

Ontology based search and storage of market information

Wolf ENGELBACH¹, Damian GAIDA², Henryk RYBINSKI², Thomas SPECHT¹
¹*Fraunhofer IAO, Nobelstr. 12, 70569 Stuttgart, Germany*

Tel: +49 711 9702128, Fax: +49 711 9702401, Email: wolf.engelbach@iao.fhg.de

²*Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland*
Tel: +48 22 6516226, Fax: +48 22 6516222, Email: hrb@ii.pw.edu.pl

Abstract: This paper describes an ontology supported software application that combines the advantages of existing Internet search engines with modern text analysis functionalities and an intelligent storage system for results documents and knowledge items. An ontology assists the user in query definition and structures the storage and retrieval of documents as well as of knowledge items. The system is implemented for the business case of SMEs that need information for their internationalisation, but it can easily be transferred to other domains just by adding an other system ontology. Its main users may be consultants and associations that are dealing with many projects that have similarly structured information demands.

1. Introduction

Increasing competition and globalisation trends are challenging companies to expand the target markets for their products and services into foreign countries. The process of internationalisation necessitates many decisions. Adequate information (e.g. relevant products and companies, or the market situation) about the specific industry niche is required to support decision making and ensure the successful implementation of the internationalisation strategy.

Such information is valid only for a particular geographical area and very niche specific. Therefore it is often not available in editorially proven commercial or public databases, but distributed over several homepages of companies, research and governmental institutions, media, and private people.

Especially SMEs (small and medium-sized companies) prefer to look for such information without investing too many resources in terms of time and money, in particular without employing external consultants. As valuable information about external business factors is readily available on the Web, what is needed is just explore the web resources properly. On the other hand the expertise of SMEs in using internet tools is rather restricted. Therefore there is a tremendous need to provide tools that would simplify the internet exploration process as a foundation for decision making in their specific internationalisation project.

For this purpose, we have developed a software prototype that combines advantages of existing Internet search engines with modern text analysis functionalities and an intelligent ontology based storage system for documents and knowledge items. It has been envisaged as a system that can perform a variety of very complex information gathering and analysis tasks. To an extent it refers to the idea of web farming systems, as defined by Hackathorn in [1]. So, one of its goals “*is the systematic refining of information resources on the Web for business intelligence*”, on the other hand it should provide means for quick exploration of new web areas.

Although the web farming idea has already been proposed more than 8 years ago, till now, there were very few attempts to practically implement it, especially for the scale of SME needs. In this sense, AMI-SME is quite a unique system combining Web exploring functionality with advanced storage system and text analysis means. One of the main features of AMI-SME is building a repository by a sequence of consecutive queries. Somehow a similar approach has been implemented with the system SensMaker, presented in [2]. In SenseMaker already clustering techniques have been used for visualizing the search results. Another approach that seems to be very close to AMI-SME was INSYDER reported in [3]. The similarity resulted from the similar goals of gathering the information sources, however the solutions proposed in [3] referred mainly to using web agents, whereas in AMI-SME we expect to tap high quality information from the existing web search engines by a simultaneous search covering a number of existing engines. So instead of concentrating on building agents, we have put more attention to developing means for cumulating local repositories, on which advanced text processing and text mining tools can be applied for knowledge extraction. In this aspect AMI-SME gets closer to SenseMaker, however we attempt to move further, especially with the use of ontologies for enriching queries, and combining multi-engine query search and applying powerful text mining tools for results analysis.

Core innovation in the system is the multi-purpose usage of the ontology, in particular (1) to support query definition, and (2) to provide means for organizing the local repository. As reported in [4], ontology can be also used for drastic improvement of clustering algorithms. The same has been confirmed for other text mining algorithms, especially for classification and categorization [5]. The documents from the repository can be annotated by ontology concepts, as well as labelled with “sticky labels” that are out of the ontology, they can be also commented by the user, and categorized (manually and automatically). Therefore reuse of information contained in the repository by both information extraction and retrieval is of much higher quality. This paper presents a description of the system architecture, the usage of the ontology and the realised user interfaces.

2. Objectives

Since the advent of WEB the problem of exploratory search was addressed in many papers, albeit not sufficiently reflected in practical applications. Its growing meaning has been recently confirmed by a number of publications collected in [6], where the issue has been exhaustively revisited. We have attempted to implement AMI-SME in line with the idea expressed in [7] and paraphrasing the famous Hamming statement to the form “*the purpose of exploratory search is insight, not data*”:

“In intelligence analysis, as in other domains, that insight comes from the process of exploration, not just from its end result. We are interested in capturing and visually representing analysts’ iterative query processes and insights to help them collect and compare information more effectively, as well as record and share the products of their analytic insights.”

