Theme Creation for Digital Collections

Xia Lin
Drexel University
Philadelphia, PA, USA
xlin@drexel.edu

Jiexun Li
Drexel University
Philadelphia, PA, USA
jiexun.li@drexel.edu

Xiaohua Zhou
Drexel University
Philadelphia, PA, USA
xiaohua.zhou@drexel.edu

Abstract

This paper presents an approach for integrating multiple sources of semantics for the creating metadata. A new framework is proposed to define topics and themes with both manually and automatically generated terms. The automatically generated terms include: terms from a semantic analysis of the collections and terms from previous user's queries. An interface is developed to facilitate the creation and use of such topics and themes for metadata creation. The framework and the interface promote human-computer collaboration in metadata creation. Several principles underlying such approach are also discussed.

Keywords: metadata creations; metadata authoring tools; topics and themes; human-computer collaboration

1. Introduction

The Internet Public Library (IPL: http://www.ipl.org) is one of the oldest digital libraries that is still actively maintained and used. Supported by a consortium of LIS schools (The IPL Consortium, n.d.), the IPL holds multiple collections of thousands of authoritative websites on various subjects. Most of these collections include sets of metadata records created by volunteering LIS students. Searching and browsing IPL collections are based on the metadata database. Because of this, it has been a priority for the IPL to create and maintain high-quality metadata within its current setting. The metadata will continue to be created by LIS students to support its mission as a teaching and learning environment for the Consortium member schools, yet high-quality metadata must be maintained to support its service to the public. It is essential for the IPL to have a powerful metadata creation tool that can be easily learned and used by professionals (or quasi-professionals) to create high-quality metadata for the digital library.

The objective of this research is to investigate how to incorporate multiple semantic sources to enhance metadata creation. Current IPL metadata consist of a set of well documented fields such as title, abstract, keywords, and subject headings. The subject headings are not a formally defined thesaurus but a set of loosely developed category terms. While the subject headings present a simple hierarchical view to the IPL collections, they do not provide strong associative and semantic relations among the headings and collections that a good thesaurus would otherwise provide. Thus, we sought solutions to build additional semantic relations among keywords, subject headings, topics and digital resources (Web pages) to enhance the IPL metadata. We particularly explored how context might be employed for metadata and how the context information might be extracted from both the semantic analysis of digital collections and the analysis of user's search logs.

The remainder of this paper is organized as follows. First, we discuss various semantic sources for metadata. We then define topics and themes and introduce a framework for metadata subject representation using multiple semantics sources. In particular, we describe the language model for semantic mapping and a bottom-up procedure of theme creation. Finally, we introduce a system we developed to integrate all the semantics sources into one rich interface.

https://doi.org/10.23106/dcmi.952109166

2. Metadata and Semantics

In computational linguistics, semantics refers to the relation between the words and sentences of a language and their meanings (Saeed, 2003). It is hypothesized that semantics can be extracted through lexical or statistical analysis of language and its structures. The meanings then can be represented by the data and structures obtained through the analysis. Similarly, semantics of metadata can be considered as the relation between metadata records and the content they represent. Metadata records are essentially "data + structures" that describe and represent various features of digital objects, including their content, context, and structures (Gill, et al. 1998). The semantics of metadata come from multiple sources. The first is the metadata standard. A metadata standard represents a consensus of how a specific type of digital objects should be described structurally. It provides a schema that specifies the metadata's namespaces, formats, required elements and allowable attributes, etc. Through naming and structuring the metadata elements, each standard provides a semantic framework that the user can "fill-in" values to create metadata records. The second source of the semantics comes from the metadata creation process. Typically, the standards do not give details on how a metadata record should be created. It is up to the metadata creator (most likely a human being) who interprets the content of the resource to be described and selects terms most appropriate for each entry of the metadata record. The human intelligence in this process provides the most significant semantic associations to connect metadata records to the content. The third semantic source of metadata is the semantics of the language. In particular, when a controlled vocabulary is used to create metadata records, the rich semantic relationships established within the controlled vocabulary enrich the semantics of metadata significantly.

