LinkedCT Live: Platform for Online Curation of Clinical Trials Data[*]

Oktie Hassanzadeh, Renée J. Miller, Fatemeh Nargesian, and Erkang Zhu

Department of Computer Science, University of Toronto
{oktie,miller,fnargesian,ekzhu}@cs.toronto.edu

Abstract. The goal of the Linked Clinical Trials (LinkedCT) project is to transform the data published on ClinicalTrials.gov into a high-quality knowledge base published as Linked Data on the Web. In this demonstration, we present the platform we have developed for both online curation of clinical trials data into linked data, and for rapid Web application development on top of this linked data. We also show a few sample applications built using this platform. We have made the project open-source and invite researchers and healthcare professionals to develop applications that will be hosted on LinkedCT.org.

Keywords: Clinical Trials, Linked Data, Semantic Healthcare Applications

1 Introduction

ClinicalTrials.gov is a large and widely used registry of international clinical studies, which is maintained by the U.S. National Institutes of Health (NIH). The registry contains detailed information about the studies including recruitment information, eligibility criteria, and clinical outcomes. This data has tremendous value to the clinical and healthcare research communities [8,9,11,12]. In 2008, we began the Linked Clinical Trials (LinkedCT) project to study how the value of this data could be enhanced by publishing the ClinicalTrials.gov data as high-quality (5-star [2]) Linked Data. Initially, the data transformation from XML to RDF was a static, manually-designed process that contained errors and resulted in quality issues and incomplete data. As a result, we later redesigned the project with the goal of having an automatic data curation process with online transformation of XML data into Linked Data. In what follows, we briefly describe this platform. We then describe our demonstration plan which includes showcasing a number of lightweight applications built on top of LinkedCT data and hosted on LinkedCT.org.

[*] This demonstration complements our full paper [4] and was supported in part by NSERC BIN.

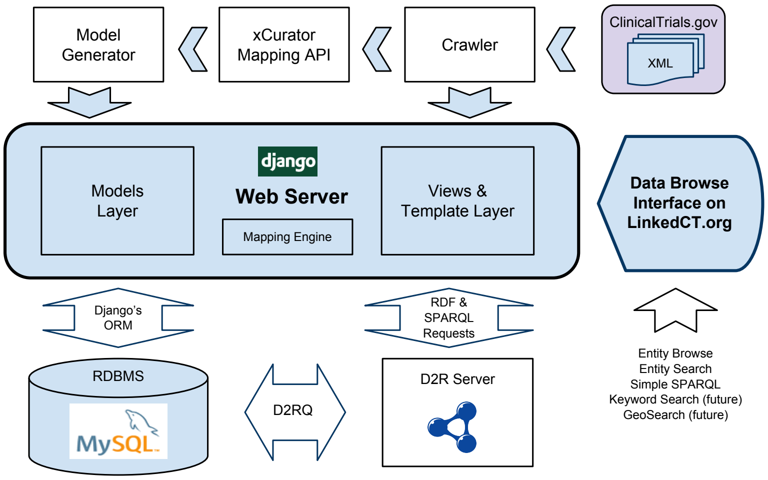

Fig. 1. LinkedCT Platform Architecture [4]

2 The LinkedCT Platform

The architecture of LinkedCT is depicted in Figure 1. Our platform provides end-to-end data curation that consumes XML data and produces current (up-to-data) Linked RDF data. This linked data can be accessed directly through links on the website or SPARQL queries, or from a number of applications that we will demonstrate.

Creating a High-Quality Knowledge Graph The creation of a knowledge graph is not simply a matter of translating XML (with its current structure) into RDF. Rather, an important part of the curation process is the creation of a structure (schema) that both accurately models the semantics of the original data and also adds value to the data by facilitating the linking of the data to external sources. The use of an automated translation tool (e.g., XSPARQL [7]) or a manual translation process such as the one originally employed in LinkedCT can result in data quality issues. In LinkedCT, we discovered many of these issues by analyzing user reports and by using LinkedCT in several LODD projects [5,6].

- Our users reported missing entity types (RDF classes) or missing attributes (RDF properties). Users may have had knowledge of classes and properties in related NIH sources or other health data sources that they expected to see represented in the same way in LinkedCT. A manual or automated translation may not always create the classes and properties users expect.
- An application or use-case for LinkedCT required that a literal property (such as the string-valued property `location_countries`) be represented as an entity type (the RDF class `Country`). This requirement may come from a desire to link data to external sources (using the URIs of the entities).

– There may be inconsistencies in the XML such as the same semantic information being represented by different XML labels (and as a result different entity types in an RDF graph derived directly from XML elements). Such inconsistencies should be reconciled to create a well-designed knowledge graph.

