Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions

Jane Greenberg*, Kristina Spurgin and Abe Crystal

School of Information and Library Science,
University of North Carolina at Chapel Hill,
100 Manning Hall, CB #3360, Chapel Hill 27599-3360 NC, USA
Fax: 01 919 962 8011 E-mail: janeg@ils.unc.edu
E-mail: kristina@infomuse.net E-mail: abe@unc.edu
*Corresponding author

Abstract: This paper reports on the automatic metadata generation applications (AMeGA) project's metadata expert survey. Automatic metadata generation research is reviewed and the study's methods, key findings and conclusions are presented. Participants anticipate greater accuracy with automatic techniques for technical metadata (e.g., *ID*, *language*, and *format* metadata) compared to metadata requiring intellectual discretion (e.g., *subject* and *description* metadata). Support for implementing automatic techniques paralleled anticipated accuracy results. Metadata experts are in favour of using automatic techniques, although they are generally not in favour of eliminating human evaluation or production for the more intellectually demanding metadata. Results are incorporated into Version 1.0 of the Recommended Functionalities for automatic metadata generation applications (Appendix A).

Keywords: automatic metadata generation; metadata applications; Dublin core; metadata experts; AMeGA project; metadata functionalities.

Reference to this paper should be made as follows: Greenberg, J., Spurgin, K. and Crystal, A. (2006) 'Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions', *Int. J. Metadata, Semantics and Ontologies*, Vol. 1, No. 1, pp.3–20.

Biographical notes: Jane Greenberg is an Associate Professor, School of Information and Library Science, University of North Carolina at Chapel Hill (SILS/UNC-CH). She teaches and conducts research in the area of Metadata and Classification, and is Director of the SILS Metadata Research Center. She holds a MLS from Columbia University and a PhD from the School of Information Sciences, University of Pittsburgh.

Kristina Spurgin is a doctoral student at SILS/UNC-CH. She received her MLS at SUNY Albany in 2003. Her interests include personal information management, organisation of information, social aspects of information, and the role of metadata in all of these areas.

Abe Crystal received his AB (Economics) degree from Princeton University in 2000, and is currently a doctoral student at SILS/UNC-CH. His research interests are in the areas of information architecture, metadata creation and use, and human-computer interaction.

1 Introduction

Metadata can significantly improve resource discovery by helping search engines and people to discriminate relevant from nonrelevant documents during an information retrieval operation. Although the importance of metadata is evident, means for efficient and effective implementation are not. Metadata implementation is complex, due to the tremendous growth in digital resource repositories and the development of many different metadata standards. Among one of the most obvious challenges is the *metadata bottleneck* (Liddy et al., 2002). It is unrealistic to depend on traditional humanly generated metadata approaches, given the massive number of digital resources requiring metadata.

Addressing this challenge is a growing body of research on automatic metadata generation focusing on digital resource content (e.g., Han et al., 2003; Liddy et al., 2002; Takasu, 2003). Research in this area is important, although examination is generally limited to selected experimental domains. Automatic metadata generation is also taking place in the operational setting via application development. These tools are being used daily to produce metadata, although they do not fully incorporate experimental research findings. This research trend reveals a *disconnect* between experimental research and application development in the area of automatic metadata generation.

Metadata generation applications, it seems, could be greatly improved by integrating relevant experimental research findings and application development activities. One way to foster this connection is through greater consultation with metadata experts (e.g., professional

cataloguers, indexers, and other persons knowledgeable about metadata creation) during application development. Metadata experts are interested in and often aware of experimental research; they are well positioned to link the *research* and the *application development* communities. Metadata experts are also knowledgeable about important bibliographic control developments that ought to be incorporated into metadata applications because they can significantly improve metadata quality (e.g., authority control). Despite this obvious source of knowledge, there is little scientific evidence of metadata expert consultation during application development.

