Semantic Reuse of Existing Metadata: A Model-based Perspective

Naimdjon Takhirov
Department of Computer and Information Science
Norwegian University of Science and Technology
NO-7491 Trondheim
Norway
takhirov@idi.ntnu.no

ABSTRACT

Existing metadata about cultural items such as movies, books and music is a valuable resource that have been created for specific purposes and managed for decades by institutions worldwide. In order to realize and increase the potential value of this metadata and make it available for new services and purposes, there is a need for proper semantic reuse and integration of this metadata. Thus, the focus of this work is on improving the quality of existing metadata. Our approach is based on model-based interpretation of existing metadata, using a conceptual domain model. The main research objectives are to support the migration of existing metadata to a new information model that enables semantic aware reuse and integration and to exploit methods to improve the reuse value of existing metadata and make it available for new types of services.

Categories and Subject Descriptors

H.2.5 [**Heterogeneous Databases**]: Data translation;

H.2.1 [**Database Management**]: Logical Design – Data models, Schema and subschema.

D.2.12 [**Software Engineering**]: Interoperability – Data mapping;

General Terms

Design, Verification.

Keywords

Semantic Web, Ontologies, Conceptual Models, Entity Matching

1. INTRODUCTION

The Internet and World Wide Web (Web) has significantly changed the way we create and distribute information. A continually increasing amount of digital information is being made available in an environment that fosters cooperation and interchange of data and services. A large portion of our cultural information is already thoroughly documented by institutions worldwide, but to realize and increase the potential value of this existing metadata, there is a need to migrate or transform this information into a representation that enables semantic reuse and integration [10,12].

However, a fair amount of valuable cultural information is not used to its full potential. This is partly because the metadata used to describe this information is usually designed to fulfill specific purposes. Therefore, it is difficult to make this information available to new types of services refraining from exploiting it to its full potential reuse value. Furthermore, unlike records in databases are identified by their primary key or other type of identifiers, this descriptive metadata about resources is usually maintained as a distinct unit and the entities that implicitly are described in these units such as an author of a book are usually identified by descriptions only. The process of discovering correct relationships between entities found in the metadata about information resources is even more challenging since there may be variations in spellings, descriptions are written in different languages, information about type of relationship is missing etc. As information found in the descriptive metadata about resources is weakly typed, i.e., entities and relationships are identified by analyzing description texts, transformation to a semantic format such as RDF will typically be inconsistent, introducing data quality problems. Therefore, to cope with inconsistencies, a correction and verification steps are needed to attain a reasonable level of consistency.

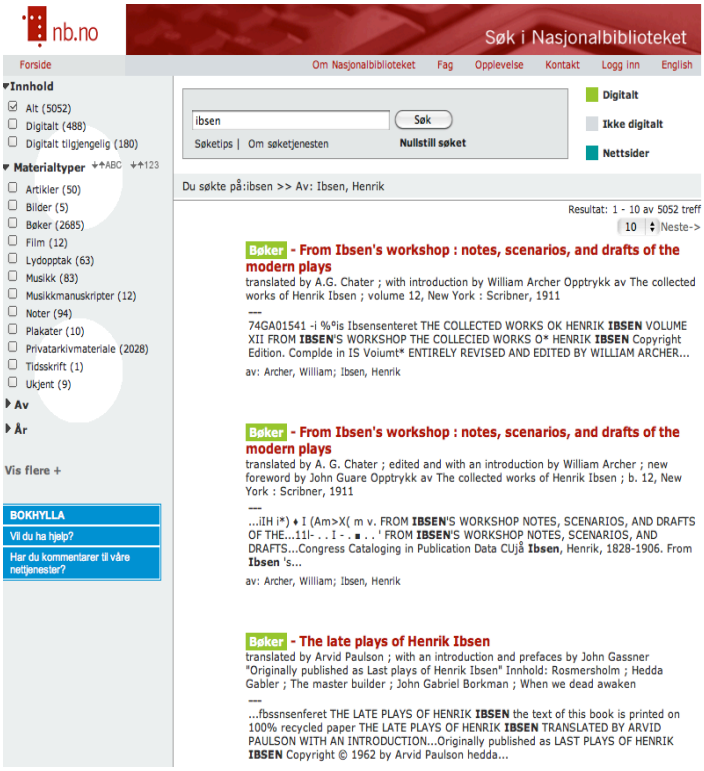
Our approach is to use a model-based interpretation of existing metadata, using an overarching conceptual domain model which provides a formal specification and facilitates integration of the information within a domain. The main research objectives are to support the migration of existing metadata to a new information model that enables semantic aware reuse and integration and to exploit methods to improve the reuse value of existing metadata and make it available for new types of services. As an example of a service, improving the reuse value can enable better exploratory search services [8] where users are presented with listings of items grouped by the type and relationships to other items to support learning about and discover the versions or editions users prefer.

