knowIT, a Semantic Informatics Knowledge Management System

Laurent Alquier
Johnson & Johnson Pharmaceutical Research & Development, L.L.C
1000 route 202
Raritan NJ 08867
+1 908 218-6800
lalquier@its.jnj.com

Keith McCormick
Johnson & Johnson Pharmaceutical Research & Development, L.L.C
1000 route 202
Raritan NJ 08867
+1 908 704-5835
kmccorm4@its.jnj.com

Ed Jaeger
Johnson & Johnson Pharmaceutical Research & Development, L.L.C
665 Stockton Dr 104
Exton PA 19341
+1 610 458-6052
ejaeger@its.jnj.com

ABSTRACT
We present knowIT, a collaborative database designed to manage the shared knowledge about Informatics Systems in a research organization. In this paper, we discuss requirements that emerged through years of use and we describe the challenges of migrating content from an existing relational database to a solution based on a Semantic MediaWiki. Finally, we review which customizations were required in order to improve user acceptance, both for editors and viewers. Our experience will serve as a case study for a pragmatic approach to knowledge management.

Categories and Subject Descriptors
K.4.3 [Computer and Society]: Organizational Impacts – computer supported cooperative work.. H.5.3 [Information Systems]: Group and Organization Interfaces – computer-supported cooperative work, Asynchronous interaction, Organizational design, Synchronous interaction, web based interaction.

General Terms
Management, Human Factors.

Keywords
Wikis, semantic web, user experiences, usability, intranets, internal communication, collaboration organizational memory, repositories,, knowledge management, knowledge transfer.

1. INTRODUCTION
At the core of our approach to knowledge management is the realization that, after decades of managing Informatics Systems, IT professionals are still relying too much on closed databases and spreadsheets to keep track of their knowledge of these systems.

WikiSym '09, October 25-27, 2009, Orlando, Florida, U.S.A.

Physical asset tracking systems are generally separated from applications inventories, which are in turn isolated from knowledge about how applications map to business processes. Each one of these systems relies on data contained in the others. The result of these silos is duplicated, obsolete data and still more spreadsheets.

The key to knowledge bases is not only data capture but also dissemination. What are we doing with knowledge we acquire? How is it shared, transferred, maintained? Many Knowledge Management initiatives fail because of lack of content, lack of participation, restrictive approval process and overly complex interfaces [1].

With that in mind, we looked for a flexible way to capture data about components in information systems, from hardware to people, and map out how they relate to one another. The result has to be accessible both by people and machines, so that relevant pieces of information can be exposed to future systems that may need them [2][3]. Because the scope of the project was initially focused on the needs of a group, we approached the design of the system in a very pragmatic way, with a goal to evolve it based on experience instead of using a formal methodology [4][5]. However, we adapted elements of such methodologies to allow the system to expand if necessary.

The purpose of knowIT is to provide a better understanding of our Informatics Systems and answer recurring questions about asset tracking, applications, systems and procedures.

Our intention in this paper is to show how a Semantic MediaWiki has allowed us to reach that goal and give a detailed review of our experience in turning a rigid website based on a relational database into a collaborative database.

2. CONTEXT AND REQUIREMENTS
This project was started in the Informatics department of a large research organization. As such, our mission is to develop and maintain systems to support pharmaceutical research. We are part of a global organization, with multiple operating companies in several countries.

The reason for creating a knowledge base in 2001 was to address the consequences of nearly constant change caused from being at the crossroad between Information Technology and Scientific Research.

Training new resources was a time-consuming scavenger hunt about 'who knows what'. There is a recurrent need to produce reports such as 'Subject Matter Experts Lists', ‘servers to upgrade’ or ‘applications portfolio’.

In addition of answering these questions, the knowledge base had multiple goals:

- Track individual components of Information Systems (Applications, Servers, Databases, Plugins)
- Map relationships between these components
- Capture the Business context (Organizations, Vendors, Geographic Location, People)
- Document known issues, procedures, processes
- Provide several ways to answer questions and generate reports about the state of Information Systems
- Provide a Self Help mechanism for individuals and other support organizations.
- Send out communications and track system outages

Flexibility was key in order to capture and retain the knowledge we need in a changing, distributed environment

2.1 Relational database

After looking at commercial products for departmental knowledge management, it was decided early on to develop a solution in-house. Products available at the time were too specialized in a certain area. Some offered a good knowledge base of articles but lacked proper asset management. Others allowed asset tracking at the cost of a simple and rigid system of articles. Most lacked flexibility and required additional cost in customizations.

Since it was launched, our initial knowledge base evolved into a tool used on a daily basis to capture information about our systems and provide an acceptable level of support with limited resources. This first attempt with a relational database gave us enough control but, over time, reached the limits of its flexibility.

