Use of Descriptive Metadata as a Knowledgebase for Analyzing Data in Large Textual Collections

Dharitri Misra, George R. Thoma, National Library of Medicine, Bethesda, Maryland, USA

Abstract

Descriptive metadata, such as an article's title, authors, institutional affiliations, keywords and date of publication, collected either manually or automatically from documents contents, is often used to search and retrieve relevant documents in an archived collection. This metadata, especially for a large text corpus such as a biomedical collection, may encapsulate patterns, trends, and other valuable information, usually revealed by using specialized data analysis software to answer specific questions. A more useful, generalized approach is to repurpose this metadata to serve as a knowledgebase to answer appropriate semantic queries

At the US National Library of Medicine (NLM), we recently archived a large biomedical collection comprising annual conference proceedings containing research findings on cholera, conducted between the years 1960-2011 under the "US-Japan Cooperative Medical Science Program" (CMSP). This program was established to address health problems in Southeast Asia and other developing countries. An R&D information management system developed at NLM, called "System for the Preservation of Electronic Resources" (SPER), automatically extracted descriptive metadata from this text corpus and built a DSpace-based archive for accessing the conference articles. SPER also used this metadata to get detailed information regarding the CMSP research community, timelines of important drugs and discoveries and international collaboration, etc., using special purpose data analysis software.

In this paper, we describe the occurrence and extraction of metadata from the CMSP document set, and present an alternative approach in which this metadata is used to build a knowledgebase to support semantic queries about the CMSP Program. Specifically, we show the OWL-based hierarchical ontology model created to represent the CMSP Program with its publications, participants and international collaboration over time. We discuss the technique used to convert the extracted metadata from relational database tables to OWL/RDF assertions suitable for supporting semantic queries. We show examples of queries performed against this CMSP knowledgebase, and discuss some scalability issues. Finally we describe how this approach could be customized for other large textual collections, including one from the Food and Drug Administration previously archived by the SPER system.

Introduction

Digital repositories often archive large document collections on specific subjects, whose contents carry important facts about these subjects and thus may constitute a useful source of knowledge for those domains. However, such context-sensitive information is often either ignored, or used simply as "descriptive metadata" to search and retrieve individual items in the collection. A major reason for ignoring this type of metadata is rooted in the labor intensive nature of their identification, capture, and dissemination. However, when such metadata is available for a document corpus, it may be further used to create a searchable knowledgebase for the collection, revealing important patterns and trends – helping researchers gain insight into the field. This is deemed especially true in the biomedical domain, comprising massive knowledge in biomedical literature – with research articles, case studies and reviews, which often carry additional information related to research, policies, participants and the contemporary understanding of health science.

One such collection is from the Joint Cholera Panels of the U.S.-Japan Cooperative Medical Science Program (CMSP), a joint commitment by the United States and Japan, founded in 1965 and continued till 2011 - to address health problems in Southeast Asia and other developing countries through an expanded, collaborative international medical research effort [1]. The complete set of publications of these Panels comprise conference proceedings from 1965 to 2011 (plus an earlier one from 1960) with research articles, lists of panelists and attendees, additional annual reports, as well as separate lists of reviewers overseeing the CMSP Program. An important goal of archiving this collection is to create a knowledge source about the timelines of various cholera-related drugs and discoveries, the CMSP research community, and factors affecting the effectiveness of the program.

The CMSP collection, held by the National Institute of Allergy and Infectious Diseases (NIAID), was archived by an R&D information management system called the "System for the Preservation of Electronic Resources" (SPER) [2] developed at the US National Library of Medicine. Using machine learning techniques, SPER identified and automatically extracted relevant metadata from the digitized text of conference articles, and various lists of contributing personnel. While the article contents and related metadata were used to build a DSpace-compatible archive [3], the combined set was used to build a knowledgebase suitable for conducting specific data analysis.

In the following sections, we provide a background to the types and occurrence of metadata in the CMSP collection and their automated extraction by the SPER system. Then we discuss the creation of an ontology for the CMSP Program, and the process of transforming the stored metadata from a flat relational database to a hierarchical OWL/RDF [4] knowledgebase with this ontology to support semantic queries - using open source tools and in-house developed software. We display the results of certain queries performed using a RESTful Web browser, and discuss some issues related to performance. Finally we outline how this approach could be customized for other large datasets, including a collection of historic medico-legal documents ("Notices of Judgment") from the FDA - labeled FDANJ [5] and previously archived by SPER, to access domain-specific knowledge contained in the dataset.

