Library & Information Science Research 31 (2009) 225–231

Contents lists available at ScienceDirect

Library & Information Science Research

Application of semi-automatic metadata generation in libraries: Types, tools, and techniques

Jung-ran Park *, Caimei Lu

College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Available online 21 July 2009

ABSTRACT

Analysis of a survey of the types and extent of tools and techniques related to semi-automatic metadata generation applied in real-world library settings indicates that practical applications in libraries seem to be at an incipient stage. More than half ($n = 149$, 52.5%) of the survey participants ($n = 285$) specify that semi-automatic metadata generation has not been utilized for metadata creation and management in their libraries. This figure becomes even higher when adding the response "don't know," constituting an additional 13.7%. The results of the survey also show that the semi-automatic metadata generation tools described by participants mostly concern metadata format conversion (38.6%) and metadata templates and forms (27%) for populating certain metadata values. Complex tools and the generation and extraction of metadata directly from the content and context of the digital objects are rarely applied in libraries. This indicates that more research is needed on the development of automatic metadata generation for semantic metadata in usable and practical settings.

1. Introduction

Metadata is essential for managing, organizing, and searching for information resources. The enormous volume of online and digital resources makes semi-automatic metadata generation a critical need. Semi-automatic metadata generation concerns metadata creation through partial reliance on software in combination with human process (Greenberg, Spurgin, & Crystal, 2005). There are some promising studies that exploit various methods for semi-automatic metadata generation for resources in different formats ranging from text to multimedia (Lu, Kahle, Wang, & Giles, 2008; Yang & Lee, 2005; Ying, Chitra, & Robert, 2005; Paynter, 2005; among others). Studies have explored automatic metadata generation not only for technical metadata (e.g., format) but also for semantic metadata by using classification schemes such as taxonomy and ontology. Metadata value is also captured through exploiting sources not only from the document (object) itself but also from context, including document usage, user profile, and metadata repositories.

2. Problem statement

As evidenced through rapidly growing digital repositories and web resources, there is a great need for semi-automatic metadata generation, especially considering the costly and complex operation of manual metadata creation. Most experimental research on automatic metadata generation claims promising results; however, feasibility and scalability have heretofore not been sufficiently evaluated in a realistic metadata creation environment. This study addresses this research gap by examining the current status of semi-automatic metadata generation in a real-world library setting.[1]

An assessment of tools and techniques used is important to help researchers and system developers identify areas in which more effort is required to improve the efficiency of metadata generation. It could, for example, lead to the development of more practicable automatic algorithm designs applicable to the metadata creation environment.

The results reported on in this paper may also guide information professionals in building a general framework that can be used for planning and evaluating the application of their own semi-automatic metadata generation vis-à-vis other institutions. For example, the study may provide information professionals with a picture of the current state of practice of semi-automatic metadata generation by informing them of the different types of functionalities and features and the most popular types of semi-automatic metadata generation tools. They can compare tools used by their own institutions with those used by others.

* Corresponding author.
E-mail addresses: jung-ran.park@ischool.drexel.edu (J. Park), cl389@glink.drexel.edu (C. Lu).

0740-8188/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.lisr.2009.05.002

[1] This study is a part of a three-year (2006–2009) project entitled "Metadata Creation and Metadata Quality Control across Digital Repositories: Evaluation of Current Practices," funded through the Institute of Museum and Library Services. The research reported on in this paper is an investigation of semi-automatic metadata generation in libraries across the United States.

The following research questions below are examined in this paper:

1. What is the current status of semi-automatic metadata generation in libraries?
2. What types of semi-automatic metadata generation techniques have been applied in libraries?

3. Literature review: semi-automatic metadata generation

As opposed to manual metadata creation, automatic metadata generation relies on machine methods to finish the metadata creation process. Greenberg (2004) distinguished between two methods of automatic metadata generation: metadata extraction and metadata harvesting. Metadata extraction in general employs automatic indexing and information retrieval techniques to generate structured metadata based on the original content of resources. On the other hand, metadata harvesting concerns a technique to automatically gather metadata from individual repositories in which metadata has been produced by automatic or manual approaches. The harvested metadata is stored in a central repository for future resource retrieval.

3.1. Descriptive and structural metadata extractions

The Automatic Metadata Generation Applications (AMeGA) project by Greenberg, Spurgin, and Crystal (2005) aimed to identify and recommend functionalities for applications supporting automatic metadata generation in the library/bibliographic control community. The project proposed two types of systems for automatic metadata generation: general content creation software together with specialized metadata generation applications. In content creation software, automatic techniques can be used to produce technical metadata, such as date_created, date_modified, and size (e.g., bytes). Automatic production of technical metadata is promising in the sense that this may increase the speed of metadata generation and facilitate more consistent metadata application.