The Automatic Metadata Generation Applications (AMeGA) project (http://ils.unc.edu/mrc/amega.htm) at the School of Information and Library Science, University of North Carolina at Chapel Hill, addresses this shortcoming by gathering data on functionalities that metadata experts would like incorporated into automatic metadata generation applications. The AMeGA project is being conducted in conjunction with the Library of Congress Bibliographic Control Action Plan that is leading information centres in this new millennium (http://lcweb.loc.gov/catdir/bibcontrol/actionplan.pdf). The goal of the AMeGA project is to identify and recommend functionalities for automatic metadata generation applications, and, ultimately, to improve the *state of the art* of these tools.

This paper reports specifically on the AMeGA metadata expert survey. The paper is organised as follows: section two provides a brief overview of automatic metadata generation; section three reviews both experimental research and application developments; section four presents the study's underlying research objective; section five reviews the study's research design and procedures; section six presents the study's results; section seven provides a contextual discussion of the results; section eight presents the study's key findings and conclusions and identifies important research areas. This paper also presents Version 1.0 of the Recommended Functionalities for Automatic Metadata Generation Applications, developed via the AMeGA project, in Appendix A.

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 uses automatic indexing techniques to mine resource content and produce structured ('labelled') metadata for object representation (e.g., Jones and Paynter, 2002; Yilmazel et al., 2004). 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 (Anderson and Perez-Carball, 2001). 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

3.1 Experimental research and digital resource content

The awesome growth of digital resource repositories provides an abundance of digital 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 et al., 1999). For example, document genre can inform textual density that can be used to predict metadata extraction algorithm performance for certain types of documents (Greenberg, 2004b).

Document genres often dictate structure, including the placement of semistructured metadata (e.g., document 'title', 'author' and 'author affiliation' generally appear as content header information in research papers). Semistructured metadata is amenable to automatic metadata generation. In fact, vector analysis experiments exploiting document structure have been fairly successful (e.g., Han et al., 2003; Takasu, 2003). Han et al.'s (2003) research has focused on the semistructured metadata found in the content header of research papers. Their use of a Support Vector Machine (SVM) algorithm, including the use of 'word' and 'line' extraction, resulted in fairly high precision and recall ratios for metadata based document retrieval. They found that the SVM algorithm outperformed the Hidden Markov Model (HMM), which was employed for the same document set of research papers. Takasu (2003) used a Variable Hidden Markov Model (DVHMM) and syntactical rules to extract bibliographic attributes from a set of journals and transactions (conference proceedings) that had first been processed via Optical Character Recognition (OCR). Bibliographic references, existing as distinct documents, were also processed via OCR for the same resources, and an error pattern recognition algorithm was run against the metadata generated via the DVHMM.

Takasu concluded that the two approaches together, which take advantage of semistructured metadata, can reduce the cost of preparing data for rule based metadata generation systems.

3.1.2 Knowledge representation systems

Digital technology has greatly increased the electronic availability of thesauri, ontologies, classification schemes, authority files, and other knowledge representation systems. This development and the web's global framework have led to the construction of metadata registries specifically for open access to multiple knowledge representations systems. Registry examples can be found for:

- thesauri (e.g., Lutes, 1999)
- ontologies (knowledge system laboratory (KSL) ontology server, Stanford university: http://www-ksl-svc.stanford.edu:5915/doc/ ontology-server-projects.html)
- descriptive metadata schemes (SCHEMAS registry: http://www.schemas-forum.org/registry/; Dublin core metadata registry: http://dublincore.org/dcregistry/).

Various types of algorithms and mapping resource content to appropriate knowledge representation systems provide a means of automatic metadata generation research.

Patton et al. (2004) provide an example of research in this area through an automatic name authority control procedure that matches names found in document content with names recorded in the LC name authority file. Liddy et al. (2002) provide another example of research in this area, using a natural language processing algorithm and resource content to generate metadata according to 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. Shortcomings exist, however, in that, testing is generally limited to specific subject domains, resource types, resource formats, and metadata elements. Researchers recognise the limitations of algorithms developed for domain vocabulary however, and have begun to develop prototype tools to employ different ontologies for metadata generation (Hatala and Forth, 2003). More research is needed to determine which approaches would be broadly applicable in metadata applications.