There are many other semantic sources that have not been considered and incorporated into current metadata practice. One that seems to be obvious is the computerized semantic analysis of terms in a text collection. The semantic analysis can extract rich semantic relationships of terms over the whole collection to form "semantic metadata" (Haase, 2004). While such semantic metadata is still a lack of precision, incorporating selected terms from the list to enhance the standard-based metadata was considered a practical trend (Al-Khalifa, 2006).

Another useful source of semantics is the user's terms and usage patterns collected over time. How users search and interact with digital collections can provide valuable semantics for metadata creation. It could be an iterative process to improve the metadata with usage statistics. The more users use the collections, the more usage patterns will be collected and the better the metadata would be when the patterns are used appropriately.

Different sources can capture the semantics of a collection from a different perspective. It would be useful to integrate multiple semantic sources abovementioned to enhance the metadata creation process. In this research, we attempted to develop a framework and an authoring tool that would incorporate semantic mapping and usage patterns as semantics sources for metadata creation.

3. A Hierarchical Framework for Subject Representation in Metadata

3.1. Topics, Themes, and the Framework

Topic Maps (ISO13250, 2002) provide a new approach to represent knowledge and create associations among subjects and digital resources. As an established standard technology that includes well defined syntaxes, structures and the underlying reference model, Topic Maps describe knowledge structures through topics, associations, and occurrences in a formal model (Pepper, 2000). In this model, a theme is also defined as "a member of the set of topics comprising a scope within which a topic characteristic assignment is valid." The theme here is only used to define scopes for topics. However, as Pepper & Gronmo (2002) pointed out, both scopes and themes are the means to "putting context into topic maps."

https://doi.org/10.23106/dcmi.952109166

We believe that there is a potential to expand the role of themes. A theme should be considered as a special topic or as "a topic in context." It can be used to provide contextual links to topics. It can become a higher level of subject indicators along with keywords, subject headings, and topics. In this research, we simply view a Topic as a subject with a name and multiple slots of properties. The properties may include different types of keywords generated from multiple sources either manually or automatically. Then, we view a Theme as a special type of topics that unites several topics around a theme. Unlike in Topic Maps where the center of the universe is "topic," we are exploring to have the "theme" as the main unit that can have its own descriptive metadata and let topics to be "characteristics" of the theme. Our assumption is that users would be more interested in "topics in context" than topics. When a searcher sends a query to a collection, the searcher will likely be more satisfied to retrieve themes that provide specific topics and resources relevant to the query than to retrieve only the resources that match the user's query directly.

Figure 1 illustrates the hierarchical framework for subject representation in metadata. As mentioned above, the properties to describe topics and themes can be from multiple sources. In this study, we only include three sets: a set of topic signatures automatically generated from a collection, a set of keywords manually assigned, and a set of keywords identified and selected from the previous user's query terms. Details on topic signature generation will be introduced in Section 4. Each set of keywords can have different weights for the purpose of indexing and retrieval. The topic can also include properties of relevant resources (occurrences), either manually selected or automatically retrieved and inserted by search engines. For the IPL, two types of resource URLs are included as properties of a topic: URLs within the IPL and those outside the IPL.