2. MOTIVATING EXAMPLE

To illustrate a problem, consider a user who is interested in listings of all works of Henrik Ibsen -- a major 19th-century Norwegian playwright. Naturally, one would search the collection

JCDL'11, June 12–17, 2011, Ottawa, ON, Canada.

of Norwegian national bibliography. Having used a fair amount of time reformulating the query, the answer could not be found (Figure 1). On the other hand, since Henrik Ibsen is a well-known writer, the list of major works of Henrik Ibsen can be found on the Wikipedia article about him[1] where information about his works are arranged in a higher abstraction level. However, not all creators are described in Wikipedia. For instance, Hjalmar Christensen is described in a Wikipedia article, but this article contains only one sentence with no listings of Christensen's work.



The main difference between the representation of information found in the national library catalog (also in description of products sold by e-commerce websites) and in the Wikipedia knowledge base is that often Wikipedia articles describe a resource at higher level of abstraction. As an example, one of the main sections in the Wikipedia article about Ibsen is listings of his works while in national library catalog we obtain a list of results mentioning Ibsen in descriptive metadata.

A first step towards a solution would be to expressing existing metadata describing potentially useful resources in RDF [14]. However, this comes with limited advantages. Simply transforming a resource description to a corresponding RDF representation which is a syntactic approach, only makes this information available for tools unable to process and does not contribute to the machine-interpretation of the implicit meaning in the records. There are currently many formats in use and each would require a specific namespace to be able to distinguish between tags and codes that have different meaning in different formats.

[1] http://en.wikipedia.org/wiki/Ibsen

3. RESEARCH QUESTIONS

Given the research challenges described above, the main question this dissertation attempts to answer is:

How can we improve the reuse value of existing metadata when making it available as semantic metadata?

This main question calls for solving several challenges, which are described in the following decomposed research questions:

- *How can we recognize and identify entities and relationships described implicitly in an existing metadata?* This question basically deals with the interpretation of a correct set of entities that is described in a metadata including the type of entity. We are looking for entities and types that are defined in a model. In order to solve this issue, a correct set of attributes must be chosen that indicate uniqueness of an entity as often entities in metadata are identified by the descriptive text (e.g., combining the title and the author of a book). Having recognized an entity, we need to determine if two or more entities in the same collection or in another are the same. Additionally, we need to identify the type of the relationship and the target entity or entities for each relationship.
- *What techniques can be employed to improve the results of a transformation?* Since entities in metadata are identified by description, we address the inconsistencies that will cause errors such as falsely identified entities or relationships, missing entities or relationships, etc. To tackle this issue, we need to discover potential errors and verify the correctness of a transformed metadata, a verification step may be needed, since potential errors may only become apparent after applying transformations.
- *How can we evaluate the quality of semantic metadata?* Detecting quality issues in semantic metadata is of significant importance to ensure high quality metadata. To evaluate the quality of transformed semantic metadata, our approach is based on adapting and using existing quality metrics, e.g. from conceptual model domain [13].

