

Related to the range of different organizations was the variety of people involved in metadata activities, from metadata and library professionals (persons with a credentialed library degree) to information/Web architects, nonprofessionals/technicians, resource authors, volunteers, and a number of participants falling into the "other" category (Table 10). This is not surprising, given the large quantities of digital resources that organizations are trying to make accessible, manage, or control in other ways. It seems the explosive growth of information on the Web has made it necessary to share the labor of metadata generation among a range of people, beyond professionally trained catalogers and indexers. Moreover, involving a range of people is valuable, not just to spread the workload, but to involve as many relevant perspectives and types of expertise as possible.

A further comparison of metadata creators in *libraries* and the *nonlibrary* environment showed that, although libraries depend heavily on metadata professionals and nonprofessionals/technicians for metadata creation, this task is more evenly split among a greater range of people in the nonlibrary environment (Tables 11 and 12 for comparison). This likely results from the fact that nonlibrary environments have fewer metadata professionals available. In the nonlibrary environment a metadata professional is often a "solo act" and thus charged with a greater variety of tasks than a metadata professional in the library environment, where the work is generally distributed among a number of metadata experts with similar skills and knowledge. Despite these observations, the library environment seems to be making greater use of a range of other people, including volunteers and persons classed as "other" (e.g., researchers), than is the nonlibrary environment. This may be a result of libraries having greater capacity to train and review metadata created by these less experienced people. What is most important about these findings, in relation to the goals of this study, is that metadata generation application development must consider the range of people involved in metadata generation and their ability to either contribute unique expertise or to free up more skilled persons to perform other tasks.

Developers of automatic metadata generation applications must also consider the range of metadata standards in both the library and nonlibrary environments. Although a large portion of the metadata expert survey focused on the Dublin Core, the research discovered a wide range of metadata standards in use (see Figure 4 and Tables 13 and 14) and the production of metadata to support a number of different metadata functions in addition to resource discovery (Table 15). Although information centers (libraries and nonlibrary environments) have always been involved in resource administration, authentication, and other tasks beyond resource organization and discovery—particularly for special materials—the Web seems to have increased this demand (Greenberg, 2003). This finding helps explain the variety of functions that metadata researchers have identified over the last several years (Greenberg, *in press*, Table 1).

Perhaps most telling of the current limitations with metadata generation software was the wide range of systems being used to generate metadata (Table 16 and Figure 5). Approximately half of the participants (94 participants, 51.1%) reported using one system, with the other half using one or more systems and with more than 91 distinct mentions of different systems, ranging from metadata generation software to word processing tools and editors. The distribution illustrated in Figure 5 shows a few very popular, predominating systems, with a long tail of less common systems. The survey question about software use was open-ended, and many participants volunteered commentary about dissatisfaction and frustration with current software use. These results are considered evidence of the need for improving current tools, although they were not analyzed in detail because the survey was not a usability study.

The final participant and organizational practice component of the survey focused on QC activities. It was encouraging to find that more than half of the organizations (85 organizations, 63.0%, Table 18) had a formal QC operation. Even so, most organizations (47 organizations, 69.1%, Table 19) depend solely on manual processes, and a very small percentage of organizations (7.4%) used fully automatic QC methods, mainly to check links in metadata records. More research is needed in this area in order to understand how participants perceive QC and how automatic metadata generation applications can aid QC.

7.2.2 Dublin Core Element Rankings

The central portion of the metadata expert survey focused on the application of automatic metadata generation methods and ranking Dublin Core elements by anticipated accuracy, appropriate application level, and funding allocation.

As reported in the results section, greater accuracy was anticipated for technical metadata (e.g., *ID*, *language*, and *format*) than for metadata requiring intellectual discretion (e.g., *subject* and *description*) (see Figure 6), although none of the elements received the ranking of “very accurate” with a score of “3.” These results are reasonable, given that automatic processing has not been proven to be error-free. Automatic indexing and related processes (e.g., automatic abstracting, indexing, and classification) have not been shown to consistently assign accurate *subject* or *description* metadata across multiple domains or for general-domain collections covering a range of topics. Nevertheless, progress has been made with the development of domain-specific automatic indexing (e.g., Nadkarni, Chen, & Brandt, 2001). Rankings given for more intellectually demanding elements could likely change in the future if automatic metadata generation applications were to incorporate domain-specific algorithms, through either interactive or automatic means. The results may also vary if both metadata creators and application designers were more aware of research progress applicable to general domain collections, such as automatic abstracting research by Johnson (1995) and automatic classification work by Losee (2003).

Creator and *publisher* metadata were given a “moderately accurate” to “not very accurate” ranking. These elements are not as intellectually challenging as, perhaps, *subject* and *description* metadata, although accurate production of these elements via automatic means is not as easy as the production of certain types of technical metadata (e.g., *date modified* and *format*). Automatic metadata generation research experimenting with semi-structured metadata (e.g., Han et al., 2003; Takasu, 2003) could likely improve the rankings for these elements. Implementing this approach in an operational setting requires means for identifying document types via human and/or automatic processes. For example, a conference paper generally contains *author* metadata in the content header, while a digital book will contain *author* metadata on a digital title page. More research is needed to further identify semi-structured metadata patterns for selected document types, although current applications should take advantage of research already conducted in this area.

Participants expected the least degree of accuracy for the *relationship* element. This element deals with intellectual bibliographic-like relationships that can be complex. It seems that developments such as Functional Requirements for Bibliographic Records (FRBR) (1998) and research in this area (e.g., Smiraglia & Leazer, 1999; Tillet 1991, 1992; Vellucci, 1997; and Weinstein, 1998) may improve the overall score for this element.

Results for appropriate metadata generation levels were similar to the accuracy level rankings in that there was much greater support for automatic processing with technical metadata and machine-readable metadata (e.g., *language*), as opposed to metadata requiring more intellectual discretion (Figure 7). Even so, participants were not unanimously in favor of automatic processing for any single element, although semi-automatic processing was found to be fairly desirable across all elements. These results and commentary following the scoring indicate that participants wish to take advantage of automatic techniques, but are aware of limitations. In general, participants want to be able to evaluate and have some control over what is generated. This type of flexibility is important to the design of metadata generation applications employing automatic techniques.

The final Dublin Core survey question related specifically to "resource allocation" and elicited a *fundamental tension* between metadata *usefulness* and *feasibility* as reported in the results section. Participant commentary highlighted the greater need for contextual understanding of metadata and the metadata creation process. It is not always evident which elements are most useful to users. Many participants stressed the importance of resource discovery and information retrieval, but the range of different metadata schemes being used in the digital world clearly indicates that different elements are useful for different functions. Results of traditional transaction log analyses for online library catalogs and even Web searching logs provide insight into the value of metadata in different contexts. Additionally, research by Lan (2002) examining metadata relevance for resource discovery and research by Hearst et al. (2002) on metadata facets and interface design provide useful methodologies for understanding the value of metadata elements in different contexts. In the metadata expert survey, one participant pointed out that "without services to exploit the metadata…it can be hard to describe its use and therefore prioritize where efforts should be spent," continuing that we need to "keep in mind what our public is demanding and expecting." In sum, research is needed to identify the types and metadata elements that are most useful in specific contexts. We must enhance our understanding of how users employ metadata for resource discovery and other functions. Ultimately, it would be most valuable to then direct automatic generation efforts to elements that are most valuable to users.

7.2.3 Additional Functionalities

The final section of the metadata expert survey found that participants considered it very important, and in some cases critical, to support automatic metadata generation for nontextual resources (Table 21). The Web is a visually rich environment, and we are a visual society. Never before in history has there been such an enormous capacity to share images for research and scholarship, teaching and learning, and other purposes. Thus it is understandable that participants showed tremendous support for improving metadata in this area and for employing automatic techniques to images wherever possible. As noted above, NISO Z39.87 (2002) provides a foundation for automatic generation of technical metadata for images, and we are likely to see greater development in this area over time. In sum, the baseline data on nontextual resources emphasizes the need to incorporate such developments into metadata generation applications.

