H-Maps: An Efficient Approach for Graphical Visualization and Navigation of Topic Maps and Data Representation in Computational Biology

Markus Leber, Thilo Klöpfer, and Steffen Hüttner

Hölle & Hüttner AG, Life Science Department, Derendinger Str. 40,
72072 Tübingen, Germany
{markus.leber,thilo.kloepfer,steffen.huettner}@h-net.com

Abstract. Nowadays semantic networks become more important in industry and modern science. For this reason there is a need for an efficient tool allowing preparation and navigation through complex data spaces. We present the novel Topic Map tool H-Maps, which offers a web based navigator providing an enhanced utilization and graphical visualization of semantic networks. Present applications transform existing content to subject-centric Topic Map representations in a one-way fashion. Preserving the advantage of transferring diverse back ends into an unified topic map in existing applications the H-Maps approach goes one step further. The resulting topic map structure can be examined and extended using an iterative procedure. Especially in science this procedure leads to a considerable improvement of flexibility and performance as can be shown within projects of computational biology.

1 Introduction

Constructed for high amounts of information and fast data access, conventional database systems often exhibit a lack of clarity and flexibility. Semantic networks use a rather intuitive approach of knowledge representation. Complex relationships can be transformed into a graph based format describing associations of different subjects with distinct properties. First used in certain areas of computer science, e.g. artificial intelligence or knowledge representation, semantic networks became popular in a wide range of applications. Especially the Topic Map standard has become an important component for knowledge transformation and semantic network representation. This standard is not restricted concerning the data source and any kind of data can be processed. Unlike conventional database systems Topic Maps are capable to describe and merge knowledge from different and distributed domains or resources.

In computational biology Topic Maps allow an optimal representation of biological data [1]. Metabolic processes, signal transduction, or cell cycle control are just few examples where biological systems form highly complex networks. Nevertheless, presently Topic Maps is rarely used for biological data. One explanation can be the fact, that current Topic Map tools often have very limited

Maicher, L.; Garshol, L. M. (Eds.): Linked Topic Maps. Fifth International Conference on Topic Maps Research and Applications, TMRA 2009 Leipzig, Germany, November 12–13, 2009 Revised Selected Papers. Leipziger Beiträge zur Informatik. ISBN 978-3-941608-06-1

or no resources for graphical visualization. As a consequence the advantages of Topic Maps are not always visible for scientists and other users. Within the interdisciplinary project PASSAGE we developed the novel Topic Map tool H-Maps, whereas H represents the company logo. H-Maps allows interactions and extensions by the user and can easily be adapted to any kind of function.

Within PASSAGE the tool is used to edit and visualize gene expression data derived from the new high throughput sequence technology. H-Maps is able to import data from different sources. The input data are converted by the Topic Maps mapping framework according to application specific ontologies and user defined mapping rules. The generated Topic Maps are managed by the H-Maps core engine. Residing on that core engine a Topic Maps server module adds application specific logic such as authentication features, preparation of data, searching capabilities and the like. Finally the Topic Maps can be visualized within a browser based Silverlight navigator serving as web application layer.

2 Application of H-Maps in Computational Biology

Biological data often exhibit a networked structure, which is not always known or visible to the respective scientific user. In this context a semantic representation of the data via Topic Maps may help to clarify the relations within the biological system and the underlying concept. Although H-Maps can be used for any kind of networked data, the program was basically developed to edit and visualize biological data schemes. H-Maps is integrated into the projects Endoplas and PASSAGE.

2.1 Endoplas – H-Maps Representation

Endoplas is a project to develop new technologies to prevent contaminations caused by medical instruments. The objective was to establish a comprehensive knowledge base with research and query possibilities, which should be made publicly accessible via a browser-based web application. The knowledge base contains information about the substances, which can cause fever, their classification and the corresponding literature.

Pyrogenes, which can cause inflammation, are an important subject of the project. Within the first phase of the project currently existing literature was investigated according to the material properties of pyrogenes, plasma processes and detection methods concerning distinct classes of pyrogenes. A Topic Map containing the application ontology was created as a basis for the classification of the results gained from the literature study. The study was mainly carried out as internet research, whereby the content was collected with the Zotero tool [2]. Zotero is a freely available add-on for the Firefox browser, which enables the users to collect, manage, and cite research from all types of sources. The collected references can be exported in RDF format.