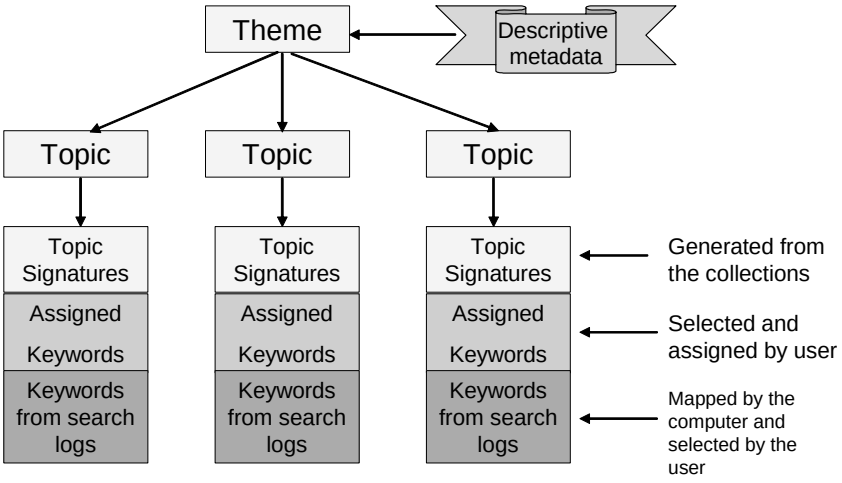

FIG. 1. A hierarchical framework for subject representation in metadata.

3.2. Collection-based Topic Signatures and Semantic Profiles

In the proposed framework of subject representation, topic signatures are a key concept. The topic signatures can be automatically generated through a topic signature model we developed (Zhou et al. 2006). The model is based on semantic mapping through a language modeling approach and a context-sensitive semantic smoothing method (Zhou et al., 2007a). Two types of mappings are created in this language model (Figure 2). One is called topic signatures that map from any term, w, in the collection to a list of topics, t's (represented by keywords, subject

https://doi.org/10.23106/dcmi.952109166

headings, or other indexing terms). The other is called semantic profiles that map a specific topic (t) to a set of terms (w's) that are most likely to co-occur with t in the collection, C.

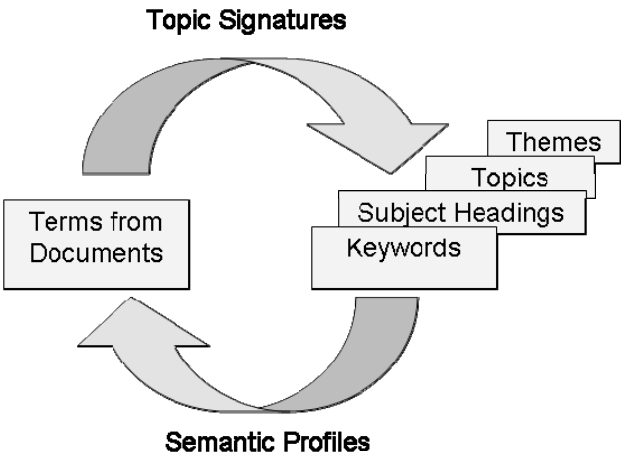

FIG. 2. Two types of semantic mapping: topic signatures and semantic profiles.

To create semantic profiles for keywords in a collection (*C)*, we first index all documents with individual terms and topics. For each topic t_k, we approximate its semantic profile using the terms *w*'s in the document set D_k containing t_k, ranked in the descending order of the conditional probability $p(w \mid t_k, C)$. We assume that the terms appearing in D_k are generated by a mixture model:

$$p(w \mid t_k, C) = (1-\alpha)\, p(w \mid t_k) + \alpha p(w \mid C)$$

where $p(w \mid t_k)$ is a topic model that represents the conditional probability of term *w* co-occurring with topic t_k. $p(w \mid C)$ is a background model describing the global distribution of terms in the collection *C*, and α accounts for the background noise. Not only does this mixture model capture the semantic associations between topics and terms in the topic model, but it also takes into account the overall term distribution of a collection in the background model. The model for t_k can be estimated using an expectation-maximization (EM) algorithm. Details of the model can be seen in (Zhou et al., 2006; Zhou et al., 2007a).

