Linked Data, an opportunity to mitigate complexity in pharmaceutical research and development

Bo Andersson
AstraZeneca R&D
221 87 LUND
+46 46 337621
bo.h.andersson@astrazeneca.com

Kerstin Forsberg
AstraZeneca R&D
431 83 MöLNDAL
+46 31 7065318
kerstin.l.forsberg@astrazeneca.com

ABSTRACT
In the paper we present opportunities and challenges in applying linked data principles to improve the research utility of data to mitigate complexity in pharmaceutical research and development. Association and interpretation of the data needed for research and development in a pharmaceutical company has become too complex task for individual scientists, or even teams to handle. Scientists need data to be organized for associations, prepared for not yet defined use and ready for automation where computers can work alongside the scientists.

Keywords
linked data, semantic web, pharmaceutical research

1. INTRODUCTION
During the WWW2007 conference a breakthrough of the Linked Data idea happened in a session where web experts demonstrated the power of a new generation of the web, a web of data. For us attending the session it was hard to imagine the full potential on what this idea would mean for individual scientists and for a pharmaceutical company. In the Large Knowledge Collider (LarKC) project scientists working in early clinical development have identified [1] how successful research and development rely on scientists having access to data and ability to utilize it, to handle uncertainty and to prepare for the unexpected. To meet these challenges we have identified three key enablers: linked data principles, semantic web standards and open ontologies.

One activity in the W3C interest group for semantic web in Health Care and Life Science (HCLS) has been to link publicly available datasets [2] such as Drug Bank and ClinicalTrial.gov. These datasets are now part of the growing Linked Open Data cloud. An important learning from the collaborative work across health care organizations, pharmaceutical companies and academia groups is that we all need the same datasets, however for different decisions and different types of applications.

2. LINKED DATA FOR PHARMA
A federated knowledge repository of data, built with Semantic Web standards, is a new generation of knowledge base beyond separated document repositories.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2011 ACM 978-1-4503-0608-9/11/03 ...$10.00.

Gaining value from the federation of knowledge requires data to be available using a standard model, of so called RDF triples. Furthermore, access to data in a standard model is not enough, but relationships among data must also be made available as RDF triples to create a web of data. This collection of associated datasets can also be referred to as Linked Data.

A typical case of a large linked dataset is DBpedia, which makes the structured content of Wikipedia available as RDF triples. The beauty of this is not only that Wikipedia data becomes available as linked data, but also that it connects to other datasets publicly available such as Drug Bank. For example, the entity identified with the URI – http://dbpedia.org/resource/Omeprazole, is linked to the descriptions of the same entity in Drug Bank, also made available as linked data. This means that facts available in the two different datasets can be combined. For example the name Losec, and the 70+ other brand names for Omeprazole, from Drug Bank can be integrated by different applications and can also be directly queried, e.g. when searching for brand names for drugs categorized in Wikipedia as essential medicines by WHO.

3. OPPORTUNITIES FOR PHARMA
Scientists routinely utilize publicly available documents as a complement to data in internal and licensed sources. Most of the scientists have preferred sources they utilize to gain knowledge from and to associate with the data relevant for their decisions and questions. The knowledge mining is often done for individual tasks, and in many cases done manually by scientists together with informatics and library experts.

Imagine an environment where internal data easily connects with external datasets from different sources. The associations made,
do make use of relationships developed by experts, in terms of so-called ontologies, and thereby help scientists find answerers not obvious today. While broaden the scientific spectrum where individual scientists find their answers, this would also stimulate more cross scientific sharing and establish a foundation for improved computational support in research and development.

We therefore propose the prospective use of linked data principles as a component of the information infrastructure for pharmaceutical research and development, to make sure internal data connects to external data. This is a step towards an information infrastructure where data is prepared for association, for not yet defined use, and for facilitating computers to function alongside scientists to mitigate the complexity.

4. CHALLENGES FOR PHARMA

From internal semantic web projects, by participation in the LarKC’ and in W3C HCLS since 2006 we have built experience in linking data. Although integration of disparate datasets today is easily done by experts, there are still more to do. We need to reach a maturity level where datasets is routinely published in a way that automatically associates them with other datasets.

The Linking Open Drug Data (LODD) team in W3C HCLS revealed various challenges during the work to connect public available datasets such as DrugBank and ClinicalTrial.gov. Many of them relate to technology but a key aspect is peoples’ willingness to constructively face the tension of opposing needs to improve the research utility of shared datasets.

Relationships are a significant challenge for shared datasets in pharmaceutical research and development. We often have a strong prevalence of terminology conflicts, synonyms, and homonyms. For instance, the word ‘drug’ can refer to the whole pharmaceutical product, or just the active ingredient. Biomedical ontologies represent what entities exist in portions of the biological and clinical reality, and how they relate to each other. One such ontology for translational medicine (TMO) [3] is developed by a W3C HCLS team with representatives from pharmaceutical companies, health care providers, National Centre for Biomedical Ontology (NCBO) and academia groups. TMO structure the different meanings of ‘drug’, by identifying different types of entities and making relationships between them explicit: ‘Molecular entity’ for single molecule kinds, ‘Active ingredient’ as a role played by a chemical substance which has the disposition to treat a certain disease and is part of a ‘pharmaceutical formulation’ for biologically active chemicals. A ‘Pharmaceutical product’ is a formulated pharmaceutical that has been approved.

The difficulty in using precise and consistent defined relationships in ontologies, such as the TMO, for improved research utility should be balanced with utilization of more mundane and relaxed relationships as for example from DBpedia, a source humans are comfortable to use.

The research utility of shared datasets will also require explicit provenance information to make it possible for scientists and computers to make trust judgments about the datasets. Linked datasets also needs to be described with a vocabulary that allow for automation of discovery, selection and association. The Vocabulary Of Interlinked Datasets (void) standard allows formally describing shared datasets and thereby improving the research utility.

A pragmatic approach for linked data management is needed that motivates people to constructively face the tension between different needs. This must be an iterative process where we continuously improve the research utility of shared datasets by applying three key enablers:

- Linked Data principles, including establishing global namespaces of URI:s
- Semantic web standards, also for provenance data and dataset descriptions
- Open ontologies, with the required level of precision and consistency

And, also a process where computers are used to work alongside scientists to automatically:

- Identify entities and assign URI:s
- Structure data and capture provenance data
- Mine for causal relationships and inconsistency

REFERENCES


http://www.larkc.eu/
http://www.w3.org/2001/sw/hcls/
http://www.w3.org/2005/Incubator/prov/wiki/Main_Page
http://vocab.deri.ie/void