Universität Leipzig
Faculty of Mathematics and Computer Science
Department of Computer Science

Patient-centered Drug Management based on Linked Open Data

(Arbeitstitel)

Master's Thesis

Leipzig, March 2013

Submitted by
Nitzschke, Marcus
Master's programme Computer Science

Thesis Supervisors:
Prof. Dr. Klaus-Peter Fähnrich
Dipl. Inf. Romy Elze

abstract

Keywords Semantic Web, Linked Open Data, Drug Management

Contents

1 Introduction

1.1 Subject

The prescription and application of drugs in health care is one of the most important parts in the healing process of a patient. If this process is not managed accurately consequences from financial losses to human harms are possible. In Germany estimations state that annually 25.000 patients die because of medication errors [1] (status 2010). Managing drugs means to support the different phases drugs passes during the healing process. The *World Health Organization* (WHO) published a Drug Management Cycle which defines four different phases [2]. Details about this Drug Management Cycle are given in section 2.1. An important base to implement a proper drug management and therefore to avoid the problem of medication errors is that all available knowledge about drugs and their characteristics is freely accessible and usable. This process of publishing medical – drug related – data has already started in the last years with projects like *Linking Open Drug Data* (LODD) [3] or BioPortal [4]. LODD is a project that links the knowledge of several vocabularies by recognizing common entities and freely provides this linked data. More details about this approach are given in section 2.3. BioPortal offers a platform where people or projects serve their biomedical ontologies and knowledge bases. These knowledge bases can in turn be mapped so that the same entities in different vocabularies are identified. Although it is possible to build third-party applications on top of this data, such applications are very rare at this time. In the case of LODD only small projects exist like DiseaseCard [5] or Pharmer [6].

Besides these third-party projects mainly clinical applications implement specific drug related functionalities. For example a *Patient Information System* stores all the details about the prescribed drugs of a certain patient. Therefore a nurse could be supported in the process of drug application by offering information about the route of application or the maximum dosage. Another application, like a *Computerized Physician Order Entry System* that selects and prescribes drugs, could check interactions of the possible drugs. And for both of the applications the side effects of a certain drug may be of interest. These examples show that many different medical applications implement drug related functionalities. And many of them also share the same requirements because there is a set of common questions, like *"What are the side effects of drug X?"*. But ignoring these shared requirements, many of the applications provide their own – mostly proprietary – knowledge bases.

A comparable situation on the data level of another domain was recently solved. The Wikidata project "centralizes access to and management of structured data" [1]. It solves the problem that many different Wikipedia provided common data about the same entities, e.g. the population of countries. Now Wikidata centrally provide this data through a common interface and the different Wikipedia refer to this source. In consequence there exist only one place where the data has to be updated and maintained.

1.2 Problem

In section 1.1 the diversity of applications that implement drug related functionalities was mentioned. This goes along with the fact that multiple knowledge bases are used which actually should provide the same data. This leads to several problems.

The first problem – insufficient integration – follows from the different scopes and the different expressivity of the knowledge bases. This means there are many knowledge bases for special domains like side effects or alternative medicine. But if one knowledge base wants to use the data of another special domain this is usually not possible, for instance because of an insufficient semantic integration. This means that term x in knowledge base X does not implicitely refer to the same object as term x in knowledge base Y. Therefore the knowledge bases have to provide the data on their

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

own.

This is the starting point of the next problem: redundancy. Many drug related knowledge bases include a set of common facts that are essential for this domain, like the code of the *Anatomical Therapeutic Chemical Classification System* (ATC). So this is a common *"Don't repeat yourself"* problem. A side effect of this problem is the higher error rate that a given statement of a knowledge base is wrong.

So the third problem is the correctness of the knowledge. Proprietary knowledge bases generally have to deal with a strong limited number of people validating the data. In contrast open projects like Wikipedia are considered as "self-healing" information systems because of the high number of volunteers that report and correct data errors. To transfer this phenomenon to medical data it is unimaginable that everyone could edit such important information, but reporting errors would be a strong impact nonetheless.