It is noted that the model represents an effective semantic mapping based not only on the content but also on the context. Due to the different focuses of different collections, the metadata used to describe the same terms or objects may vary from collection to collection. Our language model is able to capture the different semantic associations among topic signatures in different collections. Table 1 shows two topic signatures for the same topics in two different collections. For example, we conducted the semantic mapping on two of IPL collections: the IPL general collection and the IPL collection for Youth and Teens. The mapping results show strong "context interpretations." For example, the topic "reading" in the general (adult) collection is closely associated with "classics," "review," "humor," "literary," etc., while the same topic in the collection for teens is closely associated with "children's literature," "stories," "folklore," etc. Similarly, the topic signatures show that for the health issue, adults are more concerned about "mental health," "health care," "disease," etc. and the teens are more concerned about "fitness," "exercise," "nutrition," and "stress." When creating metadata for different collections, such suggestions of collection-based related terms would be very useful.

https://doi.org/10.23106/dcmi.952109166

TABLE. 1. Examples of automatically-extracted topic signatures in different collections.

Topic	**Reading**		**Health**	
Collection	*IPL*	*IPL-Teens*	*IPL*	*IPL-Teens*
Topic signatures	**reading** classics review humor literary comic Cartoons Comics Censorship Rules FOIA biology Manga Native Americans Insurance Scientists Indian ……	**reading** children's literature stories folklore magazine story books author biographies social studies children instruments fantasy Games paleontology biography	**health** mental health health care disease activism disorders public Health psychology mental Illness reproduction medicine Safety Therapy prevention Pregnancy medical Nutrition Drugs ……	**health** fitness exercise nutrition stress tic panic medicine attention deficit ADHD depression add Teaching teens

Furthermore, for each keyword, the language model creates a semantic profile by mapping the keyword to a set of related terms in the specific collection. For example, as shown in TABLE 2, in the IPL Teens collection, the semantic profile of the keyword "stress" contains a list of highly associated words, including "depression," "anxiety," "health," "disorder," "eat," etc. The number attached to each term gives $p(w|\theta_{t_k}, C)$, the conditional probability of term w co-occurring with keyword t_k in collection C. Such a semantic profile can help us better understand the keyword in a particular context, and further decide whether to include the keyword in the metadata.

TABLE 2. An example of a semantic profile for the term "stress" in IPL Teens Collection.

Semantic profile	**Probability**
stress	0.0591
depress	0.0339
teen	0.0292
anxiety	0.0286
health	0.0284
disorder	0.0264
eat	0.0226
mental	0.0221
…	…

https://doi.org/10.23106/dcmi.952109166

3.3. A Theme Creation Procedure

To create themes for the IPL, we developed a bottom-up procedure of theme creation and tested through a group of volunteers and students in selected classes. The process starts with examining users' search logs, reviewing the suggested topic signatures, and further identifying topics and themes as metadata. Figure 3 shows an example of instructions given to the theme creator. Figure 4 shows some examples of themes created by students following the procedures. Notice the themes are specific to the IPL collections and IPL users. Collectively, they indicate the content of the IPL from users' perspectives.

A. **Explore users needs.** You have access to the list of top 1000 query terms that the users used to search IPL most frequently in the past three months.
1) Browse through the list first.
2) Select terms to form groups as potential topics.
3) Identify some recurrent themes in the groups you identified.
4) Decide a theme you would like to work on

B. **Explore the collections**.
1) Use the topic signature and semantic profile tool to explore and collect relevant terms for the topics.
2) Use IPL search engines to identify relevant resources to the topics.
3) Use Google or other tools to identify relevant resources outside IPL.
4) Check if there are any pathfinders or spotlight within IPL that are relevant to the topics (both are related IPL finding-aids).

C. Create the theme
1) Describe the theme using the given metadata schema (i.e., complete the title, description, subject headings, and other descriptive fields).
2) Identify topics associated with the theme and generate the topic signatures.
3) Go over the automatically generated topic signatures and select the most relevant terms as keywords. Add your own terms as necessary.
4) You may need to go back to step A and B to complete the process.

FIG. 3. A procedure of theme creation for IPL collections.