4. RESEARCH OBJECTIVES

The main objective of this thesis is to improve the quality of existing metadata by transforming it to a semantic format and addressing the problems inherent in the approach such as implicitness of information, verification and correction and quality. The ultimate goal is to enable semantic aware services such as explorative search. Note that our focus is on enabling this kind of services rather than the service itself. Therefore, the main research objectives can be summarized as follows:

- to explore and understand the requirements for semantic metadata to enable semantic aware services;
- to propose a method for automatic recognition and discovery of equivalent and related entities in local context and in external collection(s);

- to explore techniques and solutions for automatic discovery of relationships from an ambiguous information source;
- to propose techniques and solutions for verification and correction of semantic metadata;
- to evaluate semantic metadata by adapting existing methods and metrics.

5. RESEARCH METHODOLOGY

Each question this PhD project is trying to find an answer to consists of different parts that together form a method that is similar to design science methodology [9]. This methodology includes a systematic process that involves acquiring the required background knowledge, identifying challenges, formulating hypothesis, finding evidences to the hypothesis, performing an experiment, and analyzing the results. As a case, we are looking at bibliographic records as well as FRBR as a domain specific semantic model.

6. STATE OF THE ART

The main focus of this project is to improve the reuse value of existing metadata by bringing existing metadata into a format that enables innovative semantic aware services. These services, such as semantic search, enable better support for exploiting the potential value of existing metadata. However, simply transforming existing metadata into a semantic format only is the first step, as it does not automatically enable semantic services. The transformed semantic metadata must be reinterpreted, verified and corrected in order to achieve better results. To reach this goal, our idea is to use entity matching techniques for the verification of entities discovered in existing metadata across different data sources (e.g. catalogs or databases) as well as knowledge bases (e.g. Wikipedia or Freebase). Consequently, this work is at the edge of several research fields as illustrated in Figure 2.

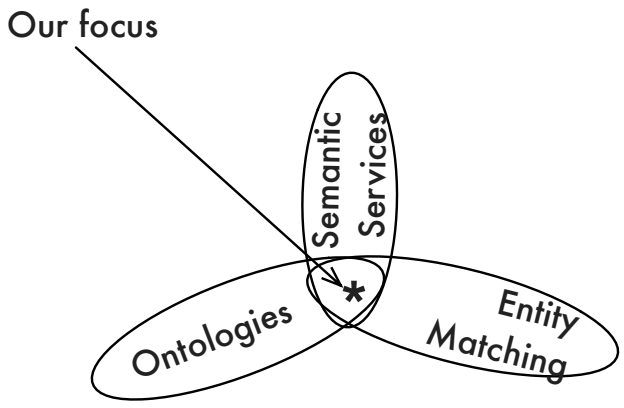


Figure 2. Combining Ontologies and Entity Matching to enable Semantic Aware Services.

6.1 Domain Models and Ontologies

A conceptual domain model can be described as an abstraction of reality according to a certain conceptualization [6]. This abstraction usually consists of an abstract representation of specific aspects of domain entities, referred to as concepts at a general level. When a conceptual model is represented as a concrete artifact, it facilitates the communication and analysis about important aspects of the domain concerned. A conceptual model often serves as a vehicle for reasoning and problem solving as well as for acquiring new knowledge about a domain. To be able to use it, the conceptualization should be expressed in a specific representation: a "language", a clearly defined specification, etc. This expression must be unambiguous and help users solve a real world problem. An example of a reference model is the CIDOC CRM [4], a high-level reference model to enable information integration for cultural heritage data and their correlation with library and archive information. The CIDOC CRM may be used for analyzing and designing cultural information systems. Another example is the FRBR model, which was initially intended as a conceptual framework for bibliographic data, but the report gives a detailed description of entities, relationship and attributes that may be used to define type-vocabularies. One of the first RDF vocabularies based on the model was published in 2005[2].