Background and Related Work

Knowledge extraction from structured sources in a machine readable/interpretable format, an area of active research, has benefited in recent years with the publication of several standards and availability of reliable open source tools [6]. The W3C specifications on Resource Definition Format, RDF [7], SPARQL query language [8] and Web Ontology Language, OWL [4] have facilitated the creation of knowledgebases and retrieval of information therein. Large relational databases, storing valuable information about various domains, may thus be transformed to Web accessible knowledgebases [9] for obtaining information not easily available otherwise in those fields. This mechanism offers an interesting avenue to make context-sensitive information in a document corpus, stored as "descriptive metadata" in relational database tables of digital archives, accessible for gaining further knowledge in corresponding fields.

However, it is often non-trivial and prohibitively expensive to manually acquire potentially useful context-sensitive metadata. We developed the SPER system to identify, locate and extract such metadata cost-effectively from the contents of semi-structured text using machine learning. SPER has been used earlier to perform automated metadata extraction (AME) and to archive the FDANJ collection, and recently, the CMSP collection. To meet the needs of NIAID for quantitative analysis of CMSP, SPER also extracted additional metadata from the CMSP document set to determine patterns and trends related to cholera, vaccine developments, therapies, and the characteristics of its research community - using special data analysis software.

Furthermore, it was deemed useful to transform the metadata in the CMSP archive to a knowledgebase for semantic query by researchers and policy makers interested in CMSP activities. This was accomplished by using the knowledge extraction techniques mentioned above, implemented by developing a pipeline process using selected open source tools. A prototype Web application was also created to receive query requests, perform semantic search on the knowledgebase, and return the results in graphical form. The techniques and tools, used for the CMSP dataset and extensible to others, are discussed in the following sections.

Metadata Extraction from CMSP Document Corpus using SPER

CMSP Metadata

The CMSP document corpus consists of annual conference proceedings in the form of short presentations and full papers (both types referred to here as articles) on cholera-related research, a set of five-year annual reports discussing the overall progress in the field, and a roster of Study Section Reviewers responsible for reviewing and funding different research areas over the years. The metadata required to perform an analysis of the CMSP program, determine the degree of international research collaboration, and identify the most active cholera researchers are in the following three categories:

1. Publication metadata (Article level) - titles, authors, institutions, and subject keywords from research articles.
2. Investigator metadata (Conference level) - name, role, designation and affiliation of panelists and attendees from the conference proceedings rosters.
3. Study Section metadata (CMSP Program level) - names and affiliations of CMSP Program reviewers from separate Study Section rosters.

Table 1 presents statistics related to the CMSP corpus. The term "instances" refers to all occurrences (not necessarily unique values) of a specific data type in the document set.

Table 1: Statistics related to CMSP Document Set (1960-2011)

Number of Conference Proceedings	57
Number of Articles	2,812
Instances of Author	13,437
Instances of Panelist	610
Instances of Conference Attendee	4,723
Instances of Participating Institution	4,416
Instances of Study Section Reviewer	3,110

Automated Metadata Extraction using SPER

Because of the large number of metadata elements to be captured for the CMSP collection, SPER was used to automatically identify and extract metadata from the OCR'ed text of the document pages. Since the CMSP documents were generated over a span of more than 50 years, the format, layout, font and legibility of their contents varied widely. Some examples of the metadata layouts are given below. Each box in Figure 1a shows the location of metadata fields in the title page of an article; Figure 1b shows location of contributors (panelist/attendee/reviewer) in rosters.