- **Literature**
  - American Authors
  - American Literature
  - Banned Books
  - Shakespeare
- **History**
  - History of Military Conflicts
  - The American Revolution
  - Presidents of the United States of America
- **People**
  - Influential Americans
  - American Leaders
  - U.S. Presidents in Context
- **Teens**
  - Raising and Nurturing a Teenager
  - Teen Entertainment
- **Social Issues**
  - Race & Ethnicity
  - Public Policy Issues
  - Internet Popular Culture
- **Sciences**
  - Science Fair Projects
  - Topics in Science
  - Grade school research project
- **Technology**
  - Computers and Libraries
  - Personal Internet Entertainment
  - Web Service Hubs
  - Emerging Technologies
  - Entertainment/Social Networking
  - Cars
- **Environment**
  - The World We Live In
  - Environmentalism
  - Geography and World Locations
- **Health**
  - Physical Fitness
  - Your Body, Your Mind, Your Health
  - Diabetes prevention
  - Health Disorders and Prevention

FIG. 4. Sample titles of themes created for IPL collections.

The procedure and the process of theme creation highlight several principles we are developing and testing:

- Themes and topics can be created through an integrated process of both manual and automatic processes. It seems that the manual process could focus on the higher levels and the automatic process on the detailed and lower levels of the subject representations.

https://doi.org/10.23106/dcmi.952109166

Each higher level of representations provides or enhances associative relationships of the lower level ones.

- Themes and topics can be represented dynamically based on the semantic analysis of the collections to be represented and on the analysis of previous user's interactions with the collections (for example, the search logs analysis). As the collections and the user's needs change over time, the meaning and the representation of the themes and topics may change dynamically.
- Context-sensitive themes do not need to be defined uniquely. Different users or systems can define same themes from different perspectives and with different names or sets of properties. Their similarities can be measured by their sharing properties (keywords and URLs, etc.). Similar themes will likely show a high degree of similarities.
- Themes and topics can be used to index and describe digital resources in multiple levels. The retrieval process can take advantages of such multi-level representations to provide search results in different granularity.

4. Implementation

To apply the framework and the procedure, we are developing an integrated system called "IPL KnowledgeStore." This section introduces the tools we adopted and the major features available in the rich interface we developed.

4.1. Semantic Mapping using Dragon Toolkit

While several complex natural language processing and statistical algorithms are needed to generate topic signatures and semantic profiles, Zhou et al. (2007b) also developed an open source toolkit to facilitate the mapping process. The toolkit, called the Dragon Toolkit, is a Java-based development package for language modeling and information retrieval, including text classification, text clustering, text summarization, and topic modeling. The toolkit is freely available for academic use at http://www.dragontoolkit.org. It provides many tools to map text collections with various representation schemes including words, phrases, ontology-based concepts and relationships. Specifically, in this research, we make use of the Dragon Toolkit APIs for the semantic mapping between terms and keywords (i.e., topic signatures and semantic profiles) in different collections.

4.2. An Integrated Interface

We are developing the IPL KnowledgeStore system using Adobe's Rich Internet Application (RIA) development environment, FLEX 3. Figure 5 shows a sample interface of the application. The interface functions as a "semantic aggregator" and a collaborative authoring workspace that provides access to multiple semantics sources, including the IPL metadata, topic signatures, semantic profiles created by the Dragon Toolkit, and the list of most frequently used search terms. The tool allows the user to create multiple types of digital objects such as subject terms, topics, themes, and metadata for IPL resources; each object itself is also described by a metadata. The user can create, modify, retrieve, and save these objects to a database or to XML files based on defined schemes. The interface provides rich interactive functions and links. Each source of semantics can be used separately or linked together. When a term in the center work space is selected, both sides of mappings (from the resource collections and from the user's terms) will be done automatically. Such mappings allow better use of associations hidden in the collections and in the user's interactions with the collections. The user can easily drag-and-drag terms from one slot to another and edit or select automatically generated terms to enhance the representations.

https://doi.org/10.23106/dcmi.952109166

FIG. 5. A sample screen of the integrated interface.

