OpenDrugWiki – Using a Semantic Wiki for Consolidating, Editing and Reviewing of Existing Heterogeneous Drug Data

Anton Köstlbacher[1], Jonas Maurus[1], Rainer Hammwöhner[1],
Alexander Haas[2], Ekkehard Haen[2], Christoph Hiemke[3]

[1] University of Regensburg, Information Science
Universitätsstr. 31, 93053 Regensburg, Germany
[2] University of Regensburg, Clinical Pharmacology, Department of Psychiatry
Universitätsstr. 31, 93053 Regensburg, Germany
[3] University of Mainz, Neurochemical Laboratory, Department of Psychiatry
Untere Zahlbacher Str. 8, 55131 Mainz, Germany;

anton.koestlbacher@sprachlit.uni-r.de; rainer.hammwoehner@sprachlit.uni-r.de;
jonas@maurus.net; ekkehard.haen@klinik.uni-regensburg.de;
hiemke@mail.uni-mainz.de

Abstract. The ongoing project which is described in this article pursues the integration and consolidation of drug data available in different Microsoft Office documents and existing information systems. An initial import of unstructured data out of five heterogeneous sources into a semantic wiki was performed using custom import scripts. Using Semantic MediaWiki and the Semantic Forms extension, we created a convenient wiki-based system for editing the merged data in one central application. Revised and reviewed data is exported back into production systems on a regular basis.

Keywords: drug database, medical information system, semantic wiki, data conversion

1 Introduction

PsiacOnline[1], a drug interaction database for psychiatry in German speaking countries, was released in 2006. As of 2010 it contains over 7000 drug interactions with comprehensive information on pharmacological mechanisms, effects and severity of each interaction. Strong emphasis lies on guidance how to handle interactions in practice. [1]

[1] *PsiacOnline* is an online service offered by *SpringerMedizin*: http://www.psiac.de

Built on top of the component-based and event-driven prado[2] framework PsiacOnline features an easy to use authoring tool for drug data. It also provides a simple XML interface for reusing data in other information systems, particularly Laboratory Information Systems (LIS).

After the system's introduction, we identified several additional data sources that are frequently used at the affiliated research institutes[3], providing content for PsiacOnline. These data sources consisted of different document types like Microsoft Excel sheets, Word documents, CSV data and relational databases. For example, biological pathway information for psychiatric drugs was kept in a Word document of which new versions were distributed to the lab staff via email. The staff also used relational databases with pharmacokinetic data that were part of a LIS system used for managing lab workflow. [2]

Other examples were manually edited Excel sheets with brand names and drug names or international non proprietary names (INN[4]) and ATC[5] code tables. They were distributed through an informal email based workflow in the lab.

Analysis of the various documents and their content showed that all content should be integrated into the existing dataset of PsiacOnline. This was the starting point of the OpenDrugWiki project, which now combines the converted and imported data sources with the existing PsiacOnline dataset. It offers an easy-to-use interface for collaborative editing of the unified data in one place.

The article shows a use case for semantic wikis in production environments in the field of professional pharmacological information in psychiatry. It describes why we chose a semantic wiki, how the import of the data is done, what the editing and review process looks like and how the data can be reused in existing and future software systems.

2 Why Use a Semantic Wiki?

Instead of trying to convert the data and thus expanding PsiacOnline and importing the data directly, we chose an approach that uses a semantic wiki as an intermediate system. This semantic wiki also provides a full replacement for PsiacOnline's authoring system. The decision for a semantic wiki was in fact not a single one, it was based on three decisions to the following questions: Why use a wiki? Why use semantic web technologies? And why use the combination of both?

---

[2] http://www.pradosoft.com

[3] University of Regensburg, Clinical Pharmacology, Department of Psychiatry; University of Mainz, Department of Psychiatry; University of Regensburg, Department for Information Science; Regional Hospital Kaufbeuren, Department of Psychiatry

[4] INN: International Non Proprietary Name: Generic name of a pharmaceutical ingredient issued by the World Health Organisation (WHO).

