Enabling semantic search in a bio-specimen repository

Shahim Essaid* Carlo Torniai and Melissa A. Haendel
Oregon Health & Science University, Portland, OR, USA

ABSTRACT

We present our effort to enhance a bio-specimen repository search application with semantic search capability. We describe the nature of the original data, the application of text processing tools, and the leveraging of existing terminologies and ontologies to build an application ontology that supports the bio-specimen search system. We also describe few of the difficulties we have encountered and possible ways for addressing them in our future work.

1 INTRODUCTION

Biomedical research relies on the use of human tissues for the analysis of the pathogenesis and genetic basis of human disease. Translational pathology has been revolutionized by the explosion of genetic information available, by the development of high-throughput platforms to examine gene expression, and by new bioinformatics tools. Numerous experimental models for the molecular profiling of disease are available, including animal models, human cell lines, and tissue specimens. To date, there are few mechanisms to relate biospecimens to this wealth of information, making searching for them very difficult. In order to take best advantage of these specimens and models of disease, it is imperative that we improve our ability to locate such resources. Modern semantic technologies can be used to link biospecimens via anatomy, disease, phenotype, genes and gene expression.

In this paper we report on ongoing effort to enhance the functionality of an existing bio-specimen repository search application, the "Biolibrary" at Oregon Health & Science University. The Biolibrary application aggregates raw data from local bio-specimen banks, performs simple preprocessing and indexing, and provides a web-based search interface to help researchers find relevant bio-specimens.

There are several limitations in the current Biolibrary application that include: the high variability of the data structure depending on its source, different encodings for similar data values, simple string matching for free text search over a large portion of the data, and the complexity of the application development due to the maintenance of a separate search UI for each data source.

The goal of our work was to analyze the existing data sources and investigate how a semantic ontology-based approach would improve the capabilities of the Biolibrary search application with a longer term goal for supporting more sophisticated classification and retrieval of bio-specimens. In this paper we present the results of our preliminary work, discuss the issues we have encountered thus far, and identify steps for our next iteration.

2 METHODS AND ANALYSIS

2.1 Overview of the initial data sources

At the time we started our project there were two main sources of bio-specimen data in the Biolibrary: the Department of Pathology repository and the Knight Cancer Institute repository. Additional data was obtained from a cancer registry to provide additional metadata for the specimens recorded in the two main sources.

The data from the department of pathology is in the form of free text for the final diagnosis section of pathology reports and it contains approximately 600,000 reports. This data also assigns one ICD-9 code per entry that represents the main diagnosis. This minimal level of coding is a significant limitation for searching these records and therefore the search application relies on a text index to perform free text searches of these records. This dataset had been augmented in the Biolibrary with additional metadata from the cancer registry by matching medical record numbers and dates of service. However, the additional coded data still did not fully capture the richer nature of information present in the free text. For example, one record was coded as:

- Malignant neoplasia of pancreases NOS
- Duodenum, normal tissue
- Stomach, normal tissue
- Pancreases, normal and adenocarcinoma

However, this pathology report also included finer grained descriptions such as:

- Low grade pancreatic intraepithelial neoplasia
- Extensive perineural invasion
- Acute and chronic cholecystitis
- Bile duct tissue with chronic inflammation
- Chronic pancreatitis
- Acute gastric serosisties

These additional specimen features are only accessible by free text search that does not accommodate synonymy or other syntactic variations. Also, any additional data from the cancer registry was only relevant for the malignant pathology specimens.

[*] To whom correspondence should be addressed: essaids@ohsu.edu

The data from the cancer center was more structured but it only represented a fraction of the data in the current search application and did not capture pathology data from the original pathology reports. Therefore, a link to additional structured representation of pathology data would also add value to this dataset.

The text of the pathology reports had been preprocessed as part of the development of the Biolibrary search application. The preprocessing extracted the main or final diagnosis sections of the original pathology reports and removed any protected health information. This process had introduced some level of fragmentation and irregularities to what otherwise would be considered natural language. Also, the final diagnosis statements that were made in these reports were usually not recorded in natural language, but rather with lengthy noun phrases with multiple punctuations, and prioritization of certain nouns over others with the goal of writing semi-structured phrases that follow specific patterns. These patterns also varied based on the specific type of report, the date a report was created, and the author of the report. This non-natural language, and irregularly structured text, complicates and limits the application and accuracy of text processing tools that are trained for natural language.

2.2 Text processing

We limited our work for this project to the surgical pathology reports in the pathology repository. The other types of pathology reports (cytology, immunology, etc.) contained a significant amount of non-natural language content and would have required significant preprocessing and customization of the tools. Also, the surgical reports will more likely have corresponding entries in other information systems such as the cancer center and registry.