5. Conclusions

Semantics of metadata might be considered as the relation between metadata records and the content they represent. In this paper, we examined two important considerations to enhance the semantics. One consideration is how to enhance content representation with context, and the other is how to integrate multiple sources of semantic sources for the purposes of metadata creation. We showed that, topics and themes could be created with a combination of automatic semantic mapping and human interpretation. The semantic mapping utilizes both content and context when suggesting topic signatures and semantic profiles for subject terms. The human users can take the suggestions to create topical themes or metadata with additional association and context interpretations. We also developed an integrated interface for metadata creation. The interface allows users to select subject terms from various semantic sources, including terms used in existing metadata records, terms from the subject headings, terms from topic signatures and semantic profiles created by the automatic semantic mapping, and terms from the search logs.

Much more research needs to be done to address the need for semantic enrichment of metadata. We plan to continue developing our system to examine several principles discussed in this paper and test the effectiveness and usability of the interface as a metadata authoring tool for the IPL. Finally, we hope to further refine the concepts of topics and themes and use them for multi-level subject indexing (such as keyword indexing, topic indexing, and theme indexing) for metadata and digital collections.

References

Al-Khalifa, Hend S., and Hugh C. Davis. (2006). The evolution of metadata from standards to semantics in e-Learning applications. *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, Odense, Denmark, 2006*, (pp. 69-72).

Cruse, Alan. (2004). *Meaning in language: An introduction to semantics and pragmatics* (2nd ed.). Oxford: Oxford University Press.

DCMI. (2008). *Dublin Core Metadata Element Set, version 1.1*. Retrieved March 30, 2008, from http://dublincore.org/documents/dces/.

Gill, Tony, Anne Gilliland, and Murtha Baca. (1998). *Introduction to metadata: pathways to digital information*. Los Angeles, California: J. Paul Getty Trust. Retrieved March 30, 2008 from http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html.

Haase, Kenneth. (2004). Context for semantic metadata. *Proceedings of the 12th Annual ACM international Conference on Multimedia, New York, USA, 2004,* (pp. 204-211). Retrieved from http://doi.acm.org/10.1145/1027527.1027574.

The IPL Consortium. (n.d.) *The Internet Public Library*. Retrieved March 30, 2008, from http://www.ipl.org.

https://doi.org/10.23106/dcmi.952109166

ISO/IEC (2002). *ISO/IEC 13250 -Topic Maps* (2nd ed.). Retrieved March 30, 2008, from, http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0129.pdf.

Manning, Christopher D., and Hinrich Schütze. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.

Pepper, Steve. (2000). *The TAO of Topic Maps.* Retrieved March 30, 2008 from http://www.ontopia.net/topicmaps/materials/tao.html.

Pepper, Steve, and Grønmo, Geir Ove. (2002). *Towards a general theory of Scope*. Retrieved March 30, 2008, from http://www.ontopia.net/topicmaps/materials/scope.htm.

Saeed, John I. (2003). *Semantics*. *Introducing liguistics, 2*. Malden. MA: Blackwell Pub.

Zhou, Xiaohua, Xiaohua Hu, Xiaodan Zhang, Xia Lin, and Il-Yeol Song. (2006). Context-sensitive semantic smoothing for the language modeling approach to genomic IR. *Proceedings of the 29th Annual international ACM SIGIR conference on research and development in information retrieval*, *Seattle, Washington, USA, 2006,* (pp. 170-177). Retrieved from http://doi.acm.org/10.1145/1148170.1148203.

Zhou, Xiaohua, Xiaohua Hu, and Xiaodan Zhang. (2007a). Topic signature language models for ad hoc retrieval. *IEEE Transactions on Knowledge and Data Engineering, 19*(9), 1276-1287.

Zhou, Xiaohua, Xiaohua Zhang, and Xiaodan Hu. (2007b). Dragon Toolkit: Incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2007, Patras, Greece* (pp. 197-201).

https://doi.org/10.23106/dcmi.952109166