Metadata Creation Methods: A Study

Tanmay Mondal
Resources Management Group
Indira Gandhi Centre for Atomic Research
Kalpakkam-603102

N. Madurai Meenachi
Resources Management Group
Indira Gandhi Centre for Atomic Research
Kalpakkam-603102
Email:meenachi@igcar.gov.in

M. Sai Baba
Resources Management Group
Indira Gandhi Centre for Atomic Research
Kalpakkam-603102

Abstract

Metadata makes information meaningful. Creation of metadata records is a continuous and sequential task. Assignment of metadata values manually is time consuming as it depends on theskill(s) of the individuals involved. Metadata extraction can be availed with the help of machine enabled methods. Semi-automatic metadata creation is a reliable method that is being used effectively for resource description. Automatic metadata extraction tools correctly retrieve various technical values from objects. This paper discusses about different metadata creation methods available in the literature.

Keywords

metadata extraction, automatic metadata creation, machine enabled methods

Electronic access
The journal is available at www.jalis.in

J·A·L·I·S
Journal of Advances in Library and Information Science
ISSN: 2277-2219 Vol. 7. No.1. 2018. pp.177-182

INTRODUCTION

Metadata is everywhere. Effective information management depends on the metadata which is considered as the backbone for successful information dissemination. Metadata aidsin how a resource can be discovered, located and accessed. As per ISO (“ISO/TS 23081-2:2007”, n.d.), metadata is defined as data describing the context, content and structure of records and their management through time. According to Berners-Lee (1977), metadata is machine understandable information about web resources or other things. The goal of the metadata is to make information in structured form, enabling the machines as well as humans to identify, discover and use. Creation of metadata record is an important requirement for effective information management. Traditionally, library cataloguers, indexers, experts, authors produce metadata records manually. Different classes of persons can create metadata records using various tools and techniques. Greenberg (2003) classified metadata producers as:

i) Professional metadata creators
ii) Technical metadata creators
iii) Content creators
iv) Subject or community creators

Catalogue cards in library are examples of metadata records. Information becoming available more and more in the digital form necessitates adopting newer approaches for metadata creation and it is desirable to have machine enabled methods. Values for metadata can be assigned either manually or extracted automatically. Two methods for obtaining metadata have been defined by Tonkin& Muller (2008) which are: as:

i) Mechanically extracted from pre-print
ii) Manually by the author or digital librarian

Metadata records created by different entities can be broadly classified as manual, semi-automatic and automatic. Guy, Powell & Day (2004) categorized metadata creation process as:

i) Automated
ii) Automated and improved by the author of the document
iii) Automated and improved by an author and by an information specialist
iv) Created manually by an author and improved by an information specialist

v) Created by an information specialist

Automated metadata extraction improves quality of metadata records. Misra, Chen &Thoma (2009) suggested a cost-effective method that can identify and extract metadata in an automated way that requires machine learning and string pattern search techniques.

2. Metadata Elements

Collection of information referred to metadata records that contain a set of metadata fields and their values. Elements or fields in metadata records contain different values that distinguish similar items from each other. Source of metadata values can be classified as internal and external. Metadata may be embedded either in the resource itself or in a record separated from the item. Content generation software (e.g., text processing software) automatically generates many values at the time of content creation process. Examples of internal metadata are: date of creation of the document, file format etc. External metadata (for content description) describes contextual information about the content and it includes: title, author, keywords, journal name, conference name and place, abstract of the article, affiliation of the author(i.e., university, institute, company etc.), acknowledgement, preface, references. Publication related aspects like publication note, related work, research projects, data sets are some other important information in metadata records. Many web based information providers (e.g., CiteseerX) extract specific part of any document. Citation, table, figure, algorithm, acknowledgment values are also considered in metadata records (Williams et al., 2014). There are more than 17000 unique metadata fields reported by Ardö (2010).

3. Manual Metadata Creation

Manual assignment of metadata values is a reliable method to describe information resources. Manual metadata generation requires trained professional. Form based editor or template based fields are used in manual metadata generation. Understanding various properties of the information object(s) is the first task in this method. Values can be provided either as a free-text or chosen from controlled vocabularies. Usually, experts or library staffs assign values following a given set of guidelines or rules. Two types of metadata namely a *priori* (expert-created metadata like traditional) and *post-hoc* (socially-constructed metadata like tagging) have been identified by Alemu& Stevens (2015). Dorbeva, Kim & Ross (2013) proposed some guidelines for manual metadata creation that follows as:

i) Visual scanning of the document (identification the objects’ nature i.e., text or video or any other object)
ii) Mental analysis to identify metadata types and their values (differentiation of values i.e., 'date' value is different from 'title' value)
iii) Entry of value(s) in correct form (correct values assignment corresponding to metadata fields)

