*International Journal of Advanced Culture Technology Vol.6 No.2 117-122 (2018)*
*DOI https://doi.org /10.17703/IJACT.2018.6.2.117*

IJACT18-2-16

# Automatic Extraction of Metadata Information for Library Collections

Gi-Chul Yang* and Jeong-Ran Park

*Department of Convergence Software Mokpo National University*
*College of Computing and Informatics Drexel University*

*gcyang@mokpo.ac.kr, p365@drexel.edu*

## *Abstract*

*As evidenced through rapidly growing digital repositories and web resources, automatic metadata generation is becoming ever more critical, especially considering the costly and complex operation of manual metadata creation. Also, automatic metadata generation is apt to consistent metadata application. In this sense, metadata quality and interoperability can be enhanced by utilizing a mechanism for automatic metadata generation. In this article, a mechanism of automatic metadata extraction called ExMETA is introduced in order to alleviate issues dealing with inconsistent metadata application and semantic interoperability across ever-growing digital collections. Conceptual graph, one of formal languages that represent the meanings of natural language sentences, is utilized for ExMETA as a mediation mechanism that enhances the metadata quality by disambiguating semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context. Hence, automatic metadata generation by using ExMETA can be a good way of enhancing metadata quality and semantic interoperability.*

**Keywords:** *Digital library, Automatic metadata generation, Semantic interoperability, Conceptual graph*

## 1. Introduction

The rapid proliferation of digital library projects has raised various critical issues and concerns regarding metadata implementation [1]. Initiating new digital collections from the ground up leads to a series of challenging metadata decisions, including selection of metadata schemes, semantics, content rules, controlled vocabularies and metadata creation workflow. In addition to the resulting interoperability issues in the aggregated environment, metadata creators must also pay attention to the newer functions of administration, provenance, rights management, and preservation. As a result, the quality of metadata in this increasingly complex environment is based on a much broader set of functional requirements beyond the application of established rules and standards. While “good” metadata reflects the degree to which it is fit for the intended functional purpose of supporting common user tasks and services, the types of required data elements can be defined only through specification of several specific, agreed-upon dimensions for operationalizing the measurement of metadata quality. Park [2] finds that completeness, accuracy, and consistency are the most commonly used criteria in measuring metadata quality.

As evidenced through rapidly growing digital repositories and web resources, automatic metadata

Manuscript received: June 20, 2018 / revised: June 21, 2018 / Accepted: June 22, 2018
Corresponding Author: gcyang@mokpo.ac.kr
Tel:+82-61-450-2355, Fax: +82-61-450-2358Author’s affiliation
Department of Convergence Software, Mokpo National University, Korea

generation is becoming more critical, especially considering the costly and complex operation of manual metadata creation. The wide range of metadata types encompassing technical (e.g., format, date), descriptive (e.g., title) and semantic metadata can be generated through a variety of methods and sources. Metadata value may be captured through exploiting sources not only from the document (object) itself but also from its context including document usage, user profile and metadata repositories.

Automatic metadata generation is apt to consistent metadata application. In this sense, metadata quality and interoperability can be enhanced by utilizing a mechanism for automatic metadata generation. This study aims at presenting a mechanism of automatic metadata extraction in order to alleviate issues dealing with inconsistent metadata application and semantic interoperability across ever-growing digital collections.

## 2. Metadata mapping

The flexibility and complex structure of natural language allows for the representation of a concept in various ways. In natural language, mapping between word forms and meanings can be many-to-many. That is, the same meaning can be expressed by several different forms (e.g., synonyms) and the same forms may designate different concepts (e.g., homonyms). In addition, the same concept can be expressed by different morpho-syntactic forms (e.g., noun, adjective, compound noun, phrase and clause).

The word uses of synonyms (e.g., author, writer, creator), homonyms (e.g., bank [building] vs. bank [river]) and polysemy (multiple related meanings of a word that are enumerated in alphabetical order in a typical dictionary entry) in face-to-face human interactions add the richness and creativity of natural language. Any ambiguities and misunderstandings that are engendered are usually resolved smoothly through communication cues provided during social interactions such as repetition and elaboration, social context and non-verbal cues (e.g., facial expressions and gestures).

However, in an information retrieval setting these same semantic ambiguities bring about lowered recall and reduced precision. In addition, these linguistic phenomena may engender confusion in the sense that different communities may use dissimilar word forms to deliver identical or similar concepts, or may use the same forms to convey different concepts. In order to attain quality metadata and semantic interoperability accordingly, a mediation mechanism that provide contextual relations among metadata elements and their corresponding definitions and usage is essential. Conceptual graphs have good potential to facilitate proper interpretation of metadata concepts and accurate and consistent usage of the data elements. A conceptual graph is one of formal languages that represent the meanings of natural language sentences [3]. Conceptual graph can be utilized as a mediation mechanism that enhances the metadata quality by disambiguating semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context. Hence, it is beneficial to use conceptual graphs as a language to describe formal knowledge source that the metadata information for library collections can be extracted automatically.