The system featured a web interface accessible to anyone, with a self-help area and search results that could be bookmarked.

It allowed us to capture relationships between applications, servers, databases, people and vendors, as well as documentation on procedures and known issues.

However, we had partial success at including content from other groups as the web interface provided rigid forms and few tools for reviews and discussions.

Changes to the data model translated into an increasing effort to change the interface accordingly. Real world scenarios eventually tend to deviate from pre-defined data models. The way the system responds to exceptions from the established data model is crucial to keep up with changing business environments, organizational structures or the constant flow of tools and versions.

This limited flexibility and the lack of collaborative capabilities forced us to reconsider that approach and look for alternatives.

2.2 Collaborative databases

Lessons learned from the relational database helped us draw a better picture of the capabilities required for a good Informatics Knowledge Management system.

- Flexible yet structured content management
- Enable complex relationships between entities
- Modular and extensible design
- Supports both search and query mechanisms
- Based on collaboration (discussions, comments, community editing)
- Automation of maintenance tasks
- Tools to improve quality of content (monitor and resolve inconsistencies)
- Track and roll back changes
- Import of content from several sources
- Export of content to several formats

Content Management Systems (CMS) provide many of the desired requirements in terms of customization, modularity and version tracking. However, very few support the creation of structured content.

Over the year, we evaluated several products such as Plone [6], Confluence [7], Microsoft Office SharePoint Server [8] or Drupal [9]. We finally decided to use MediaWiki [10] as a platform for our new system, based on evidence that it is used successfully for several popular knowledge repositories starting with Wikipedia.

Since 2006, the Semantic MediaWiki (SMW) project has grown into a solid solution to provide structured data to MediaWiki [11][12]. One of its major strengths is a flexible annotation system that can be queried and displayed in multiple formats.

3. KNOWIT

3.1 Migration from a relational database

Mapping a relational database into a Semantic MediaWiki data model is ultimately a positive change of thinking from a central design to a shared responsibility among contributors. It requires careful planning as well as some compromises due to the differences in structures between the two systems [13].

Our relational database had a classical design – one table per entity and additional tables for relationships with multiple attributes (cross-reference tables). Simple relationships were mapped using foreign keys.

By contrast, Semantic MediaWiki treats everything as a Page.

Pages have unique names, with constraints on the use of case and special characters. These constraints forced us to rename some entities as they were imported and to follow basic naming conventions to avoid ambiguities between pages with identical names.

Attributes need to be defined on each page using a Property and a Value. For example, to declare that a plugin is designed for a particular application, one would have to write the following on the Page created for that plugin:

[[Has host application::some application]]

Like Column Names in a table, Semantic Properties have a defined type (Page reference, date, string, text or even geographic location). Going from a foreign key to a Page reference was a simple process. Other types, such as text fields, had to be adjusted since SMW does not assign a size to a text string.

Pages can be grouped into Categories, loosely following the definition of Tables in the relational database. Categories represent what a page 'is'.

Complex relationships require creating a Category and pages with the corresponding properties.

An extension to SMW, Semantic Forms, was quickly identified as a practical way to standardize on the common properties each page should include. MediaWiki's system of Template was used to standardize how properties were presented to user.

Planning for the migration made it clear that a review of the data model was a necessary step. Attributes in some tables had become obsolete or had inconsistent naming conventions.

Entities from the relational model were organized into a hierarchy of categories to define 'what' we wanted to track in the system.

Semantic MediaWiki provided the necessary flexibility to build a detailed ontology progressively, using real data and use cases to drive the data model and avoiding the cost of a long data modeling effort. In turn, this ontology served as a basis to automate the creation of categories, templates and forms in the Semantic MediaWiki data model.

We decided to import as much content as possible, both for current and retired systems in order to preserve historical data. Tables related to main inventories, such as applications or servers, were imported in one step, while some tables, such as procedures or persons, required too much review. Special forms were created to import them into the wiki on a case-by-case basis.

The migration of data between the two systems followed a typical Extract-Transform-Load (ETL) process. We used scripts to extract data one table at a time. The content of each table was reviewed and corrected as necessary, and used to generate text files with page titles and source code in wiki format. A MediaWiki script (bot) was then used to create pages into the wiki from each text file.

This careful approach ensured a smooth transition to the new system within a matter of months without downtime, as data was immediately usable after being imported.

3.2 Customizations of Semantic MediaWiki

Initial feedback from both editors and viewers of the knowledge base suggested the default user interface needed to be streamlined. For example, search results and recent changes showed too many details and pages tend to become unreachable without a rigorous design. Formal wiki usability studies confirmed issues with navigation, hyperlink management and coordination between contributors [14].

Fortunately, several extensions to MediaWiki are available to significantly improve usability.