To address these issues, we needed to go beyond automated translation to create a schema and mapping discovery system. xCurator [10] is an end-to-end system for transforming semi-structured data into a well-designed linked knowledge base. LinkedCT was the main evaluation platform for xCurator. In this demonstration, we will use a new light-weight web application that uses the output of the mapping generator component of xCurator and can run on modest hardware or virtual machines. This component uses the *Crawler* module (Figure 1) that monitors ClinicalTrials.gov for new trials and for updates to existing trials. In the *Model Generator*, the mappings created by the xCurator mapping generator are translated into an intermediate Object-Relational Mapper (ORM) model definition (implemented in the Django framework [1]). We are then able to use a reliable relational backend to store our data while using D2R Server [3] to generate RDF. Moreover, the Django framework facilitates rapid application development and provides an infrastructure to host applications in a common platform and where the data resides.

3 Demonstration Plan

Our goal in this demonstration is three-fold. First, we will demonstrate the online transformation of XML data into high-quality Linked Data. To this end, we will be using a secondary server with an empty database and start submitting XML data to the server to show the transformation process. We will then make changes to an already processed XML file and show how the system manages different changes in the input data. Our goal is to illustrate the importance of having a dynamic (or online) mapping discovery process that can adapt to changes in the data and its structure.

Second, an important goal of LinkedCT is to study how the value of open data can be enhanced by publishing it as curated Linked Data. In the demonstration, we will show examples from the ClinicalTrials.gov source and its keyword search interface. We will contrast these with results from a search over LinkedCT to show how high-quality Linked Data can enable and enhance various exploration and analysis tasks.

Our final goal is to show how new applications can be developed on top of our platform. We will use our simple "Live Statistics" page (`http://linkedct.org/stats/`) as a sample "Getting Started" application and show how it can be written and deployed in only a few minutes. We will also show more advanced applications that are currently under development by our team: 1) a Geo-Search interface that shows clinical studies involving a given condition or medication in the proximity of a given location, and 2) a fuzzy look-ahead semantic keyword search interface powered by SRCH2 (`http://srch2.com`).

We have made the platform open-source and an important goal of the demonstration is to encourage the community to implement new applications that can be hosted on LinkedCT.org to further add value to the Clinical Trials data.

References

1. Django Web Framework, https://www.djangoproject.com/. [Online; accessed 07-07-2015].
2. Tim Berners-Lee. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006. [Online; accessed 07-07-2015].
3. C. Bizer and R. Cyganiak. D2R Server - Publishing Relational Databases on the Semantic Web. In *Proc. of the Int'l Semantic Web Conference (ISWC)*, 2006.
4. O. Hassanzadeh and R. J. Miller. Automatic Curation of Clinical Trials Data in LinkedCT. In *Proc. of the Int'l Semantic Web Conference (ISWC)*, 2015.
5. A. Jentzsch, B. Andersson, O. Hassanzadeh, S. Stephens, and C. Bizer. Enabling Tailored Therapeutics with Linked Data. In *Proceedings of the WWW2009 workshop on Linked Data on the Web (LDOW2009)*, 2009.
6. Anja Jentzsch, Jun Zhao, O. Hassanzadeh, Kei-Hoi Cheung, Matthias Samwal, and Bosse Andersson. Linking Open Drug Data (Triplification Challenge Report). In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS'09)*, 2009.
7. A. Polleres, T. Krennwallner, N. Lopes, J. Kopecký, and S. Decker. XSPARQL Language Specification. `http://www.w3.org/Submission/xsparql-language-specification/`. [Online; accessed 07-07-2015].
8. A. P. Prayle, M. N. Hurley, and A. R. Smyth. Compliance with Mandatory Reporting of Clinical Trial Results on ClinicalTrials.gov: Cross Sectional Study. *BMJ*, 344, 2012.
9. J. S. Ross, T. Tse, D. A. Zarin, H. Xu, L. Zhou, and H. M. Krumholz. Publication of NIH Funded Trials Registered in ClinicalTrials.gov: Cross Sectional Analysis. *BMJ*, 344, January 2012.
10. S. Hassas Yeganeh, O. Hassanzadeh, and R. J. Miller. Linking Semistructured Data on the Web. In *Proc. of the Int'l Workshop on the Web and Databases (WebDB)*, 2011.
11. D. A. Zarin, T. Tse, and N. C. Ide. Trial Registration at ClinicalTrials.gov between May and October 2005. *New England Journal of Medicine*, 353(26):2779–2787, 2005.
12. D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide. The ClinicalTrials.gov Results Database–Update and Key Issues. *New England Journal of Medicine*, 364(9):852–860, 2011.