Evaluating NDF-RT

Jonathan Mortensen
Department of Biomedical Engineering
Case Western Reserve University, Cleveland, OH, USA
Jonathan.Mortensen@cwru.edu

August 6, 2010

Abstract

The National Drug File - Reference Terminology (NDF-RT) presents a large set of drug classes which can potentially be used in clincal settings. Using semantic web technologies such as XML (eXtensible Markup Language) and RDF (Resource Description Framework), the drug classes in NDF-RT were enriched for the purpose of comparison to themselves and to drug classes in other terminologies. Methods and results of this comparison will be presented in a paper submitted to SMBM 2010. In addition, the enhancement process will enable the refactoring of a webservice currently in development for NDF-RT, replacing the backend with a more intuitive database sytstem. Finally, the enrichmed version of NDF-RT was used by another group of students in their data integration project that focused on linking drugs to pharmacogenomic information. The processes used to enrich and query NDF-RT serve as a model for linking data and terminologies in new ways.

1 Introduction

Clinical Decision Support presents an opportunity to improve patient outcomes and reduce healthcare costs. Furthermore, with "meaningful use" requirements and Electronic Health Records (EHR), relevant data will soon be digital, and clinical decision support can soon be added to these systems. Clinical terminologies are the backbone of any EHR and clinical decision support system, as they provide structured clinical knowledge. The National Drug File - Reference Terminology (NDF-RT) is a drug terminology produced by the Veterans Health Administration (VA) and is recommended as the standard in e-prescribing systems[1]. However, the drug classes in NDF-RT have not been recently examined. Work completed during the Medical Informatics Summer Training Program focused around NDF-RT, including enriching NDF-RT, comparing the drug classes in NDF-RT and enhancing a webservice for NDF-RT. A formal presentation of the drug class comparison component is presented in the Appendix and will be submitted to Semantic Mining in Biomedicine 2010. The following is preliminary work.

1.1 Background on NDF-RT

NDF-RT was created based upon the National Drug File, a listing of medications produced by the Department of Veteran Affairs[2]. It serves as a reference standard for a variety of medical situations related to drugs and medications. NDF-RT is a description logic-based model available in Web Ontology Language (OWL) and XML formats. It includes 9 kinds of information: Cellular or Molecular Interactions, Clinical Kinetics, Diseases Manifestations or Physiologic States, Pharmaceutical Preparations, Physiological Effects, RxNorm Dose Forms, Therapeutic Categories, and VA Drug Interactions. Pharmaceutic Preparations contains drugs organized into three categories: Products by Generic Ingredient Combination, Products by VA Class and External Phar-

macologic Classes (EPC). There are 485 VA Drug Classes organized into a basic hierarchy. A drug normally belongs to only one class. In addition, there are 425 EPCs. Differing from the VA Classes, the EPCs have a nearly flat hierarchy and are defined by various properties such as physiologic effect, therapeutic category, ingredient and mechanism of action. It is expected that these classes have a greater granularity. Yet, in the standard distribution of NDF-RT, the clinical drugs lack an explicit membership to External Pharmacologic Classes.

2 Methods

To compare the two classes in NDF-RT, VA Classes and EPC, clinical drugs needed to be classified into EPCs. To do this, preliminary work of translation and enrichment was performed. This was done using a pipeline process summarized by Figure 1. Methods of this process were piloted by [3] Details of each step follow.

2.1 XML Translation to OWL

To do classification, description logic (DL) is required. Description logics required an OWL formatted terminology. Therefore to do classification of Drugs into EPCs, an OWL format of NDF-RT was necessary. Because NDF-RT OWL was released at irregular intervals, but NDF-RT XML releases were consistent, a method of translation from XML to OWL was created. For this translation, an extensible style sheet language transform (XSLT) was developed which converted the given XML into the RDF/XML specification of OWL. XSL is a generalized language for transforming XML documents. Using [4] and OWL examples of NDF-RT, a conversion was developed between XML elements, translating the XML version into OWL.

2.2 Validation of XML translation

To ensure the translation to OWL was complete and reflected the versions produced by the VA, four different forms of validation were used. The first was an OWL validator, which validates the structures of an OWL file. The second method was through a direct file diff. A version of NDF-RT with both the XML and OWL version was chosen. Another OWL version was then generated using the transform we created. The two files were then compared line by line after a case-sensitive sort. The third method utilized an XML-Diff utility, comparing the generated and original file. The XML-diff utility compares two XML files based on the data within elements and their children, but does not consider order of elements. Because the type of OWL files compared are RDF/XML, such a utility can be used. Thus, this method validates that the two files contain the same data, even though two elements may not be on the same line in a file. The final method utilized an XML-Diff utility, comparing the generated and original file. The XML-diff utility compares two XML files based on the data within elements and their children, but does not consider order of elements. Because the type of OWL files compared are RDF/XML, such a utility can be used. Thus, this method validates that the two files contain the same data, even though two elements may not be on the same line in a file.

2.3 OWL Enrichment

Next, NDF-RT was enriched in three ways. First, the EPCs were be converted from primitive classes to defined classes. This was necessary so that during classification, the classifier found that the conditions which defined an EPC were sufficient to classify a drug, which also had those properties, as a member. Because has_TC (has therapeutic category) is only part of EPCs and not ingredients, the condition was left of out the equivalentClass property, the property which makes an EPC a defined class. Next, the property has_Ingredient was made equivalent to has_ChemicalStructure. This was done because EPCs are defined by chemical structures whereas drugs are defined by ingredients. For the purpose of this classification, we considered the two properties to be equivalent. Finally, to ensure drugs which only treat or prevent were classified as a member of a class which treats or prevents, such as an antimalarial, the property may_treat_or_prevent was added. Next,

may_treat and may_prevent were changed to subproperties of may_treat_or_prevent. Furthermore, all occurrences of may_treat and may_prevent in EPCs were changed to may_treat_or_prevent. Each of these three enrichment processes (XSLTs) were performed separately and sequentially, for interchangeability, updating, and the ability to remove one or more components. Such programming design is essentially to allow for rapid development.