3.2 Automatic metadata generation applications

Growing recognition of the importance of metadata helps to explain the development of tools known as *metadata generation applications*, which are tools designed specifically, and only, to output metadata records. These applications are primarily, although not exclusively, for digital resource content representation. A list of applications following the Dublin Core metadata standard is found at: http://www.dublincore.org/tools/. Content creation software (software used to create resource content, such as Microsoft Word or a web editor) and the cataloguing module of Integrated Library Systems (ILSs) also support metadata creation for digital resources, and include automatic functionalities to enhance and maintain metadata quality.[1]

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, 2003).

The increased availability of metadata generation applications is exciting because of the potential to vastly improve the efficiency and effectiveness of metadata production for digital 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 standardisation of access points) and element qualification (DCMI Metadata Terms, 2004), which can facilitate the production of high quality standardised metadata.
- Automatic techniques are rarely exploited. It seems that experimental research findings – specifically, the development of sophisticated automatic indexing algorithms focusing on resource content, semistructured metadata, and knowledge representation systems – have yet to be fully incorporated into the current automatic metadata generation applications. Moreover, a wide range of discipline-specific 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 Organisation (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 and Greenberg, 2005; Greenberg et al., 2003). However, there is little evidence of rigorous review of application usability.

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

4 Research objective

The AMeGA project was founded to identify and recommend functionalities for applications supporting automatic metadata generation for digital resources, and ultimately, to improve the state of the art for metadata applications. A variety of research techniques underlie the AMeGA project, and are outlined in the final report: http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf. The underlying objective of the research reported on in this paper was to identify the functionalities that metadata experts desire in automatic metadata generation applications. Metadata experts are knowledgeable about the range of important bibliographic control functions that facilitate the creation of high quality metadata, and their input is important to the design of more sophisticated and robust automatic metadata generation applications.

5 Methodology

The *survey* approach was used to identify system functionalities desirable for automatic metadata generation applications. The research was exploratory. The survey portion reported on in this paper was informed, in part, by the Consortium to Develop an Online Catalog (CONDOC, 1981). CONDOC was an ad hoc consortium formed in 1980, which conducted a survey in order to identify key features for online library catalogues, specifically for small to medium sized college and university libraries. Although the scope of AMeGA is much broader than this project, CONDOC's underlying rationale of *pooling expertise* because "collectively, the knowledge and skills of participants 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 in creating metadata or administering metadata/cataloguing activities. The survey gathered data on the participants and their metadata/cataloguing experience, participants' knowledge and opinions about automatic metadata generation following the *Dublin Core Metadata Element Set, Version 1.1: Reference Description* (http://dublincore.org/documents/dces/), and participants' opinions about automatic metadata generation and desired system functionalities.

The study was restricted to *digital document like objects* (DDLOs), defined as "primarily textual resource[s] … accessible through a web browser" (Greenberg, 2004a). DDLOs may contain images, sound, and nontextual formatting, but they must contain textual content (e.g., HTML/XHTML resources, Microsoft Word documents, Adobe Acrobat PDF documents). The restriction was implemented because of research resource constraints. The survey was designed using SurveyMonkey.com and included both structured and open ended questions.[2] 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 pilot tested prior to being officially launched.

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

Table 1 Electronic distribution for AMeGA recruitment message

Electronic mailing lists	*Blogs*
1 AutoCat (autocat@listserv.acsu.buffalo.edu)	1 Catalogablog (http://catalogablog.blogspot.com/2004_06_27_catalogablog_archive.html #108861938970165296)
2 METS listserv (mets@loc.gov)	2 Infomusings (http://www.infomuse.net/blog/archives/2004_06.html#000794)
3 Dublin core general listserv (dc-general@jiscmail.ac.uk)	3 Bibliolatry (http://www.bibliolatry.net/2004/07/meta.html).
4 Open archives initiative general interest list (oai-general-request@openarchives.org)	
5 CIC (Big Ten) academic libraries OAI list (oai-cic-l@listserv.uiuc.edu)	
6 CIC library metadata listserv (cic-lib-metadata@cic.net)	
7 Serialst (serialst@list.uvm.edu)	
8 OLAC listserv (olac@listserv.acsu.buffalo.edu).	