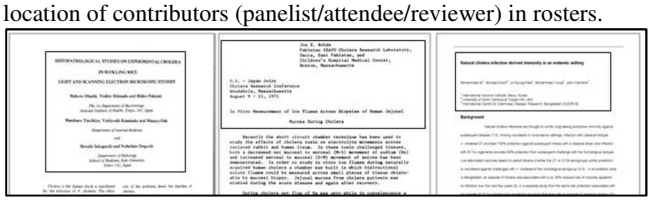


Figure 1a. *Metadata location in title pages in sample CMSP Articles*

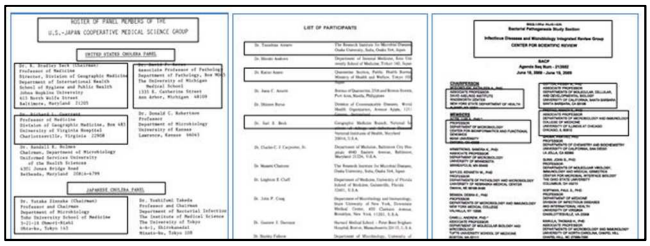


Figure 1b. *Metadata location in sample CMSP Rosters*

SPER used keyword matching and layout analysis techniques to identify different types of documents from the OCR'ed document pages. Three metadata models were developed, using a combination of Support Vector Machine [10] and Hidden Markov Models [11], to handle AME for articles, panelist/attendees and Study Section reviewers respectively. These models were then used to classify the text lines in each page by recognizing the *named entities* such as a person's name, address or affiliation, and determine the bounding box for each item. Regular expression matching and gazetteer look-up was used to identify individual metadata elements (article title, author name, institution, address etc.) within a bounding box. Once the metadata elements for a

batch of items were extracted through AME, they were reviewed and validated by an operator to correct for poor document quality and model errors, if any. The validated metadata for the CMSP articles was then stored in a MySQL database. This metadata, along with the scanned images and PDF derivatives of the articles, was used to build a DSpace-based repository with standard search/retrieval capability for individual articles.

Metadata Post-processing

Post-processing was needed to clean up the metadata and store certain static information to support analysis of the CMSP Program. The steps were:

a) Disambiguation of investigator/reviewer names so as to uniquely identify the contributions of a single individual over the years – especially difficult since the same name was expressed in many different ways in the documents.
b) Determining the country of an institution when not explicitly specified.
c) Building a static table assigning a "group number" to each country, based upon its gross domestic product or GDP - to assess international collaboration by such groups. (For example, all developing countries have group number 3.)

This updated relational database was then used, by customized data analysis modules, to determine various patterns and trends useful for assessing the success of CMSP. This has been presented in a separate paper [12].

Development of a Knowledgebase from the CMSP Metadata

Modeling and Generating a Knowledgebase

Creation of a knowledgebase (KB) from structured or unstructured data related to a domain is a multi-step process. The sequential steps we have followed to generate a KB from a relational database are shown in Figure 2a, with the numbered boxes 1-4 showing the resultant data structure after each step. A brief description of these processing steps is given below.

1. **Developing the Domain Ontology** - The first and most critical step is the selection of entities, their attributes and relationships, and the associated rules that would form the foundation of the KB. The relations may be hierarchical where one entity is a subtype of another (parent) entity. This conceptual design is then transferred to a machine readable and machine-interpretable form, along with the rules and restrictions pertaining to the entities and their relations, and is called the *ontology* (or informally, the *taxonomy*) for the domain [13]. The ontology for a domain, therefore, provides an explicit specification of conceptualization for that domain. It is often expressed using the W3C OWL specification and referred to as an *ontology model*.
2. **Generation of the RDF Graphbase -** In the next step, the corresponding relational database (RDB) is transformed to a representation consisting of RDF graphs or triples of *Subject, Predicate, and Object.* First, an RDF schema representing an Entity-Relation model [14] is generated from the corresponding RDB schema, using the RDB table and column names and their properties. Next, using this RDF schema, each row of an RDB table is converted to the corresponding set of RDF triples (Figure 2b), resulting in an Entity-Relation-based structure (with subjects and objects as the entities and predicates as the relations) known as an *RDF graphbase*.
3. **Creation of the Concepts database -** The third step transforms the triples in the RDF graphbase to a hierarchical form based upon the ontology model created in step 1, and in compliance with the rules and restrictions therein. This output structure, also expressed in RDF, is called the *Concepts* or *Assertion database,* where each graph constitutes an assertion.
4. **Adding the Inferences -** In the last step, additional RDF assertions are derived from the base assertions by applying inference rules in the ontology model through an inference engine or reasoner. This complete set of assertions then constitutes the knowledgebase for the specific domain. Note that this step is usually performed at runtime, when subsets of inferences, corresponding to individual queries, are generated dynamically. Generating all inferences statically offline and inserting them into the existing assertions could be both error-prone and expensive, especially for large, complex datasets.