A CG can be constructed by assembling percepts. In the process of assembly, concept relations specify the role that each percept plays, and concepts represent percepts themselves. A concept can be generic or individual. The function referent maps concepts into a generic marker * or a se I = {#1, #2, #3, …}, the elements of which are individual markers. The function *type* maps concepts into a set of type labels. A concept c with *type* (c) = t and *referent* (c) = r is displayed as [t : r] in the linear form. The function type also can be applied to relation. A relation r with *type* (r) = t is displayed as ( t ) in the linear form. Types of concepts (type labels) are organized into a hierarchy called type hierarchy. Type hierarchy forming operations include conjunction and disjunction operations, so the type hierarchy will be a formal lattice. The type hierarchy constitutes a partial ordering and becomes a type lattice when all the intermediate types are introduced. There are many knowledge handling CG applications such as in [4] and [5].

A CG can be represented in three different forms. There is a graphic notation called the *display form* (DF), a more compact notation called the *linear form* (LF) as well as a concrete syntax called the *conceptual graph interchange form* (CGIF), which has a simplified syntax and a restricted character set designed for compact storage and efficient parsing. Both DF and LF are designed for communication with humans or between humans and machines. For communication between machines, the CGIF has a simpler syntax. Hence, we will develop an efficient search engine for SD's represented in CGIF in this paper.

Following is the CGIF of a conceptual graph that represents the prepositional content of the English sentence *Tom is going to New York by car*.

[Go *x] (Agnt ?x [Person: Tom]) (Dest ?x [City: New York]) (Inst ?x [Car])
or (Agnt [Go] [Person: Tom]) (Dest [Go] [City: : New York]) (Inst [Go] [Car])

CG is one of a good formal language for representing the meanings of natural language sentences. Hence, it is a good idea to use CGs as a language to describe formal knowledge source that the metadata information for library collections can be extracted automatically.

## 3. Automatic extraction of metadata information

A way of automatic extraction mechanism of metadata information for library collections called ExMETA is introduced in this section. Unlike the current human intervened way of generating standard metadata, ExMETA will generate input data for digital collection management software (such as CONTENTdm) [6] directly from the raw data in order to produce outputs tagged by a standard metadata such as Dublin Core (DC) [7].

The DC metadata standard is a simple yet effective element set for describing a wide range of networked resources. The Dublin Core standard comprises fifteen elements and, followings are some of the examples.

**Contributor:** The DC element used to designate Person(s) or organization(s) in addition to those specified in the CREATOR element who have made significant intellectual contributions to the resource.
**Creator:** The person or organization primarily responsible for creating the intellectual content of the resource.
**Title:** The DC element used to designate the name given to the resource.

The DC examples presented in HTML format are as follows.
< META NAME ="DC.Title" CONTENT=" Gone with the Wind ">
< META NAME ="DC.Creator" CONTENT=" Margaret Mitchell ">
< META NAME ="DC.Date" CONTENT="1936">

Figure 1 depicts the current way of generating standard metadata and a way of using ExMETA to generate standard metadata.