[5] ATC: Anatomical Therapeutic Chemical Classification System, used to classify drugs and other medical products, controlled by WHO Collaborating Centre for Drug Statistics Methodology (WHOCC).

2.1 Why Use a Wiki?

The main reason for favoring a wiki is that we will invite more institutes and authors to contribute to PsiacOnline, therefore supporting collaboration and versioning is of great importance. Wikis are well known to be of great use for distributed text editing and reviewing. This also applies to scientific communities. [3][4][5]

We wanted to replace the various inconvenient email based workflows by one structured storage and workflow system. We anticipate time savings in the participating organizations by centralizing and streamlining the editing and reviewing process. Time savings are already confirmed by users and are mostly achieved by discarding the inefficient email based workflows and by the possibility of editing all data in one place as well as the ability to instantly see changes made by other users.

Eventually, the affiliated research institutes do not only want to use the new system to publish information on drug interactions which is both, necessary and useful for psychiatrists or family doctors, but they also need a central platform for exchange of the underlying pharmacokinetic mechanisms which is important to motivate and execute further research.

2.2 Why Use Semantic Web Technologies?

Semantic Web technologies provide standards-based data exchange (RDF/XML) and storage methods (Triple Stores) and a powerful query language (SPARQL). This makes it easy to export parts of the semantic data back into production systems like PsiacOnline, LIS or other medical information systems (MIS) after they passed a review process.

The possible integration of data from other existing pharmaceutical or biomedical ontologies, for example the Open Biological and Biomedical Ontologies[6] was another reason to favor semantic technologies. Having the ability to provide data for services like Linked Life Data[7] or Linking Open Drug Data (LODD[8]) was additionally convincing.

2.3 Why Use a Combination of Both?

Based on the above arguments, the decision for a semantic wiki was identified as the best way to go forward. Both, wiki and semantic web technologies in combination, together with using a triple store connected to the wiki support each other in achieving the goals described above. Using a semantic wiki offers the possibility to extend the underlying data model at any time without trouble. This is important, as extensions will be needed when new data relevant for research becomes available.

---

[6] http://www.obofoundry.org

[7] http://linkedlifedata.com/

[8] http://esw.w3.org/topic/HCLSIG/LODD

3 Semantic Wiki Evaluation

We evaluated four of the mature semantic wiki engines: IkeWiki [6], Semantic MediaWiki [7] (with the Halo extension [8]), OntoWiki [9] and AceWiki [10]. Finally Semantic MediaWiki was chosen to be the product best-suited for our purposes, mainly because of its usability and the underlying MediaWiki [11] software, known to have a broad developer and user base and a big variety of available extensions. Further results of the evaluation, presented in a more general article, can be found in [12].

4 Implementation Details

OpenDrugWiki is based on Semantic MediaWiki and extensively uses templates, magic words[9], and various extensions. These include the Parser Functions extension[10], Semantic Results Format[11] and Semantic Forms[12] for convenient editing. Attached to the wiki we use the basic triple store for Semantic MediaWiki provided by Ontoprise [13]. It is based on Jena [14] and allows querying the semantic data that is stored in the wiki via a SPARQL [15] webservice. Having all semantic data available through a standards-based remote interface makes it easy to integrate new applications and gives us a way to bring it back into production systems.

4.1 Data Conversion and Importing

Converting and importing the various data sources proved not to be a trivial task. While reading the SQL databases and Excel sheets is a solved problem, the Microsoft Word documents are converted to Excel sheets using an add-in for Microsoft Word that was developed for this specific task.

After preprocessing the various data sources into Excel files and relational databases, the main import job is performed by a PHP CLI application. This application processes all data by matching known terms to semantic classes (brand names, drug interactions, INN etc) and merges duplicate entries coming from the different sources (see Fig. 3). It also computes redirections for sameAs-relations, based on the information provided in the legacy data sources. The import application then generates articles which are directly imported into MediaWiki using the its command line interface.