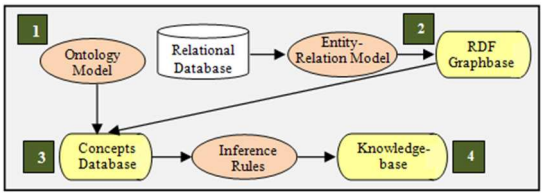


Figure 2a. *Transformation of a Relational database to a Knowledgebase*

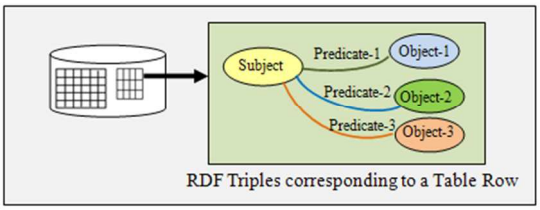


Figure 2b. *Conversion of a Relational database table row to RDF triples*

Building the CMSP Knowledgebase

The CMSP KB was created following the steps 1-4 discussed above. The three types of metadata (listed earlier), extracted from the CMSP document corpus and stored in the CMSP's DSpace-based repository, served as the critical input for this function.

However, some data reformatting was necessary as all metadata fields for the cholera articles were stored in one database table (coded with item IDs, metadata field IDs and their values) in the repository. This structure does not map directly to an entity-relation model, which requires each metadata field to be represented in its own column so that an explicit relationship between an article and that metadata field could be created and the corresponding triples (such as: *Article, hasAuthor, Author*) be generated for searches on that field. Hence, using an SQL script, a new MySQL database was created from the original CMSP database, with a restructured metadata table. Furthermore, only the subset of original tables required to build the CMSP KB were included in its schema. This streamlined database was called the *CMSP Entity-Relation (E-R) database*, since it directly converts to the corresponding RDF entity-relation graphbase.

Implementation of the Data Transformation Framework

The KB generation steps discussed above and depicted in Figure 2a are implemented through the Java-based framework shown in Figure 3. It consists of a set of processes, which operate in a pipeline fashion and execute steps 1-4 below, to transform the CMSP E-R database to the corresponding knowledgebase. The rectangular boxes in the figure represent these processes and the ovals correspond to the data they operate upon. These processes are built with reliable open source tools and in-house modules (for mostly domain-specific tasks) and compliant with W3C standards.

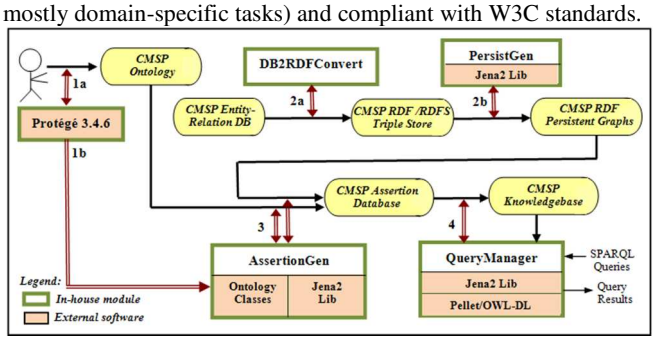


Figure 3. *Framework for building the CMSP Knowledgebase*

1) CMSP Ontology creation: This crucial step (1a) is performed manually, using *Protégé 3.4.6 Ontology Editor* [15], to create the CMSP ontology in OWL format, presented in the next section. It is followed by the generation of a set of Java classes (as Java Beans) corresponding to this ontology, using the *code generation* function in Protégé 3.4.6 (step 1b). These classes are used to build the CMSP OWL assertion database later in step 3.