The machine learning method has been widely applied to metadata extraction. Han, Giles, Manavoglu, and Zha (2003) developed a support vector machine classification-based method for metadata extraction from the HTML header part of research papers. They used a feature extraction technique based on domain-specific databases and an iterative line classification process based on contextual information.

Takasu (2003) presented a method for extracting bibliographic attributes from reference strings captured using optical character recognition (OCR). Based on this method, he also proposed an extended hidden Markov model called dual and variable length output hidden Markov model (DVHMM). A DVHMM can align a pair of strings and parse a string. Using this model Takasu discussed two approaches to bibliographic attribute extraction: reference alignment and reference parsing. In reference alignment, attribute values are associated with the bibliographic record to enrich the vocabulary of the bibliographic database. In reference parsing, attribute values are extracted from OCR-processed references for bibliographic matching. Experiments show that useful attribute values can be extracted from OCR-processed references.

For scanned scientific journal entries, Lu, Kahle, Wang, and Giles (2008) proposed a system that automatically generates structural and descriptive metadata. The generated metadata were identified at the volume, issue, and article level based on a set of extracted features. These include style features (e.g., alignment), structure and contextual features (e.g., paragraph beginning and line space), font features (e.g., character width and word height), and semantic and linguistic features (e.g., names and word count). Rule-based pattern matching and machine learning approaches were used in the process of automatic metadata generation.

At the volume level, journal title, volume number, and issue number may be extracted. At the issue level, issue number, starting and ending page number, and number of articles may be extracted. At the article level, automatically extracted metadata include article title, author name and volume and issue number. Experimental results showed that the proposed system is efficient in generating metadata with relatively high recall and precision.

The studies mentioned above all use document structure to generate descriptive and structural metadata. However, there is an inherent limitation of descriptive and structural metadata vis-à-vis satisfying the needs of resource organization and retrieval.

3.2. Semantic metadata extraction

Generating semantic metadata requires more intellectual discretion than does descriptive or structural metadata. Yang and Lee (2005) categorized semantic metadata creation (also known as semantic annotation) into two models. The first model, called ontology-driven semantic tagging, may be used to generate a set of semantic tags that describe the original content of the resource at different structural levels. The second model, called semantic metadata generation, is aimed at generating metadata that semantically describe the content of the annotated resource. A system following this second model may define its own ontology or adopt a predefined ontology.

Saini, Ronchetti, and Sona (2006) also used ontology for semantic metadata extraction for learning objects. The goal of their research was to classify a collection of learning objects to the most appropriate topic based on a given domain ontology. The collection of learning objects was first classified through an expectation maximization process using the naive Bayes classifier. The classifier was initialized based on the descriptive keywords and lexical terms used for labeling each node in the ontology. After classification, each learning object was automatically associated with semantic metadata.

For automatic semantic annotation, Yoldas and Nagypal (2006) also used ontology. Their method started with an information extraction step, which extracts named entities from the document text. The extracted named entities were transformed to the initial semantic metadata based on a given ontology. The initial metadata was further extended into ontology entities not explicitly stated in the document text by exploiting various ontology-based heuristic rules. Evaluation shows that the results of this method are greatly improved through application of the heuristic rules.

Ontology functions as an effective tool for automatic metadata generation; however, its content and structure need to be keep updated as knowledge evolves. As such, an ontology-independent approach for automatic semantic annotation is also explored in the literature. Yang and Lee (2005) proposed a semantic metadata extraction method that requires no predefined ontology. Their proposed method incorporates semantic annotation and ontology creation processes based on machine learning approaches. The collection of training web pages was first clustered by the self-organizing map algorithm that generates a feature map called the keyword cluster map (KCM). Following this step, a semantics extraction process was applied to the KCM to identify a set of keywords that label the main theme of each page. An ontology was then developed based on the extracted keywords and their relationships. The generated ontology was used for creating metadata for a web page.

Different from other ontology-dependent approaches, this approach creates its own ontology based on training data. Accordingly, the quality of the generated ontology and the metadata of the web pages to a great extent rely on the quality of the training data. In this sense, depending on the size and quality of the training data, the

Table 1
Electronic mailing lists for the survey.