Participants were almost as in favor of support for automatic metadata generation for foreign language resources as they were for nontextual resources (Table 22). This observation is likely the result of the impact of the Web's global scope and the fact that participants were working with foreign language materials or serving multilingual populations. Less enthusiastic, however, was support for translating metadata records into different languages (Table 23). Participants' responses presented in the results section related to practical matters. Another related reason for limited support may be standard cataloging practices, whereby bibliographic records for foreign language resources are generally not translated. Participants' opinions revealed a pronounced split on the need for machine translation of metadata records into different languages. Thus, this functionality may not be a high priority for automatic metadata generation applications. This perception may change over time, however, given that many digital library projects and other initiatives strive for interoperability on a global scale. The fact that the Dublin Core has been translated into more than 30 languages (http://www.dublincore.org/resources/translations/) may, potentially, have an impact on this issue. In fact, Van Duinen's recent research (2004) on the André Savine collection demonstrates the importance of being able to translate traditional bibliographic records from Russian to English and vice/versa, and highlights the value of Dublin Core translations as a valuable framework that can enhance access to materials in the digital world.

Workflow option results (Figure 8) clearly reveal support for automatic metadata generation, although most participants (203 of the 212 who answered this question, 96.2%) were unwilling to recommend fully automatic techniques. These responses pertain to the Dublin Core element rankings and participants' knowledge that automatic processing has not been proven to be fully error-free, particularly across domains or in the general domain environment in which many participants work.

It is possible that the very limited participant support for fully automatic metadata generation (eight participants, 4.0%) stems from fear of job loss—at least for some participants. Participants may feel slightly threatened by the notion of machines taking over their jobs; however, participant commentary recorded throughout the survey provided no evidence of this reaction. This consideration (feeling threatened) is also negated by participants' overwhelming desire to incorporate automatic techniques into the metadata generation workflow (Figure 8) and the strong desire to integrate metadata examples, content guidelines, and schemes into applications (Figure 9 and results from open-ended responses). One exception is a very small percentage of participants (1.4%), who stressed the need for fully manual metadata generation. Despite these findings, the impact of automation on the psyche of the individual and the social fabric of the workplace cannot be underestimated (e.g., Zuboff, 1998). It is recommended, therefore, that research be pursued on metadata experts' perceptions of automation and its impact on their current worth. Research specifically addressing automation in the library environment (e.g., Dakshinamurti, 1985), even on a more general level, can provide more insight into this issue.

8. Recommended Functionalities for Automatic Metadata Generation Applications, Version 1.0

The research presented in this final report provides data for the identification of recommended functionalities for automatic metadata generation applications. Influential bibliographic control models such as Weintraub's (1979) four functions underlying bibliographic control (finding, listing, identifying; gathering; collocating; and 4. evaluating/selecting), based on Cutter's objectives (1904), and ongoing research on conceptual models of the metadata creation process stemming from the Metadata Generation Research project (http://ils.unc.edu/mrc/mgr_index.htm) also provided a useful framework for presentation of recommended functionalities. The recommendations are identified as Version 1.0 because it is likely that they will be enhanced and modified over time, with greater input from the larger bibliographic control/metadata community. The recommendations are organized as follows:

- System Goals
- General System Recommendations
- System Configuration
- Metadata Identification/Gathering
- Support for Human Metadata Generation
- Metadata Enhancement/Refinement and Publishing
- Metadata Evaluation
- Metadata Generation for Nontextual Resources

1. System Goals

Automatic metadata generation applications exploit automatic techniques in order to improve the efficiency and effectiveness of metadata generation. Intelligent use of automatic techniques can allow human resources to be directed to metadata creation and evaluation activities that automatic processing cannot adequately complete. Automatic metadata generation is considered more efficient, more consistent, and less costly than human metadata generation. These conclusions are based primarily on automatic indexing research. The recommended functionalities presented here are based on these premises.

The recommended functionalities are mainly restricted to DDLOs, a limitation of the AMeGA project. A portion of the recommendations are, however, applicable to other resource formats, and automatic metadata generation for nontextual resources is briefly addressed in Section 8 of the recommendations.

Additional limitations caused by practical research constraints, including the AMeGA's project restriction to a one-year investigation, are as follows:

- The recommendations focus specifically on the *metadata generation task* and do not address resource selection, authenticity, or value, which are collection development activities.
- The recommendations do not consider resource acquisition, circulation, or other types of functions that ILSs (integrated library systems) generally support.
- The recommendations emphasis is on descriptive metadata and the Dublin Core, and do not consider other types of metadata (e.g., administrative, usage, structural, and provenance metadata)
- The recommendations do not distinguish between different types of DDLOs (e.g., Webpages, WORD documents, PDF documents, etc.), and optimize metadata generation for each type.
- With the exception of the recommendations regarding flexibility for metadata harvesting and extraction from different levels of a resource, these recommendations do not address the complex and compound relationships that DDLOs can have (see, for example, the World Wide Web Consortium's initiative on compound document formats (http://www.w3.org/2004/CDF/)).

2. General System Recommendations

2.1. System should be transparent to individuals who want to know what algorithms are being used. In other words, selected organizational employees or users should be able to view underlying algorithms or any other documentation guiding the metadata generation activity.

2.2. System should automatically generate *meta-metadata* to track the metadata creation process. (Meta-metadata is metadata about the metadata. For example, the *name* of the person who created the metadata, or the *date* the metadata was created.) A profile should be established to determine exactly what the organization would like tracked

(Section 3 covers profiling). Among activities that the system should be able to automatically track are the following:

2.2.1. What algorithms and automatic processes are employed to produce specific metadata elements.

2.2.2. Who intervened to produce metadata (if a person is involved).

2.2.3. When (e.g., date/time) each metadata element was generated.

2.2.4. When (e.g., date/time) a metadata element is revised.

2.2.5. What algorithms and techniques, including human intervention, are employed to revise a metadata element or record.

2.2.6. Version tracking for metadata elements and completed metadata records.

2.3. System should support flexible field lengths for textual metadata elements (e.g., *title* and *description*).

2.4. System should support metadata element repeatability.

2.5. System should ensure that mandatory metadata is captured by either automatic or human processes before a metadata record is published (e.g., default values can be assigned to mandatory elements, or a catch page can be presented to a person).

2.6. System should be usable by multiple types of metadata creators. (Different interfaces may be designed for different user classes, e.g., metadata experts and resource authors.)

3. System Configuration

The system should allow for the configuration of profiles, including metadata element settings. System should be able to automatically integrate all profiles into the metadata generation operation.

Rationale: Automatic application of profiles during metadata generation will inform creation of high-quality metadata in an efficient and effective manner.

3.1. System should be able to store the following types of ***profiles***.

3.1.1. **Resource type** (e.g., research reports, Web documents, journal articles). *DCMI Type Vocabulary*: http://www.dublincore.org/documents/dcmi-terms/#H5 (Section 5, *DCMI Metadata Terms*, 2004) can assist with resource type profiling. System should support automatic detection of resource types using stored profiles.

3.1.1.1.Resource type knowledge should be used for the extraction of semi-structured metadata (e.g., Han et al., 2003; Takasu, 2003).

3.1.1.2.Resource type knowledge should be used to determine which, if any, automatic indexing algorithm(s) should be implemented (Greenberg, 2004b).

3.1.2. **Web resource levels**. System should allow Web resource levels to be predetermined for execution of metadata harvesting and extraction. (How many levels into the main domain should metadata be extracted or harvested from? The main domain is understood as the top URL for a resource.) System should support different level determinations for different resource types.

3.1.3. **Content standards** for topical domains/disciplines and named entities.[5] System should automatically identify topical domains/disciplines or named entities by matching resource content with stored content standards, and suggest standard values from these tools.

3.1.3.1.1. Examples of topical domain/discipline content standards include subject classification and code systems, controlled vocabularies, and ontologies.

3.1.3.1.2. Examples of named entity content standards include name authority files and geographic indexes.

3.1.4. **Metadata standards** (e.g., Dublin Core, Encoded Archival Description). System should be able to detect if a resource has metadata and if it follows a registered metadata standard. The system should be able to automatically read Resource Description Format (RDF) representations, and link to registered element definitions and application profiles.

3.1.5. **Cross-walks** (e.g., Woodley, 2000). System should store cross-walks that will automatically convert existing metadata records to preferred representation standards and facilitate interoperability and metadata exchange.

3.1.6. **Syntax standards and preferences** (see Greenberg, 2003).

3.1.6.1.System should allow for the storage of content syntax standards. (Examples: Date metadata may follow the World Wide Web Consortium Date and Time

[5] The term content standard is used in these recommendations to represent controlled vocabulary tools, classification schemes, ontologies, authority control tools, and other types of schemes that provide content value. These types of tools have been labeled in many different ways (e.g., attribute value schemes, knowledge representation schemes).

Formats (http://www.w3.org/TR/NOTE-datetime) of YYYY-MM-DD, or personal name ordering preference may be *surname, forename*.)