2) CMSP Graphbase generation: This is a two part process; in the first one (2a), the Entity-Relation model (RDFS schema) and the RDF triples are created from the CMSP Entity-Relation database, using an in-house tool called *DB2RDFConvert*. This tool is derived from an open source tool, *DB2RDF* [16], modified to be scalable for large databases. Next, the RDF dataset is converted to persistent storage (as backend SQL tables), for faster performance, by *Jena2* [17] invoked from the wrapper module *PersistGen (2b).*

3) CMSP Assertion database creation: This is implemented by the in-house module *AssertionGen*, using lower level Jena modules and the CMSP ontology-specific Java classes created in Step 1b. Data is accessed from the persistent graphbase using Jena2 API, transformed into OWL-based structures using the Java classes and CMSP specific logic, and then stored as assertions in the output dataset in OWL format.

4) CMSP Knowledgebase generation and usage: At query time, the QueryManager module accesses the data in the OWL-based assertion dataset using Jena2. Jena2 supports several reasoners, and either *Pellet* [18] or OWL-DL may be used by the QueryManager with Jena2. Based upon the query, additional assertions are generated dynamically by the reasoner, creating a transient, memory-resident CMSP knowledgebase.

There are additional tools/features we found useful during the development phase. For example: using the Protégé 3.4 platform, one can test the validity of the assertion database by running a reasoner, and also by issuing SPARQL queries on it. Similarly, the SPARQL query server Fuseki (earlier name: Joseki) [19], may be used to query the assertion database from a Web browser.

CMSP Ontology Structure

The class hierarchy in the CMSP OWL ontology model, depicting the CMSP *Concepts*, is shown in Figure 4a, with the class "Thing" (superclass of all OWL classes) omitted for simplicity. Figure 4b displays main relationships between those classes, omitting superclass-subclass (:isa-a) relations.

