Automatic Curation of Clinical Trials Data in LinkedCT*

Oktie Hassanzadeh[1,2] and Renée J. Miller[1]**

[1] Department of Computer Science, University of Toronto
[2] IBM T.J. Watson Research Center

Abstract. The Linked Clinical Trials (LinkedCT) project started back in 2008 with the goal of providing a Linked Data source of clinical trials. The source of the data is from the XML data published on ClinicalTrials.gov, which is an international registry of clinical studies. Since the initial release, the LinkedCT project has gone through some major changes to both improve the quality of the data and its freshness. The result is a high-quality Linked Data source of clinical studies that is updated daily, currently containing over 195,000 trials, 4.6 million entities, and 42 million triples. In this paper, we present a detailed description of the system along with a brief outline of technical challenges involved in curating the raw XML data into high-quality Linked Data. We also present usage statistics and a number of interesting use cases developed by external parties. We share the lessons learned in the design and implementation of the current system, along with an outline of our future plans for the project which include making the system open-source and making the data free for commercial use.

Keywords: Clinical Trials, Linked Data, Data Curation

1 Introduction

The clinical research community and the healthcare industry have well recognized the need for timely and accurate publication of data related to all aspects of clinical studies, ranging from recruitment information and eligibility criteria to details of different phases and the achieved results [8,15,17,19,20]. ClinicalTrials.gov is currently the main mechanism of achieving this goal. Maintained by U.S. National Institutes of Health, it is the largest and most widely used registry of clinical studies with registered trials from almost every country in the world. There has been a significant increase in the number of registered trials

* The data source is publicly available at `http://linkedct.org`. Data dumps available at `http://purl.org/net/linkedct/datadump`. Please note scheduled maintenance down times on our Twitter feed `https://twitter.com/linkedct`. Resource URIs validated as proper Linked Data by `http://validator.linkeddata.org/` ("All tests passed"). Part of LOD cloud. Registered on `http://datahub.io`.

** Partially supported by NSERC BIN.

as a results of a mandate by FDA and requirement from various journals that a trial needs to be registered before it can start or the results can be published [24,25]. There is also an ongoing effort in the community to increase the quality of the data, and require publication of the results after the completion of the registered trials.

The Linked Clinical Trials (LinkedCT) project started in 2008 with the goal of publishing the ClinicalTrials.gov data as high-quality (5-star [4]) Linked Data on the Web. Our inspiration for the project came from a simple experiment on matching patients with clinical trials as a part of the LinQuer project [10]. A simple task of retrieving all the trials on a certain condition along with all their attributes turned into a laborious Extract-Transform-Load process. The process involved studying the schema of the XML data, writing code to crawl the data and load them in IBM DB2 on a local server to be able to query the data using DB2's pureXML features. Our initial goal was to simply publish the result of this transformation as Linked Data using D2R server [5], with the main challenge being discovering links to external sources [11]. Initial user feedback and work done as part of the LODD project [13,14] on developing use cases over the data made it apparent that the transformation process was not only laborious, but also error-prone. The errors along with the slow and static transformation process called for a new solution that replaces the manually designed transformation process with a mostly automated *curation* [7,23] of the clinical trials data.

The following section describes some of the challenges faced in designing an automated data curation process and our solution using xCurator [23]. We then describe a number of interesting applications and usage scenarios of LinkedCT. We finish the paper with a number of future directions which includes making the data available free for commercial use, and making the platform open-source to facilitate development of applications hosted on `LinkedCT.org`.

2 Data Curation in LinkedCT

In this section, we describe the end-to-end curation process we have designed to construct an up-to-date high-quality Linked Data source out of the XML data published by ClinicalTrials.gov. Figure 1 shows the overall architecture of the system. In what follows, we describe different components of the system.