Electronic mailing lists
1. AUTOCAT: AUTOCAT@LISTSERV.SYR.EDU
2. Dublin Core listserv: DC-LIBRARIES@JISCMAIL.AC.UK
3. Metadata librarians listserv: metadatalibrarians@lists.monarchos.com
4. Library and Information Technology Association listserv: lita-l@ala.org
5. OnLine Audiovisual Catalogers electronic discussion list: OLAC-LIST@LISTSERV.ACSU.BUFFALO.EDU
6. Subject Authority Cooperative Program listserv: SACOLIST@LISTSERV.LOC.GOV
7. SERIALST: SERIALST@LIST.UVM.EDU
8. Text Encoding Initiative listserv: TEI-L@LISTSERV.BROWN.EDU
9. Electronic Resources in Libraries listserv: ERIL-L@LISTSERV.BINGHAMTON.EDU
10. Encoded Archival Description listserv: EAD@LISTSERV.LOC.GOV

generated ontology may be primitive. Thus although this approach does not require a predefined ontology, a high-quality training dataset with sufficient size is necessary for this approach.

3.3. Metadata extraction systems

There are systems and frameworks extracting both semantic metadata and other descriptive and structural metadata. Paynter (2005) developed some of the first automatic metadata creation tools for a virtual library called INFOMINE. The library has a large collection of scholarly resources gathered from the Internet and includes automatically generated metadata for the collected resources using tools from the iVia Virtual Library Software package. The automatically created metadata includes not only descriptive metadata such as title and creator, but also some complex semantic metadata such as keyphrase and subject.

The tools may also be used to automatically classify resources according to Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH). The methods applied in these tools include syntactic processing and machine learning algorithms such as support vector machines and logistic regression. An important feature of this automatic metadata assignment system is that it incorporates a metadata evaluation tool to support an iterative development process. In this way, the system can be adjusted and improved based on the evaluation results on the quality of the generated metadata.

Paynter assigned and evaluated six required metadata fields: title, creator, description, keyphrase, LCSH, and INFOMINE category. Different metadata assignment methods and evaluation criteria were adopted for each metadata field. For instance, title metadata was created by harvesting the HTML tags (e.g., title, dc.title) and evaluated against three criteria: exact match accuracy, content-word precision, and recall. For more complicated semantic metadata, other approaches were used. For instance, to assign value for description, AutoAnnotator, which is based on sentence and paragraph scoring, was applied. The evaluation criteria for the description field include stemmed content-word precision and recall, in addition to unstemmed content-word precision and recall. The system may also assign LCSH automatically. However, this process is limited by the availability of expert-assigned training data. Overall, the work introduced a comprehensive system for automatic metadata generation and evaluation. The quality of metadata extracted by this system is evaluated by human experts and compared to the metadata created manually.

Metaextract is a system for metadata extraction in the domain of math and science education for grades K-12 (Yilmazel, Finneran, & Liddy, 2004). The system was designed to extract Dublin Core (DC) and Gateway to Educational Materials (GEM) metadata on both the item and collection levels using natural language processing techniques. Collection-level metadata is generated based on a collection-specific configuration. The item-level metadata is extracted from the content of educational documents using three extraction modules: eQuery, HTML-based modules, and a keyword generator module.

For learning objects, Cardinaels, Meire, and Duval (2005) identified four main categories of sources for metadata extraction: document content analysis, document context analysis, document usage, and composite document structure. In the proposed framework, learning object metadata was derived from two sources: the learning object itself and the context in which the learning object is used. The central component of the framework is the simple indexing interface (SII), which consists of two major groups of classes that generate the metadata: object-based indexers and context-based indexers. Based on this framework, an automatic metadata generator was designed and implemented for the Blackboard Learning Management System.

In terms of learning objects in both text and video formats, Ying, Chitra, and Robert (2005) presented the IBM Magic System. The system includes various content analytic modules for metadata generation: a) audiovisual analysis modules that recognize semantic sound categories and identify narrators and informative text segments; b) text analysis modules that extract title, keywords and summary from text documents; and c) a text categorizer that classifies a document according to a pregenerated taxonomy. This system can facilitate content reuse and repurposing, improve interoperability and engender more timely registration of content by course developers and authors.

There is a rapidly growing amount of multimedia data online for scientific research, learning, and education. Mechanisms for automatic metadata generation for multimedia resources are needed to facilitate data retrieval and management. Nontextual multimedia resources present special challenges in terms of content representation and metadata generation. Multimedia data contain an infinite amount of semantic information, making it impractical to index multimedia contents at once.

Griffioen, Yavatkar, and Adams (1996) introduced a system called Modeling Object-Oriented Data Semantics (MOODS), which identifies metadata dynamically for multimedia resources. The system does not identify and record all the semantic information contained in a multimedia object at once. Instead, meaningful metadata depending on certain domain and information needs are first extracted, indexed, and stored in the database. When a user issues a query, the system first returns the matching objects based on metadata initially stored in the database. If no matching objects are found, the system can run a processing engine to dynamically search for multimedia objects that may contain the semantic information in quest. The located objects are further processed to extract the semantic information. In addition, the system also allows users to define a knowledge base of semantic inference rules so that users may search multimedia data based on high-level semantic concepts. In this way, the MOODS system may be used to dynamically extract a wide range of semantic metadata from multimedia resources.