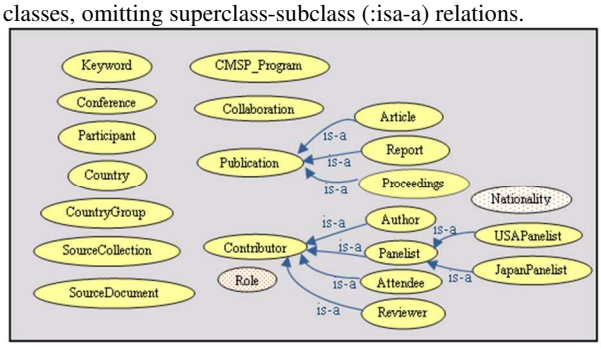


Figure 4a. *CMSP Concepts hierarchy as OWL Classes and Subclasses*

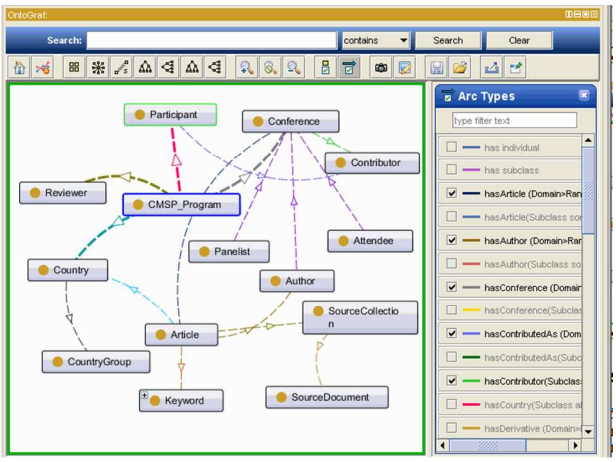


Figure 4b. *Relationship between main CMSP OWL Classes*

CMSP Semantic Queries and Results

Semantic queries were issued against the CMSP knowledgebase, described above, using SPARQL (V1.0). The text box below provides a simple example of a SPARQL query performed on the CMSP knowledgebase to retrieve information on all CMSP articles published by each country.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX cmsp: <http://nlm.ceb.sper/ontologies/cmsp-prog#>
SELECT ?article ?year ?countryName  ?countryGroup
FROM <file:/C:/CmspDataset/CmspConf_assert.owl>
WHERE {?article cmsp:hasParticipatingCountry ?country.
          ?article cmsp:hasPublicationYear ?year.
          ?country cmsp:hasCountryName ?countryName.
          ?country cmsp:isInCountryGroup ?countryGroup.
      } ORDER  BY ?countryName
```

The statement using the relationship cmsp:*hasParticipatingCountry* fetches all Articles and the set of Country objects associated with each Article, whereas cmsp:*isInCountryGroup* retrieves the CountryGroup object for a specified Country (corresponding to the country's assigned group number, explained under the section "Metadata Post-processing").

To make the CMSP knowledgebase accessible to users over the Web, a prototype Web application was created and run under Tomcat. The graphical rendering of query results was performed by using the PrimeFaces [20] library within the application.

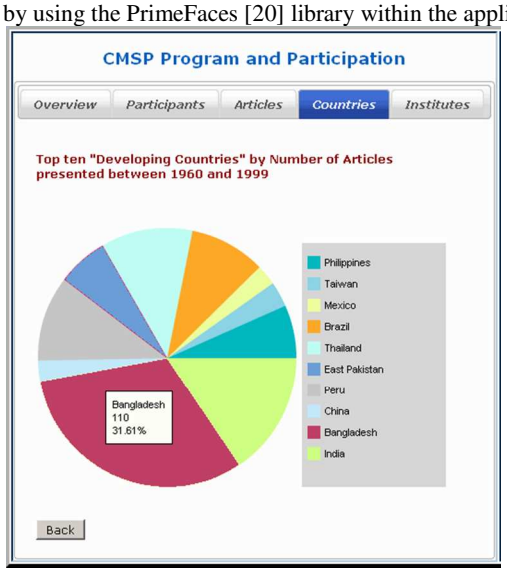
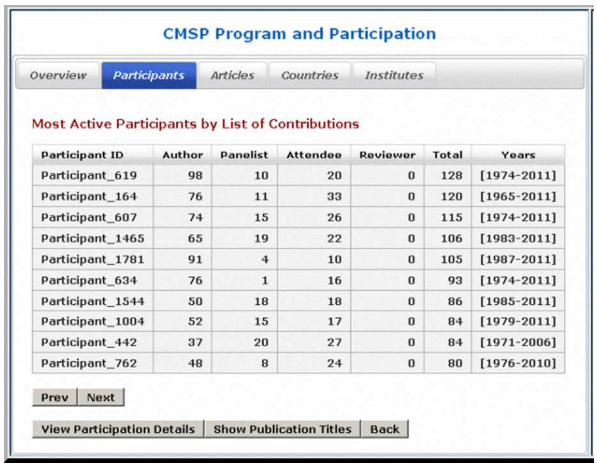


Figure 5a. *Web display of semantic query results on most active Countries*



Participant ID	Author	Panelist	Attendee	Reviewer	Total	Years
Participant_619	98	10	20	0	128	[1974-2011]
Participant_164	76	11	33	0	120	[1965-2011]
Participant_607	74	15	26	0	115	[1974-2011]
Participant_1465	65	19	22	0	106	[1983-2011]
Participant_1781	91	4	10	0	105	[1987-2011]
Participant_634	76	1	16	0	93	[1974-2011]
Participant_1544	50	18	18	0	86	[1985-2011]
Participant_1004	52	15	17	0	84	[1979-2011]
Participant_442	37	20	27	0	84	[1971-2006]
Participant_762	48	8	24	0	80	[1976-2010]

Figure 5b. *Web display of semantic query results on CMSP Participants*

The results of two sample queries, along the lines of queries done for the CMSP data analysis task mentioned earlier, are presented in Figures 5a and 5b. The first one indicates the research activity of the top ten countries in the "developing world" in the 20th century - highlighting international participation in the program during that period, as a pie chart. (The underlying SPARQL query is similar to the one shown in the example.) Similarly, Figure 5b shows the most active participants, their contribution in different roles and their periods of participation in a tabular form. (Note that we have displayed Participant IDs rather than their names since the CMSP repository is not yet in the public domain.) More specific queries, such as the development timeline of a particular vaccine, or the degree of collaboration between different country groups can be conducted using "Advance Query" forms, not shown here.

Performance and Scalability

The retrieval time for SPARQL queries against an OWL knowledgebase is dependent upon the number of assertions in the dataset as well as the structure of the query itself – although in general it is slower than equivalent SQL queries against a relational database. However, efficiency may be improved under a newer release of SPARQL (e.g. V1.1) with subqueries and count features, and other optimizations [21].