2.1 From XML to RDF Knowledge Graph

Although there are mapping tools and systems for transforming XML into RDF (e.g., XSPARQL [2]), a key challenge in building a high-quality linked knowledge base is construction of a target data model that accurately describes the entity types and their relationships that exist in the original data, and that also facilitates knowledge discovery and linkage to external sources. As stated earlier, our initial manually-designed transformation of the data into relational and RDF resulted in a number of quality issues reported by users and discovered through working on usage scenarios as a part of the LODD project [13,14]. For example:

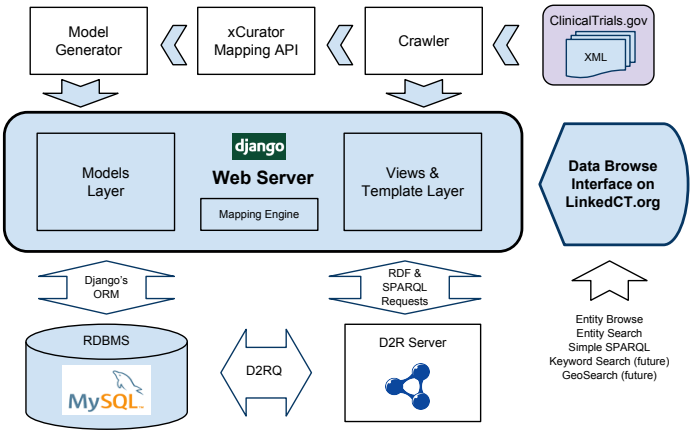

Fig. 1. LinkedCT Platform Architecture

- Users familiar with the original NIH data pointed out missing information in the form of either missing entity types (RDF classes) or missing attributes (RDF properties) from entities of a certain type. Such data can easily be missed in a manual mapping process.
- Use cases required a literal property (e.g., location country represented as a string-valued property) to be represented as an entity (e.g., Country being an RDF class). For example, linking data is typically done only over entities (with URIs that can be linked).
- Users found inconsistencies with the original XML, and we were unable to verify the reason (programming error vs. an update in the source) due to lack of provenance information or caching of the original NIH data.

Figure 2 shows a sample XML tree from the data that can help explain the reason behind some of the problems in a manual mapping design.

- Nodes with label `mesh_term` contain string values. Only a careful study by an expert can reveal they describe entities of type Drug.
- A simple approach of making a type (class) per each non-leaf node in the tree (which is similar to the common RDB2RDF approach of creating a class per each table) will result in entities of type `id_info` whereas this node is simply grouping a list of identifiers and is not representing an entity. Such extraneous types can make the data hard to query and understand.
- There are nodes in the tree such as `collaborator` and `lead_sponsor` that have different labels but represent the same type of entity, i.e., an agency.

The above challenges are addressed in xCurator [23], an end-to-end system for transformation of semi-structured data into a linked knowledge base. Our experience in LinkedCT has been one of the main use cases for evaluation of the accuracy of the mapping discovery in xCurator. The xCurator mapping gener-

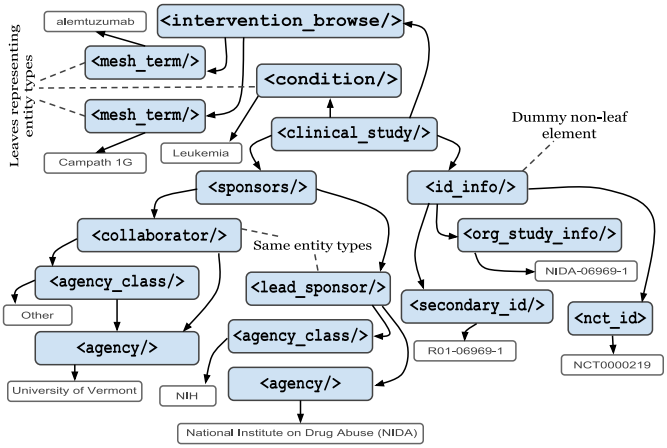

Fig. 2. Sample XML from ClinicalTrials.gov

ator uses a number of heuristics based on statistical measures and other data properties over a large enough sample of instances to automatically construct a set of classes that refer to real-world entity types. A main criteria for identification of entity types in xCurator is the ability to link instances of the derived type with entities in external knowledge bases, which will in turn result in a higher quality linked data sources in terms of linkage to external sources. We refer the reader to Hassas et al. [23] for a detailed description of the mapping discovery and evaluation using LinkedCT data. The results clearly show the superiority of xCurator's automatic mapping discovery to the initial manual mapping, even if only a small random subset of the data is used to generate the mappings.