In a second step, the RDF export from Zotero was mapped into a Topic Map using the H-Maps converter. Topics Maps allows a better visualization of

the data in comparison to the RDF format. Based on the application-specific ontology, which is defined at the beginning, a set of mapping rules to transform the RDF constructs to Topic Maps constructs can be configured interactively with the converter's GUI. The rule set can be stored in a separate XML file. As a consequence, mappings can be repeatedly applied to the source data as often as their content has changed. Slight variations of the mapping rules can also be achieved by storing the rules in distinct files. Finally, the resulting XTM file can be imported into the Topic Maps engine and rendered by the Topic Maps navigator.

Both the H-Maps converter and the navigator were specifically aimed at be user-friendly. Even inexperienced users, who were not familiar with the concept of Topic Maps, were able to operate the H-Maps suite without any problems.

2.2 PASSAGE – H-Maps Representation

The project PASSAGE is based on the high throughput sequence technology. In contrast to conventional sequencing systems, which just can determine the nucleotide sequence of one single DNA molecule, the new high throughput sequencing systems are able to analyze hundreds of thousands up to several million DNA molecules. As a consequence the high throughput sequencing systems are used for a wide range of methods in molecular biology. Within PASSAGE the new technology is used to generate expression profiles. Expression profiles allow conclusions about the genes activated in the cells under the respective experimental conditions. If a gene is activated, copies of the gene are generated, which consist of mRNA molecules. The new high throughput sequencing systems are able to display the whole transcriptome including almost all mRNA molecules via sequencing.

Our approach uses certain methods of experimental probe generation and evaluation of the sequencing results. Another important unique feature is the close collaboration between the program PASSAGE and H-Maps. The topic map generated by H-Maps is successively extended depending on the stage of expression profile analysis. In the beginning a small topic map is induced. This map gives an overlook about the most important statistical data of the different experiments. In the next step the data of different experiments are compared by PASSAGE. The result is a statistic about the similarity and differences of the experiments. These statistics are stored as new topics in the Topic Map connecting the experiments to a large network.

Figure 1 shows a section of a network. The topic in the center represents an experiment. The occurrences of a topic are displayed as a list within a panel. The occurrences describe the parameters and conditions, under which the experiment was performed. The browser window to the right contains a diagram, which belongs to the topic in the center. PASSAGE creates diagrams, which are stored as images in PNG format. During the transfer into the topic map the respective path to an image is stored as occurrence. Selection of the respective occurrence activates the diagram and the image is displayed in a browser window. The

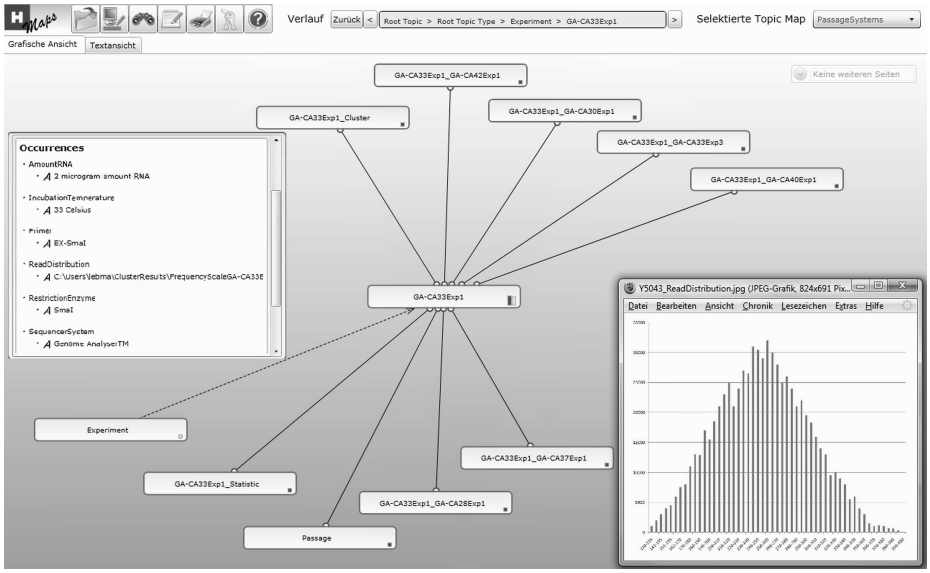

Fig. 1. The Topic Maps navigator H-Maps depicts experiments with their environment as topics. The occurrences of a topic are listed in the left panel. Occurrences can contain references to images, which can be displayed in a browser window

connections to other experiments are described by further topics located in the boundary area.

Since it is not reasonable to display all the sequencing data and every statistical value, interesting data are summarized to histograms providing an overview. As a consequence, a good balance between knowledge and information layer is achieved. Additional information can be gathered from attached resources whenever required.