3.1.6.2. System should allow for the storage of element ordering preferences. (Example: A primary author and a secondary author determined by their contribution to a resource.)

3.1.6.3. System should support the storage of preferred encoding standards, including their syntaxes (e.g., MARC, XML)

3.1.7. **Creators**. System should store metadata creator profiles and preferred automatic processing sequences for individual metadata creators. System should automatically detect metadata creator status (e.g., via login identification code) and use status to implement the sequencing of automatic processing during metadata creation. (Example: An organization may want a metadata professional to have more opportunity to review and revise metadata during the creation process than a resource author does.)

3.1.8. **Digital signatures**. System should maintain a profile of digital signatures for trusted metadata generation organizations or people, or link to a trusted metadata evaluator (e.g., if a registry for trusted digital signatures is established). Profiles for digital signatures could help determine the level of automatic harvesting that should be employed. (A profile of digital signatures for poorly producing metadata sources may also be kept so that metadata from such affiliations is not harvested.)

3.1.9. **Metadata element settings** for standard and default values should be stored.

3.1.9.1. System should store standard values for specified metadata elements. (Example: An organization may always require the same value/information for *rights* metadata.)

3.1.9.2. System should store default metadata values. (Example: An organization may want a specific metadata value assigned to an element [e.g., *format* value of *html/text*] if the automatic application or human metadata creator does not assign an element value.)

3.1.10. **Harvesting/extraction sequencing**. System should store profiles for preferred harvesting and extraction sequences. (Example: The emphasis might be placed on metadata extraction for resources without a digital signature.)

3.1.11. **Confidence ratings.** System should employ automatic processing to measure overall metadata record *quality* (emphasizing *accuracy of representation*) and individual metadata element quality. (See section 7.1 of these recommendations.)

3.2. System should support profiles matching the items listed in Section 2 of the recommendations, and other items that will facilitate automatic metadata generation.

3.3. System should allow profiles to be added, deleted, and revised over time.

4. Metadata Identification/Gathering

System should use automatic capabilities to identify and gather any metadata associated with a resource.

Rationale: Automatic functionalities should be exploited as much as possible to detect any existing metadata (structured or semi-structured) associated with a resource for economic purposes.

4.1. Deriving, harvesting, and extraction activities should be guided by established Web resource levels, if a profile has been established.

4.2. **Deriving metadata** (creating metadata based on system properties)

4.2.1. System should automatically generate metadata using stored system properties, such as *date_created* and *date_modified.*

4.3. **Harvesting metadata** (gathering existing metadata).

4.3.1. System should automatically detect if metadata is associated with a resource.

4.3.2. System should read digital signatures, according to established profiles, to determine the degree to which metadata should be harvested (or perhaps should not be harvested) from an existing source.

4.3.3. System should harvest existing metadata associated with a resource (or harvest metadata required according to accepted profiles). Several sources that provide data for harvesting include content creation software, HTML/XHTML and XML MetaTags, custom databases, and bibliographic tools such as EndNote and ILSs.

4.4. **Extracting metadata** (pulling metadata from resource content).

4.4.1. System should extract semi-structured metadata according to resource type profiles.

4.4.2. System should extract keywords from resource content. Extraction algorithm implemented can be informed by resource type.

5. Support for Human Metadata Generation

System should use automatic techniques as much as possible to aid human metadata generation.

Rationale: Using automatic functionalities to assist humans during metadata generation will improve the efficiency of human metadata generation.

5.1. System should dynamically link to content standards, stored in profiles or made accessible via network protocols, to aid humans creating subject and named-entity metadata.

5.2. System should have word processing functionalities such as automatic spell checking, automatic terminology corrections, and other common text processing features to assist humans during metadata generation.

5.3. System should allow for macros to be developed so that standard metadata values can be easily created. Macros should also support acronyms and type-ahead functions stored in a profile.

5.4. System should have customizable input templates for users with different skill levels and responsibilities.

5.5. System should support collaborative metadata creation for different types of creators (for example, a resource author and a professional metadata creator).

5.6. System should track metadata record status by automatically generating *meta-metadata* to document who worked last in creating the metadata, what changes were made, etc., to aid this process (see item 2.2 in the recommendations).

6. Metadata Enhancement/Refinement and Publishing

System should employ automatic techniques to enhance and refine both automatically generated and manually generated metadata.

Rationale: Employing automatic techniques to enhance and/or refine metadata will improve the quality and overall functionality of the metadata.

6.1. System should dynamically link to content standards, and verify that topical/subject and named-entity metadata is authorized, when possible.

6.2. System should automatically support metadata qualification and encode qualifiers.

6.2.1. System should automatically qualify metadata that matches content standards (schemes).

6.2.2. System should automatically qualify metadata refinements and other schemes. Dublin Core qualifiers provided from the *DCMI Metadata Terms* (2004) may aid with qualification.

6.3. System should support word processing functionalities such as automatic spell checking, automatic terminology corrections, and other common text processing features to run against all metadata (also stated as item 5.2 in these recommendations).

6.4. System should verify that metadata produced follows the preferred metadata standard (e.g., Dublin Core).

6.5. System should support automatic linking of metadata records representing related items through authorized *relation* qualifiers, or other metadata elements such as uniform title and creator. Records linking preferences should be set up in a profile. For example, if a profile is set up on the basis of Functional Requirements for Bibliographic Records (FRBR), relationships should be automatically linked to follow this model.

6.6. System should automatically convert metadata to appropriate or preferred syntaxes (content, ordering, and encoding syntaxes [see item 3.1.5 in these recommendations]).

6.7. System should support translation of metadata element values or full metadata records into different languages with appropriate diacritics.

7. Metadata Evaluation

System should use automatic techniques to evaluate metadata quality and provide a statistical rating score. Examples of criteria are given below in 7.1.1 to 7.1.6.

Rationale: Automatic metadata evaluation techniques will improve the efficiency of metadata evaluation, enable human resources to be directed to metadata evaluation that automatic processing cannot adequately perform, and ultimately improve metadata quality.

7.1. System should use a range of criteria to determine metadata quality. Statistical data gathered via the underlying criteria should be used to generate a confidence rating of the metadata record's overall quality and quality of the metadata per element. (An organization may not want to spend human resources evaluating metadata records given high confidence ratings, but rather direct resources to metadata records given lower confidence ratings.) Examples of evaluation criteria follow in question format:

7.1.1. How much metadata was harvested? Was a digital signature associated with the metadata, and if so, was it registered as a trusted source?

7.1.2. How much metadata was extracted?

7.1.3. What extraction algorithm was used, and what is the overall confidence rating of the algorithm?

7.1.4. How well did the automatically generated metadata match content standards used to assign metadata values? (Example: A direct match [e.g., matching "Web commerce" to "Web commerce"] should receive a higher ranking than a partial match [e.g., matching "Web commerce" to "Web business]). Information retrieval techniques such as term stemming, removing stop words, and term flipping need to be considered here.

7.1.5. How well did the humanly generated metadata match content standards used to assign metadata values? (Scoring examples from 7.1.4 directly above apply.)

7.1.6. How complete is the metadata record in terms of matching a standard metadata scheme?

***Note, for items 7.1.1. to 7.1.6:** Each organization will need to identify its criteria for evaluation and create a profile that will enable a score to be generated. Bruce and Hillmann's (2004) discussion of metadata quality can aid in further establishing evaluation criteria.

7.2. System should filter and flag problems (e.g., syntax, authority control, encoding problems). They should be filtered first, subjected to automatic revision, then flagged for human review, if automatic revision does not improve the confidence rating to an acceptable level.

7.3. System should automatically route problematic metadata records that cannot be corrected via automatic processes, to appropriate persons, according to the problem (e.g., metadata experts or resource authors) for review.

8. Metadata Generation for Nontextual Resources

Automatic techniques should be used as much as possible to create metadata for nontextual resources (e.g., visual resources, geospatial resources, moving images).

Rationale: A variety of technical metadata is generated automatically when nontextual digital resources are created. This metadata is valuable, and a human should not spend time recreating it, when it can be harvested from nontextual resources' source code.

8.1. Profiles can be set up over time to determine what metadata can be reasonably harvested from such sources. The *Data Dictionary: Technical Metadata for Digital Still Images* standard, National Information Standard Organization (NISO) Z39.87 (2002) identifies technical metadata that is generated automatically by image capture software and can be harvested for metadata record creation.

8.2. Metadata standards for nontextual resources need to be incorporated into system profiles to facilitate harvesting of technical and descriptive metadata (both system and human generated) that is useful.