Huge numbers of records have already been described manually. Most of the library activities like cataloguing, users’ registration, keywords’ assignments, etc. take manually assigned values following a set of rules. But it is a time consuming process. It is often seen that managing various digital objects manually is not consistent from organization to organization. Maintaining quality and consistency of metadata records are some concerns as it is person dependent. Manually created systems are quite useful but require a lot of initial effort to create and difficult to maintain (Chen &Dumais, 2000). Manso, Wachowicz&Bernabé (2010) described manual metadata creation as a monotonous, harsh and resource consuming assignment. Incomplete or simply put metadata values return inconsistency in records that retrieve poor search results and eventually reduce the usability of the collections. It is reported(“BA Insight”, n.d.) that in the normal course, majority of the values are tagged with the first term from the dropdown list picked as the quickest choices.

The following example describes of a metadata record in Dublin Core (DC).

```
<HTML>
<HEAD>
<TITLE> Manual Metadata Creation </TITLE>
<meta name=”DC.Title” CONTENT=“Learning Metadata”>
<meta name=”DC.Creator” CONTENT= “Mondal, Bikram”>
<meta name=”DC.Type” CONTENT= “document”>
</HEAD>
<BODY><P> Metadata is an important aspect for information management.</p>
</BODY>
</HTML>
```

Values corresponding to different fields are assigned to describe the resource object.

4.Semi-Automatic Metadata Creation

Semi-automatic metadata creation is carried out involving machine assistant, in addition to manual effort. In this method, some values are automatically controlled while the remaining values are provided manually. Greenberg, Spurgin&Crysta (2006) pointed that most professionals prefer an application which will execute automatic algorithms and afterwards allow a human to evaluate and edit the results. This process can be considered as a hybrid method for metadata creation. Text processing softwaregenerate values like ‘date of creation’, ‘file size’, ‘language’ automatically while author or creator provides other required values (i.e., ‘title’, ‘subject’ etc.). Semi-automatic metadata generation tools are mostly used for metadata cross-walking or schema conversion (HTML meta tags to DC elements, DC to METS, MARCXML to DC record etc.). It is also used in populating certain metadata values in the metadata template, meta tag harvesting, content extraction (text summarization, key phrase extraction), automatic indexing, text and data mining, extrinsic data generation, social tagging (Park &Brenza, 2015; Park& Lu, 2009). It helps in data migration process, data correction process, data communication processes.

MarcEdit(www.marcedit.reeset.net) is a popular semi-automatic tool which is used to convert one record into different formats (MARC to MARC/XML). Dublin Core Viewer, a Firefox add-on can convert HTML meta-tags into Dublin core elements (Lauke, n.d.). Harvard’s‘JHove’(http://jhove.openpreservation.org/) can recognize different kinds of textual, audio and visual file formats. Integration of semi-automatic tools with daily workflows in technical services of libraries can enhance the quality of metadata records for digital collection. To develop, maximize and sustain semi-automatic metadata generation workflows, administrative support for finance, human resources, training are required (Park et al, 2015). Technical knowledge is also required for employing tools in semi-automatic metadata creation methods.A semi-automatic metadata creation process for an item available on the internet archive (https://archive.org/details/ideology00mcle) is given in the Table 1.

Table 1: Semi-Automatic Metadata Creation

Manual	Automatic
Ideology by **McLellan, David** Publication date **1986** Topics **Ideology** Publisher **Minneapolis, MN**:**University of Minnesota Press** Collection **printdisabled; inlibrary; browserlending; internetarchivebooks; China** Digitizing sponsor **Internet Archive** Contributor **Internet Archive** Language **English**	Bookplateleaf0004 BoxidIA144316 Camera Canon EOS 5D Mark II City Minneapolis, MN Donor bostonpubliclibrary ExtramarcNYU Bobcat Identifier ideology00mcle Identifier-ark ark:/13960/t46q33c3r Isbn0816615225 Lccn86001383