[9] http://www.mediawiki.org/wiki/Manual:Magic_words
[10] http://www.mediawiki.org/wiki/Extension:ParserFunctions
[11] http://www.mediawiki.org/wiki/Extension:Semantic_Result_Formats
[12] http://www.mediawiki.org/wiki/Extension:Semantic_Forms

For generating the articles we created boilerplates for each article class (brand names, pharmaceutical ingredients, drug interactions, others) consisting of various MediaWiki templates. Article boilerplates and templates were defined manually after analysis of the available data. A simple mapping between columns in the relational databases, the word and excel tables and the semantic properties is performed by the import tool.

Using this method, we generated and imported about 15,000 articles, consisting of about 3,000 brand name entries, 4,000 entries on pharmaceutical ingredients (INNs), 7,000 drug interactions and 1,000 other articles. All articles have data-type information and semantic properties which results in about 150k RDF triples. Reading and processing the data, generation of wiki articles and importing them into SMW take about one hour altogether.

4.2 Editing and Reviewing

After having successfully imported all data, the semantic wiki is used for editing and reviewing by the PsiacOnline authors as well as carefully selected associate authors. Since there is no Semantic MediaWiki extension available which supports a collaborative peer reviewed editing process, we are forced to manually track all changes made to the wiki articles. This is done by the core authors of PsiacOnline who immediately review all edits made by other authors. To control which data is exported back to production systems we store the user name and the revision date of each article as semantic properties. Only articles which were approved by one of the core authors are imported into production systems. This means, that a reviewer who checks edits of a normal author has to resave the edited articles, even if he himself made no changes. This process works perfectly for the moment, but cannot be seen as a long term solution.

4.3 Querying the Wiki

As a proof of concept and a useful application for the staff in the lab we created a simple Ajax powered web interface to retrieve data from the wiki. Given a list of drugs or brand names, it shows all drug interactions, the biological pathways involved and the citations on which the displayed information is based (see Fig. 2). This tool demonstrates the possibility to query the triple store attached to the wiki and can be used completely independent from the wiki itself. Being in private beta phase at the moment, it will be made publicly available, when the content is completely reviewed and double checked.[13]

The tool uses a PHP proxy script which queries the SPARQL endpoint, preprocesses returned data, and delivers it back to the Ajax application using JSON. Preprocessing consists primarily of dealing with the returned XML and character-set related quirks.