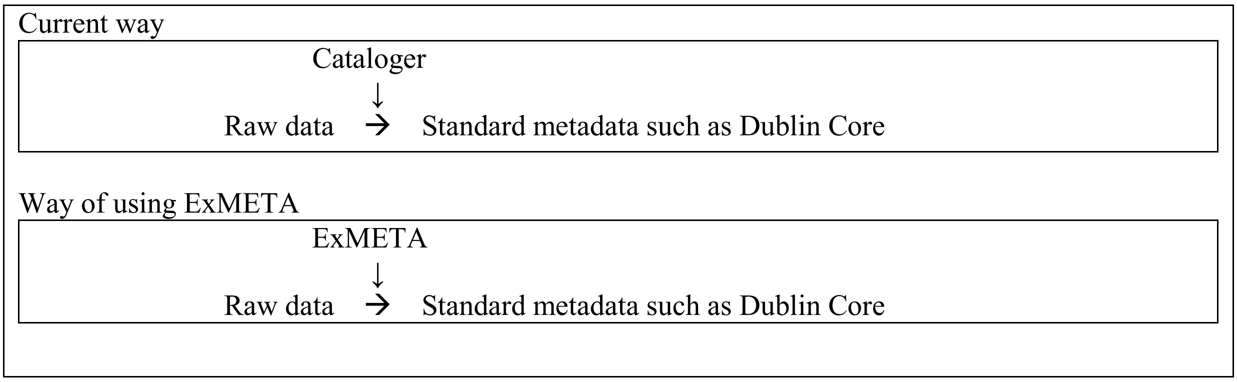
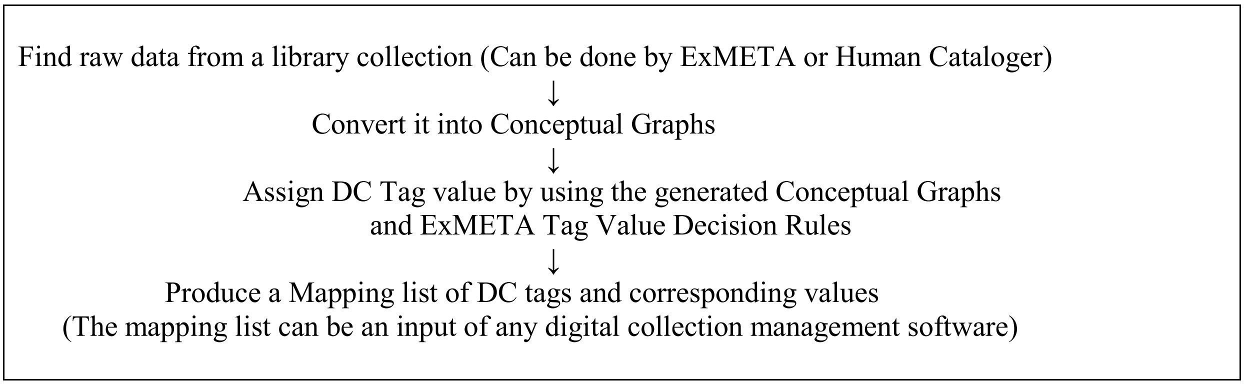

Current way
Cataloger
Raw data
Standard metadata such as Dublin Core
Way of using ExMETA
ExMETA
Raw data
Standard metadata such as Dublin Core

**Figure 1. Ways of Generating Standard Metadata**

As shown in Figure 1, ExMETA may enhance the semantic interoperability by facilitating consistent metadata generation.

Figure 2 describes a brief workflow of ExMETA.



Find raw data from a library collection (Can be done by ExMETA or Human Cataloger)
↓
Convert it into Conceptual Graphs
↓
Assign DC Tag value by using the generated Conceptual Graphs
and ExMETA Tag Value Decision Rules
↓
Produce a Mapping list of DC tags and corresponding values
(The mapping list can be an input of any digital collection management software)

**Figure 2. Workflow of ExMETA**

As shown in Figure 2, the raw data is the input of ExMETA and the raw data is written in any natural language. Of course, different CG converter programs are needed to handle different natural language sentences. For example, a raw data ‘Gone with the Wind is a romance novel written in 1936 by Margaret Mitchell’ can be converted into a CG such as

```
[Write] -
          (AGNT)->[PERSON: Margaret Mitchell]
          (OBJ)-> [NOVEL: gone with the wind]->(ATTR)->[romance]
          (PTIME)-> [YEAR: 1936]
```

Next step of ExMETA is assigning DC Tag values by using the generated CG and ExMETA Tag Value Decision Rules (TVDR). The generated CG contains the information ‘the agent of WRITE is [PERSON: Margaret Mitchell] ‘ and ‘the object of WRITE is [BOOK: gone with the wind]’. However, digital collection management software such as CONTENTdm needs DC Tag information such as CREATOR, TITLE and so on. TVDR is necessary to find DC Tag information from the generated CG which does not contain explicit DC Tag information. Some example TVDRs are as follows.

Rule 1: AGNT of (Write, Draw, Make, Complete, Compose, ...) is become a value of CREATOR tag.

Rule 2: OBJ of (Write, Draw, Make, Complete, Compose, …) is become a value of TITLE tag.

Rule 3: AGNT of (Support, Donate, Help, Edit, Transcribe, Illustrate, …) is become a value of CONTRIBUTOR tag.

…

According to TVDR rule 1, ExMETA can generate a pairs (CREATOR, Margaret Mitchell). Also, a pair (TITLE, gone with the wind) can be generated by using the rule 2. Finally those pairs are collected in a list like ((CREATOR, Margaret Mitchell), (TITLE, gone with the wind),…) to submitted as an input data for a digital collection management software.

Figure 3 depicts the architecture of ExMETA.