9. Conclusions and Future Research Directions

This report presents results of both a *features analysis* and a *metadata expert survey*, identifies recommended functionalities for automatic metadata generation applications, and highlights important research needs in the area of automatic metadata generation. The last section of this report (Section 10) recommends next steps for the Library of Congress.

The features analysis identified the metadata functionalities supported by content creation software commonly used to create digital resources. All the applications support at least three descriptive metadata elements, and many elements could be conceptually mapped to the Dublin Core. In general, these metadata elements support file organization and searching, and are useful for resource discovery. With the exception of Dreamweaver, all the applications harvest metadata from stored system properties, and five of the seven applications use metadata extraction techniques, although the methods were rudimentary. Finally, both EndNote and Winamp harvest metadata from networked resources.

These findings are important because they verify that content creation software supports metadata generation. This can provide an important data source for metadata generation applications. In fact, some metadata generation applications already harvest metadata associated with digital resources (see the review of DC-dot in Greenberg, 2004b). Despite these findings, little is known about the frequency of content creation software metadata features use—specifically, the features requiring human input. Moreover, there has been little study of the quality of metadata that content creation software generates via automatic means, or the quality of metadata produced by different types of metadata creators (e.g., resource authors, information/Web architects) using content creation software features.

Research is needed to address these shortcomings to better understand how to leverage metadata generated via content creation software or associated with digital resources. It is likely that current metadata features are not fully used in content creation software, and it is important to study how to better implement metadata features. Researchers must be mindful that, in some cases, there is a desire to erase any metadata to eliminate all aspects of document accountability. In fact, Microsoft recently released a metadata removal tool to "scrub leaky metadata from documents" in response to news items about Iraq's security and intelligence and issues of metadata tracking (Libbenga, 2004). Future research must, therefore, also address questions about metadata trust and validation.

The other main research component of this report, the *metadata expert survey*, identified desired system functionalities for automatic metadata applications. Results indicate that metadata experts favor using automatic metadata generation, particularly for metadata that can be created accurately and efficiently. However, participants generally did not favor eliminating human evaluation or production for the more intellectually demanding metadata (e.g., *subject* metadata). Even so, most participants agreed that automatic processes should be employed to aid humans creating metadata—including metadata requiring intellectual discretion. Two metadata functionalities strongly favored by participants are:

- Running automatic algorithm(s) initially to acquire metadata that a human can evaluate and edit.
- Integrating content standards (e.g., subject thesauri, name authority files) into the metadata generation applications.

Support for the first functionality requires the integration of research findings in the areas of automatic indexing, abstracting, and classification. It is suggested that metadata generation applications can be improved by taking advantage of algorithms developed via:

- Domain-specific automatic indexing research (e.g., Nadkarni et al., 2001).
- Automatic abstracting research (e.g., Johnson, 1995).
- Automatic classification research (e.g., Losee, 2003).
- Document genre research (Toms et al., 1999).
- Automatic metadata generation research experimenting with semi-structured metadata (e.g., Han et al., 2003; Takasu, 2003).

The second functionality requires that metadata applications leverage current information infrastructure developments. Already, there is the Resource Description Framework (RDF) (http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.) and the World Wide Web Consortium (W3C) Ontology Markup Language (OWL) (http://www.w3.org/TR/owl-features/), which permit interoperability and sharing of content standards. As noted above, the Web's global framework has led to construction of metadata registries specifically for sharing knowledge representations such as thesauri and ontologies (e.g., Lutes, 1999; Knowledge System Laboratory (KSL) Ontology Server, Stanford University: http://www-ksl-svc.stanford.edu:5915/doc/ontology-server-projects.html) and metadata schemes (SCHEMAS Registry: http://www.schemas-forum.org/registry/; Dublin Core Metadata Registry: http://dublincore.org/dcregistry/). Many of these developments also support Semantic Web Construction (Heery & Wagner, 2002). Finally, there are applications such as the Library of Congress' Catalogers Desktop (http://desktop.loc.gov/) that integrate many important bibliographic tools. Automatic metadata generation applications providing access to useful resources, in an intelligent manner, will be able to greatly enhance metadata quality.

The features analysis and the metadata expert survey highlight research areas important to the development of automatic metadata generation applications. Four other important research areas identified in this report include the following:

- Improving the use of automatic methods to assist with QC of metadata.
- Studying the contextual need for metadata (e.g., which metadata elements are important for which functions and which classes of users).

- Incorporating automatic metadata developments from nontextual resource communities such as the image community.
- Examining the psychological and social impacts of automation on metadata experts.

Application development results from research, although scientific evaluations of application functionality may not always be conducted because of limited resources and pressures to produce a product. Even so, the best applications draw from research and incorporate research findings. For libraries to *take leadership* in bibliographic control of Web resources, they must become more involved in the development of superior and more robust automatic metadata generation applications, and application designers must incorporate research findings. The Library of Congress can lead the development of metadata efforts by building a metadata application that incorporates the functionalities recommended in this report, continuing to facilitate and foster automatic metadata generation research efforts, and expanding communications beyond the library environment to content creation software vendors and other communities in which metadata plays a vital role.

10. Recommended Next Steps for the Library of Congress

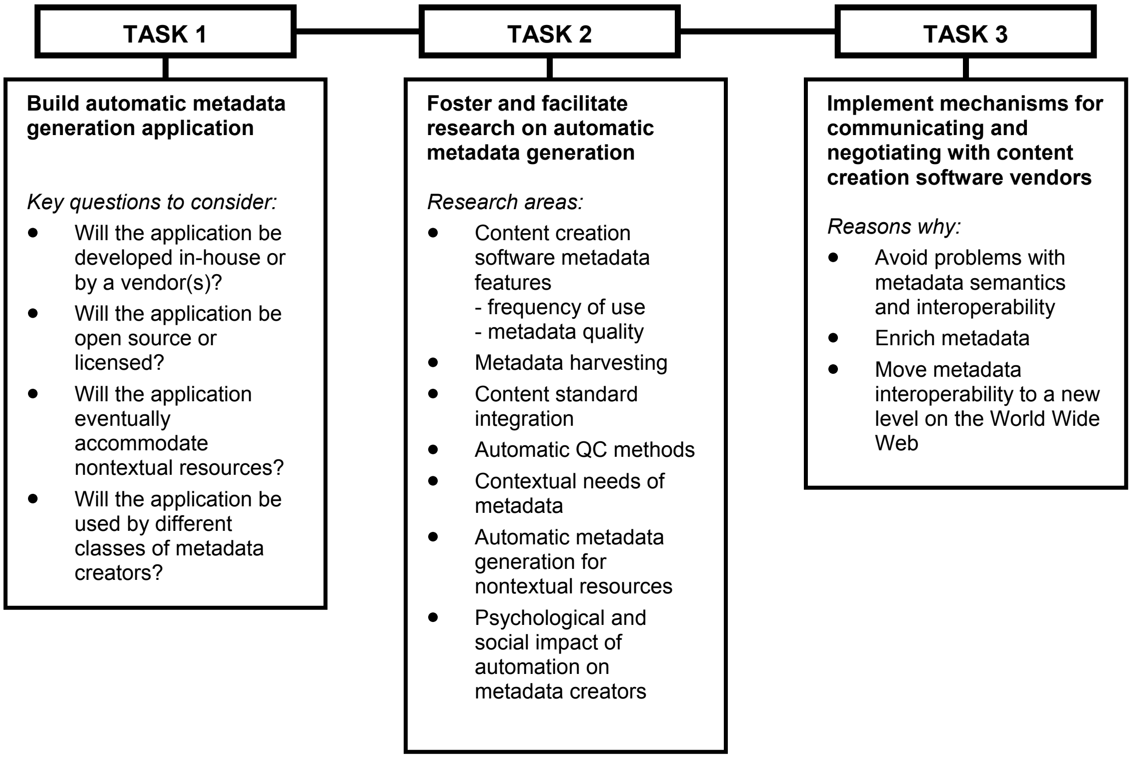
The *Bibliographic Control of Web Resources: A Library of Congress Action Plan* (LC Action Plan) recognizes the need to improve the bibliographic control of Web resources, and Section 4.0 specifically targets the development of automatic tools to address this need. The AMeGA project was established to address need and identify recommended functionalities for automatic metadata generation applications. The recommendations are presented in Section 8 of this report. This section outlines recommended next steps for LC.