Our long-term goal is to be able to extract structured data from the reports that represents various descriptions of the individual specimens being described in the reports, and then use a semantic structure to relate them to facilitate their search and linking to other relevant biomedical entities. An initial step towards these goals, that has a more immediate benefit for the current search application, is to use tools that map the text to a standard vocabulary to address synonymy issues, and that allow us to operate on the content of the text at a higher semantic level. There are several tools available for such a task. The Unified Medical Language System (UMLS) developed by the NLM provides the MetaMap named entity recognition tool that is specifically trained for biomedical text and is based on the content of the UMLS. MetaMap was chosen as the preferred option given its relationship to the UMLS and the anatomy and pathology related ontologies contained within it (such as FMA and SNMOED-CT.)

Table 1 shows a summary of the mappings that we identified in the surgical pathology reports from the pathology repository for few high level categories.

Category	Reports	Concepts
Anatomy	154,830	5580 (1924)
Histology	63,572	349 (182)
Abnormal anatomy	23,448	695 (200)
Pathology & disease	137,161	5393 (1877)
Physiology & function	50,002	472 (161)

Table 1. Main concept categories, unique count of concepts (with concepts mapped over 10 times), and number of pathology reports containing one or more of these concepts.

Use of the MetaMap tools on the surgical pathology reports presented some challenges:

- The reports were partial and only included the “final diagnosis” section. The more natural language text portions of the reports (such as gross descriptions) were not available. This limited the identification of concepts such as gross phenotypic descriptions, negation, related medical conditions, etc., because these elements are less likely to occur in the final diagnosis section.
- The non-natural language patterns of the final diagnosis sections affected the output from MetaMap. For example, an anatomical organ might be referenced and then followed by a punctuation mark and an anatomical subpart or some other qualifier. This affected how MetaMap chunked the text, identified noun phrases, and mapped concepts to the text segments. MetaMap has few configuration options that helped to address some of these issues to some degree. However, a more detailed examination of the text patterns used in surgical pathology reports, with additional preprocessing, can further improve the mapping process.
- This project had limitations on access to the pathology reports due to their sensitive nature. Access and storage of any data was limited to a remote database and this added significant overhead for integrating and using MetaMap in such an environment. This also made it difficult to aggregate the different outputs generated by different MetaMap configurations that were used to improve its performance.
- Visualizing the results from MetaMap tools is somewhat difficult. There are no good GUI tools to brows the mapping results, especially after they have been saved in a customized fashion in the remote database.

2.3 Leveraging ontologies

In addition to mapping the text to a set of UMLS concepts to address synonymy issues, inter-entity relationships are also very useful for information retrieval. They allow a user to expand their search by following certain relationships; here we limited ourselves to subtype and parthood relationships. The UMLS itself provides high-level relationships in the form of the semantic network; however, we used the more specific relationships from source terminologies for this purpose. These terminologies were used to create an application-specific OWL representation of the entities and relationships that supports our use case.

For each mapped entity, its supertype closure was obtained from the UMLS hierarchy table to organize the entities into a single taxonomy, with top-level classes being the UMLS's semantic network classes for each top-level non-semantic network entity. For the anatomical entities, an effort was made to map them to the closest FMA entity when there was no direct UMLS mapping. For this, non-FMA entities that had a SNOMED-CT anatomical mapping were linked to FMA entities by navigating the SNOMED-CT hierarchy until a nearby FMA entity is found.

The UMLS relationships table, which contains relationships from each of the source vocabularies, was then examined to identify a set of relationships that have a strong parthood nature even if the parthood is not a strict structural parthood. The goal was to retrieve as many parthood-like relationships that would be useful for information retrieval purposes. These relationships were then added to the anatomical entities described above.

The next step involved using SNOMED-CT disease descriptions to relate abnormal morphology and disease to anatomical entities. SNOMED-CT provides grouped binary relationships (i.e. a binary representation of n-ary relationships) between disease, abnormal morphology, and anatomy. These relationships were used to relate the disease and abnormal morphology entities mapped in the pathology reports to the anatomical entities built above. Also, there was some overlap between the anatomy and abnormal morphology hierarchies in SNOMED-CT that had to be addressed to produce two disjoint hierarchies.

The goal of the above steps was to build an OWL application ontology that allows a user to query along various abstractions of anatomical entities to find related anatomical, morphological, and disease entities. The retrieved set of entities can then be used to query for the pathology reports that have been mapped with the UMLS codes for those entities by the MetaMap tool, thereby retrieving the relevant bio-specimens.

2.4 Application integration

The approach we took for an initial and simple integration is to append the concept mappings to the end of the text of the reports in the form of "concept id, start location, end location" entries that will be indexed by the text indexing system of the current application. This enables performing a concept-based search in the form of a free text search as long as the concepts identifiers are used as the keywords for the search. The concept mapping location information can be used as an indicator for where the concept matched the text and it can also support text highlighting.

This integration has been done in a separate instance of the indexing engine. Our next steps will be: 1) To setup a test instance of the existing application to make this enhanced content available to end-user evaluation. 2) Enhance the UI to help a user browse the application ontology in order to support concept selection and expansion for the search.