3 Representation of Topic Maps via H-Maps

Based on the experience we made with our leadoff version of a Topic Maps browser on building subject-centric application systems [3], we reworked crucial parts of the software evolving in a refined design. While the first version provides technically matured navigation capabilities, but does not offer any graphical visualization, the second version is focused on the visualization of the presented knowledge structures to further improve navigation and the data retrieval. In this context a change of technology from the traditional combination of HTML and JavaScript to Silverlight was chosen. Silverlight is designed to develop rich internet applications and delivers improved performance and interactivity.

The import of data is enhanced by a new Topic Maps converter, which is able to integrate content from multiple data sources. The integration of heterogeneous

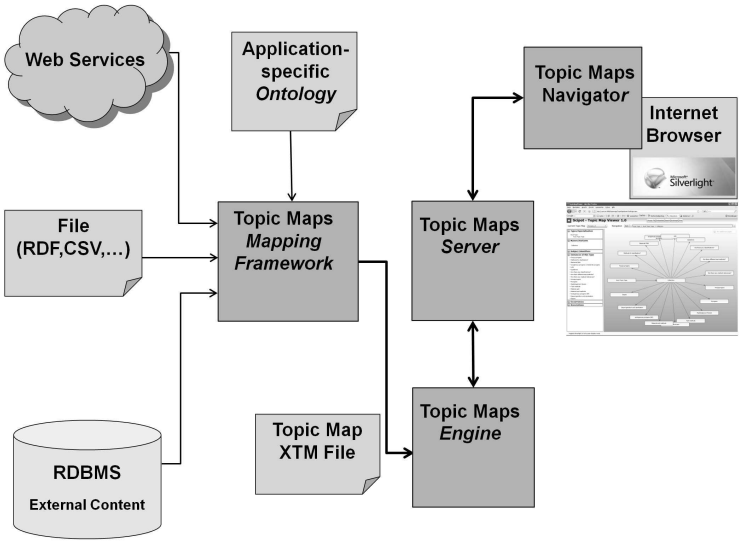

Fig. 2. The schema describes the data flow starting from the converter module to integrate heterogeneous content systems through the Topic Maps engine. Finally the results are visualized in a browser

content allows for a unified view on the information space and serves as a key feature to build a more coherent and robust knowledge management tool.

3.1 Architectural Overview of H-Maps

The main design objective for the H-Maps suite was a highly modular architecture. As a result the program suite is both easily extensible and adaptable to different scenarios. Figure 2 illustrates the core components of H-Maps demonstrating the data flow from data integration of external resources to knowledge presentation.

The Topic Maps core engine supports TMAPI [4], which is the de facto standard for accessing and manipulating Topic Maps. TMAPI is a set of Java interfaces which was adopted and ported to the C# .NET programming language. The implementation of the interfaces has been aligned with the TMDM ISO standard [5]. The Topic Maps server is designed as middleware infrastructure. It adds application specific functionality to the core engine, such as authentication, and provides a set of services, which can be accessed from clients, especially the browser-based Topic Maps navigator.

The following sections describe two of the main Topic Maps modules of the H-Maps suite, the mapping framework and navigator. These two application modules build up the endpoints, providing interfaces, which are operated by the end users.

3.2 Silverlight Navigator as Intuitively Usable Web Frontend

The TMDM [5] specifies a graph-based data model for visualizing semantic networks. Consequently, one of the main design goals of the navigator was to visualize the knowledge structures of discourse in such a graph. Providing data as graph is most intuitive as it has proven in wide-spread mind maps [9].

Work is in progress on standardizing a graphical notation to define ontologies and represent Topic Maps instance data. Topics and associations are the two fundamental Topic Maps elements. Within graph visualization these elements are represented by symbols. Topics are represented as nodes, whereas associations are rendered as edges between the topic symbols. The graph is enriched by further information such as association roles, which are played by topics in the association, and a significant name for each topic.

Besides the graph visualization additional list and tree views containing characteristics of the focused topic are provided. Occurrences may either be a value inside the topic map or an external information resource. Text values can be viewed directly, with embedded links being parsed and rendered in the same way as in common browsers. External addressable links can be shown directly in a separate browser window.

Information is supplemented by properties of the relevant knowledge models in order to anchor the topic's subject in the underlying ontology. In particular this comprises the complete type hierarchy of the topic, the topic's names with variants, and its subject identifiers.

The coexistence of listings and the visualization graph side by side yields in a good survey of the displayed Topic Map section. Clicking on a symbol or item would retrieve all data of the topic in question. This offers a multitude of access paths to navigate through the displayed knowledge model.