A three-pronged approach to developing an automatic metadata generation application is recommended. This approach is comprised of multiple components. The three main tasks are to:

1. Build an automatic metadata generation application.
2. Foster and facilitate research on automatic metadata generation.
3. Implement mechanisms for communicating and negotiating with content software creation vendors.

These tasks and the accompanying subtasks are listed in Figure 10, and are discussed below.

Figure 10: Three-Pronged Approach for the Development of Automatic Metadata Generation Applications for the Library of Congress



TASK 1

Build automatic metadata generation application

Key questions to consider:

- Will the application be developed in-house or by a vendor(s)?
- Will the application be open source or licensed?
- Will the application eventually accommodate nontextual resources?
- Will the application be used by different classes of metadata creators?

TASK 2

Foster and facilitate research on automatic metadata generation

Research areas:

- Content creation software metadata features
  - frequency of use
  - metadata quality
- Metadata harvesting
- Content standard integration
- Automatic QC methods
- Contextual needs of metadata
- Automatic metadata generation for nontextual resources
- Psychological and social impact of automation on metadata creators

TASK 3

Implement mechanisms for communicating and negotiating with content creation software vendors

Reasons why:

- Avoid problems with metadata semantics and interoperability
- Enrich metadata
- Move metadata interoperability to a new level on the World Wide Web

Task 1. Build an Automatic Metadata Generation Application

The recommended functional requirements outlined in Section 8 of this report can serve as a request for proposals (RFP) for construction of an automatic metadata generation application. Although several of the functionalities require further research (see Task 2 below), most recommendations can be addressed now by mustering appropriate resources.

LC can pursue construction of an automatic metadata generation application by leveraging in-house expertise and/or working with vendors and researchers. LC may also consider disseminating the recommendations to the greater metadata community, in a request for comment (RFC) format, to help identify additional functionalities or to prioritize them. Given LC's leadership role, it is recommended that the institution move ahead swiftly to develop an application while making users aware that the application will be enhanced over time. Some immediate decisions are important for moving forward, including following:

- Will the application be open source or licensed, perhaps with a tool such as the Catalogers Desktop?
- Will the application be for textual resources and/or digital resources in multiple formats?
- Will the application be developed for metadata experts and/or the larger community of metadata creators?

The first question is a policy matter beyond the scope of this report. Regarding the second and the third questions, it is recommended that the initial application focus on DDLOs, with recognition of the organic nature of applications to anticipate incorporating functionalities for other formats over time. It is also recommended that the application be developed for the larger community of metadata creators (e.g., metadata experts, resource authors, information/Web architects) so that numerous classes of metadata creators can benefit from LC's leadership.

Task 2. Foster and Facilitate Research on Automatic Metadata Generation.

A number of action items identified in the LC Action Plan have cultivated research efforts related to metadata, including the research presented in this report. It is anticipated that LC, as a leading institution, will continue to advance research efforts in the area of metadata and incorporate findings into the development of automatic metadata generation applications. The following list includes key research questions identified through the AMeGA project.

- Content creation software questions:
    - How frequently are content creation software metadata features used? If not, why not?
    - How can the use of content creation software features be improved?
    - In cases where content creation software metadata features are used, who is using these features and what is the quality of metadata they produce?

- How can state-of-the-art metadata creation applications be improved to evaluate and harvest high-quality metadata associated with digital resources originally produced with content creation software?
- How should content standards (e.g., subject thesauri, name authority files) be integrated, via automatic means, into metadata generation applications?
- How should automatic methods be employed to assist with QC of metadata?
- What are the different contextual needs of metadata (e.g., which metadata elements are important for which functions and which classes of users)?
- How should metadata developments taking place in the image community and other related developments for nontextual resources be incorporated into automatic metadata generation applications?
- What is the psychological and social impact of automation on metadata creators?

The recommended functionalities in Section 8 of this report address some of these questions, suggesting methods and offering preliminary approaches for addressing them. Nevertheless, further investigation of these questions is sure to improve the development of automatic metadata generation applications.

Task 3. Implement Mechanisms for Communicating and Negotiating with Content Creation Software Vendors to Improve Metadata Functionalities

Research on content creation software metadata features presented in this report is exploratory. Nevertheless, the results demonstrate that content creation software supports metadata generation and can provide an important data source for automatic metadata generation applications. It is recommended that LC lead efforts to communicate and negotiate with content creation software vendors regarding current and future metadata functionalities. LC and the overall the library community have immense bibliographic control expertise, and are poised to apply that expertise. LC is an important institution and can provide a vital link to the metadata generation continuum through communicating with content creation software vendors. Means of communication to consider include conference panels, reports, workshops, and other forums whereby interested parties can meet and discuss metadata issues. LC should also negotiate with content creation software vendors to improve their metadata features.

LC can look to developments with ONIX (Caplan, 2001), the metadata standard developed in response to the enormous growth in online book sales. The library community was initially absent from this development, and ultimately faced a rich metadata source that contained many semantic problems and hampered interoperability. Communication has improved over time, and in fact the library community is able to enrich their metadata at times, based on developments associated with ONIX. In an account of metadata developments at the Library of Congress Bibliographic Control Conference in 2000, Caplan (2001) highlighted this development and then energetically stated that we need to "work proactively with publishers to establish enough commonality between our respective rule sets to allow meaningful exchange and reuse of metadata." LC is in a strong position and should heed this advice in relation to content creation software, moving metadata interoperability to a new level on the World Wide Web.

11. References

Anderson, J. D., & Perez-Carball, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval - Part I: Research, and the nature of human indexing. *Information Processing & Management 37*(2), 231–254.

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice* . Chicago, IL: ALA.

Caplan, P. (2001). International metadata initiatives: Lessons in bibliographic control. In P*roceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web*, Library of Congress, Washington, D.C., November 15-17, 2000. Retrieved January 5, 2005, from http://www.loc.gov/catdir/bibcontrol/caplan_paper.html.

CONDOC. (1981). Revisiting CONDOC: A new look at the online catalog sponsored by the Ala Catalog Use Committee. Available at: <listserv@listserv.buffalo.edu>. FTP Request: "CONDOC Report.

Crystal, A., & Greenberg, J. (*in press*). Usability of a metadata creation application for resource authors. *Library and Information Science Research*, *27*(2).

Cutter, C. A. (1904). *Rules for a dictionary catalog* (4th ed.). Washington, D.C.: Government Printing Office.

Dakshinamurti, G. (1985). Automation's effect on library personnel. *Canadian Library Journal*, *42*, 343-351.

DCMI Usage Board. (2004). *DCMI metadata terms*. Retrieved January 5, 2005, from http://dublincore.org/documents/2004/09/20/dcmi-terms/ .

International Federation of Library Associations and Institutions. (1998). *Functional requirements for bibliographic records: Final report*. Retrieved January 5, 2005, from http://www.ifla.org/VII/s13/frbr/frbr.pdf .

Greenberg, J. (2003). Metadata and the World Wide Web. In M.S. Drake (Ed.) *Encyclopedia of library and information science* (2nd ed.) (pp.1876-1888). New York: Marcel Dekker, Inc.

Greenberg, J. (2004a). Definitions of terms used in the AMeGA Survey. Retrieved January 5, 2005, from http://ils.unc.edu/mrc/amega_survey_defs.htm.

Greenberg, J. (2004b). Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, 6(4), 59–82.

Greenberg, J. (*in press*). Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 41(3/4). Also to appear in R. Smiraglia (Ed.), *Metadata: A cataloger's primer*. New York: Haworth Information Press.

Greenberg, J., Crystal, A., Robertson, W. D. & Leadem, E. (2003). Iterative design of metadata creation tools for resource authors. In Sutton, S. Greenberg, J., and Tennis, J. (Eds.). *Proceedings of the 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and* Applications, Seattle, Washington, September 28-October 2, 2003. Retrieved January 5, 2005, from http://www.siderean.com/dc2003/202_Paper82-color-NEW.pdf.

Gunter, B., Nicholas, D., Huntington, P., & Williams, P. (2002). Online versus offline research: Implications for evaluating digital media. *Aslib Proceedings, 45*(4), 229–239.

Han, H. C., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E.A. (2003). Automatic document metadata extraction using support vector machines. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 37 – 48). New York: ACM Press.

Hatala, M. & Forth, S. (2003). System for computer-aided metadata creation. In *Proceedings of 12th International Conference of the World Wide Web Consortium (WWW2003)*, Budapest, May 20-24, 2003.