The main reason for attempting to integrate the results into the existing application is that the existing search application also allows the user to limit a text search by additional criteria obtained from the cancer registry and other sources of clinical information. It also has few other important workflow features that are required by end users. And, to evaluate the enhanced content, it will have to be used in a similar fashion to the non-enhanced version of the application.

3 DISCUSSION

The preliminary results described above highlight differences between the resulting mapped concepts and the initial structured data. The pathology reports are linked to matching cancer registry records, which used approximately 310 anatomical codes from the ICD-O oncology terminology. However, most reports had only one or two corresponding registry records that captured the main anatomical sites for a malignant pathology. The pathology reports, however, frequently reference many anatomical sites and this data was not previously accessible. Also, reports for non-malignant pathology have no corresponding cancer registry records. The limitations are similar for histology, abnormal morphology, and disease related information. The mapping related each report to approximately 4.7 anatomical (normal and abnormal), 1.5 histological, 3 disease, and 1.5 physiological entities. Also, the following table compares the preexisting data to the mappings for two randomly selected malignant and benign reports:

Preexisting data	Concept mappings
Pancreas head, stomach greater curvature, mucinous adenocarcinoma, GI stromal sarcoma	Common bile duct, duodenum, pancreas, right and left vagus nerves, stomach, mucinous adenocarcinoma, papillary neoplasm, splenic lesion, nodular fibrosis, GI stromal tumor, cytologic atypia.
Chondrodystrophy	Achilles tendon, bone of calcaneum, nodule, synovial membrane, bone tissue, cartilaginous exostosis, granuloma annulare, rheumatoid nodule

Table 2. An informal comparison of preexisting structured data vs. generated mappings for a malignant and a benign pathology report.

A manual review of the two reports in table 2 shows most mappings to be correct. However, there are redundant mappings and, for example, the “rheumatoid nodule” should be negated because it was referenced in a sentence that referred to a staining procedure that differentiated the pathology to be a “granuloma annulare nodule” pathology as opposed to a “rheumatoid nodule” pathology. Refining the mapping process to address these issues is a difficult process but it is one of our goals for our next iteration.

The application ontology that was developed was also used to evaluate the benefits of semantic entity expansion. The various classes in the ontology were marked to indicate how they were mapped to the text, and the results of several Description Logic (DL) queries were manually reviewed to predict the returned results if such a query was performed in the production system. The results of these queries showed the benefits of the concept mapping (as opposed to text matching) and the logical relationships derived from the source terminologies. These results were also briefly reviewed with the main stakeholders of the current system and there was clear interest in incorporating such functionality in a future version of the system. The results were more sensitive than specific. However, the sensitivity was valuable for our use case because the current search results are usually small or empty especially for a search that is constrained by other medical and demographic criteria.

Although the results are promising, there is a set of issues that need to be addressed in the next iteration of the project. First, the output of the MetaMap mapping process had to be persisted in a relational database. This made it difficult to experiment with different representations of the mappings or with different mapping tools since it frequently required a database schema modification and a modification of related queries and scripts. Second, we used the UIMA framework in the second part of our work, which allowed for exploring other text processing tools in a more flexible way. However, this amplified the difficulties of working with a relational database as the only form of persistence of the results.

Working in a relational database also complicates both the integration of generated data with preexisting data, and the later exploration of the results. Having an ability to integrate the text reports, their mappings, and the generated ontology in an RDF triplestore would have significantly simplified the workflow and evaluation in addition to the other benefits of an RDF data representation. While this was not a viable option for the exploratory phase of the project, it will be considered in our future work. The next iteration of our project will develop tools that address some of these limitations.

This preliminary work has demonstrated to our stakeholders the benefits of enhancing the existing system with semantic capabilities through the use of text processing tools, terminologies, and ontologies. We plan to refine these results for our use case, and to use the knowledge we gained during this process and the derived data we are generating to develop reusable artifacts for other related bio-specimen projects. A longer term goal is to develop an ontology based model for bio-specimen data that could support sophisticated data exchange and integration with other biomedical resources.

ACKNOWLEDGEMENTS

This work has been funded by the Medical Research Foundation of Oregon through a New Investigator award for "Semantic Linking of Biospecimen Resources", and the support of Robert Schuff and Allison Grossblatt-Wait from the Oregon Clinical & Translational Research Institute.

REFERENCES

Aronson, A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program in Proc AMIA Symp 2001. 17–21.

International Health Terminology Standards Development Organization (IHTSDO). Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) http://www.ihtsdo.org.

Rosse, Cornelius, and José L V Mejino Jr. “A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy.” Journal of Biomedical Informatics 36, no. 6 (December 2003): 478–500.

Unified Medical Language System (UMLS). Available from: https://uts.nlm.nih.gov/