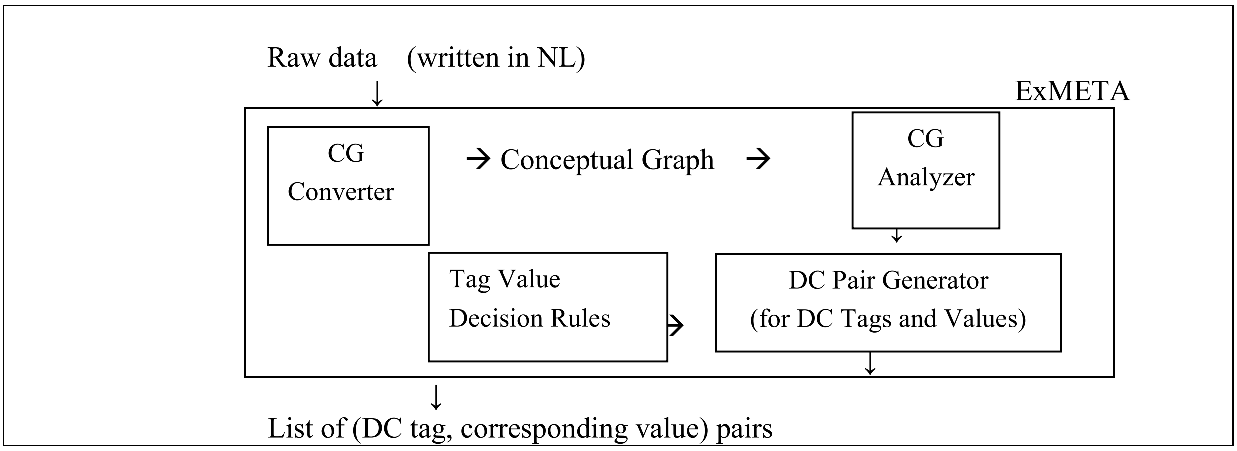


**Figure 3. Architecture of ExMETA**

The CG converter first performs syntactic analysis on an input sentence and constructs a corresponding CG. The CG analyzer disassembles the generated CG to find out the (Relation, Concept) pairs and the pairs passed to DC Pair Generator to produce the final result. The final result is a list of (DC tag, corresponding value) pairs which can be used as an input for any digital collection management software.

## 4. Conclusions

Utilization of automatic metadata generation is becoming imperative owing to rapidly growing digital resources and costly operation of manual metadata creation. In this study, we introduced the automatic metadata extractor ExMETA which is designed by utilizing conceptual graphs (CG) as the internal representation. ExMETA is capable of analyzing natural language sentences and of generating descriptive and structural metadata. We believe that this will contribute to enhancing the speed of metadata generation as well as metadata quality and semantic interoperability.

Current human intervened way of generating standard metadata may cause the inconsistent metadata assignment problem as well as the semantic interoperability problem in metadata mapping. A Way of Automatic Extraction of Metadata Information for Library Collections is introduced in this article to solve those problems. Introduced automatic metadata extractor, ExMETA, allows the elimination of human (i.e., cataloger) intervention that often cause the incorrect assignment and inconsistent mapping of metadata in the process of producing standard metadata from raw data of digital collections. ExMETA handles natural language sentences and use CGs as the internal representation. CG is a good formal language to represent the meaning of natural language sentences. Hence, automatic metadata generation by using ExMETA can be a good way of enhancing metadata quality and semantic interoperability.

Future study lies in testing usability and in application of ExMETA to real-world information settings such as libraries to assess its potential for improving the efficiency of metadata generation. Automatic indexing techniques are used in metadata extraction to generate descriptive and structured metadata based on the content of resources. There is a need to extend ExMETA to semantic metadata generation which requires a more sophisticated algorithm than that of descriptive and structural metadata generation.

## 5. Acknowledgement

This Research was supported (in part) by Research Funds of Mokpo National University in 2017.

## References

[1] Caplan, Priscilla. *Metadata Fundamentals for All Librarians*. Chichgo: American Library Associa tion. 2003.

[2] Park, Jung-ran. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Special Issue on **Metadata and Open Access Repositories (**M.S. Babinec and H. Mercer Eds.). *Cataloging and Classification Quarterly,* Vol. 47(3): 213-228. 2009.

[3] Sowa, J. Conceptual Structure: Information Processing in Mind and Machine, Addison Wesley, Massachusetts, 1984.

[4] Uta Priss, Simon Polovina, Richard Hill, Conceptual Structures: Knowledge Architectures for Smart Applications, Lecture Notes in Computer Science, Volume 4604, 2007.

[5] Yang, Gi-Chul & K.S. Choi, Construction of a Knowledge Base by using Korean Text, AAAI96- Fall Symposium, MIT, U.S.A. 1996.

[6] CONTENTdm, 2013. http://www.contentdm.com/

[7] Dublin Core Metadata Initiative. 2005. DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/