Closely related to conceptual modeling, a somewhat different approach to capturing "knowledge" has been taken from the artificial intelligence domain, which is called ontologies [18]. The distinction between conceptual models and ontologies is not clearly defined, however as a general assumption ontologies are used to *precisely* represent concepts and the main focus is on discovering new knowledge by means of automated reasoning and formal semantics. Ontologies are one of the main building blocks of the Semantic Web [1] and often defined as a formal, explicit specification of a shared conceptualization of a domain [7]. Conceptualization conveys an abstract model, while explicit means that the elements must be clearly defined. The formal property indicates that the specification should be machine processable. An ontology is the representation of the knowledge of a domain, where a set of objects and their relationships is described by a vocabulary. More specifically, a domain ontology defines a set of entities and relationships which is helpful for modeling a domain with a common vocabulary.

One of the areas ontologies are often used is integration of information from heterogeneous sources to meet the demand of complete access to available information [21]. In order to achieve semantic interoperability in a heterogeneous information system, the meaning of the interchanging information must be made explicit. The heterogeneity nature requires the integration process to tackle problems with existing metadata, which can be implicit/ambiguous since the quality of metadata in different sources may vary extensively which may result inconsistency during interpretation. The success of the Semantic Web to a great extent depends on easy creation and management of semantic metadata, as pointed out by [19]. A few works have been dedicated to tackle the problem of creating a semantic metadata

[2] http://vocab.org/frbr/core.html

from existing information, such as [3,19]. The main idea is to bring existing resources from unstructured or semi-structured, static or dynamic and a local database or Web-based resources to a more ontology-based representation often stored as RDF triples.

The approach in this project differs from the ones mentioned above in several ways. First, by understanding the requirements to semantic metadata we address the potential problems in the early stage to avoid them as much as possible. Second, we focus on the inconsistencies of the results of transformations to a semantic metadata and verification of the quality of the semantic metadata. Third, we study the application of existing evaluation metrics from conceptual modeling and data integration fields to measure the quality of semantic metadata.

6.2 Entity Matching

One of the important and challenging tasks in data integration and data cleaning is entity matching (also referred to as duplicate identification, record linkage, entity resolution). The main goal is to identify equivalent entities (objects, instances). This is particularly difficult if the data is of heterogeneous character or of limited (poor) quality. Table 1 depicts three duplicate entries and the task of entity matching system is to identify these entities that correspond to the same real-world publication.

Title	Author	Venue	Year
The merge/purge problem for large databases	M.A. Hernandez, S.J. Stolfo	Proceedings of the ACM SIGMOD international conference	
The Merge/Purge problem for Large Databases	A.H. Mauricio, J.S. Stolfo	Proc. of the 1995 ACM SIGMOD conference on management	1995
The merge/purge prob- lem for Large Databases	M. Hern	Proceedings of the 1995 ACM SIG-MOD conference on management	1995

Table 1. An Example of Multiple Reference to the Same Real-world object. Example is taken from [11].

As can be seen from the example, there are variations in the way information appears in these three entries. The ultimate goal of any entity matching strategy is to decide whether these three entries correspond to the same real-world entity. Entity matching has its roots from the database community as a common task of the migration of legacy data from multiple sources into a new one [5].

The problem of entity matching can be formalized as follows. Given the sets of entities $A \in E_a$ and $B \in E_b$ a particular entity matching approach must find all corresponding entities $A \times B$. The result of matching, sometimes called mappings, is often assigned a similarity score s to a particular correspondence, which is between [0,1].

Transforming metadata from a representation based on a legacy data structure to a semantic format is in no way a trivial task and the metadata is often subject to false interpretation due to data quality problems. Therefore, an interesting application of entity matching is to help us understand the degree we are able to verify the correctness of this transformed semantic metadata. To accomplish this task, a general assumption is to compare the set of properties of local entities against an external data source, which is typically a knowledge base such as DBpedia or Freebase.

6.3 Semantic-aware Services

Better exploitation of existing metadata can potentially lead to many benefits [20]. A general approach to achieve this goal is to interpret entities and relationships in an existing metadata and make the semantics explicit, which will consequently enable semantic reuse and integration. Examples of this kind of services are semantic and explorative search, which will be described in the rest of this section.