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 fora in addition to those listed in Table 1.

6 Data analysis

The metadata expert survey results reported on in this paper focuses on participants' knowledge and opinions about automatic metadata generation of Dublin Core metadata, participants' opinions about automatic metadata generation, and the functionalities that they would like to see incorporated into automatic metadata generation applications. Background information on participants is also given. Additionally, Version 1.0 of the Recommended Functionalities for Automatic Metadata Generation Applications, based on this research and other aspects of the AMeGA project, is presented in Appendix A.

6.1 Participant profiles

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 started the survey; approximately one third of these participants did not complete it, mainly because they found it was beyond the scope of their work experience. Research has shown that paper invites/paper surveys have a significantly higher response than both paper invite/web surveys and email invite/web surveys (Hayslett and Wildemuth, 2004). Even so, researchers have found that online surveys do yield a "higher response quality" than do self completion postal surveys and other offline methods (Gunter et al., 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 2. (Percentages for Tables hereafter do not all add up to exactly 100% because of rounding to the one point decimal). The largest proportion of participants providing information on their professional role were identified either as administrators/executives (51 participants, 29.5%) or cataloguers/metadata librarians (49 participants, 28.3%). Among the five persons identified as *other* were a Freedom of Information Officer, a consultant, scientific assistant, a person holding a master's degree, and a bioinformatician. The largest percentage of participants (70 participants, 40.7%) providing institutional affiliation information were active in academic library environment, although participants were also from college or university settings (beyond the library), government libraries, government agencies, nonprofit organisations, corporate libraries, corporations/companies, and public libraries. Three quarters of participants (161 participants, 75.2%) had three or more years of cataloguing and/or indexing experience. Table 3 summarises participants' years of experience involved in cataloguing/indexing. Finally, the majority of participants (192 participants, 90.1%) were involved in metadata generation of DDLOs. Many participants were also involved in other metadata activities, such as administration/supervision and record maintenance.

Table 2 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 3 Participants' cataloguing/indexing experience

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

$n = 214$.

6.2 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 4). With the exception of one participant who skipped this question, all of the participants had at least heard of the Dublin Core, and a little over three quarters of the participants (174 participants, 80.6%) had worked with the Dublin Core (Table 4, summation of the last four rows). Approximately one third of the participants (32.9%) had worked extensively with Dublin Core and thirteen participants were involved in the development of this metadata standard (summation of the last two rows).

Table 4 Dublin core knowledge/experience

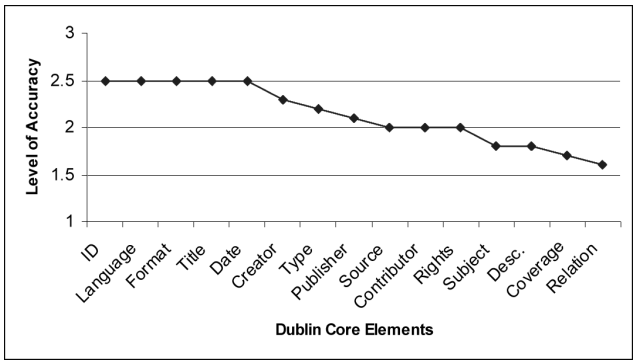
Knowledge/experience	*# of Participants*
Heard of DC, but not familiar with it	17(7.9%)
Read DC standard and/or have had DC training, but not worked with it	25(11.6%)
Worked w/DC a little	90(41.7%)
Worked with DC extensively	71(32.9%)
Involved in DC development	6(2.8%)
Worked with DC extensively and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

n = 216.

The feasibility/usefulness analysis focused on *expected accuracy* and *appropriate 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 1. 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).

Open ended comments on accuracy ratings were analysed and revealed a number of themes, the most prevalent of which was a perceived scepticism about the accuracy of automatic techniques for the generation of metadata requiring intellectual discretion (primarily *subject metadata*). A number of participants emphasised the value of controlled vocabulary and were sceptical 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 [without] 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 incorporate metadata extraction into the 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.