In the prototype version of our CMSP Web application (which uses SPARQL V1.0), no effort was made to improve the retrieval time though special query optimization. Nevertheless typical retrieval times for some queries (conducted on a developmental Windows XP computer with a 2.67 GHz CPU) are presented in Table 2. Note that the last entry involves ten separate SPARQL queries to the KB to obtain the desired result set.

Table 2. Retrieval Statistics for CMSP Knowledgebase

Number of instances of all classes	65,727
Number of statements	409,689
Number of CMSP Participants	7,853
Time to initialize OWL model for querying	15.53 sec
Overhead of each query to the model	0.31 sec
Time to retrieve all countries for all Articles (# of instances: 3597)	0.54 sec
Time to retrieve all contribution data for all Participants (# of instances: 21,829)	1.02 sec
Time to retrieve contributions of 10 selected Participants by role (# of instances:1001)	3.86 sec

It is expected that benchmarks on a "server quality" machine would improve performance by an order of magnitude, and storing certain static data to minimize number of queries for a search would enhance retrieval speed. However, scalability could still be an issue for very large datasets. In such cases other alternatives, such as converting SPARQL queries to SQL [22] may be pursued.

Another scalability concern is related to the performance of different reasoners in dealing with large datasets. For example, we encountered memory problems in using Pellet against the full CMSP dataset, while it worked fine for smaller test sets.

Application to Other Collections

The technique of converting the context-sensitive metadata of a text collection to a knowledgebase, described in this paper, may be extended to other collections to discover patterns and trends, independent of the structure of such metadata. The components that need to be developed specifically for each collection are:

a) Domain-specific ontology in OWL representing the dataset.
b) Java classes corresponding to the ontology, usually produced by a code generator such as Protégé 3.4.
c) The module to interface with Jena2 (in Figure 3) to convert the RDF graphbase to an OWL-based assertion database.

The FDANJ Collection

As an example, we discuss below the specific case of the FDANJ collection [5] archived by SPER, with a single metadata category (as opposed to three for CMSP). It comprises a set of 70,000 published Notices of Judgment (NJ) on court cases for

adulterated and misbranded foods, drugs, and cosmetics, released by the FDA between 1906 and 1964. The metadata fields include:

- Case title, product name, issue (publication) date
- Defendant name
- Adjudicating court where the case was prosecuted
- Seizure date and locations
- Locations from where the product was shipped, and to where it was being shipped

The ontology model developed to describe the relation between an NJ and its metadata elements is shown in Figure 6.

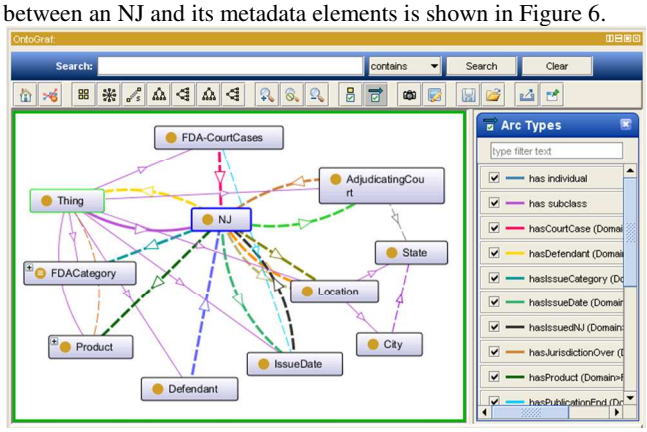


Figure 6. *Relationship between FDANJ OWL Classes*

The FDANJ knowledgebase built from its metadata set, using the procedure discussed in this paper, would be valuable in understanding the types of foods/drugs that were most often misbranded and/or adulterated in the USA in earlier periods, the routes of illegal inter-commerce trades, peak periods of such illegal activities, traders/drug companies involved in those activities, and the courts where such cases were prosecuted..

Conclusion

In this paper, we have shown the usefulness of context-sensitive descriptive metadata from large textual collections in revealing important facts about a collection, not generally available otherwise – with the CMSP collection as a specific example. We have outlined how the CMSP metadata was located and extracted from the contents of the documents in a cost-effective manner using machine learning. We have discussed our pipeline process and useful public-domain tools in transporting the metadata from a relational database to a knowledgebase for conducting semantic searches to find useful patterns and trends. Finally, we have discussed how this process could be applied to other collections to find useful domain related information.