Hayslett, M. M., & Wildemuth, B. W. (2004). Pixels or pencils? The relative effectiveness of Web-based versus paper surveys. *Library and Information Science Research*, *26*(1), 73–93.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K. P. (2002). Finding the flow in web site search. *Communications of the ACM, 45*(9), 42–49.

Heery, R., & Wagner, H. (2002). A metadata registry for the semantic web. *D-Lib Magazine*, *8*(5). Retrieved January 5, 2005, from http://www.dlib.org/dlib/may02/wagner/05wagner.html.

Heyman, B. L. (1981). In line to get on line: A background report on CONDOC (The Consortium to Develop an On-line Catalog). *Colorado Libraries, 7*(4), 10-13.

Ji, Y. G. & Salendy, G. (2002). A metadata filter for intranet portal organizational memory information systems. *International Journal of Human-Computer Studies*, *56*(5), 525 – 537.

Johnson, F. (1995). Automatic abstracting research. *Library Review, 44*(8), 28 - 36.

Lan, W. C. (2002). From document clues to descriptive metadata: Document characteristics used by graduate students in judging the usefulness of web documents. Doctoral dissertation, University of North Carolina at Chapel Hill.

Libbenga, J. (2004). Microsoft releases metadata removal tool. *The Register*. Retrieved January 5, 2005, from http://www.theregister.co.uk/2004/02/02/microsoft_releases_metadata_removal_tool .

Liddy, E. D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N. E., Diekema, A., McCracken, N. J., Silverstein, J., & Sutton, S. A. (2002). Automatic metadata generation & evaluation. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 11-15, 2002, Tampere, Finland (pp. 401–402). New York: ACM Press.

Losee, R. (2003). Adaptive organization of tabular data for display. *Journal of Digital Information 4*(1). Retrieved January 5, 2005, from http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Losee/.

Lutes, B. (1999). Web thesaurus compendium. Retrieved January 5, 2005, from http://www.ipsi.fraunhofer.de/~lutes/thesoecd.html.

Nadkarni, P., Chen, R., & Brandt, C. (2001). UMLS concept indexing for production databases: A feasibility study. *Journal of the American Medical Information Association, 8*(1), 80–91.

National Information Standards Organization. (2002). *Data dictionary: Technical metadata for digital still images*. Proposed NISO standard Z39.87. Retrieved January 5, 2005, from http://www.niso.org/standards/resources/Z39_87_trial_use.pdf.

Patton, M., Reynolds, D., Choudhury, G. S., & DiLauro, T. (2004). Toward a metadata generation framework: A case study at the John Hopkins university. *D-Lib Magazine*, *10*(11). Retrieved January 5, 2005, from http://www.dlib.org/dlib/november04/choudhury/11choudhury.html.

Research Libraries Group. (2003). *Automatic exposure: Capturing technical for digital still images*. Retrieved January 5, 2005, from www.rlg.org/longterm/ae_whitepaper_2003.pdf.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Schwartz, C. (2002). S*orting out the web: Approaches to subject access*. Westport, Connecticut: Ablex publishing.

Smiraglia, R. P., & Leazer, G. H. (1999). Derivative bibliographic control relationships: The word relationship in a global bibliographic database. *Journal of the American Society for Information Science*, *50*(6): 493–504.

Takasu, A. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 49 – 60). New York: ACM Press.

Tillet, B. (1991). A taxonomy of bibliographic relationships. *Library Resources & Technical Services*, *35*(2), 150 – 158.

Tillett, B. B. (1992). Bibliographic relationships: An empirical study of the LC machine-readable records. *Library Resources & Technical Services*, *36*(2), 162 – 88.

Toms, E., Campbell, D., & Blades, R. (1999). Does genre define the shape of information: The role of form and function in user interaction with digital documents. *Proceedings of the 62nd American Society for Information Science Annual Meeting,* pp. 693-704.

Van Duinen, R. S. (2004). New discoveries in the André Savine collection: Examining the author-generated metadata contained in the bibliographic and biographical record of André Savine. Unpublished Master's Paper, School of Information and Library Science, University of North Carolina at Chapel Hill. Retrieved January 7, 2005, from http://hdl.handle.net/1901/121.

Vellucci, S. L. (1997). Bibliographic relationships. Paper presented at the *International Conference on the Principles and Future Development of AACR*, Toronto, Canada. Retrieved January 5, 2005, from http://collection.nlc-bnc.ca/100/200/300/jsc_aacr/bib_rel/r-bibrel.pdf.

Weinstein, P. C. (1998). Ontology-based metadata: Transforming the MARC legacy. In *Proceedings of the 3rd ACM International Conference on Digital Libraries,* June 23-26, Pittsburgh, PA (pp. 254 – 263). New York: ACM Press.

Weintraub, K. D. (1979). The essential of the bibliographic record as discovered by research. *Library Resources & Technical Services*, *23*(4), 391-405.

Woodley, M. (2000). Metadata standards crosswalks. In Baca, M. (Ed.), *Introduction to metadata: Pathways to digital information*. Los Angles, CA: Getty Information Institute. Retrieved January 5, 2005, from http://www.getty.edu/research/conducting_research/standards/intrometadata/3_crosswalks/index.html.

Zhang, Y. (2000). Using the internet for survey research: A case study. *Journal of the American Society for Information Science*, *51*(1), 57-68.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. Oxford: Heinemann Professional.

Appendix A – Part 1: By Michelle Cronquist

A summary of observations on OPAC features that involve or could potentially include automatic processes. Submitted January 2004.

Martha M. Yee and Sarah Shatford Layne, *Improving Online Public Access Catalogs*
Recommended features in OPACs:

- Linking to related works (works with the same author, subject, uniform title)
- Indexing of uniform titles and titles proper together, so that users can retrieve a given work regardless of whether its uniform title is equal to its title proper
- Partitioning of author-title search results into editions of the work, works about the work, and works related to the work
- Sophisticated global change feature allowing controlled fields to be changed and transcription fields to be left alone
- Automatic truncation of searches (perhaps only as an option after a search fails)
- Automatic matching of names, titles, and subjects with similar headings (e.g., a search for Taylor, Jonathan brings up works by Taylor, J., or a search including "&" automatically searches also for "and") (again, perhaps only as an option after a search fails)
- Fuzzy matching when a search fails

Automatic features in specific systems
ENCompass Solutions (http://www.endinfosys.com):

- Integrates local and remote OPACs, e-books, electronic databases, e-journals, and digital collections, allowing patrons to search all of them at once
- Automatically integrates e-journal holdings from different providers; automatically produces links to these holdings
- Automatically creates administrative metadata (e.g., creator or editor of a record, date and time created or edited)
- Includes a Universal Catalog (union catalog) for consortia, which is automatically updated when local records are added or changed
- ImageServer: provides access to digitized collections (images, maps, books, etc.)
    - "ImageServer automatically creates USMARC records and stores them in the Voyager database"
    - "allows staff to describe images without having to understand USMARC"
    - "offers a template to input descriptive information"
    - automatically creates 856 field to link to the digital object
    - objects can be organized into folders; system automatically cross-references related folders (not clear how "automatic" this really is)
- LinkFinderPlus: provides links to full-text articles from online databases
    - "automatically provides links or search choices to the library OPAC and popular Internet resources such as Google"
    - provides access to related materials, such as abstracts, reviews, Internet search engines.

Additional notes:
Automatic features supported by OPACs:

- Global update (such as updating subject headings that have changed)
- Collocation of records that have the same author, subject, uniform title.
- Checks headings against authority file
- Replaces nonauthorized headings (usually done by an authority vendor)
- Checks validity of MARC format
- Consistency checking (e.g., checking language code against initial articles)
- Automatically generated fixed field (005) giving information on the creation and editing of the record
- Strips certain characters from search strings
- Orders titles and headings by number or date, rather than character by character
- Creates keyword index
- Checks for duplicate records
- Creates local authority records when necessary
- Creates reports of headings automatically changed

Appendix A – Part 2: By Amanda Wilson

Amanda Wilson
28 April 2004
AMeGA ENCompass Report

ENCompass Report
(Interview with Drusilla Simpson of the State Archives)

Overview

The State Archives uses the ENCompass Solution product which, according to ENCompass, has integrated metadata functionalities into its system. However, the State Archives has learned that while the system may incorporate various metadata schemes (including MARC, EAD, and Dublin Core), many of the traditional advantages of OPACs (including global updates, authority files, and capacity for tens of thousands of records or more) was not lost. To retain those functionalities, an institution would need to also purchase the Voyager product.