One of related techniques to our work is the emerging field of semantic search. The main goal of semantic search is to improve search accuracy by understanding the intent of the searcher expressed as a query and context [15]. To calculate relevance, a semantic search technique makes use of semantics to produce highly relevant results [2]. Semantic search approaches employ several techniques based on searching triple-stores (data encoded in RDF/OWL), natural language processing, clustering and classification, text-mining [22] and ontology-based search [16].

Our approach is related to semantic search but there are several factors that differentiate our work from semantic search. One of the main problems we are focusing is the quality of semantic metadata. While application of this metadata is important during indexing stage of semantic search, the quality of this metadata has profound impact on how this index is constructed and searched. Another difference is that we address the problem of ambiguous entities after the transformation process by correcting and enhancing semantic metadata. Furthermore, we apply techniques to verify that the enhancements are actually correct.

7. EXPECTED CONTRIBUTIONS

The expected contributions of this work are based on the results of outcome of the research questions. In addition to general understanding of the requirements for semantic metadata, this dissertation will contribute with improving the quality of transformation and migration of existing metadata.

- recognition and discovery of equivalent entities in the local and external collections: one of the main problems with metadata is that it consider- ably differs from the structures represented using relational databases. As an example, in database tables, there are foreign key constraints that can “strongly” link two tuples from separate tables using the primary key of a table, while with metadata we often need to construct identity of an entity from a descriptive text (e.g., using titles/names as key);

- relationship discovery for entities with ambiguous semantics. The practice of metadata management differs and varies greatly in terms of quality. Consequently, entities are often represented in the metadata but with implicit semantics. The large body of existing metadata requires a post-transformation effort that will discover the correct relationships for entities with ambiguous semantics;
- verification and correction of semantic metadata. Transforming existing metadata into a semantic format is the first step towards a shared semantic model. The subsequent step involves assessing and improving the quality of this se- mantic metadata. An automatic verification process must therefore perform several operations to achieve this goal, i.e., lookup the entity in the local and external sources, measure the similarity based on attributes of an entity, etc;
- evaluate the quality of semantic metadata. To evaluate the quality of semantic metadata, we consider applying existing metrics to assess and various techniques (e.g. verification on LOD) to improve the quality of the semantic metadata.

8. PRELIMINARY RESULTS

We have designed a format for exchange of MARC-based information that makes the entities and relationships of the FRBR model explicit [17]. Furthermore, a substantially improved version of this framework has been proposed and is currently under submission. As a result, the RQ1 (see Section 3) has partially been addressed and in order to completely solve it, we have to perform more experiments.

We are currently working on challenging task (RQ2), which includes verification of entities represented in the transformed data. Our assumption is that by discovering equivalent entities in other sources, we can improve the quality of existing metadata. As an example, if entities in the original record have missing attributes or have ambiguous semantics, this missing information can be fetched from other sources and the enhanced version of the record will be obtained as a result of the transformation. The techniques we are experimenting with is discovering entities in the LOD cloud by applying semantic matching techniques on entities' attributes as well as looking up other data sources via specific protocols (e.g., z39.50).

As far as the RQ3 is concerned, we plan to apply existing metrics to evaluate the quality of semantic metadata. Our preliminary study shows, existing data integration metrics such as *minimality* or *completeness* cannot be directly applied in entity-oriented transformation scenario. In such a case, we are typically interested in evaluating the quality of the transformation (e.g. Does the entity have all of its attributes? Have all of its relationships been discovered correctly? etc.).

9. CONCLUDING REMARKS

A vast amount of metadata related to cultural items has been created by memory institutions in the last decades using different standard or ad hoc schemas, and a main challenge is to make this existing metadata accessible as reusable semantic metadata. This metadata is already thoroughly documented by memory institutions and to make this metadata available as reusable semantic metadata, it must be interpretable in a context that makes sense to machines. Simply transforming the data does necessarily enable semantic interoperability as the metadata must be reinterpreted as well as transformed. Additionally, the transformation typically introduces data quality problems, which calls for evaluation methods applicable in this scenario. These issues will be addressed in this project.

10. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. Scientific American, 33(4), 2001.

[2] T. Brasethvik. Conceptual modeling for domain specific document description and retrieval - An approach to semantic document modeling. PhD thesis, Norwegian University of Science and Technology, 2004.