The approaches for generating structural metadata and basic descriptive metadata are relatively straightforward, while the systems for extracting semantic metadata such as subjects and keywords mostly rely on sophisticated natural language processing,

Table 2
Job titles of participants.

Job titles	Number of participants
Other	135 (44.6%)
Cataloger/cataloging librarian/catalog librarian	99 (32.7%)
Metadata librarian	29 (9.6%)
Catalog and metadata librarian	26 (8.6%)
Head, cataloging	26 (8.6%)
Electronic resources cataloger	17 (5.6%)
Cataloging coordinator	15 (5.0%)
Head, cataloging and metadata services	15 (5.0%)

$N = 227$.

Table 3
Professional activities specified in "other" category.

Professional activities	Number of participants
Cataloging and metadata creation	31 (10.2%)
Digital project management	23 (7.6%)
Technical services	17 (5.6%)
Archiving	16 (5.3%)
Electronic resources and serials management	6 (2.0%)
Library system Administration Other	6 (2.0%)

$N = 99$.

machine learning, and text mining techniques. In general, the metadata extraction systems introduced above achieve satisfactory performance. However, the quality of automatically extracted metadata is questionable.

The systems reviewed above mostly rely on human evaluation, which limits the scalability of metadata evaluation. If the value of a metadata element generated by a system is not accurate, it is further refined manually or through other semi-automatic processes. Metadata evaluation should be an important component of an automatic metadata generation system. Methods and metrics for effective metadata evaluation may enable system designers to detect the limitations of the system in terms of quality metadata generation and thereby improve the system performance.

As shown, various methods and techniques for automatic metadata generation have been explored for resources ranging from text to multimedia. Those methods and techniques have been applied to generate not only technical metadata but also semantic metadata. In addition to standard sources such as object content, metadata value is captured through exploitation of various sources encompassing document usage, user profile and metadata record assemblies, and domain ontologies. However, evaluation of the feasibility and scalability of research results has not been studied sufficiently in a realistic metadata creation environment. To address this research gap, in the following sections, the extent and types of tools and techniques of semi-automatic metadata generation in a real-world library setting will be examined.

4. Research procedures

A web survey was designed using the online survey tool WebSurveyor (now Vovici) and included both structured and open-ended questions. The survey was extensively reviewed by members of the advisory board (a group of three experts in the field) and pilot-tested prior to being officially launched. Participants were recruited through survey invitation messages and subsequent reminders to the electronic mailing lists of communities of metadata and cataloging professionals (Table 1).

Individual invitations were also issued, and flyers were distributed at selected metadata and cataloguing sessions during the 2008 American Library Association (ALA) midwinter conference held in Philadelphia.

Table 4
Participants' job responsibilities.

Job responsibilities	Number of participants (%)
General cataloging (e.g., descriptive and subject cataloging)	171 (56.4)
Metadata creation and management	153 (50.5)
Authority control	147 (48.5)
Non-print cataloging (e.g., microform, music scores, photographs, video-recordings)	133 (43.9)
Special material cataloging (e.g., rare books, foreign language materials, government documents)	126 (41.6)
Digital project management	101 (33.3)
Electronic resource management	62 (20.5)
Integrated library system management	59 (19.5)
Other	51 (16.8)

$N = 303$.

Table 5
Use of semi-automatic metadata generation tools.

Response category	Response rating
Yes	97 (34.0%)
No	149 (52.5%)
Don't know	39 (13.7%)

$N = 285$.

5. Results

During the 62-day period from August 6, 2008, to October 6, 2008, a total of 303 completed responses were received by the survey management system. The survey attracted a large number of initial participants ($n = 1371$). Among the initial participants who started the survey, a total of 303 (22.1%) people completed it. Incompletion of the survey may stem from the fact that the survey subject matter may have been outside the scope of the participants' job responsibilities. The length of the survey may also have been a factor in the incompletion rate.

5.1. Participant profiles

Less than half ($n = 121$) of the survey participants provided institutional information. This is mostly due to the fact that the question was optional, following a suggestion from the Institutional Review Board at Drexel University. According to the institutional background from 121 responses, the majority of participants ($n = 91$) are from academic libraries; followed by participants ($n = 21$) from public libraries and from other institutions ($n = 9$).

Table 2 illustrates the job titles of the participants.