3.3 Format Converter as Highly Extensible Connector Framework

One of the great features of Topic Maps is its standardized merging process. Using the merging process, several approaches were made towards unified and integrated knowledge maps [7, 8]. The result is a much more coherent and robust knowledge management infrastructure, instead of using singular user interaction layers of different data sources.

H-Maps provides a set of converters, which can read data from distinct sources in different formats. The converters are highly configurable and adaptable to varied user scenarios.

In a first step an application-specific ontology has to be created by the user. Based on that ontology, an individual user defined mapping from the external source data onto Topic Maps constructs can be specified in a second step. Afterwards the topic map can be persisted via the standardized exchange format XTM [6]. The converter allows to automatically generate unified knowledge spaces, which can be extended by heterogeneous content systems and can be configured according to the user's preferences. The resulting knowledge space is not static but can be updated and replenished by subsequently inserted data amendments.

4 Conclusions and Further Work

We presented an approach for a new highly modular application for both an extensible integration mechanism on the back end and an efficient graphical visualization and navigation on the frontend. The approach has been used to manage and visualize biological data with excellent results. H-Maps has become an important feature of the projects Endoplas and PASSAGE. It allows a perspicuous view on the relationship of scientific literature and reveals common properties of different experiments.

Due to the graphical visualization our navigator results in benefits at the user interface level. Providing a multitude of access paths, the users can intuitively navigate through the topic maps without having any background knowledge of the subjects of discourse.

On the back end our combined mapper & converter module provides a way to readout the content of multiple data sources, which are in distinct formats. Thus a unified view on multiple content resources of interest can be achieved. By the use of the converter in various application scenarios it has provided proof of its universal applicability in practice. The navigator was handled by users, who were not familiar with the concept of Topic Maps, without any problems. Thereby the ease of operations of the navigator and the fast accessibility of complex data could be shown. As a consequence, the easy accessibility of the expert results for non-experts was proven.

The project is currently in progress but has yet reached an advanced degree of maturity. For now, the navigator displays the whole characteristics of exactly one focused topic without any restrictions. Within the next design step we will focus on data reduction and consolidation. As a promising approach we plan to rank the Topic Maps constructs by relevance. The ranking is based on weightings, which will be derived as a combination of computed and user defined values. By focusing on the most relevant topics, the navigator will be enabled to display a wider area of the Topic Map clearly represented. Changing the weightings will lead to different filtering of constructs in each case. The result is an improvement of different views on the rendered knowledge structures. As a consequence, non-obvious coherences can become apparent. Quick retrieval of relevant subjects will further be enhanced by providing intelligent and efficient searching, sorting, and filtering capabilities. On the other hand the converter will be successively extended to support additional formats, whereupon we regard supporting relational data bases as a must.

Furthermore H-Maps will be enhanced by editing functionality to not only display networked knowledge areas but also manage knowledge repositories in a graphical-interactive manner.

5 Acknowledgments

We would like to thank Peter Kuhn of Hölle & Hüttner AG for triggering work on the H-Maps suite. Further thanks go to our partners of the PASSAGE project

Dr. Kai Sohn and Christian Grumaz of the Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, and the coworkers of the GATC Biotech AG for providing gene sequence probes, and giving support in biochemical and genetic issues.

References

1. Stümpflen, V., Gregory R., Nenova K.: From Biological Data to Biological Knowledge, Lecture Notes in Computer Science, Volume 4438, Springer, pp. 62 – 66 (2006)
2. Zotero. A personal research assistant, http://www.zotero.org
3. Weber, G.E., Eilbracht, R. and Kesberg, S.: Topic Maps as Application Data Model for Subject-centric Applications. Fourth International Conference on Topic Maps Research and Application, TMRA 2008, placeCityLeipzig, country-regionGermany, October 16-17, 2008 TMAPI project. Topic Maps API, http://www.tmapi.org
4. ISO/IEC 13250-2, Information Technology – Topic Maps – Part 2 – Data Model. International Organization for Standardization (2006)
5. ISO/IEC 13250-3, Information Technology – Topic Maps – Part3 – XML Syntax. International Organization for Standardization (2007)
6. D. Bogachev: Enterprise Knowledge Maps: Towards subject-centric computing. Topic Maps 2007, Oslo, Norway (2007)
7. Hänninen, H.: Enterprise Knowledge Integration Using Topic Maps. Topic Maps 2008, Oslo, Norway (2008)
8. Farrand, P.; Hussain, F.; Hennessy, E.: The efficacy of the mind map study technique. Medical Education 36(1), pp. 426–431 (2002)
9. GTM. Graphical Topic Maps notation. http://www.isotopicmaps.org/gtm