3.2.1 Simplified interface

The first area of improvement was about simplifying the interface and making it more familiar to users.

The look and feel of the interface was changed using a MediaWiki skin and style sheet customizations to provide the same standardized interface as other collaborative applications on our intranet.

Category pages were customized based on content. Instead of an alphabetical index of geographical pages, maps are used to display locations. A timeline and a list of dates in chronological order are used for events. Articles and procedures are organized by topics.

The Dynamic Pages List extension was used to create simplified views of 'Recent changes' or 'Popular content'. This extension is a good complement to Semantic queries as it provides a way to query pages on non-semantic properties (such as count of visits to a page, author name or creation date).

We also replaced the default search by Sphinx, an open source search engine that integrates well with MediaWiki. Sphinx allows narrowing down search results by category. It provides control over how search results are displayed and brings up results closely related to search terms (not necessarily exact terms).

3.2.2 Improved navigation

A lot of attention was given to providing users with clear instructions about what they can do with the system. To that effect, pages are organized in three main tasks – Search, Explore and Contribute.

An index of categories displays a hierarchy that provides a simple view of the ontology used to categorize content.

Also available at the top of each page, a search portlet provides access, from a single form, to several search engines used by our research organization.

Links to Index pages are available at the top of every page. The Semantic Drill-down extension provided us with a clean way to browse content, with optional filters to narrow down results. A custom made A-Z index provides a table of content while at the same time, hiding administrative pages or templates. Finally, a Glossary provides domain dependant definitions for common acronyms.

In an attempt at facilitating contributions, shortcuts to resources for contributors and a link to create new pages are available from all pages.

3.2.3 Aggregation pages

A major challenge using Wikis is the risk of leaving content unreachable. We provided several ways to highlight content and make it more visible.

We used multiple semantic properties to create user defined vocabularies. Instead of a uniform system of tags, Editors can add Keywords, Purpose or Topics to most pages. In turn, these properties are used to group pages in ways that users are most familiar with. For example, Topics are used to group Articles together into Handbooks.

Similarly, we made extensive use of semantic queries to group together pages into dynamic collections. For example, event based pages provide timelines about content. We also created architectural views to replace traditional static diagrams and highlight relationships between different components of our systems. Similarly, Master articles or Systems pages combine hand-picked selection of pages with inline queries to provide overviews on a topic.

3.2.4 *Machine readable*

It is not enough to make content accessible to users, content should also be reachable from other systems as well.

To this end, we exposed the underlying semantic structure of the content with browsable links to the underlying structure using the Resource Description Framework (RDF) format. A local triple store and daily snapshots of the full ontology are available to Semantic Data Integration systems. We also expose our Vocabulary with simple lists of page names for each category.

Additionally, Application Programming Interfaces (API) to the MediaWiki internal structure are available in multiple programming languages and can be used to expose data to advanced automation tasks.

These features make knowIT ready for future integration with other systems.

4. USER EXPERIENCE

knowIT was officially launched after only few months of experimentation and customizations. A year later, it is used to prepare routine communications, outage reports and applications inventory reports for other groups. Growth rate has been steady over that period.

The benefits of the new system for the three major tasks identified vastly outweigh the few limitations we observed.

4.1 Search

Replacing MediaWiki default search mechanism with an open source solution allowed us to enhance traditional search results with semantic content. Our customized interface groups results by categories and provides semantic summaries depending on the category. For example, search results on Persons display contact information instead of the traditional excerpt.

In the future, it would be helpful to analyze the quality of results based on search patterns and to include more Semantic Search techniques, such as faceted browsing [15][16].

4.2 Explore

Users seem satisfied with customizations added to improve navigation and make the system consistent with other collaborative sites on intranet.

knowIT provides notification systems (Real Simple Syndication or RSS feeds, watchlists) but to date, these have been mostly used by more experienced users.

Although we are providing multiple ways to reach relevant content, information can still be buried at times. The need to build aggregation pages to bring up content is an ongoing battle when working with wikis.

4.3 Contribute

Unlike Wikipedia, anonymous vandalism of pages is less of a concern in a professional environment.

MediaWiki is not a document management system and it requires an external workspace for sensitive data. Understanding these limitation helps define the scope of content we want to accept in this knowledge base.

Quality control of content is an ongoing challenge. Naming conventions, consistency of labels, case sensitiveness and synonyms require that editors follow best practices at the expense of acceptance of the system. Advanced knowledge management tools from MediaWiki such as redirection of pages or lists of wanted links are invaluable tools for quality control, but they do not replace the need for data curators in addition to contributors.

Some users regretted the lack of Rich Text editor and a visual annotation system. This is something we will investigate in the future, notably through the use of the more advanced version of Semantic MediaWiki, SMW+ [17].