2.2 Web Application Design

Although xCurator provides an end-to-end solution for mapping discovery and creation, along with publication of the resulting knowledge base in RDF following the Linked Data principles, we chose to use only the mapping generator component due to a number of reasons. First and foremost, xCurator is designed as a generic tool that can be used for any semistructured data whereas the strict focus on clinical trials data can help us better tune the system and algorithms to improve the outcome. Moreover, we have been able to tune LinkedCT's implementation to make a relatively light-weight web application that unlike xCurator can run on modest hardware or virtual machines. As shown in Figure 1, the web application is extended with a *Crawler* module that continuously checks for new trials on ClinicalTrials.gov and also checks for updates in existing trials. The xCurator mapping generator component uses the crawled data in a one-time process to generate mappings. These mappings are translated into an intermediate Object-Relational Mapper (ORM) model definition used by the web application

which is implemented in the Django framework [1]. This is done in the *Model Generator* module which is a Python code generator that can directly be used in the web application.

In addition to the ORM models, the Django web application handles HTTP service requests with a set of templates that provide the HTML and RDF view for the data browse web interface accessible at `http://linkedct.org`. In addition, it provides various APIs called by the crawler for addition and update process. Linkage to external sources such as DrugBank, PubMed, GeoNames, and DBpedia are performed using pre-defined linkage rules embedded in the *mapping engine* module in the Django web application and called during addition and update procedures. The mapping engine also performs duplicate detection in a similar way using pre-defined rules. The rules are defined using the results of our previous study on the quality of various linkage techniques [10,11].

2.3 Data Backend and SPARQL Endpoint

Ideally, the web application can work on top of a reliable RDF store for storage and querying. Unfortunately, there are no active projects on RDF support over Django Web Framework, and no non-commercial RDF stores capable of handling the very large number of updates and queries that LinkedCT needs. The alternative option is using a relational backend. We are currently using a MySQL database hosted on a secondary server. For an RDF view and SPARQL endpoint, we use the D2R server with D2RQ mappings [5] that are similarly generated automatically out of xCurator mappings by the model generator module. For scalability, we have to put a limit in D2R server configuration that limits the number of results returned and so the SPARQL endpoint is only useful for basic querying with small result sets. This makes it feasible to keep the web application lightweight despite the large load and large amount of data. Clearly, the limit on the SPARQL endpoint is far from ideal and one of the main shortcomings of our framework that we wish to address in the future as pointed out in Section 4. For applications requiring full SPARQL support, we make NTriples data dumps available regularly (once a month) and on demand.

3 Applications and Usage Statistics

Although ClinicalTrials.gov provides a relatively powerful "advanced search" feature, there is still a clear benefit in using LinkedCT even for basic data discovery and semantic search over the data. For example, using simple SPARQL queries or even on the Linked Data HTML browse interface on `http://linkedct.org`, one can quickly find a field named `is_fda_regulated` for entities of type Trial.[3] The ClinicalTrials.gov web pages and its advanced search do not include this field, and a keyword search for this field name yields no answer (likely because their

[3] See: `http://data.linkedct.org/resource/trial/fields/is_fda_regulated/` - at the time of this writing, there are 58,122 trials with `is_fda_regulated` set to `Yes` and 110,889 trials set to `No`.

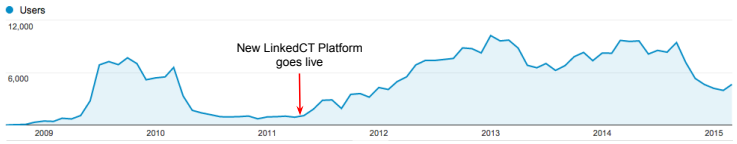

Fig. 3. Number of Unique LinkedCT.org Website Users

search is only over data, not metadata). Another benefit of publishing the data as Linked Data and providing a SPARQL endpoint is facilitating application building. One such application is the mobile faceted browsing application developed by Sonntag et al. [21] that uses LinkedCT and other interlinked sources to assist clinicians with various patient management activities.