[13] Project Website: http://www.opendrugwiki.org/wq

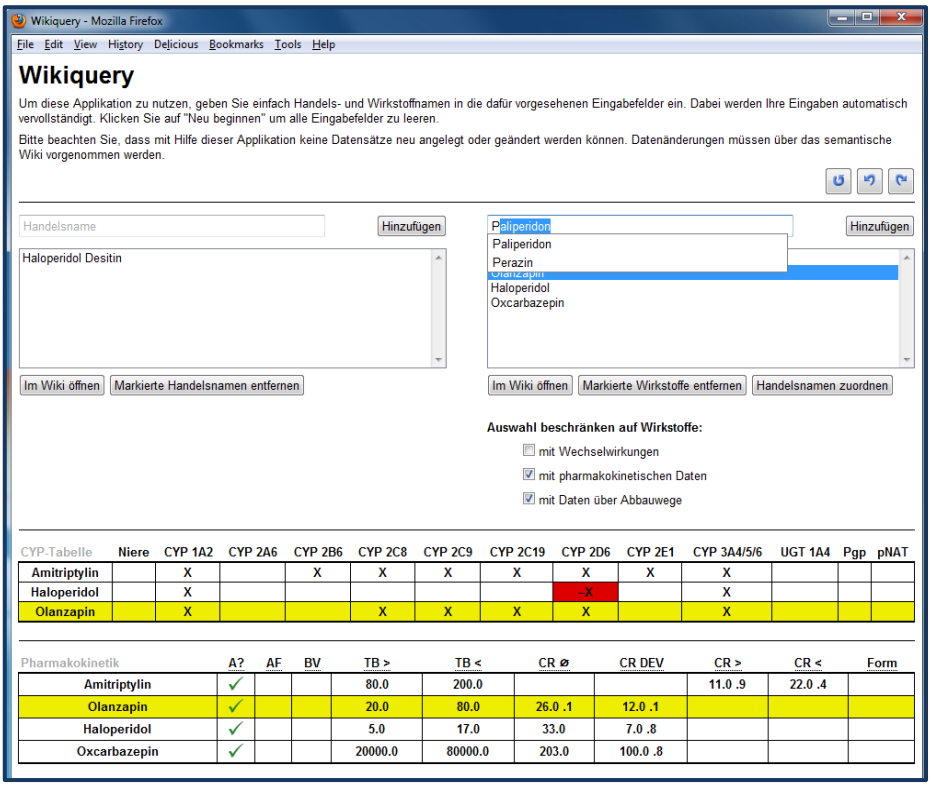

Fig. 1. Screenshot Wikiquery-tool (german)

4.4 Exporting Data

One of the crucial requirements for the project is the possibility to export data back into production systems like PsiacOnline and the labs' LIS software. Initial results were easily achieved by retrieving data from the wiki using ASK or SPARQL via JSON and XML interfaces. This results in structured data which is then synchronized with the data in production systems (see Fig. 3).

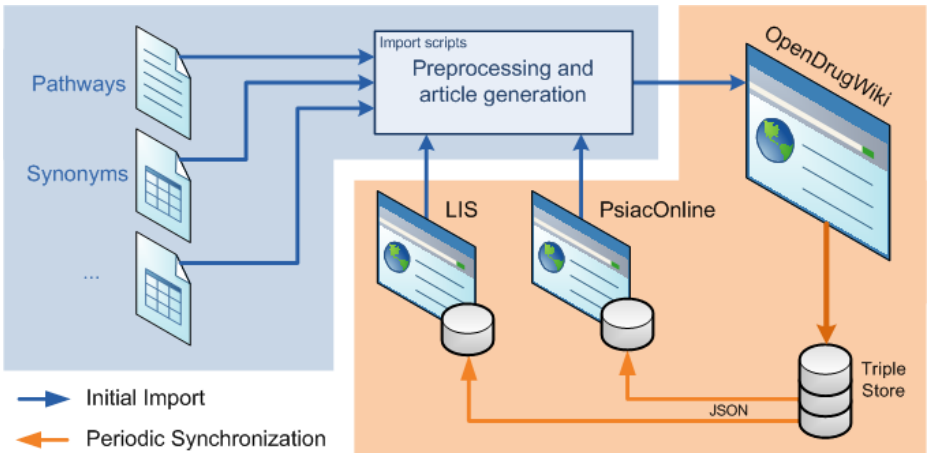

Fig. 2. Graphical overview of the import and export process

5 Conclusion and Prospects

With the approach presented in this article we wanted to show that a semantic wiki is an appropriate tool for consolidation of data with heterogeneous structure, sources and quality. The next step in this ongoing research project is the evaluation of the wiki's suitability to support continuous editing and reviewing processes in the different organizations, especially from a usability standpoint. We are anticipating good results, as Semantic Forms provides an easy-to-use interface for most purposes.

Since Semantic MediaWiki by default only provides semantic data for the latest revision of an article there is currently no easy way for integrating review processes. As we are preparing to open the wiki up to more and more research organizations for editing and contributing information on drugs and drug interactions, being able to have a reviewed and officially approved state of an article is currently the most important missing feature.

We would like to see the MediaWiki extension Flagged Revisions[14] integrated with Semantic MediaWiki, since this would help us to implement review processes and subsequently have reviewed and approved semantic data available in the triple store.