Section 1.1 presented two third-party applications using open data as their knowledge sources. A common problem using these data sources is the extensive schema knowledge that is required to use the data efficiently. Besides the information about the usage of the given interface the developers are often forced to understand the structure of the data to perform their respective queries. This can be a deal-breaker for developers and therefore for innovative applications.

1.3 Motivation

If the given problems of section 1.2 could be solved this would lead to a better data quality in the first place. Data quality contains in this context the expressivity, completeness and correctness of the knowledge. Medical applications would benefit from this improvement by getting a proper knowledge foundation. Following from that this would lead to unified, more well-grounded decisions for drug related questions. This would hopefully avoid some of the prescription errors that are made during the healing process of a patient.

Solving the problem that developers often have to know the schema of a certain knowledge base the consequence would be a reduced starting barrier for application developers. Instead of knowing about the whole schema of a knowledge base, only the knowledge about using the given interface would be required. Probably this would lead to many interesting applications on top of the given data.

1.4 Objectives

Emerged from the motivation, this thesis shall proof that Linked Open Data is an appropriate knowledge source for drug management tools. For this purpose a web-based *Application Programming Interface* (API) will be provided. This API will support the process of answering drug related questions according to the WHO drug management life cycle. This is achived by offering several API endpoints – e.g. for drug-drug interactions – that gather information from several semantic knowledge bases and return the merged results. One design goal is the simplicity of the API to reduce the starting barrier of developing new applications based on the given interface. More details about the implementation process are described in chapter 3.

This thesis shall also provide two use cases where the usage of the API will be demonstrated and evaluated. The first use case is the integration in Dispedia – an information system in the complex field of rare diseases [7]. The main purpose here is the integration of the drug-drug interaction endpoint. The second use case is a personal drug management portal built around this API to support privat persons managing their prescribed drugs. This use case shall demonstrate the other available endpoints of the provided interface. Chapter 4 will offer more information about this evaluation process.

2 Preliminaries

2.1 Drug Management

To get a better understanding what the term Drug Management stands for it is essential to clearify the term Management at first. Management comprises the non-executing tasks of a certain domain... The term Drug Management was adopted by the WHO in their publication about managing drugs at health centres [2]. The term management is defined in their publication as follows: TODO:

> Management is the act or art of being responsible or in charge and conducting or supervising something (e.g. a health centre pharmacy, business, public undertaking) with a degree of skill and address. It is the judicious use of means to accomplish an end (i.e. public health).

With this definition in mind the question can be answered why it is useful to manage drugs. Refering to the WHO there are three main reasons:

1. "Firstly, drugs are part of the link between the patient and health services."
   This is the most trivial reason that addresses the direct influence of drugs to the patients health state. If drugs are not managed properly the success of the treatment process is imperiled.
2. "Secondly, poor drug management [...] is a critical issue, but major improvements are possible that can save money and improve access."
   Besides the medical benefits, a proper drug management can also result in financial advantages. These can be achived by an improved selection and procurement of the required drugs.
3. "Finally, drugs are no longer the responsibility of health workers only."
   blah TODO:

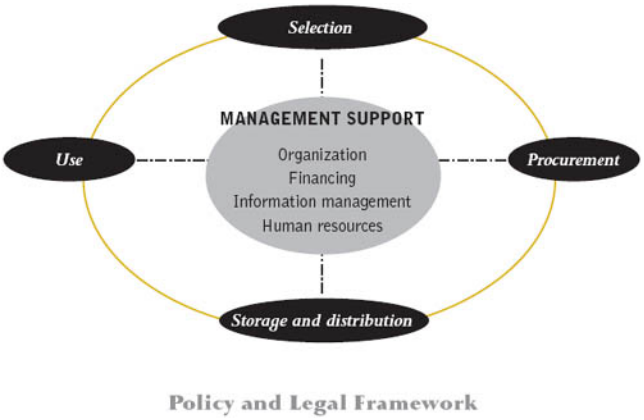

Figure 2.1: WHO drug management life cycle

The WHO proposes a drug management life cycle which is the foundation to implement a proper drug management at a certain institution. This life cycle which is shown in figure 2.1 contains four phases which are described shortly in the following listing: TODO:

- Selection
- Use
- Procurement
- Storage and Distribution

According to the WHO these phases "are interlinked and are reinforced by appropriate management support systems (i.e. tools)". The API which is developed by this thesis is settled up in this category of support systems.