Another basic advantage of publishing high-quality Linked Data is an implicit and important yet undervalued effect on the visibility of the data and search engine rankings. Figure 3 shows the number of unique visitors to LinkedCT.org website since the start of the project. The initial website receives a large number of visitors after its initial announcement and being indexed by Web search engines, but then the number goes down to under 1,000 by May 2011 when the new platform goes live. Our analysis of the access logs show that the main decrease is from search engine referrals. This completely changes after the new platform described in Section 2 goes live, which happens quietly without any public announcements of the new platform. Again, our analysis shows that a large portion of the increase is the result of search engine referrals, but this time the number remains high. This can be attributed to both the dynamic update of the trials in the new platform, and the higher quality and quantity of links to external sources. Again, achieving this without any effort on search engine optimization or public announcements on the project shows an interesting side outcome of following Linked Data principles.

Apart from the above-mentioned applications and basic advantages of publishing Linked Data, there are several very interesting healthcare applications that rely on LinkedCT as one of their primary sources. Examples include:

- Zaveri et al. [26] perform a very interesting study on research-disease disparity. The study shows that there is a large gap between the amount of resources spent on a disease (in terms of clinical trials and publications) and the disease fatality (death rate). LinkedCT is one of the three core data sets used in this study, a study which required links to external sources.
- The Linked Structured Product Labels (LinkedSPLs) project aims at publishing FDA's drug label information as Linked Data on the Web [6]. A main use case in the project is discovering missing Adverse Drug Reactions (ADRs) through linkage to finished trials on LinkedCT and their associated PubMed articles.
- PAnG project [3] that aims at discovering patterns in knowledge graphs, uses LinkedCT data to find clusters of strongly related studies, drugs, and diseases. An example of an interesting pattern is one that shows the drug

"Varenicline", a drug used to treat nicotine addiction, has recently been linked to treating alcohol use disorders. At the time of this writing, this information is absent from the Wikipedia article on Varenicline.

4 Conclusion & Future Directions

Despite the relatively large user base and various applications built on top of LinkedCT, the project is far from complete. In Section 2, we presented an honest description of the current system architecture, including a few shortcomings of the platform such as a three-layer process for generating RDF, and a SPARQL endpoint that only allows simple queries with small output. A list of known issues is available on our issue tracker at `https://code.google.com/p/linkedct/issues`. Some of these shortcomings resulted in duplicate efforts in the community, for example Bio2RDF's inclusion of ClinicalTrials.gov data in its latest release [9], which is based on a static one-time processing of the XML source as in our initial release [11], although the mapping seems to be of a high quality. As a result, we recently made LinkedCT's Web server code open-source to not only build a community to maintain the project, but also to expand the features in the Web server and build in-house user-contributed applications. The code is available on GitHub at `https://github.com/oktie/linkedct`. We will also maintain a list of projects contributed by users and application scenarios, and will be open to new proposals. Examples of applications include a geographical search interface showing trials on a certain condition in a given proximity on Google Maps, and a fuzzy keyword search interface powered by SRCH2 (`http://srch2.com`).

Without a doubt, various healthcare applications that rely on LinkedCT data are critical to the success of the project. An important use case of the data is facilitating matching of patients with clinical trials. Previous work has shown promising results, but using custom transformations and the original XML data [12,16,18]. It would be interesting to see how LinkedCT can be used in such scenarios and with real patient data. Use cases requiring reasoning may also need an extension of the ontology or its mapping to an existing domain ontology such as Ontology of Clinical Research [22]. Another interesting study that becomes possible as a result of LinkedCT data is a longitudinal study over trials using the RDF data dumps that are published monthly since 2013. To further facilitate commercial applications and use cases developed by commercial entities, we have changed the data license from CC-BY-SA-NC to Open Data Commons Open Database License (ODbL) that allows non-restricted commercial use.