A benefit wikis and semantic web technologies offer is the possibility to create a multilingual information system by using interwiki links and semantic relations. A future task will be the creation of multiple wikis that will allow us to connect terms, drug names and drug interactions in different languages. Mapping classes and properties to standard drug and biomedical ontologies is therefore an important task either. In the near future we will integrate more data from new sources as they become available and begin connecting other production systems to the wiki to provide an efficient tool for researchers for editing drug data in one place.

[14] http://www.mediawiki.org/wiki/Extension:FlaggedRevs

6 References

1. Köstlbacher, A., Hiemke C., Haen E., Eckermann, G., Dobmaier, M., Hammwöhner R., PsiacOnline - Fachdatenbank für Arzneimittelwechselwirkungen in der psychiatrischen Pharmakotherapie. In: Osswald, A., Stempfhuber, M., Wolff, C. (Hrsg.). Open Innovation. Proc. 10 Internationales Symposium für Informationswissenschaft. Konstanz: UVK, 321-326. (2007)
2. Köstlbacher, A.: Information Management In A Neurochemical Laboratory, In: Kuhlen, R.r (Hrsg.) (2009). Information: Droge, Ware oder Commons? Proc. 11 Internationales Symposium für Informationswissenschaft. Konstanz: UVK, 567-570.
3. Leuf, B., Cunningham, W.: The Wiki Way: Quick Collaboration on the Web. Addison-Wesley, New York, 2001.
4. Hoffmann, R.: A wiki for the life sciences where authorship matters. Nature Genetics, Vol.: 40/9: 1047-1051 (2008)
5. Baumeister, J., Reutelshoefer, J., Nadrowski, K., Misok, A.: Using Knowledge Wikis to Support Scientific Communities. In: Proceedings of 1st Workshop on Scientifc Communities of Practice (SCOOP), Bremen, Germany (2007).
6. Schaffert, S.: IkeWiki: A semantic wiki for collaborative knowledge management. In: Tolksdorf, R., Paslaru Bontas, E., Schild, K., editors, 1st Int. Workshop on Semantic Technologies in Collaborative Applications (STICA'06) Manchester, UK, (2006)
7. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: Proceedings of the 5th International Semantic Web Conference (ISWC'06). LNCS, vol. 4273, pp. 935-942. Springer, Heidelberg (2006)
8. Friedland, N.S., Allen, P.G., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., Staab, S., Moench, E. Oppermann, H., Wenke, D. Israel, D. Chaudhri, V., Porter, B., Barker, K., Fan, J. Chaw, S.Y., Yeh, P., Tecuci, D.: Project halo: towards a digital Aristotle, AI Magazine, Winter 2004, (2004)
9. Auer S., Dietzold S., Riechert, T.: OntoWiki – A tool for social, semantic collaboration. In Yolanda Gil, Enrico Motta, Richard V. Benjamins, and Mark Musen, editors, Proc. 5th Int. Semantic Web Conference (ISWC'05), number 4273 in LNCS, pages 736–749. Springer, (2006)
10. Kuhn, T.: AceWiki: A Natural and Expressive Semantic Wiki. In: Semantic Web User Interaction at CHI 2008: Exploring HCI Challenges (2008)
11. MediaWiki contributors: MediaWiki, The Free Wiki Engine, http://www.Mediawiki.org/w/index.php?title=Mediawiki&oldid=65192 (accessed March 3, 2010)
12. Köstlbacher, A., Maurus, J.: Semantische Wikis für das Wissensmanagement. Reif für den praktischen Einsatz? In: DGI e.V./M. Heckner, C. Wolff (Ed.). Information Wissenschaft und Praxis. Mai/Juni 2009, Dinges & Frick GmbH, Wiesbaden, pp. 225-231 (2009)
13. Ontoprise GmbH (Ed.): Basic Triplestore, http://smwforum.ontoprise.com/smwforum/index.php/Help:Basic_Triplestore (accessed March 3, 2010)
14. McBride, B.: Jena: A Semantic Web Toolkit. In: IEEE Internet Computing November/December 2002, pp. 55-59. (2002)
15. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/ (accessed March 3, 2010)