One point has to be remarked regarding the scope of the WHO recommendations. Althought the cited publication addresses health centres, recommendations like the life cycle are easily adoptable to other areas, e.g. private healthcare.

2.2 Semantic Web

The ongoing growth of digital data in the past years has revealed many shortcomings that occur with these amounts of data. Fields of research like *Big Data* also show the importance of managing these amounts of data efficiently. One key issue since years is

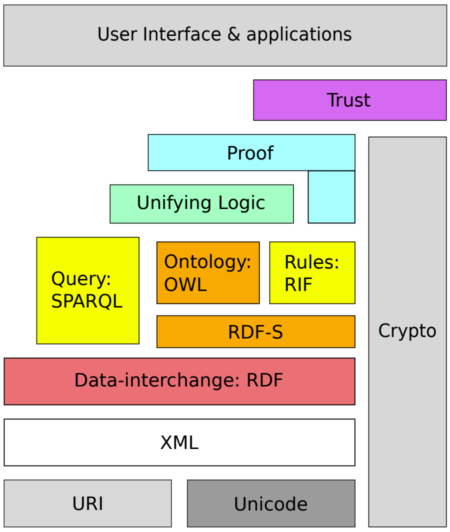

Figure 2.2: Semantic Web Stack

the distinction between syntactically and semantically processable data. Tim Berners-Lee proposed in 2001 a model to describe specific data statements in a better way. This approach was called *Ressource Description Framework* (RDF) and is one important part of the Semantic Web Stack. This stack combines several technologies and standars that are necessary to transfer the syntactic web to a semantic web. The Semantic Web Stack is illustrated in figure 2.2. The advantages of a semantic web are numerous. The main advantage is that the information are not only machine readable but machine processable. This entails other improvements like reasoning that transfers implicit knowledge into explicit.

The following paragraphs will describe the most important technologies and standards of the Semantic Web Stack that are necessary for this thesis:

At the lowest level of the stack there are the Uniform Resource Identifier (URI). A URI identifies a specific resource, e.g. a car. For several other domains there already exists such concepts as URIs. For example the Uniform Resource Locator (URL) identifies web sites or an International Standard Book Number (ISBN) identifies books. The URI itself is just a set of characters that can be additionally divided into five segments: scheme, authority, path, query and fragment. An example of such a URI would be: `http://example.com/Alice`. This URI would describe the resource “Alice”.

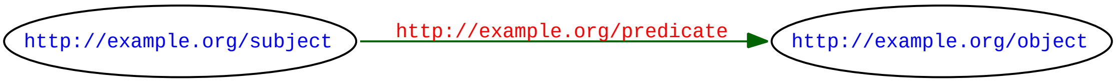
http://example.org/subject
http://example.org/predicate
http://example.org/object

Figure 2.3: Example of a RDF triple

Based on the identification of resources, the already mentioned Resource Description Framework has the goal to describe these resources. Therefore RDF is one of the most important concepts of the Semantic Web Stack. The approach of RDF to model knowledge is to express statements as triples. This is tight to how natural language is mostly constructed. Obviously that is why the components of the triple use the well known names of *Subject*, *Predicate* and *Object*. Figure 2.3 shows an example triple. The subject of such a triple has to be always an URI, same belongs to the predicate. An exception is the object which can be additionally a literal. A literal is a direct information typed in a specific way, e.g. as date or simply text, which refers to no other resource. For more advanced models it is possible to describe the subject and object in a anonymous way, without a concrete entity. These anonymous resources are called blank nodes, but will be without broader relevance for this thesis. At the end one important distinction has to be made about RDF. Often RDF is getting confused with RDF/XML – a notation of RDF statements. While RDF itself only describes how to model information, several RDF notations exist that express these information in several ways. Examples for such notations besides the already mentioned RDF/XML is N-Triples, Turtle or JSON-LD.

On top of RDF the Semantic Web Stack levels SPARQL, an RDF query language. SPARQL enables comprehensive queries over a knowledge base built on RDF. SPARQL itself uses many well known keywords from SQL, but because of the graph structure of RDF and therefore of the triple stores the languages naturally differ.