The largest proportion of participants ($n = 135$, 44.6%) chose the "other" category instead of choosing one of the given job titles. Following the survey request, participants who chose the "other" category further specified their job titles. These job titles are classified based on common characteristics of the professional activities. This, in part, was informed by the matrices of job responsibilities developed by Park, Lu and Marion (2009) and Park and Lu (2009) for job description analyses of cataloging and metadata professionals.

Most of the job titles given as other are associated with one of the professional activities listed in Table 3.

As can be seen in the table above, cataloging and metadata creation and management comprise the bulk of the job activities of survey participants. The question related to job responsibilities further shows this important characteristic of the participant profile. Table 4 further illustrates this.

Job responsibilities in Table 4 show that survey participants engage with the core activities of cataloging such as descriptive and subject cataloging, metadata creation and management, authority control, nonprint and special material cataloging, electronic resource and digital project management and integrated library system management (see Park & Lu, 2009; Park, Lu, & Marion, 2009, for details on job

Table 6
Types of semi-automatic metadata generation tools.

Types	Response rating
Metadata format conversion	38 (38.6%)
Templates and editors for metadata creation	26 (27.0%)
Automatic metadata creation	16 (16.7%)
Library system for bibliographic and authority control	15 (15.6%)
Metadata harvesting and importing tools	8 (8.3%)

$N = 96$.

Table 7
Metadata format conversion.

- MarcEdit (mentioned by 19 participants)
- Extensible Stylesheet Language Transformations (XSLT) (mentioned by 6 participants)
- Ad hoc scripting and conversion tools (mentioned by 2 participants)
- Local program for converting from Qualified DC to TEI
- Tools offered by LC for MARC to Metadata Object Description Schema (MODS) creation
- Tools for converting Filemaker Pro databases and other databases to MODS
- Automated crosswalks from MARC to DC, MODS, and Metadata Authority Description Schema (MADS) developed at NUL
- Using DC to MARC crosswalk to get ETD's from our institutional repository into our catalog
- Tools for transforming MARC to Extensible Markup Language (XML)
- Scripts to convert data from MARC records into MODS; to convert data from local database into Visual Resources Association (VRA)
- Proprietary (AMCon Research) software, converts RTF to Encoded Archival Description (EAD), XML Document Type Definitions (DTD)
- Locally created crosswalk programming

responsibilities and competencies of cataloging and metadata professionals).

The "other" category encompasses activities such as collection development, department/personnel management and supervision, public services, acquisitions, preservation and conservation, digital library projects design and development, metadata schema, and system development and archival processing. As discussed in related studies (Park & Lu, 2009; Park, Lu, & Marion, 2009), there are a wide array of responsibilities including collection development and public services incumbent on cataloging and metadata professionals.

In terms of work experience, more than half of the respondents ($n = 170$, 58%) reported over five years of experience in cataloging and metadata creation: 6 to 15 years ($n = 92$, 31.6%); over 16 years ($n = 78$, 26.4%). Approximately one third of respondents ($n = 102$, 34.5%) reported that they have worked as a cataloging/metadata librarian for 1 to 5 years. The rest ($n = 24$, 8.1%) reported experience of less than a year.

5.2. Application of semi-automatic metadata generation

In order to understand the current status of the application of semi-automatic metadata generation, the participants were first questioned on whether they and their fellow catalogers/metadata librarians use any tools for semi-automatic metadata generation. Table 5 shows the percentage of participants ($n = 285$) answering "yes," "no," or "don't know" to this question.

Table 8
Templates and editors for metadata creation.

- CONTENTdm's template creator (technical metadata) (mentioned by 5 participants)
- Template creator (allows metadata creator to populate default values for similar items) (mentioned by 4 participants)
- Macro Express (a Windows-based application that allows automation of routine functions, such as filling out web forms) (mentioned by 3 participants)
- NoteTab (NoteTab XML editor scripting language and clip libraries create a straightforward data-entry process to eliminate typos and other inconsistencies in XML code) (mentioned by 2 participants)
- oXygen (XML editor) (mentioned by 2 participants)
- Auto-lookups in vocabularies like TGN
- Homegrown extraction of MIME Required fields in the metadata generator online form
- Templates that already have certain local information for theses and locally produced AV materials
- Media Management Tool automatically generates some date fields and administrative fields. It also allows for filling in some metadata fields for a whole batch
- Pre-populated forms
- The Data Creation Template adds the level of description and date of description
- Athena easy entry (Athena widget set)
- Use templates Interwoven TeamSite software to create minimal, basic DC metadata for e-resources on website

Table 9
Automatic metadata creation tools.