Acknowledgement

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

We also thank Dr. Robert Hall and Ms. Susan Payne of NIAID for their collaboration in processing the CMSP documents and conducting the data analysis.

References

[1] The U.S.-Japan Cooperative Medical Science Program (http://www.niaid.nih.gov/topics/globalResearch/region/eastAsiaPacific/usjapan/Pages/history.aspx).

[2] D. Misra, S. Mao, J. Rees, G. Thoma, Archiving a Historic Medico-legal Collection: Automation and Workflow Customization, Proc. IS&T Archiving Conference, Washington DC, pg 157-161. (2007).

[3] DSpace, MIT (http://www.dspace.org).

[4] OWL Web Ontology Language Overview (http://www.w3.org/TR/owl-features/).

[5] FDA Notices of Judgment Collection, 1906-1964 (http://archive.nlm.nih.gov/fdanj/).

[6] LOD2: Report on Knowledge Extraction from Structured Sources (http://static.lod2.eu/Deliverables/deliverable-3.1.1.pdf).

[7] RDF Primer (http://www.w3.org/TR/2004/REC-rdf-primer-20040210/).

[8] SPARQL Query Language for RDF (http://www.w3.org/TR/rdf-sparql-query/).

[9] Database-to-Ontology Mapping Generation for Semantic Interoperability (http://www.academia.edu/181720/Database-to-Ontology_Mapping_Generation_for_Semantic_Interoperability).

[10] C.Cortes, V. Vapnik, Support-vector Network. Machine Learning. Vol. 20, pages 273-297, (1995).

[11] L.R. Rabiner, B.H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall. (1993).

[12] D. Misra, R.H. Hall, S.M. Payne, G.R. Thoma, Digital Preservation and Knowledge Discovery Based on Documents from an International Health Science Program, Proc. 12th ACM/IEEE-CS JCDL, pg 23-26 (2012).

[13] Ontology (http://www-ksl.stanford.edu/kst/what-is-an-ontology.html).

[14] P.P. Chen, MIT. The Entity-Relationship Model, Toward a Unified View of Data (http://csc.lsu.edu/news/erd.pdf).

[15] Protégé, Stanford Center for Biomedical Informatics Research (BMIR) at the Stanford University School of Medicine (http://pretege.stanford.edu).

[16] DB2RDF: (http://db2rdf.sourceforge.net/).

[17] Apache Jena (http://jena.apache.org/about_jena/about.html).

[18] Pellet: A practical OWL-DL reasoner (http://www.sciencedirect.com/science/article/pii/S1570826807000169).

[19] Fuseki: serving RDF data over HTTP (http://jena.apache.org/documentation/serving_data/index.html).

[20] PrimeFaces: Next Generation Component Suites (http://code.google.com/p/primefaces/).

[21] SPARQL 1.1 Overview (http://www.w3.org/TR/2012/PR-sparql11-overview-20121108/).

[22] Translating SPARQL queries into SQL using R2RML (http://antoniogarrote.wordpress.com/2011/01/10/translating-sparql-queries-into-sql-using-r2rml/).

Author Biography

Dharitri Misra is a Staff Scientist at the U.S. National Library of Medicine. Her work focuses on digital preservation of biomedical collection and information extraction from textual data, including development of frameworks and tools to support such work. She received her M.S. and Ph.D. degrees in Physics from the University of Maryland.

George R. Thoma is a Branch Chief at an R&D division of the U.S. National Library of Medicine. He directs R&D programs in document image analysis, biomedical image processing, animated virtual books, and related areas. He earned his B.S. from Swarthmore College, and his M.S. and Ph.D. from the University of Pennsylvania, all in Electrical Engineering. Dr. Thoma is a Fellow of the SPIE, the International Society for Optical Engineering.