5. Automatic Metadata Creation

In automatic metadata creation, values corresponding to various terms, acronyms, keywords, meta-tags from digital documents (or corpus of documents) are extracted with the help of machine enabled process. It helps to overcome human variations that occur in manual metadata creation process. Different methods and techniques are used for digital resources which are in different formats ranging from text to multimedia. Park &Brenza (2015) defined automatic metadata generation as a process that relies on machine learning language. It is considered more efficient, more consistent and less costly than human metadata generation (Greenberg et al., 2006). Greenberg (2004) proposed two methods namely harvesting and extraction for metadata generation. Harvesting is based on alreadyavailable metadata records whereas extraction process has to identify various metadata values and extract those from digital resources. Components of OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) can automatically harvest information from data sources. Automated metadata extraction depends on digital object format, genre and metadata quality requirements (Dobreva,Kim & Ross, 2008). Automated metadata extraction focuses on the following tasks:

i) Classification of objects’ nature
ii) Identifying the required metadata values from the objects based on appropriate techniques

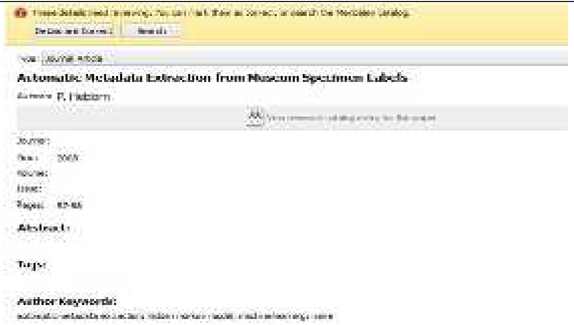
iii) Extraction of those values precisely

At first, identifying documents' structure and the position of metadata values are to be correctly located. Usually first few pages of documents (e.g., journal paper, thesis, report) contain most of the metadata values like title, author, journal, volume, publisher, etc. Both supervised and unsupervised machine learning techniques are used for automatic metadata extraction. Some of the methods used for metadata generation are natural language processing, machine learning, text analysis, domain-specific taxonomies, rules based, pattern matching, feature extraction, controlled vocabularies, folksonomies, ontology (Park et al., 2015; Huynh & Hoang, 2010; Mitchell, 2006). Some of the popular machines learning techniques used are Support Vector Machines (SVM), Hidden Markov Models (HMM), Conditional Random Fields (CRF). Features used for automatic metadata extraction are style features such as font face, size and style, semantic and linguistic features, structure and context features, font features, stylistic analysis, use of knowledge bases (Flynn et al., 2007; Kovacˇevic, Ivanovic, Milosavljevic&Konjovic, 2011; Mitchell, 2006). Some of the systems used to extract different values automatically from resources are listed in Table 2.

Table 2: Systems for Automatic Metadata Extraction Task

Sl.no	System	Details
1	Apache Tika (https://tika.apache.org/)	A framework which can automatically extract many values from different objects.
2	Mendeley (https://www.mendeley.com/)	Reference management software that can automatically extract many values like title, author, date, page number form articles.
3	Metadata Extraction Tool (http://metaextractor.sourceforge.net)	It automatically extracts preservation-related metadata from a range of file formats like PDF documents, image files, sound files Microsoft office documents, and many others (developed by the National Library of New Zealand).
4	CiteSeerX (http://csxstatic.ist.psu.edu/ about)	An archive for computer and information science related documents extracts Dublin Core (DC) metadata from digital objects automatically.
5	Metaextract (Yilmazel, Finneran, Liddy, 2004)	In the domain of math and science, it is designed to extract Dublin Core (DC) and Gateway to Educational Materials (GEM) metadata using natural language processing.

KEA, MrDlib, Alchemy, ParsCit are some other tools used for automatic metadata extraction (Casali, Deco &Beltramone, 2016). Figure 1 shows automatic metadata extraction in Mendeley which uses SVM. It correctly extracts name of the article, author name (thoughanother author name is not retrieved), year of publication, page numbers of the item and author provided keywords. This method makes metadata record accurate and reliable. There is a provision for modifying the record manually, if required. .



[illegible]

Type: Journal Article

Automatic Metadata Extraction from Museum Specimen Labels

Authors: P. Heidorn

Journal:

Year: 2008

Volume:

Issue:

Pages: 57-68

Abstract:

Tags:

Author Keywords:

[illegible]

Figure 1: Automatic Metadata Extraction in Mendeley

6. **Emerging Areas in Creation of Metadata**

Describing complex object requires good descriptors. Learning objects are complex and need different approaches for identification. Saini, Ronchetti&Sona (2006) used ontology for classifying learning objects automatically into a given taxonomy. Ontology can be effectively used for automatic metadata generation (Park et al., 2015). Hatala& Richards (2003) created a mechanism using ontology and rule based approach for suggesting the most relevant values for selected metadata fields. For semantic metadata generation (automatic semantic annotation), ontology can be used implicitly or explicitly (Yang & Lee, 2005). Data science requires dynamic metadata that can describe life cycle of data from creation, capture, storage, and preservation to complex process like data use, reuse, repurposing, and modification. Greenberg (2017) proposed social metadata repository on storing or linking social data and identifying bibliographic relationships for making social data more valuable and reusable for searching and retrieval. Metadata will play an important role in providing semantics information for the design of digital library architectures in future. Table 3 shows different features of metadata creation tasks.