2.4 Classification of Enriched OWL

After the enrichment was performed, NDF-RT was classified using HermiT, a DL OWL classifier developed at Oxford University [5]. This is a robust classifier that performs well on large, complex terminologies such as NDF-RT. The classifier produces additional subClassOf and equivalentClass axioms which are merged with the original OWL file using a graph union operation in jena [6].

The result of this enrichment was the classification of clinical drugs into the External Pharmacologic Classes.

2.5 Storage and Presentation

The enriched, classified version of NDF-RT was loaded into a Virtuoso endpoint [7] for SPARQL (SPARQL Protocol and RDF Query Language) querying, which allowed for evaluation of the Drug Classes and availability of the terminology to the NDF-RT API.

3 Results

The completion of this preliminary work enabled two things: the comparison of NDF-RT drug classes to themselves and to other drug terminologies, and the reformulation of the NDF-RT API (Application Programming Interface). The comparison is discussed in the other work[8]. The NDF-RT API webservice backend is currently shoehorned into a relational database structure. With the translation of NDF-RT to OWL/RDF, NDF-RT can be uploaded instead to an RDF-triple store backend. The NDF-RT API is being reworked to support the simpler triple store database. In specific, the queries for obtaining data from NDF-RT were rewritten in SPARQL.

4 Discussion

The enrichment processes in NDF-RT and the application of the result provide an example of the power of linked data. By using a classifier, drugs are automatically classified into drug classes. This is powerful alone, as it simplifies classification. Furthermore, placing NDF-RT onto a SPARQL endpoint enabled connections to large data sources, including the UMLS and Uniprot (developed by Brian Kirk and Emily Doughty). This example modeled the potential of linking previously disparate sources, a great opportunity not yet fully developed. Methods by which NDF-RT was exploited may serve as a model for increased data migration to the semantic web environment.

From the preliminary exploration of classification, one point was noted: the External Pharmacologic Classes may not be suitable for clinical use as many classes are defined by physiologic effect and not therapeutic intent, leading to inadequate classification of drugs from the clinical perspective. Automatic classification cannot be performed without curation. In addition one struggle throughout the process involved fine tuning the RDF-Store, Virtuoso, for performance and reliability with large datasets.

5 Conclusions

NDF-RT is a national standard as a drug terminology. Its enrichment and presentation in a format appropriate for the semantic web enables meaningful data use, providing an opportunity for drug class comparisons and data linkage in an intuitive open fashion.

Acknowledgments

I would like to thank Dr. Bodenreider for the guidance on this project, May Cheh for Coordination of the Medical Informatics Program, Brian Kirk and Emily Doughty who were collaborators on various components in the lab, and Dr. McDonald and the Lister Hill Center for the support.

References

[1] R. A. Miller, "Clinical decision support and electronic prescribing systems: A time for responsible thought and action," *Journal of the American Medical Informatics Association*, vol. 12, no. 4, pp. 403–409, 2005.

[2] M. J. Lincoln, S. H. Brown, V. Nguyen, T. Cromwell, J. Carter, M. Erlbaum, and M. Tuttle, "US department of veterans affairs enterprise reference terminology strategic overview," *Medinfo*, vol. 11, no. Pt 1, p. 391395, 2004.

[3] O. Bodenreider, F. Mougin, and A. Burgun, "Automatic determination of anticoagulation status with NDF-RT,"

[4] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, and B. Parsia, "The national cancer institutes thesaurus and ontology," *Journal of web semantics*, vol. 1, no. 1, p. 7580, 2003.

[5] "HermiT."

[6] "Jena semantic web framework."

[7] "Virtuoso universal server."

[8] J. Mortensen and O. Bodenreider, "Evaluating pharmacologic classes in NDF-RT,"

Figures

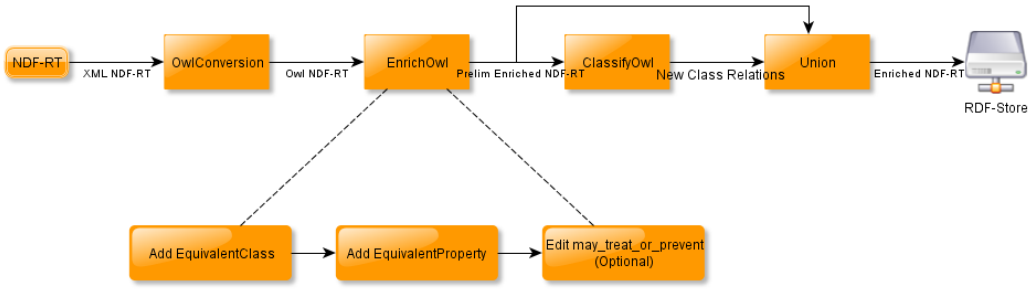

Figure 1: Pipeline of the enrichment process on NDF-RT From XML to Database Endpoint

Appendix

An attached Paper "Evaluating Pharmacologic Classes in NDF-RT" will be provided before submission to SMBM 2010