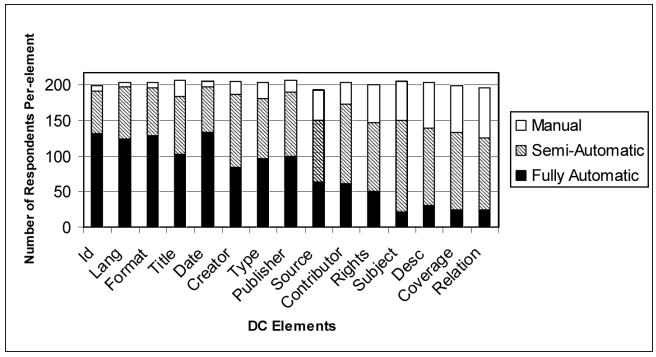
Figure 1 Expected accuracy level for automatic generation of Dublin core



n varied slightly per metadata element.

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 2. 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 is 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 2 parallel, to some degree, the accuracy expectancy results shown in Figure 1.

Figure 2 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 choose '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 [*sic*] 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 centred 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 sceptics – 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 sceptics, who argued for focusing resources on areas where full automation is feasible, particularly 'physical' fields such as *identifier* or *format*. Sceptics 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 sceptic; 'a total waste', said another. The sceptics 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.

6.3 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.

6.3.1 Automatic metadata generation for nontextual and foreign language resources

Although the survey emphasised 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 5 indicate that participants thought it was very important to develop automatic or semiautomatic methods of generating metadata for nontextual content, although many emphasised that this was a difficult task. Several respondents indicated that it may be even more important to develop automatic methods for nontextual resources because of the absence of text for indexing. One respondent said "There is only the metadata to rely on for resource discovery rather than full text indexing". Another added that automatic metadata generation for nontextual resources "will be more important in the long run than for textual resources since multimedia resources 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".[3]

Table 5 Automatic metadata generation for non-textual 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 likely 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 interrelate 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 semiautomatically 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 6. Most participants indicated that this function was 'somewhat important', with many more indicating that it was 'very important' (95 participants, 44.8%) as opposed to 'not important' (15 participants, 7.1%).

Table 6 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 participants' 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 [*sic*] foreign language functionality would be extremely useful.

Table 7 shows 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 (Table 7). 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%). Participant comments related 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".

Table 7 Automatic metadata generation for machine translation

Importance value	*Response rating*
Very important	48 (22.7%)
Somewhat important	112 (53.1%)
Not important	51 (24.2%)

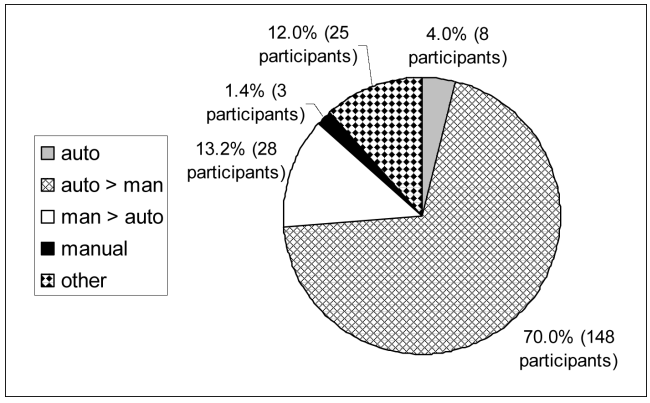
n = 211.

6.3.2 Additional functionalities for automatic metadata generation applications

The final portion of the survey examined workflow involving automatic metadata generation (or not), the integration of cataloguing examples and tools, 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. Responses to this question are shown in Figure 3. 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 three 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 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 3 Metadata generation workflow