- JHOVE for generating technical metadata (mentioned by 3 participants)
- iVia tool from Data Fountains at University of California Riverside (mentioned by 2 participants)
- Interwoven MetaTagger
- Author contact information is generated from our personal information database
- Madison Digital Image Database (MDID) to generate DC
- Automatic generation of Metadata Encoding and Transmission Standard (METS) and Extensible Text Framework (XTF)
- Digitization equipment (scanners, cameras, etc.) creates technical preservation metadata
- Homegrown unqualified DC generator
- Technical information about the digital file is automatically generated by the software we use
- Screen scraping of data from websites and then batch uploading into DSpace

As shown above, one third of participants ($n = 97$, 34.0%) confirmed that they and their fellow catalogers and metadata librarians use tools for semi-automatic metadata generation. Over half the participants ($n = 149$, 52.5%) indicated that they do not use any tools for semi-automation metadata generation.

To examine the types of automatic metadata application tools used in libraries, the participants were further asked to describe tools and any applications relating to semi-automatic metadata generation. One third of the participants ($n = 96$, 34%) provided meaningful information. For the purposes of the study, the descriptions of tools were classified into the following five categories based on common characteristics. When there is an overlap, the most prominent features and characteristics of the description are used for classification. The categories are:

1. Metadata format conversion.
2. Templates and editors for metadata creation.
3. Automatic metadata creation.
4. Library systems for bibliographic and authority control.
5. Metadata harvesting and importing tools.

Table 6 illustrates the frequency of use of the above-mentioned categories.

As shown, the most frequently mentioned semi-automatic metadata generation tools concern converting metadata formats ($n = 38$, 38.6%). MarcEdit is the most widely used tool in this category. Some participants reported homegrown tools (e.g., "locally created crosswalk programming"). Table 7 illustrates some of the metadata format conversion tools described by survey participants. (Tools mentioned by more than two participants are indicated.)

Following metadata format conversion, templates and editors ($n = 26$, 27%) appear to be the most widely used tools among surveyed participants. Templates for populating technical metadata is the most frequently mentioned tool. For instance, CONTENTdm, a digital collection management software package, allows the creation of a template for generating technical metadata. This is the most frequently mentioned tool in this category. Some of the templates and editors described by survey participants are illustrated in (Table 8). (Tools mentioned by more than two participants are indicated.)

Table 10
Integrated library systems and authority control.

- OCLC Connexion
- OCLC Cataloging System
- Use an OCLC macro to generate authority name headings
- IPAC generates authority files automatically based on the bibliographic files
- LMS (Library Management System) for authority control
- Programs that generate authority records from various programs for NAR & and LCSH submission to LC
- Aleph 500

Table 11
Metadata harvesting and importing tools.

• National Library of New Zealand metadata harvester (for technical and some preservation metadata)
• Our institutional repository software includes a batch loader that we can use to load metadata from Medline records for journal articles
• Import of existing MARC records
• We use a system based on DSpace and MS Office apps to pull metadata from groups of submitted digital materials and feed metadata into our instance of DSpace

A small number of participants ($n = 16$, 16.7%) described tools for automatic metadata creation. Some of these tools include: Harvard University's JSTOR Harvard Object Validation Environment (JHOVE), iVia (a component of the Data Fountains suite from University of California Riverside), and Interwoven MetaTagger. A few participants also described homegrown tools (e.g., "Homegrown unqualified Dublin Core generator") for automatic metadata creation. Table 9 illustrates some of the comments and tools described by the participants in this category. (Tools mentioned by more than two participants are indicated.)

Some participants ($n = 15$, 15.6%) described the cataloging module of the Integrated Library Systems (ILS) as a type of automatic metadata generation tool. These include Online Computer Library Center (OCLC) Connexion, Internet/Intranet Public Access Catalogue (IPAC), and Aleph 500, among others. The ILS cataloging module may provide high-quality authority and bibliographic control based on its automatic capabilities (Greenberg, Spurgin & Crystal, 2005).

Some of the comments and ILSs described by survey participants are illustrated in (Table 10).

Metadata harvesting and importing tools ($n = 8$, 8.3%), the last category reported on semi-automatic metadata generation, are used for gathering and importing metadata records between repositories. For instance, one participant described the National Library of New Zealand metadata harvester for gathering technical and some preservation metadata for his/her institution. Other participants noted the type of metadata they import from Medline and Machine Readable Cataloging (MARC) records.

Table 11 illustrates some of the tools in this category and the comments of participants.

6. Discussion

Data analysis shows that slightly over half ($n = 149$, 52.5%) of the survey participants specify that semi-automatic metadata generation has not been used for metadata creation and management in their libraries. This figure becomes even higher when including the response "don't know," which constitutes 13.7%. This indicates that semi-automatic metadata generation in libraries has not yet been fully exploited. The results of the survey also show that semi-automatic metadata generation tools described by participants mostly concern metadata format conversion (38.6%).