ENCompass' use of metadata accommodates institutions that have different types of materials requiring different metadata schemes to manage collections. Each scheme has an associated repository in which records for materials are housed. ENCompass then requires a mapping script in order to search across repositories. Besides an incomplete knowledge of the utility of metadata schemes such as EAD (elements are listed in alphabetical order, which confounds the intellectual organization of the scheme), the repository organization is advantageous (please see diagram attached). However, the system is probably best suited for a small institution, or one just beginning to use an automated collection management system because restructuring data in existing databases may be cost-prohibitive cost for larger, more established institution such as the State Library.

Questions and Answers

1. Which ENCompass Solutions product do you use: ENCompass Remote Access, ENCompass for Digital Collections, or ENCompass?

The State Archives uses ENCompass. They had an in-house system to manage their collections and wanted a system that would give them a graphical interface for end-user searching of the database.

2. ENCompass Solutions is advertised as an added-feature product that operates independent of Endeavor's principle OPAC product Voyager. Do you use Voyager?
    a. If so, can you envision some challenges others would face based on your experience?
    b. If not, what product do you use and what were the challenges implementing the product?

The State Archives does not use Voyager, but the State Library does. The Archives uses a Visual Basic database as an intermediate collection management tool and then periodically uploads new records into ENCompass. ENCompass is being used as a publishing tool primarily and for user-defined data entry capabilities.

All of the features available in the Voyager OPAC—including patron registration, authority files, global changes, and reports—are not incorporated into ENCompass.

3. Would you describe your general usage of ENCompasss Solutions product?
    a. Used as a search tool?
    b. Used as a cataloging tool?
    c. What are the different types of resources added to the system?
    d. If records for images, sounds, or photographs are kept, are special arrangements needed?

ENCompass is used as a search tool for users. However, because of the storage infrastructure for different types of records, cross-searching problems arose with the metadata. (Please see attached sheet for diagram of the State Archives' collection management system structure.) Essentially, mapping across fields was not as efficient as the homegrown system MARS and the information in MARS, would have to be reorganized to be stored in ENCompass. One of the positives about ENCompass, which is only 2 years old, is that the system attempted to embrace EAD. Conversely, the input form for EAD documents lists all elements alphabetically, thereby removing the intellectual organization of the metadata scheme. The State Archives has coped with this and is hoping for change in later versions of the system.

4. XML is used for structured requests (SR) and receipt of information (ROI) in ENCompass Solutions. For ROI portion, DC elements are used as input fields.
    a. Are any of the fields automatically generated?
        i. Endeavor purports that administrative metadata is generated automatically. What have been your experiences?
            1. Which DC elements do you use?
                a. Is the DC qualified or unqualified?
                b. If qualified, which elements receive refinements?
            2. Are some elements generated better than others?
            3. Is any additional human editing required?
    b. Are there (additional) fields you would like to see automated to assist in the use of metadata?

The State Archives does not use this portion of the system. In fact, the Archives discovered that for many functions of the system that were demonstrated by the representative or purported by ENCompass, the Voyager component is needed.

5. Do you use the ImageServer or LinkFinderPlus?
    a. For ImageServer (if you use it):
        i. Does ENCompass create USMARC records accurately?
            1. When describing images, do staff necessarily need to know USMARC?
            2. Is the template to input descriptive information intuitive?
        ii. How well does ENCompass organize objects into folders automatically?

Again, the State Archives does not use this feature. The reason for not using ImageServer is that the capacity for the system is not sophisticated enough to handle the volume of images that the State Archives has 50,000+ images. Druscie Simpson thinks that ImageServer may be phased out in favor of the ENCompass for Digital Libraries which is better equipped to handle large collections.

6. ENCompass Solutions uses a single interface to search print, e-journal, e-databases, and digital collections. What automatic or semi-automatic metadata, if any, is produced when integrating these resources?

The State Archives does not use ENCompass for that service.

Next Steps

The next step is to talk to Wake Forest University about their use and experiences with the ENCompass system. Perhaps that institution uses more of the product's metadata features; maybe the same problems that the State Library witnessed are experienced at the university as well.

My Observations

State Archives' use of automatic or semi-automatic features in ENCompass:

1. Mapping across repositories (semi-automatic)

ENCompass' stated support of automatic metadata generation features, but not used at State Archives:

1. Integrates all materials libraries have access to (e.g., e-books, e-journals, OPACs, digital collections), which enables cross-searching.
2. Automatically integrates e-journal holdings from different providers and produces links to these services.
3. Automatically creates administrative metadata (e.g. creator and editor of records, time and date information about creation or editing of records).
4. Universal catalog (union catalog) for consortia and updates when records are added locally.
5. ImageServer:
    a. Automatically creates USMARC records.
    b. Automatically creates 856 field to link to image.
    c. System cross-references related folders in which objects are stored.
6. LinkFinderPlus:
    a. Provides links to OPAC and internet search engines.
    b. Provides access to related materials (e.g. abstracts, reviews, etc.).

My thoughts on metadata creation activities that OPACS could do automatically

1. Mapping across controlled vocabularies (when searching across different systems)
2. Search scaffolding (return materials related to search terms in specific subject areas)
3. ILL forms for materials not housed in library (extension of serials solutions)

4. Other recommended features from Yee and Layne's *Improving Online Public Access Catalogs* include:
    a. Partitioning author-title search results into editions of work, works about the work, and works related to the work;
    b. Linking of related works

Diagram

ENCompass Solutions
Collection Management System

Search/ End User Interface

Repository	Repository	Repository	Repository
EAD	Dublin Core	MARC	Microsoft Word
<title>	<title>	<245>	
Materials: Finding Aids	Materials: Publications	Materials: Collections	Materials: Table of Images

The ENCompass system collocates records for different materials into repository collections by institution specifications. The State Archives chose to have materials divided based on description scheme. To enable searching across all collections, elements in each scheme had to be mapped to elements in each of the other schemes (**title** element is illustrated above).

Issues that may arise from this organization come from institutions that have kept their records in different intellectual organization. For example, the State Archives kept their information in a home-grown database, MARS, which organized their materials from material type to item: 3 = Map Collection (group); 3.1= County Maps (series); 3.1.1= box (box); 3.1.1.1= Alamance (folder); 3.1.1.1.1= actual map (item.). Reorganizing from this organization to ENCompass' may require a great institutional effort,

Appendix B -- The Metadata Expert Survey

SurveyMonkey.com
because knowledge is everything

Privacy | Contact Us | Logout

Home | New Survey | My Surveys | List Management | My Account | Help Center

Sunday, December 26, 2004

Design Survey

Show All Pages and Questions

<< Back | Preview

To change the **look** of your survey, select a choice below. Click 'Add' to create your own custom theme.

Theme: Blue Ice | Add

AMeGA Survey

Edit Title | Edit Numbering | Edit Logo

Add Page

1. AMeGA Survey

Edit Page | Delete Page | Copy/Move | Add Logic

Generating good quality metadata in an efficient manner is essential for organizing and making accessible the growing number of rich resources available in today's information environments.

One method of making metadata creation more efficient is to automate its generation.

This survey is being conducted to determine which aspects of metadata generation are most amenable to automation and semi-automation for digital document-like objects.

More information about this project can be found at the project web site.

Add Question | Add Page

2. Participant Information

Edit Page | Delete Page | Copy/Move | Add Logic

Information about you and your professional metadata activities may help us determine if there are different automatic metadata generation needs in different environments.

Providing this information is optional; however, be assured that any information you give will be kept confidential and your anonymity will be preserved in any reports of survey results.

Add Question | Add Page

Edit | Delete | Copy/Move

1. What is your name? ***(optional)***

Add Question | Add Page

Edit | Delete | Copy/Move

2. What is your email address? ***(optional)***

Add Question | Add Page

Edit | Delete | Copy/Move

3. What is your professional title? ***(optional)***

Add Question | Add Page

Edit | Delete | Copy/Move

4. What institution/organization do you work for? ***(optional)***

Add Question | Add Page

3. General Metadata/Cataloging Experience

Edit Page | Delete Page | Copy/Move | Add Logic

Information about your metadata/cataloging experience will provide context for our findings on automatic metadata generation needs.

Add Question | Add Page

Edit | Delete | Copy/Move | Add Logic

5. How many years of experience do you have cataloging or indexing any type of material?

□ Less than one year
□ One year
□ Two years
□ Three years
□ More than three years

Add Question | Add Page

Edit | Delete | Copy/Move | Add Logic

6. How many years of experience do you have working with metadata (cataloging, indexing, other metadata activities) specifically for digital document-like objects?