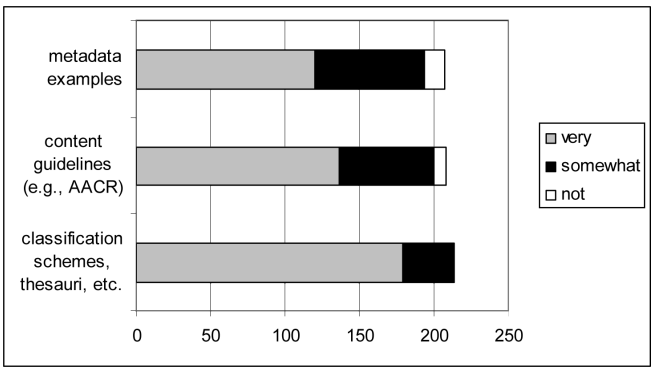


n = 212

Participants were asked about the desirability of integrating metadata/cataloguing examples, content creation guidelines,

and subject schemes into automatic metadata generation applications. These results are shown in Figure 4. Participants indicated it was generally 'very desirable' or 'somewhat desirable' to integrate any of these aids, showing the greatest support for subject schemes.

Figure 4 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 analysing the results include the following:

- system should integrate name authority files for personal and organisational names
- system should have the ability to import and export metadata in standard formats. Platform independence for formats is desired.
- system should support automatic and semiautomatic 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/organisational customisability 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 (see Appendix A).

7 Discussion of results

This study helped to identify functionalities desired in automatic metadata generation applications, while providing insight into the progress and the limitations in this area. The following discussion helps in interpreting the study's results by briefly addressing the participant population, and focusing primarily on the Dublin Core element rankings, automatic metadata generation challenges and desired system functionalities for automatic metadata generation applications.

7.1 Participants

The study confirmed that participants were metadata experts, with approximately three quarters of them (161 participants, 75.3%, Table 3) having three or more years of cataloguing/indexing experience and 90.1%, involved in metadata creation and/or other types of metadata activities (e.g., administration/supervision, maintenance/ quality control). Additionally, more than three quarters of the participants (174 participants, 80.6%, Table 4, summation of the last four rows) had worked with the Dublin Core. Participant experiences help validate their answers and the conclusions drawn.

7.2 Dublin core element evaluation

The portion of the survey focusing on Dublin Core asked participants to:

- rank each metadata element by anticipated accuracy when using automatic methods
- identify the appropriate application level for automatic metadata generation per element
- determine the appropriate resource allocation per metadata element.

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 1), 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 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 et al., 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 application designers incorporated experimental research developments that are 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 some types of technical metadata (e.g., *date modified* and *format*). Automatic metadata generation research experimenting with semistructured 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 semistructured 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 the Functional Requirements for Bibliographic Records (FRBR) (1998) and research in this area (e.g., Smiraglia and Leazer, 1999; Tillet 1991, 1992; Vellucci, 1997; Weinstein, 1998) may improve the overall score for this element.

Results for appropriate application levels for automatic metadata generation 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 2). Although semiautomatic processing was found to be fairly desirable across all elements, participants were not unanimously in favour of automatic processing for any single element. 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 survey question specific to the Dublin Core related 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. The wide range of metadata schemes being used in the digital world, however, confirms that metadata are needed for a variety of functions (e.g., administration, security, preservation, etc.). Individual metadata elements have been shown to be multifunctional (Greenberg, 2001). 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 prioritise 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.

7.3 Additional functionalities

The final section of the metadata expert survey provided insight into participants' opinions on automatic metadata generation for nontextual and foreign language resources and additional desired functionalities for automatic metadata generation applications. 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, teaching and learning. Given these circumstances, it is understandable that participants indicated that it was important and in some cases, critical, to support automatic metadata generation for nontextual resources (Table 5). Participants strongly voiced the need to improve metadata generation in this area and for the most part advocated automatic techniques for image metadata wherever feasible. 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 short, the baseline data on nontextual resources emphasise the need to incorporate such developments into metadata generation applications.

Participants were almost as in favour of support for automatic metadata generation for foreign language resources as they were for nontextual resources (Table 6). 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 7). Participants' responses presented in the results section related to practical matters. Another related reason for limited support may be standard cataloguing 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. The absence of definitive support suggests that this functionality is not currently a high priority for automatic metadata generation applications, although this opinion may change over time, 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 3) 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 and the strong desire to integrate metadata examples, content guidelines, and schemes into applications (Figure 4 and open ended responses). One exception is a very small percentage of participants (1.4%), who stressed the need for *fully manual* (*human controlled*) 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, 1988). It is recommended 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 Conclusions and future research directions

The metadata expert survey results presented in this paper help to identify recommended functionalities for automatic metadata generation applications. They also highlight important research needs in the area of automatic metadata generation.