Metadata creation templates and editors (27%), used for converting metadata records from one metadata standard to another, transform the format of metadata records rather than generate new ones. For instance, MarcEdit is used for converting MARC records to non-MARC metadata records. It is inevitable to see some degree of semantic loss of certain data elements during metadata format conversion through the crosswalk. This is mostly owing to a conceptual mismatch between source and target metadata standards (see Park, 2002, for details). The most frequently mentioned metadata standards for data format conversion include DC, Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), Text Encoding Initiative (TEI), and Visual Resources Association (VRA). It is worthwhile to note that, as discussed in the studies by Park, Lu, and Marion (2009) and Park and Lu (2009), those metadata standards also appear the most frequently in job announcements targeting catalogers and metadata professionals.

Metadata creation templates and editors are in general used for facilitating manual metadata creation. Some of the common features and functions of templates and editors include populating default values of certain metadata elements for a whole batch, eliminating typos and inconsistencies of data inputs of certain data elements, and auto-lookups of controlled vocabularies.

As discussed in Park (2009), there are a variety of metadata editors (see theDublin Core Metadata Initiative (DCMI) (1995–2009), http://dublincore.org/tools). For instance, the DC metadata editor DC-Dot (http://www.ukoln.ac.uk/metadata/dcdot/) enables automatic DC metadata generation for certain data elements and allows the generated metadata to be edited. The results of the study by Greenberg, Spurgin, and Crystal (2005) indicated that a simple one-page web template with textual guidance and selective use of features such as drop-down menus and pop-up windows may facilitate quality metadata generation.

The data analysis in the current study shows that experimental research findings have not yet been fully integrated into application of semi-automatic metadata generation in library settings. This result is consonant with the study by Greenberg, Spurgin, and Crystal (2006). They pointed out a disconnect existing between experimental research and application development in the area of automatic metadata generation. The primary reason given for the disconnection is that most research findings are only possible in an experimental environment. Specifically, most experimental research is limited to specific domains, resource types and formats, and metadata elements. When the applications developed under an experimental environment are applied in a practical setting, their usability and effectiveness tend to be in doubt. Even in an experimental environment, the applications cannot achieve total accuracy.

The tools described by the participants in the current study are mostly used for creating technical metadata and simple descriptive metadata. Automatic metadata creation tools are designed to generate metadata automatically based on the content and context of the digital objects or other sources. As discussed in the literature review section, experimental studies explored automatic metadata generation not only for technical metadata (e.g., format, date) but also for semantic metadata (e.g., subject).

However, as stated earlier, there is an inevitable limitation of descriptive and structural metadata in terms of satisfying the needs of resource organization and retrieval. Tools for semantic metadata generation are likewise rarely used in libraries. Some participants mentioned that such tools are still under development. For instance, one participant stated, "For a research and development project, we are working with a keyword extraction tool to automatically suggest topic terms (uncontrolled)."

7. Conclusion

The enormous volume of online and digital resources makes semi-automatic metadata generation a critical need. As shown by current studies, various methods and sources have been explored for automatic metadata generation. The wide range of metadata types encompassing technical (e.g., format, date), descriptive (e.g., title) and semantic metadata are generated through a variety of methods and sources. Metadata value is captured through exploiting sources not only from the document (object) itself but also from its context including document usage, user profile and metadata repositories. Existing classification schemes encompassing LCC and LCSH and other domain-specific taxonomies and ontologies are also utilized in the application of semi-automatic metadata generation.

In this study, the extent and types of tools and techniques related to semi-automatic metadata generation applied in a real-world setting (i.e., libraries) were examined. In contrast to the active

research and promising results obtained from some experimental applications for automatic metadata generation, practical applications of semi-automatic metadata generation in libraries seem to be at an incipient stage. As Greenberg, Spurgin, and Crystal (2006) noted, the development of sophisticated algorithms and techniques for automatic indexing has not yet been fully incorporated into automatic metadata generation applications. The survey participants provided detailed information on metadata elements generated through automatic metadata creation tools. However, these elements comprise mostly technical and simple descriptive metadata. Complex tools and the generation and extraction of metadata directly from the content and context of the digital objects are rarely applied in libraries.