Given the following two triples in Turtle notation:

```
@prefix ex: <http://example.com/>.
ex:anna ex:name ‘‘Anna’’.
ex:bob  ex:name ‘‘Bob’’.
```

Then the following simple SPARQL query would return the result “Bob”.

```
SELECT ?name
WHERE {
  ex:bob ex:name ?name.
}
```

With SPARQL 1.1 so called *Federated Queries* were introduced. These Federated Queries are Statements that are queried against multiple (federated) SPARQL endpoints. This is a key feature for an easy integration of several RDF graphs.

2.3 Linked Open Data

As described in the previous section RDF statements spans a graph of triples. Thereby, each knowledge base spans their own graph. Examples for large knowledge bases are DBpedia[2], LinkedGeoData[3] or Drugbank. *Linked Data* in general now describes the situation that entities of one knowledge base link to entities of another knowledge base. For example an entity of DBpedia about a drug could link to the same entity in Drugbank for further information. If the mentioned knowledge bases are freely available and usable then this is called *Linked Open Data*.

Tim Berners-Lee proposed some rules that state whether a data source is Linked Open Data, or not. These rules are:

1. Use URIs to denote things.
2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.
3. Provide useful information about the thing when its URI is dereferenced, leveraging standards such as RDF, SPARQL.
4. Include links to other related things (using their URIs) when publishing data on the Web.

Further, Berners-Lee categorizes Linked Open Data on the basis of a five star rating. Therefore the more a knowledge base follows these principals the more stars it earns.

[2] `http://dbpedia.org`

[3] `http://linkedgeodata.org`

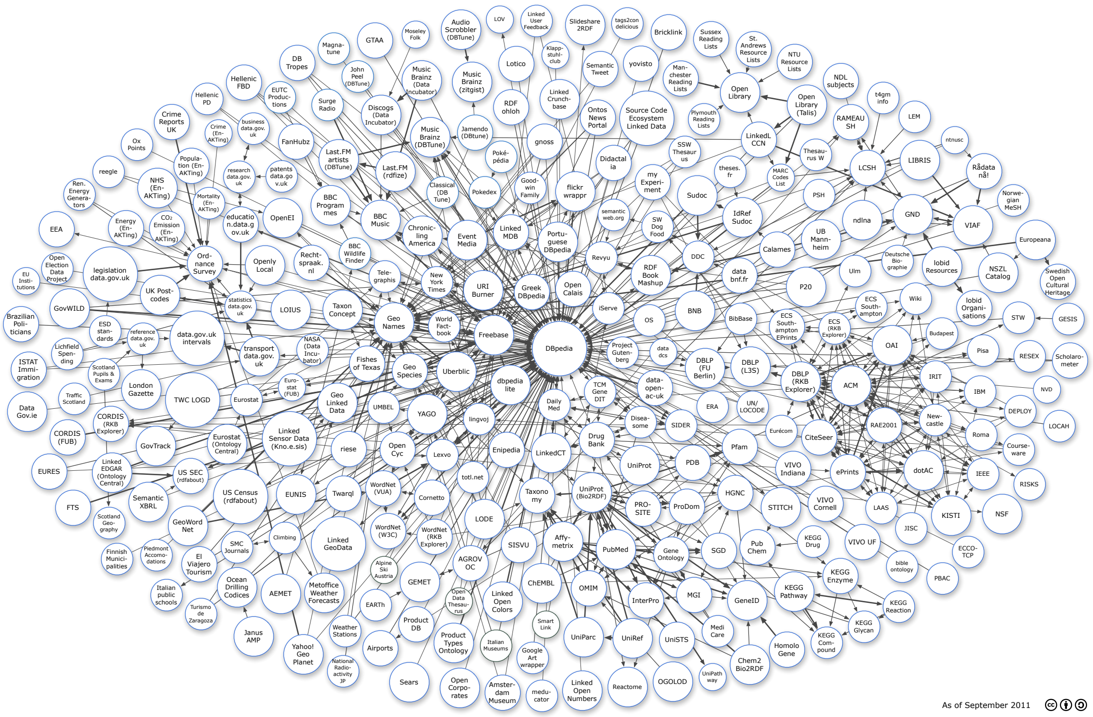

Figure 2.4: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

- **1 star** - Available on the web (whatever format) but with an open licence, to be Open Data
- **2 stars** - Available as machine-readable structured data (e.g. excel instead of image scan of a table)
- **3 stars** - as (2) plus non-proprietary format (e.g. CSV instead of excel)
- **4 stars** - All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
- **5 stars** - All the above, plus: Link your data to other people's data to provide context