Results indicate that metadata experts are in favour of using automatic metadata generation, particularly for metadata that can be created accurately and efficiently. However, participants were generally not in favour of eliminating human evaluation or production for the more intellectually demanding metadata (e.g., *subject* metadata). Nevertheless, the majority of participants agreed that automatic processes should be employed to aid humans creating metadata – including metadata requiring intellectual discretion. Two metadata functionalities which participants strongly favoured 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, etc.) 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. As noted above, the web's global framework has led to construction of metadata registries specifically for sharing knowledge representations such as thesauri, ontologies, and descriptive metadata schemes. Additionally, there are 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. Many of these developments also support Semantic web construction (Heery and Wagner, 2002). Finally, there are applications such as the Library of Congress' Catalogers Desktop (http://desktop.loc.gov/) that integrate many important bibliographic tools useful for automatic metadata generation. Automatic metadata generation applications providing access to useful resources, in an intelligent manner, will be able to greatly enhance metadata quality.

Although the research presented here is limited by the participant population, an emphasis on the Dublin Core and DDLOs, and the questions that defined the survey, the results highlight research areas important to development of automatic metadata generation applications. Research questions developed as a result of the metadata expert survey are:

- how should content standards (e.g., subject thesauri, name authority files, etc.) be integrated into metadata generation applications to support automatic metadata generation?
- 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 are the psychological and social impacts of automation on metadata creators?

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. Application designers must incorporate research findings if they are to build superior and more robust automatic metadata generation applications, and they must tap into the metadata experts as a knowledge source. Metadata experts are interested in and often aware of experimental research findings, and they are well positioned to link the *research* and the *application development* communities. Metadata experts are also knowledgeable about important bibliographic control developments that ought to be incorporated into metadata applications because they can significantly improve metadata quality. The Library of Congress (LC) recognises the strength of metadata experts and is well positioned to lead an effort to build better automatic metadata generation applications through the LC Bibliographic Control Action Plan (http://www.loc.gov/catdir/bibcontrol/actionplan.pdf). LC has demonstrated a commitment by supporting research on the identification of functionalities recommended in this report (Appendix A). Ultimately, increasing communication among all parties in the metadata enterprise will help us improve the current state of the art of metadata generation applications and gain better control of the rich world of digital information that defines the ever expanding World Wide Web.

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.

References

Anderson, J.D. and 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 and Management*, Vol. 37, No. 2, pp.231–254.

Bruce, T.R. and Hillmann, D.I. (2004) 'The continuum of metadata quality: defining, expressing, exploiting', in Hillmann, D.I. and Westbrooks, E.L. (Eds.): *Metadata in Practice,* ALA, Chicago, IL, pp.238–256.

CONDOC (1981) 'Revisiting CONDOC: a new look at the online catalog sponsored by the ala catalog use committee', *FTP Request: CONDOC Report,* Available at: listserv@listserv.buffalo.edu.

Crystal, A. and Greenberg, J. (2005) 'Usability of a Metadata Creation Application for Resource Authors', *Library and Information Science Research*, Vol. 27, No. 2, pp.177–189.

Cutter, C.A. (1904) *Rules for a Dictionary Catalog*, 4th ed., Government Printing Office, Washington, DC.

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

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

Greenberg, J. (2001) 'A quantitative categorical analysis of metadata elements in image applicable metadata schemas', *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 11, pp.917–914.

Greenberg, J. (2003) 'Metadata and the World Wide Web', in Drake, M.A. (Ed.): *Encyclopedia of Library and Information Science,* 2nd ed., Marcel Dekker Inc., New York, pp.1876–1888.

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,* Vol. 6, No. 4, pp.59–82.

Greenberg, J., Crystal, A., Robertson, W.D. and 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. and Williams, P. (2002) 'Online versus offline research: implications for evaluating digital media', *Aslib Proceedings,* Vol. 45, No. 4, pp.229–239.

