Hubble: Linked Data Hub for Clinical Decision Support

Rinke Hoekstra[1,3(✉)], Sara Magliacane[1], Laurens Rietveld[1], Gerben de Vries[2], Adianto Wibisono[2], and Stefan Schlobach[1]

[1] Department of Computer Science, VU University Amsterdam, Amsterdam, The Netherlands
{rinke.hoekstra,s.magliacane,laurens.rietveld,k.s.schlobach}@vu.nl
[2] Department of Computer Science, University of Amsterdam, Amsterdam, The Netherlands
{g.k.d.devries,a.wibisono}@uva.nl
[3] Leibniz Center for Law, University of Amsterdam, Amsterdam, The Netherlands

Abstract. The AERS datasets is one of the few remaining, large publicly available medical data sets that until now have not been published as Linked Data. It is uniquely positioned amidst other medical datasets. This paper describes the Hubble prototype system for clinical decision support that demonstrates the speed, ease and flexibility of producing and using a Linked Data version of the AERS dataset for clinical practice and research.

Keywords: Linked data · Adverse event · Clinical decision support · Health care

1 Introduction

This paper describes a prototype system for clinical decision support, Hubble, that demonstrates the ease with which (medical) legacy data can be turned into RDF, linked to other data sets, and used to support both clinical research and clinical practice. At the heart of the system lies a Linked Data version of the Adverse Event Reporting System (AERS) dataset of the Federal Drug Administration (FDA). To some extent, this exercise can be categorised under the heading 'yet another exposing of data as Linked Data'. However, the system convincingly demonstrates three important *sales pitches* of Linked Data: *interoperability*, *interlinking*, and *tool availability*. In particular, the system shows the huge difference in *usability* between the original 'dead' dataset and the 'live' Linked Data version, it shows how *quickly* this can be achieved using standard tools, and it *validates* the standard three-tier architecture that separates *data*, *application logic* and *presentation*. We validate the quality of the dataset by comparing it to results of a real use-scenario on the AERS dataset.

Clinical Decision Support. Clinical decision support (CDS) can be defined as "the use of the computer to bring *relevant* knowledge to bear on the health care

E. Simperl et al. (Eds.): ESWC 2012 Satellite Events, LNCS 7540, pp. 458–462, 2015.
DOI: 10.1007/978-3-662-46641-4_45

and well being of a patient" [2]. Clinical guidelines play a central role in CDS systems; they contain the consolidated knowledge on patient treatment. However, guidelines are *slow movers*, decided upon in periodic conferences where new evidence is weighed for updating a guideline document. The evidence itself, however, accumulates at a tremendous pace: there are numerous clinical trials and well over 10 thousand publications on breast cancer every year. Therefore, a CDS should bring together patient information, relevant guidelines and important new findings in clinical research. In this context, the key challenge is: how to ensure that the information presented by the CDS is relevant and trustworthy?

The Data. The fields of health care and life science (HCLS) have traditionally seen a lot of attention from the Semantic Web community, and vice versa: semantic web languages, and their predecessors have proven to be a convenient paradigm for representing biomedical knowledge. Vocabularies in the HCLS field are highly standardised; computer analysis, and computer-based information exchange are ubiquitous throughout the field (viz. the Humanities). As a result, many (bio)medical databases and terminologies are now published as linked data, taking up about a fourth of the Linked Data cloud. Examples are medical vocabularies such as SnomedCT, MeSH, MedDRA, and the NCI Thesaurus (all part of the Unified Medical Language System (UMLS)),[1] and datasets such as LinkedCT (clinical trials), Sider, Drugbank and RxNorm (drug information), Uniprot (protein sequences), to name but a few.

The AERS datasets is one of the few remaining, large publicly available medical data sets that until now have not been published as Linked Data. An adverse event (AE) is an adverse change in health or side effect while the patient is receiving treatment. A *serious* adverse event (SAE) is life-threatening and, amongst others, may result in death, requires hospitalisation or prolongation of existing hospitalisation and will result in persistent or significant disability or incapacity. Known chemotherapy-related SAEs in breast cancer (US only) were linked to 22 % of hospitalisations. Clearly, from a clinical perspective, serious adverse events are very important: this is where CDS can make a huge difference.

1.1 System Description

The architecture of Hubble follows a three tiered architecture: (a) a 4Store triple store,[2] containing the AERS dataset (AERS-LD), CTCAE,[3] a selection of DBPedia, Sider, and Drugbank.[4]; (b) a set of SPARQL 1.1 queries and some server-side code; and (c) a Java Smart GWT framework client interface.[5] This section briefly discusses the way we *convert* and *link* data, *annotate* documents and *present* the result to a user.