In conjunction with SPARQL this leads to large federated collections of knowledge that are comprehensively queryable. Figure 2.4 gives an overview over the Linked Open Data cloud in 2011. But this kind of figure will end soon because of the rapidly evolving number of data sets.

3 Methods

- sparql federated queries nicht überall unterstützt -> construct und dann lokales mergen
- übersicht datensätze
  - tabellarisch/grafisch
  - proprietäre quellen die nicht verwendet werden konnten
    - senger et al

4 Evaluation

4.1 Dispedia

4.2 Portal

4.2.1 Subject

- schwer für ärzte überblick über verfügbare medikamente am markt zu behalten
- oft bilden sich aus erfahrung paare von medikamenten zu gegebener krankheit die nur selten dann erneuert werden (Quelle oä?)
-
- ärzte verschreiben oft unabhängig voneinander medikamente
- patient hat möglichkeit rezeptfreie medikamente zusätzlich einzunehmen
- bpz verweisen nur grob auf mögliche interaktionen
  - oft nur gruppen von medikamenten genant, nicht medikamente konkret

4.2.2 Problem

- patienten bekommen arzneimittel verschrieben auf grund der erfahrung der ärzte nicht evidenzbasiert bzw. an aktuellen studien ausgerichtet
- mehr oder minder blindes vertrauen des patienten den ärzten gegenüber
-
- bei n arzneimittel gibt es (n^2-n)/2 mögliche interaktionen
  - 2->1, 3->3, 4->6, 5->10

4.2.3 Motivation

- transparente möglichkeit der unterstützung von drug management eines patienten
- bereicherung der persönlichen freiheit durch nachvollziehbarkeit und kontrolle der eigenen arzneimitteltherapie

5 Summary

6 Discussion

- OpenPHACTS - ähnlicher ansatz für drug discovery

Bibliography

[1] Pharmazeutische Zeitung. Medikationsfehler: 25.000 Tote jährlich. `http://www.pharmazeutische-zeitung.de/index.php?id=32557`. abgerufen am 02.12.2012.

[2] World Health Organization. *Management of Drugs at Health Centre Level - Training Manual*, 2004.

[3] Anja Jentzsch, Jun Zhao, Oktie Hassanzadeh, Kei-Hoi Cheung, Matthias Samwald, and Bo Andersson. Linking open drug data. In *Triplification Challenge of the International Conference on Semantic Systems*, pages 3–6, 2009.

[4] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2):W541–W545, 2011.

[5] José Luís Oliveira, Gaspar Dias, Ilídio Oliveira, Patrícia Rocha, Isabel Hermosilla, Javier Vicente, Inmaculada Spiteri, Fernando Martin-Sánchez, and António Sousa Pereira. Diseasecard: A web-based tool for the collaborative integration of genetic and medical information. In *Biological and Medical Data Analysis*, pages 409–417. Springer, 2004.

[6] Ali Khalili and Bita Sedaghati. Pharmer–towards semantic medical prescriptions. In *eTELEMED 2013, The Fifth International Conference on eHealth, Telemedicine, and Social Medicine*, pages 9–14, 2013.

[7] Romy Elze and Klaus-Peter Fähnrich. The dispedia framework: A semantic model for medical information supply. In *ICONS 2013, The Eighth International Conference on Systems*, pages 59–63, 2013.

List of Figures

List of Tables

Declaration

This master's thesis is the result of my own work. Material from the published or unpublished work of others, which is referred to in the thesis, is credited to the author in the text. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this thesis and the degree examination as a whole.

Leipzig, 14. March 2013 Signature