[3] V. Crescenzi, G. Mecca, P. Merialdo, U. Roma, T. Università, B. Università, and R. Tre. Roadrunner: Towards automatic data extraction from large web sites. pages 109–118, 2001.

[4] M. Doerr and P. L. Boeuf. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine. Association for the Advancement of Artificial Intelligence, 24(3), 2003.

[5] C. Drumm, M. Schmitt, and E. Rahm. Quickmig - automatic schema matching for data migration projects. In Proc. of the Sixteenth Conference on Information and Knowledge Management (CIKM, 2007.

[6] F. T. Fonseca and J. E. Martin. Learning the differences between ontologies and conceptual schemas through ontology-driven information systems. J. AIS, 8(2), 2007.

[7] T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, 1993.

[8] J. A. Gulla, H. O. Borch, and J. E. Ingvaldsen. Contextualized Clustering in Exploratory Web Search. In Emerging Technologies of Text Mining. Techniques and Applications, page 184. IGI Global, 2008.

[9] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. MIS Quarterly, 28(1):75–105, 2004.

[10] E. Hyvönen. Semantic portals for cultural heritage. In P. Bernus, J. Błażewicz, G. Schmidt, M. Shaw, S. Staab, and R. Studer, editors, Handbook on Ontologies, International Handbooks on Information Systems, pages 757–778. Springer Berlin Heidelberg, 2009.

[11] H. Köpcke and E. Rahm. Frameworks for entity matching: A comparison. Data Knowl. Eng., 69:197–210, February 2010.

[12] D. Koutsomitropoulos, G. Solomou, and T. Papatheodorou. Metadata and Semantics in Digital Ob ject Collections:

A Case-Study on CIDOC-CRM and Dublin Core and a Prototype Implementation. Journal of Digital Information, 10(6), 2009.

[13] J. Krogstie, G. Sindre, and H. D. Jørgensen. Process models representing knowledge for action: a revised quality framework. EJIS, 15(1):91–102, 2006.

[14] D. A. Rob Styles and N. Shabir. Semantic MARC, MARC21 and the Semantic Web. In Linked Data on the Web (LDOW2008), Bejing, China, 2008.

[15] G. Solskinnsbakk and J. A. Gulla. Combining ontological profiles with context in information retrieval. Data Knowl. Eng., 69:251–260, March 2010.

[16] D. Strasunskas and S. Tomassen. On variety of semantic search systems and their evaluation methods. In Proceedings of the International Conference on Information Management and Evaluation (ICIME 2010), pages 380–387, Cape Town, South Africa, March 2010. Academic Conferences Publishing.

[17] N. Takhirov, T. Aalberg, and M. Žumer. An XML-based representational document format for frbr. In 1st International Symposium on Web Intelligent Systems and Services. Springer, 2010.

[18] M. Uschold and M. Gruninger. Ontologies: principles, methods and applications. The Knowledge Engineering Review, 11(2):93–136, 1996.

[19] R. Volz, S. Handschuh, S. Staab, L. Stojanovic, and N. Stojanovic. Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the semantic web. Journal of Web Semantics, 1(2):187–206, 2004.

[20] G. Vossen, M. Lytras, and N. Koudas. Editorial: Revisiting the (machine) semantic web: The missing layers for the human semantic web. IEEE Trans. on Knowl. and Data Eng., 19:145–148, February 2007.

[21] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hu ̈bner. Ontology-based integration of information - a survey of existing approaches. In IJCAI-01 Workshop: Ontologies and Information Sharing, pages 108–117, Seattle, WA, USA, 2001 2001.

[22] W. Wei and J. A. Gulla. Sentiment learning on product reviews via sentiment ontology tree. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 404–413, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[23] J. Yuan, A. Bahrami, C. Wang, M. Murray, and A. Hunt. The a semantic information integration tool-suite. In Proceedings of the 32nd International Conference on Very Large Databases (VLDB), Seoul, Korea, September 12-15, 2006. Jos ́e.