In addition to forms, having the option to add inline annotations allows the system to handle exceptions to the data model very well. Having the option to structure content outside of preset forms allows contributions to provide suggestions to the data model and turns a knowledge base into a collaborative database.

On the administrative side of content management, the new system provides similar flexibility, both in the data model and in customization of the interface. Most customizations can be done through the web, without need to schedule a change control request or downtime.

New properties can be added to Semantic Forms as needed with little impact on the system. Unattended jobs propagate updates to pages across the wiki seamlessly. Forms also include a mechanism for auto-completion, which helps with consistency of labels.

Although automatic loading of data was used during the migration from the old system, there is a severe lack of visual tools to help update multiple pages at a time. The web interface is limited to changing one page at a time or to doing simple search and replace of text strings across pages.

5. CONCLUSION AND FUTURE WORK

Semantic MediaWiki provides a robust system for a collaborative database. Overall, we are very satisfied with the migration of content from our previous system. The challenges we faced during that process actually helped us improve the quality of our content and reduce ambiguities.

knowIT is already helping us answer questions about our Informatics systems on a routine basis :

- Because of a simple and flexible data model
- Multiple ways to search, explore and contribute content
- Semantic annotations to structure content with forms
- Solid knowledge management tools
- Several options to import / export data
- A low cost, scalable and tested platform

For the next phase of deployment, our challenges will be more Cultural than Technical. We are hoping to eventually break the cost of managing content with more advanced tools for bulk updates, an easy to use interface, the ability to quickly locate information and the reduced burden of sharing tacit knowledge about our environment.

Increased advocacy and awareness about knowIT will help reach a critical mass of content providers.

6. ACKNOWLEDGMENTS

The system presented here is the result of contributions from many people over a long period of time. We would like to thank current and past contributors for their patience, ideas and support – Jim Gainor, Brian Wegner, Deborah Yates, David Epstein, John Baum, Lisa Valetta, Mario Dolbec, Brian Johnson and Emmanouil Skoufos.

7. REFERENCES

[1] Munson, S. A. “Motivating and Enabling Organizational Memory with a Workgroup Wiki”, *WikiSym ‘08.*

[2] Berners-Lee, T. Linked Data. 2006. http://www.w3.org/DesignIssues/LinkedData.html

[3] Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. “Linked Data on the Web” , In Proceedings WWW2008, Beijing, China.

[4] Monge, S., Ovelar, R. and Azpeitia, I. “Repository 2.0: Social Dynamics to Support Community Building in Learning Object Repositories”- *Interdisciplinary Journal of E-Learning and Learning Objects* Volume 4, 2008

[5] Chalmeta, R. and Grangel, R. 2008. Methodology for the implementation of knowledge management systems. *J. Am. Soc. Inf. Sci. Technol.* 59, 5 (Mar. 2008), 742-755. DOI= http://dx.doi.org/10.1002/asi.v59:5

[6] Miller, R. “Collaboration Management with Archetypes”, In Proceedings “Second Plone Conference”, Vienna, Austria, 2004

[7] Confluence, http://www.atlassian.com/software/confluence/

[8] Microsoft Office SharePoint Server, http://www.microsoft.com/Sharepoint/default.mspx

[9] Drupal, http://www.drupal.org

[10] MediaWiki. http://www.mediawiki.org

[11] Krötzsch, M., Vrandecic, D. and Völkel. M. “Semantic MediaWiki”. In Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, Lora Aroyo, eds.: *Proceedings of the 5th International Semantic Web Conference (ISWC-06).* Springer 2006.

[12] Krötzsch, M., Vrandecic, D., Völkel,M., Haller, H., and Studer, R. “Semantic Wikipedia”. *In Journal of Web Semantics 5/2007.* Elsevier 2007.

[13] Millard, D., Bailey, C., Boulain, P., Chennupati, S., Howard, Y., Davis, H. and Wills, G. (2008) “Semantics on Demand: Can a Semantic Wiki Replace a Knowledge Base?” *New Review of Hypermedia and Multimedia*, 14 (1). pp. 95-120. ISSN 1740-7842

[14] Desilets, A., Paquet, S. and Vinson, N.G. “Are Wikis Usable ?”, *WikiSym ‘05.*

[15] Auer, S., Dietzold, S. and Riechert, T. OntoWiki - A Tool for Social, Semantic Collaboration, *5th International Semantic Web Conference,* Athens, GA, USA, November 5-9, 2006, LNCS 4273

[16] Guha, R., McCool, R., and Miller, E. "Semantic search." Budapest, Hungary: 12th Intl. World Wide Web Conference (WWW 2003), 2003, pp. 700-709. [Online]. Available: doi=10.1.1.1.2813

[17] SMW+, http://wiki.ontoprise.de/wiki/index.php/Main_Page