[1] See http://www.nlm.nih.gov/research/umls/.
[2] See http://4store.org.
[3] CTCAE, subset of MedDRA, lists AEs for cancer therapy: http://bit.ly/zOVPUt.
[4] See http://dbpedia.org, http://www4.wiwiss.fu-berlin.de/sider/ and http://www4.wiwiss.fu-berlin.de/drugbank/, respectively.
[5] See http://code.google.com/p/smartgwt/.

Data Conversion and Linking. The AERS data files are published on a quarterly basis, as zip files containing dollar separated tables. These zip files are roughly 20 MB in size, and available from the FDA website from two separate static webpages.[6] Converting this data is a five step process: (1) scrape the FDA website, download and unzip the data dump; (2) check integrity of the files, applying fixes if necessary;[7] (3) import the data into a MySQL database; (4) dump the data to RDF following a D2RQ mapping;[8] and (5) import the data into 4Store.[9] This conversion was implemented as a pipeline called through a Python provenance wrapper. This wrapper generates provenance information expressed in the PROV-O vocabulary.[10] Due to hardware limitations we had to restrict the dataset to the years 2011 and 2012 (first two quarters), resulting in a total size of 80 M triples.

The AERS dataset is uniquely positioned amidst other HCLS datasets, providing opportunities for linking to drug, location, patient and diagnosis related information. Furthermore, reports in AERS are filled in by hand. Linking out to other datasets could help in identity reconciliation (e.g. drug names, marketing names, and chemical substances) as well as detecting misspellings (e.g. in manufacturer names). We specified mappings between the UMLS, Sider, LinkedCT, Drugbank, DBPedia and CTCAE datasets using the SILK link specification language [4], resulting in over 60 K links based *only* on exact string matches.[11]

Annotations. Step two is the automatic annotation of scientific publications and clinical guidelines (available as PDF files) using the vocabularies in the repository. This process has three steps: *stripping* of PDF documents to plain text, *indexing* the plain text documents and *generating* annotations. We use the PDFBox library[12] for conversion to plain text. Each document is then divided into separate paragraphs, dubbed '*chunks*'. For each chunk we store the coordinates of its bounding box in the PDF. The chunks are then indexed for terms (including synonyms) from the CTCAE ontology, using Lucene.[13]

The Annotation Ontology (AO) is a vocabulary for annotating scientific publications and documents on the Web. AO has a lightweight provenance model, which allows storing information about the authors, curation and different versions for each annotation. We use the AO format [1] to represent an annotation for every term found, using a *prefix-postfix selector* to identify its position inside a chunk. The advantage of using this method is that annotations will persist across different manifestations of the same document (pdf, html, xml, etc.).

[6] See http://1.usa.gov/uyoAI.
[7] For instance, some rows contain line breaks in the wrong places, do not properly escape the separator character or span fewer columns than expected.
[8] See https://github.com/cygri/d2rq.
[9] Unfortunately, exposing through D2R Server turned out to be too slow.
[10] PROV-O-Matic, see http://github.com/Data2Semantics/, currently in alpha stages of development. PROV-O: See http://www.w3.org/TR/prov-o/.
[11] Using less exact matching on drug names can have unwanted consequences.
[12] See http://pdfbox.apache.org/.
[13] See http://lucene.apache.org.

We store an *image selector* that uses the chunk bounding box: this image selector is then used to highlight part of the PDF document.

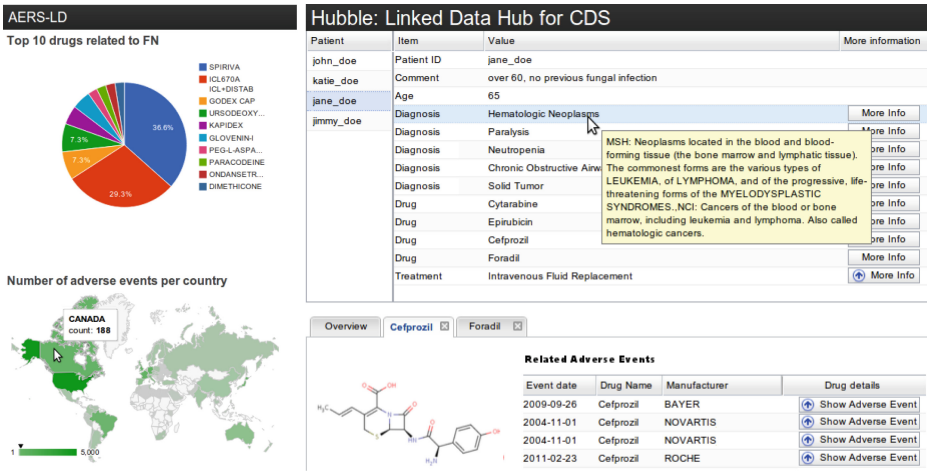

Fig. 1. Examples of possible visualisations (a) and Hubble interface (b)