References

1. Django Web Framework, https://www.djangoproject.com/.
2. W. Akhtar et al. XSPARQL: Traveling between the XML and RDF Worlds - and Avoiding the XSLT Pilgrimage. In *ESWC*, pages 432–447, 2008.

3. P. Anderson, A. Thor, J. Benik, L. Raschid, and M. Vidal. PAnG: finding patterns in annotation graphs. In *SIGMOD*, pages 677–680, 2012.
4. Tim Berners-Lee. Linked Data - Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006. [Online; accessed 04-27-2015].
5. C. Bizer and R. Cyganiak. D2R Server - Publishing Relational Databases on the Semantic Web. In *ISWC Posters and Demonstrations Track*, 2006.
6. R. D. Boyce et al. Dynamic Enhancement of Drug Product Labels to Support Drug Safety, Efficacy, and Effectiveness. *J. Biomedical Semantics*, 4:5, 2013.
7. P. Buneman, J. Cheney, W. Chiew Tan, and S. Vansummeren. Curated Databases. In *PODS*, pages 1–12, 2008.
8. R. M. Califf, D. A. Zarin, J. M. Kramer, R. E. Sherman, L. H. Aberle, and A. Tasneem. Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007-2010. *JAMA*, 307(17):1838–1847, 2012.
9. M. Dumontier et al. Bio2RDF Release 3: A Larger, More Connected Network of Linked Data for the Life Sciences. In *ISWC*, pages 401–404, 2014.
10. O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. A Framework for Semantic Link Discovery over Relational Data. In *CIKM*, pages 1027–1036, 2009.
11. O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. LinkedCT: A Linked Data Space for Clinical Trials. Technical Report CSRG-596, University of Toronto, August 2009.
12. Z. Huang, A. ten Teije, and F. van Harmelen. SemanticCT: A Semantically-Enabled System for Clinical Trials. In *KR4HC/ProHealth*, pages 11–25, 2013.
13. 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.
14. Anja Jentzsch, Jun Zhao, O. Hassanzadeh, Kei-Hoi Cheung, Matthias Samwal, and Bosse Andersson. Linking Open Drug Data. In *I-SEMANTICS*, 2009.
15. C. Laine et al. Clinical Trial Registration: Looking Back and Moving Ahead. *New England Journal of Medicine*, 356(26):2734–2736, 2007.
16. B. MacKellar et al. Patient-Oriented Clinical Trials Search through Semantic Integration of Linked Open Data. In *IEEE ICCI*CC*, pages 218–225, 2013.
17. G. D. Novack. Clinical Trial Registry–Update. *Ocular Surface*, 7(4):212–4, 2009.
18. C. Patel et al. Matching Patient Records to Clinical Trials Using Ontologies. In *ISWC*, pages 816–829, 2007.
19. A. P. Prayle et al. Compliance with Mandatory Reporting of Clinical Trial Results on ClinicalTrials.gov: Cross Sectional Study. *BMJ*, 344, 2012.
20. J. S. Ross et al. Publication of NIH Funded Trials Registered in ClinicalTrials.gov: Cross Sectional Analysis. *BMJ*, 344, 2012.
21. D. Sonntag et al. Clinical Trial and Disease Search with Ad Hoc Interactive Ontology Alignments. In *ESWC*, pages 674–686, 2012.
22. Samson W Tu et al. OCRe: An Ontology of Clinical Research. In *11th International Protege Conference*, 2009.
23. S. Hassas Yeganeh, O. Hassanzadeh, and R. J. Miller. Linking Semistructured Data on the Web. In *WebDB*, 2011.
24. D. A. Zarin et al. Trial Registration at ClinicalTrials.gov between May and October 2005. *New England Journal of Medicine*, 353(26):2779–2787, 2005.
25. D. A. Zarin et al. The ClinicalTrials.gov Results Database–Update and Key Issues. *New England Journal of Medicine*, 364(9):852–860, 2011.
26. A. Zaveri et al. ReDD-Observatory: Using the Web of Data for Evaluating the Research-Disease Disparity. In *WI*, pages 178–185, 2011.