□ Less than one year
□ One year
□ Two years
□ Three years
□ More than three years

Add Question | Add Page

Edit | Delete | Copy/Move | Add Logic

7. Which metadata-related activities are you or have you been involved in? Check all that apply.

□ metadata creation
□ supervising/administration of metadata creation
□ maintenance/quality control of metadata

□ Other Role(s) (please describe)

Add Question | Add Page

4. Please Note: Edit Page | Delete Page | Copy/Move | Add Logic

The remaining questions in this survey specifically refer to metadata for digital document like objects.

We define a digital document-like object as a primarily textual resource that is accessible through a Web browser. It may contain images, sound, and non-textual formatting, but it must contain textual content. Examples include HTML/XHTML resources, Microsoft Word documents, Adobe PDF documents, etc.

The term "digital resource" will be used hereafter to refer to digital document-like objects.

Add Question | Add Page

5. Current metadata practices for digital resources Edit Page | Delete Page | Copy/Move | Add Logic

The questions on this page gather information specifically about how your institution or organization is currently working with metadata for digital resources (digital document-like objects).

Add Question | Add Page

Edit | Delete | Copy/Move | Add Logic

8. Who else in your organization/consortium/initiative is creating metadata for digital resources? Check all that apply.

□ Metadata professionals (e.g. catalogers/indexers)
□ Other professionals (e.g. reference staff, subject specialists)
□ Information/web architects
□ Non-professional, paraprofessional, or technical organizational staff
□ Resource authors
□ Volunteers
□ Other (please specify)

Add Question Add Page

Edit Delete Copy/Move Add Logic

9. What metadata standards are you using to describe digital resources in your organization? Check all that apply.

□ MARC Bibliographic Format
□ Simple Dublin Core
□ Qualified Dublin Core
□ Dublin Core application profile(s)
□ EAD (Encoded Archival Description)
□ GEM (Gateway to Educational Materials)
□ MODS (Metadata Object Description Schema)
□ TEI Header/TEI (Text Encoding Initiative)
□ Darwin Core
□ SCORM (Sharable Courseware Object Reference Model)
□ IEEE LOM (Learning Object Metadata)
□ GILS (Government Information Locator Service)
□ Other (please specify)

Add Question Add Page

Edit Delete Copy/Move Add Logic

10. Which function(s) does the metadata you are creating for digital resources support? Check all that apply.

□ Resource discovery/information retrieval

□ Preservation

□ Internal adminstration

□ Rights management

□ External harvesting/resource sharing

□ Other (please specify)

Add Question Add Page

Edit Delete Copy/Move

11. What software/systems are you using to support metadata creation for digital resources? Please name the systems used, and comment on their effectiveness and user-friendliness.

Add Question Add Page

Edit Delete Copy/Move Add Logic

12. For metadata creation for digital resources, have you ever used a tool that has automatic capabilities such as:

- Fully automatic metadata generation
- Semi-automatic metadata generation
- Metadata extraction
- Metadata harvesting

○ Yes
○ No

Add Question Add Page

Edit Delete Copy/Move

13. Are you performing any evaluation or quality control of metadata created for digital resources? If so, please explain the process and criteria you use.

Add Question Add Page

6. Automatic Dublin Core Metadata Generation

Edit Page Delete Page Copy/Move Add Logic

In this section, we are specifically interested in your thoughts on the feasability and usefulness of automatic generation of Dublin Core metadata for digital resources (digital document-like objects).

The questions in this section list the official Dublin Core element set, which is defined at http://www.dublincore.org/documents/dces/.

Add Question Add Page

Edit Delete Copy/Move Add Logic

14. What is your level of knowledge and/or experience with Dublin Core? (Please check all that apply.)

□ I have heard of it, but am not familiar with the standard.

- □ I have read the standard and/or have had training on the standard, but have not worked with it.
- □ I have worked with the standard a little.
- □ I have worked with the standard extensively.
- □ I have been involved in the development of the standard.

Add Question | Add Page

Edit | Delete | Copy/Move

15. For each Dublin Core element, indicate the level of accuracy you would expect for resource metadata generated using *fully automatic* metadata generation techniques, based on your experience.

	Very Accurate	Moderately Accurate	Not Very Accurate
Title	○	○	○
Creator	○	○	○
Subject and Keywords	○	○	○
Description	○	○	○
Publisher	○	○	○
Contributor	○	○	○
Date	○	○	○
Type	○	○	○
Format	○	○	○
Identifier	○	○	○
Source	○	○	○
Language	○	○	○
Relation	○	○	○
Coverage	○	○	○
Rights	○	○	○

Add Question | Add Page

Edit | Delete | Copy/Move

16. Please share any general comments you have about your ratings in the previous question.

Add Question Add Page

Edit Delete Copy/Move

17. Assign what you think would be the appropriate level of automatic generation for each element in the list below.

	Fully automatic	Semi-automatic	Manual creation
Title	○	○	○
Creator	○	○	○
Subject and Keywords	○	○	○
Description	○	○	○
Publisher	○	○	○
Contributor	○	○	○
Date	○	○	○
Type	○	○	○
Format	○	○	○
Identifier	○	○	○
Source	○	○	○
Language	○	○	○
Relation	○	○	○
Coverage	○	○	○
Rights	○	○	○

Add Question Add Page

Edit Delete Copy/Move

18. Do you have any additional comments on your answers to the previous question?

Add Question Add Page

Edit Delete Copy/Move

19. Assume you have limited funds to spend on the development of an automatic metadata generation tool. How would you allocate your budget on the development of automatic techniques for the following elements?

Choose "High" if you would devote extensive funding to the element.
Choose "Medium" if you would devote moderate resources to it.
Choose "Low" if you would devote few resources to it.

	High	Medium	Low
Title	○	○	○
Creator	○	○	○
Subject and Keywords	○	○	○
Description	○	○	○
Publisher	○	○	○
Contributor	○	○	○
Date	○	○	○
Type	○	○	○
Format	○	○	○
Identifier	○	○	○
Source	○	○	○
Language	○	○	○
Relation	○	○	○
Coverage	○	○	○
Rights	○	○	○

Add Question | Add Page

Edit | Delete | Copy/Move

20. Do you have any additional comments on your answers to the previous question?

Add Question | Add Page

7. General Automatic Metadata Generation

Edit Page | Delete Page | Copy/Move | Add Logic

The questions in this section gather some of your thoughts and ideas about automatic metadata generation in general.

Add Question | Add Page

Edit | Delete | Copy/Move

21. How desirable would it be to integrate the following into an automatic metadata generation application?

	Very desirable	Somewhat desirable	Not desirable
Content guidelines (AACR2, standards, etc.)	○	○	○
Classification schemes, controlled vocabularies, thesauri, etc.	○	○	○
Examples of metadata records	○	○	○

Add Question | Add Page

Edit | Delete | Copy/Move

22. Are there other features you think would be desirable in an automatic metadata generation application?

Add Question Add Page

Edit Delete Copy/Move

23. Do you have any further comments on the previous two questions?

Add Question Add Page

Edit Delete Copy/Move Add Logic

24. How important is it for an automatic metadata generation tool to support the creation of metadata records for foreign-language resources?

Very important	Somewhat important	Not important
○	○	○

Add Question Add Page

Edit Delete Copy/Move Add Logic

25. How important is it for an automatic metadata generation tool to provide machine translation of metadata records into multiple languages?

Very important	Somewhat important	Not important
○	○	○

Add Question Add Page

Edit Delete Copy/Move Add Logic

26. How important is it for an automatic metadata generation tool to support the creation of metadata records for non-textual digital resources (e.g. multimedia)?

Very important	Somewhat important	Not important
○	○	○

Add Question Add Page

Edit Delete Copy/Move

27. Please provide any comments or concerns about automatic metadata creation for non-textual digital resources.

Add Question Add Page

Edit Delete Copy/Move Add Logic

28. What metadata creation workflow would you prefer?

○ Fully automatic

○ Initial metadata representation automatically generated, then edited by a person

○ A person enters some metadata, which is then automatically processed/encoded and/or augmented

○ A person creates all metadata

○ Other (please specify)

Add Question Add Page

Edit Delete Copy/Move

29. Please provide any observations, comments, concerns or feedback regarding this study or automatic metadata generation in general.

Add Question Add Page

8. Thank you!

Edit Page Delete Page Copy/Move Add Logic

Thank you for your time and input on completing this survey!

Click "Done" to submit your results.

Add Question Add Page

<< Back Preview