Table 3: Features of Metadata Creation Methods

Sl.no	Manual	Semi-automatic	Automatic
1	Traditional method to assignmetadata values.	Used as a complementary to traditional method.	Completely new method to implement for metadata extraction.
2	Analyzing documents is time consuming.	Interaction of manual and machine take less time than manual process.	Continuous process, carried out by machines takes minimum time.
3	Accuracy depends on personal efficiency of the person involved in this task.	Accuracy depends on tools and its application areas.	Accuracy will be ascertained before the tool being used.
4	Not useful if data are more and heterogeneous.	Useful in data conversion from one standard to another.	Useful in handling huge data.

7. Summary

Discovery of information depends on effective management of metadata. Values corresponding to different metadata fields can be extracted automatically. Tools can be integrated in metadata creation process as it helps administrators to carry out various submissions tasks effectively. Values like date of submission, file formats, subjects, source, type of object can be automatically identified by machines. Template based form with predefined values makes tagging of objects fast and accurate. Data dictionary can be integrated with metadata form to carry out such tasks. To make metadata records uniform and consistent, harvesting mechanism can be effectively used. Accuracy of metadata tools must be examined thoroughly before they are implemented in any discipline.

8. REFERENCES

[1]. Alemu, G., & Stevens, B. (2015).*An Emergent Theory of Digital Library Metadata: Enrich then Filter*. Waltham, MA :Chandos Publishing.

[2]. Ardo, A. (2010). Can we trust Web-page metadata, Journal of Library Metadata.10 (1), 58-74.doi:10.1080/19386380903547008

[3]. BA INSIGHT.(n.d). Machine-Generated Meta Data in SharePoint. Retrieved from http://bainsight.com/files/Product/Whitepaper-AutoClassifier-Machine-Generated-MetaData.pdf

[4]. Berners-Lee, Tim. (1997). Axioms of Web Architecture: Metadata. Retrieved from https://www.w3.org/DesignIssues/Metadata.html

[5]. Casali, A., Deco, C and Beltramone, S. (2016). An Assistant to Populate Repositories:Gathering Educational Digital Objects and Metadata Extraction. IEEE RevistaIberoamericana de TecnologiasdelAprendizaj.11(2), 87-94.

[6]. Chen, Hao and Dumais, Susan. (2000). Bringing Order to the Web: Automatically Categorizing Search Results, In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 01-06 April, The Hague, The Netherlands, pp. 145-152. doi:10.1145/332040.332418

[7]. Dobreva, M., Kim, Y. and Ross, S. (2008).Designing an automated prototype tool for preservation quality metadata extraction for ingest into digital repository, In: Cunningham, P. and Cunningham, M. (eds.) Collaboration and the Knowledge Economy: Issues, Applications, Case Studies, IOS Press, Amsterdam.ISBN 9781586039240. Retrieved from http://eprints.gla.ac.uk/51154/

[8]. Dorbeva, M., Kim, Y and Ross, R. (2013). Automated Metadata Generation.Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction/

[9]. Flynn, P, et al. (2007). Automated Template-Based Metadata Extraction Architecture. In 10th International Conference on Asian Digital Libraries (ICADL), 10-13 Dec, Berlin, Heidelberg, pp. 327–336.

[10]. Greenberg, J. (2003). Metadata and the World Wide Web. Dekker, Marcel (Ed.), Encyclopedia of Library and Information Science, New York, pp. 1876-1888.

[11]. Greenberg, J. (2004).Metadata extraction and harvesting: a comparison of two automatic metadata generation applications. Journal of internet cataloging, 6(4), pp.59-82.doi:10.1300/J141v06n04_05

[12]. Greenberg, J., Spurgin, Kristina and Crysta, Abe. (2006).Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions.Int. J. Metadata, Semantics and Ontologies, 1(1), pp.3-20.doi:10.1504/IJMSO.2006.008766

[13]. Greenberg. (2017). Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy between Data Science and Metadata.Journal of Data and Information Science, 2(3),pp.19-36.doi: 10.1515/jdis-2017-0012

[14]. Guy, M., Powell, A. & Day, M. (2004).Improving the Quality of Metadata in Eprint Archives.Ariadne, 58. Retrieved fromhttp://www.ariadne.ac.uk/issue38/guy/