User Interface. The Hubble interface is our prototype CDS system.[14] It lists patients, shows information about a selected patient, and presents more detailed information about specific parts of the patient information to the bottom (Fig. 1b). This is done in several steps, each consisting of a single SPARQL query. First, we retrieve a list of available patient records. Second, when the user selects a record we retrieve actual patient data: detailed patient information (diagnoses, drugs, age, etc.), enriched with information from the Linked Life Data (LLD) endpoint (e.g. a drug description tooltip).[15] At the same time, we retrieve annotations that match the patient description, and depict small snippets of clinical guidelines and relevant literature. We can drill-down from this detailed information to: more information about a *diagnosis* (taken from LLD), similar *cases* in AERS-LD, *drug information* such as its chemical structure (from LLD and Drugank) and common AEs related to the drug, *provenance* information about an annotation, and the underlying text. We have intentionally limited the amount and diversity of information presented through the interface, pending feedback from expert users.

The AERS-LD repository is publicly accessible through its SPARQL endpoint, and can be browsed through a customised Pubby browser interface.[16]

[14] See http://aers.data2semantics.org/prototypeInterface.
[15] See http://linkedlifedata.com.
[16] See http://aers.data2semantics.org for more information. Pubby: http://www4.wiwiss.fu-berlin.de/pubby/.

Arguably a more actionable presentation than dollar-separated files. The endpoint turned out to be very well suited for various visualisations of the underlying data (Fig. 1a).[17]

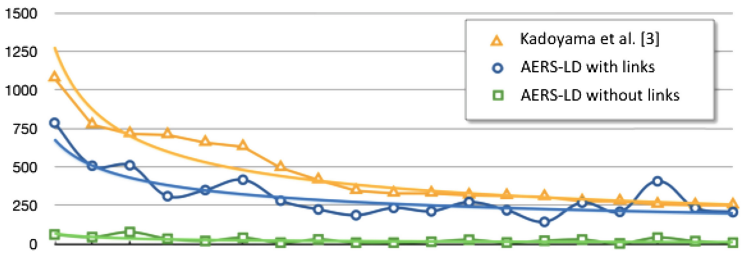

Fig. 2. Number of co-occurrences of Adverse Events and 5-FU (Y-axis). The X-axis represents the ranking of AEs based on the one in [3].

1.2 Evaluation and Discussion

The Hubble CDS prototype was built in a very short period (literally over Christmas), and is already showing real potential for clinical research. We validated the dataset by comparing results from AERS-LD with a study into the co-occurrence of AEs with two drugs (5-FU and Capecitabine) [3]. This study was preceded by a labor intensive effort to clean the dataset: consolidation of multiple names for drugs and removal of duplicate submissions and non-drug entries. We compared AE-drug co-occurrence on the same selection of AEs in [3] *with* and *without* taking advantage of the 60 K links (see above) in AERS-LD. We did not apply any other data cleaning or harmonisation and used only a limited dataset. The result, depicted in Fig. 2, although far from perfect, shows that the Linked Data cloud provides a huge bootstrap for improving the quality of results.

Future work includes publishing the full AERS-LD dataset (all 7 years), increasing both the breadth and depth of annotations through supervised annotation of guidelines, combined with large scale annotation of scientific publications, improving selection and ranking of query results in the Hubble interface based on annotations, citation indexes, and provenance related information.

References

1. Ciccarese, P., et al.: An open annotation ontology for science on web 3.0. J. Biomed. Semant. (2011)
2. Greenes, R.A.: Clinical Decision Support: The Road Ahead. AP/Elsevier Science and Technology, Burlington (2007)
3. Kadoyama, K., et al.: Adverse event profiles of 5-Fluorouracil and Capecitabine: data mining of the public version of the FDA adverse event reporting system, AERS, and reproducibility of clinical observations. Ing. J. Med. Sci. **9**, 33–39 (2012)
4. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - a link discovery framework for the web of data. In: 2nd Workshop about Linked Data on the Web (LDOW 2009) (2009)

[17] Built directly from the endpoint using Sgvizler, http://sgvizler.googlecode.com/.