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

Hatala, M. and Forth, S. (2003) 'System for computer-aided metadata creation', *The Twelfth International World Wide Web Conference* (WWW2003), May 20–24, Budapest.

Hayslett, M.M. and Wildemuth, B.W. (2004) 'Pixels or pencils? the relative effectiveness of Web-based versus paper surveys', *Library and Information Science Research*, Vol. 26, No. 1, pp.73–93.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K. and Yee, K.P. (2002) 'Finding the flow in website search', *Communications of the ACM*, Vol. 45, No. 9, pp.42–49.

Heery, R. and Wagner, H. (2002) 'a metadata registry for the semantic web', *D-Lib Magazine,* Vol. 8, No. 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*, Vol. 7, No. 4, pp.10–13.

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.

Johnson, F. (1995) 'Automatic abstracting research', *Library Review*, Vol. 44, No. 8, pp.28–36.

Jones, S. and Paynter, G.W. (2002) 'Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications', *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 8, pp.653–657.

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.

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

Losee, R. (2003) 'Adaptive organization of tabular data for display', *Journal of Digital Information,* Vol. 4, No. 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.

Metta Matters (2003) DCANZ and National Library of Australia and Dublin Core ANZ. http://www.nla.gov.au/meta/.

Nadkarni, P., Chen, R. and Brandt, C. (2001) 'UMLS concept indexing for production databases: a feasibility study', *Journal of the American Medical Information Association,* Vol. 8, No. 1, pp.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. and DiLauro, T. (2004) 'Toward a metadata generation framework: a case study at the John Hopkins university', *D-Lib Magazine*, Vol. 10, No. 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.

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

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

Tillet, B. (1991) 'A taxonomy of bibliographic relationships', *Library Resources and Technical Services*, Vol. 35, No. 2, pp.150–158.

Tillett, B.B. (1992) 'Bibliographic relationships: an empirical study of the LC machine-readable records', *Library Resources and Technical Services*, Vol. 36, No. 2, pp.162–188.

Toms, E., Campbell, D. and 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', *Proceedings of the 3rd ACM International Conference on Digital Libraries,* June 23–26, Pittsburgh, PA, ACM Press, New York, pp.254–263

Weintraub, K.D. (1979) 'The essential of the bibliographic record as discovered by research', *Library Resources and Technical Services*, Vol. 23, No. 4, pp.391–405.

Wilson, P. (1968) *Two Kinds of Power: An Essay on Bibliographical Control,* University of California Press, Berkeley, CA.

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

Yilmazel, O., Finneran, C.M. and Liddy, E.D. (2004) 'Metaextract: an NLP system to automatically assign metadata', Proceedings of *the 4th IEEE-CS Joint Conference on Digital Libraries*, June 7–11, Tuscon, AZ, ACM Press, New York, pp.241–242.

Zuboff, S. (1988) *In the Age of the Smart Machine: The Future of Work and Power*, Heinemann Professional, Oxford.

Notes

[1]A review of metadata generation features for both content creation software and ILSs is in the AMeGA project's final report: http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf.

[2]A copy of the metadata expert survey is found in Appendix A of the AMeGA project's final report: http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf.

[3]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 paper (Section 2).

[4]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 labelled in many different ways (e.g., attribute value schemes, knowledge representation schemes).

Appendix A:

Version 1.0 Recommended Functionalities for Automatic Metadata Generation Applications

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 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 inputs from the larger bibliographic control/metadata community.

The recommendations are organised 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 optimise 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 organisational 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 organisation 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 semistructured 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.[4] 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 Examples of topical domain/discipline content standards include subject classification and code systems, controlled vocabularies, and ontologies.

3.1.3.2 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. 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 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 organisation 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 organisations or people, or links 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 organisation may always require the same value/information for *rights* metadata.)

3.1.9.2 System should store default metadata values. (Example: An organisation 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* (emphasising *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 semistructured) 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 semistructured 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 customisable 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 authorised, 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 authorised *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 7.1.1–7.1.6.

"*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 organisation 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 7.1.1–7.1.6*: Each organisation 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 and 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.