The major limitation of this study derives from the participant population; thus, there is no attempt to generalize the findings. However, results indicate an impending need to exploit promising findings from experimental studies to improve the efficiency of metadata generation in a real-world library setting. As evidenced through rapidly growing digital repositories and web resources, semi-automatic metadata generation is becoming ever more critical. It is essential for system designers to apply experimental research findings and the expert knowledge of metadata professionals to their work on semi-automatic metadata generation (Greenberg, Spurgin, and Crystal, 2006). The results of this study also indicate that more research is needed on the development of automatic metadata generation for semantic metadata in usable and practical settings.

Acknowledgments

This study is supported through a research award from the Institute of Museum and Library Services. We would like to express our appreciation to the editors and reviewers for their invaluable comments.

References

Cardinaels, K., Meire, M., & Duval, E. (2005, May). *Automating metadata generation: The simple indexing interface*. Paper presented at the 14th International World Wide Web Conference, Chiba, Japan. Retrieved from http://www2005.org/cdrom/docs/p548.pdf

Dublin Core Metadata Initiative. (1995–2009). *Tools and software*. Retrieved from http://dublincore.org/tools

Greenberg, J. (2004). Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, *6*(4), 58—92.

Greenberg, J., Spurgin, K., & Crystal, A. (2005). *Final report for the AMeGA (Automatic Metadata Generation Applications) project*. Chapel Hill: University of North Carolina School of Information and Library Science. Retrieved from http://dlist.sir.arizona.edu/878/01/lc_amega_final_report.pdf

Greenberg, J., Spurgin, K., & Crystal, A. (2006). Functionalities for automatic metadata generation applications: A survey of metadata experts' opinions. *International Journal of Metadata, Semantics and Ontologies*, *1*(1), 3—20.

Griffioen, J., Yavatkar, R., & Adams, R. (1996, April). *Automatic and dynamic identification of metadata in multimedia*. Paper presented at the 1st IEEE Metadata Conference, Silver Springs, MD. Retrieved from http://www.dcs.uky.edu/~moods/metadata.ps

Han, H., Giles, C.L., Manavoglu, E., & Zha, H. (2003, May). *Automatic document metadata extraction using support vector machines*. Paper presented at the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, Houston, TX. Retrieved from http://clgiles.ist.psu.edu/papers/JCDL-2003-automata-metdata.pdf

Lu, X., Kahle, B., Wang, J.Z., & Giles, C. L. (2008, June). *A metadata generation system for scanned scientific volumes*. Paper presented at the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, Pittsburgh, PA. Retrieved from http://www.cse.psu.edu/~xlu/publications/jcdl08-lu.pdf

Park, J. r. (2002). Hindrances in semantic mapping among metadata schemes: A linguistic perspective. *Journal of Internet Cataloging*, *5*(3), 59—79.

Park, J. r. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging and Classification Quarterly*, *47*(3), 213—228.

Park, J. r., & Lu, C. (2009). Metadata professionals: Roles and competencies as reflected in job announcements, 2003–2006. *Cataloging and Classification Quarterly*, *47*(2), 145—160.

Park, J. r., Lu, C., & Marion, L. (2009). Cataloging professionals in the digital environment: A content analysis of job descriptions. *Journal of the American Society for Information Science and Technology*, *60*(4), 844—857.

Paynter, G. W. (2005, June). *Developing practical automatic metadata assignment and evaluation tools for internet resources*. Paper presented at the 5th ACM/IEEE-CS Joint Conference on Digital libraries, Denver, CO. Retrieved from http://ivia.ucr.edu/projects/publications/Paynter-2005-JCDL-Metadata-Assignment.pdf

Saini, P. S., Ronchetti, M., & Sona, D. (2006). Automatic generation of metadata for learning objects. *Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies* (pp. 275—279). IEEE Computer Society: Washington, DC.

Takasu, A. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 49—60). Washington DC: IEEE Computer Society.

Yang, H. C., & Lee, C. H. (2005). Automatic metadata generation for web pages using a text mining approach. *Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration* (pp. 186—194). Washington, DC: IEEE Computer Society.

Yilmazel, O., Finneran, C.M., & Liddy, E. D. (2004, June). *MetaExtract: An NLP system to automatically assign metadata*. Paper presented at the 4th ACM/IEEE-CS joint conference on Digital libraries, Tucson, AZ. Retrieved from http://www.cnlp.org/publications/p237-yilmazel.pdf

Ying, L., Chitra, D., & Robert, F. (2005). Creating MAGIC: System for generating learning object metadata for instructional content. *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 367—370). New York: Association for Computing Machinery.

Yoldas, U., & Nagypal, G. (2006, October–November). *Ontology supported automatic generation of high-quality semantic metadata*. Paper presented at the 5th International Conference on Ontologies, DataBases, and Applications of Semantics, Montpellier, France. Retrieved from http://www.imagination-project.org/upload/Nagypal-Yoldas-publication.pdf