Therefore, core objectives and innovations of the software are a combination of domain specific pre-structuring, automatic analysis and manual annotations and structuring:

1. A persistent storage for different search projects and their related queries and results. This allows working over a long time span on the same project with different users.
2. A multi-purpose usage of ontology to support query definition and result organisation with a complex but intuitive and expandable structure for storing of and navigation in detected documents and gathered knowledge items.
3. The integrated usage of text analysis functionalities for clustering, labelling and filtering to keep an overview in the ocean of gathered information pieces.

4. The integrated view on several individually selected sources that need to be consulted to obtain the necessary information for answering domain specific questions.

Ideally, all the phases of the process are ontology supported, reducing the time needed for the user intervention between the consecutive processes and activities, and improving the quality of the knowledge acquired.

3. Methodology

The software is being developed within an EU-supported CRAFT project by six RTD partners and seven SME partners from five European countries [8]. It is based on a detailed analysis of the SME's requirements for information management within their internationalisation activities, and on the short-comes of similar existing information systems. The specification includes prioritised use cases, design prototypes as well as data models. After implementation and integration, two testing and improvement circles with the SME user partners follow. In parallel, the SMEs have specified the service portfolio that will accompany the software product.

There is a distributed development by research partners in four countries, with clear separation of responsibility: for the graphical user interface (Fraunhofer IAO, Germany), the ontology management (GraphiTechn, Italy), information sources (CIMNE, Spain) and the overall backend system functionality for storage and persistency (Warsaw University of Technology, Poland). The software technology for the prototype is characterised by using an existing development framework (objectledge), which allows reusing many basic components and simplifying development routines [9].

4. System architecture

Main modules

The AMI-SME system is composed of several modules, which functionally can be grouped into following categories:

1. Query formulation support and query run:

These modules support the process of defining user information retrieval needs. The process is usually nontrivial and iterative. The modules discover the same or similar queries that have been run earlier under another name. They can also be used to visualize a detailed search history within a given session or a research project. Easy access to the ontology can be used for building more advanced queries. The modules performing query run translate the original query to the standard corresponding to the needs of a given service, then merge the received results. They also can use filtering masks (for filtering a domain, language, document types, dates, etc.)

2. Search result gathering, and local storage and analysis

Local storage provides the user with a possibility of iterative analysis of search results and trends identification. In particular the modules provide possibilities to visualize new items within a search result, compare queries, locate similar documents in the repository, etc. Analysis modules automate these processes with knowledge discovery algorithms. Analysis is based on a number of known Text-Mining and text analysis algorithms including clustering, classification and entity extraction.

3. Analysis results visualization and summarization

These modules provide synthetic overviews of gathered information and analysis results, allowing users to quickly prepare management level reports. The reporting tools

are statistical summaries helping the users to comprehend the results of analysis (classifications, numbers of extracted entities etc.).

Document Processing Framework

In AMI-SME the central part of the system is the framework of document analysis and storage. The framework is composed of a database of so called Search Result Documents (SRD) and interconnected components for various document transformations.

A single Search Result Document (SRD) defines one hit in the result of searching for relevant information. A result of a search may be any single resource being a part of the answer that was found *via* the AMI-SME search facilities, other search engines, and manually by browsing the Internet. Basically it represents a document – a web page, a PDF file, a MS Word file, etc. – which can be found and is reachable from Internet or by other means. It is a flexible data structure, as it represents many different types of documents and search results. Individual SRDs may have different sets of properties and may only share some of them, but all of them are identifiable, since they have an URI (in most cases URL).

Search results coming from various search engines represent Internet resources. While processing an answer, AMI-SME merges the answers and pre-selects a limited number of the hits having the highest relevancy rank, and eventually satisfying filtering criteria (if defined by the user). The reasons for decreased feasibility of direct retrieval of all resources represented by search results are following:

- Many search queries are very imprecise, rendering a large number of results.
- Most of the times, queries do not express the users' information need, this causes too many of the results to be not relevant.
- Resources pointed by the search results may take very long time to download (increasing the cost of system use, and decreasing usability).

There are cases in which documents (SRDs) are related to each other in various kinds of relations – direct and indirect. Direct, for instance image document being a part of an HTML document, indirect – two documents coming from the same information source, e.g. a monitored web portal section or search engine.

SRD is seen at the API level (application processing interface) as an object with the set of attributes and relations. For instance most of the simple attributes as dates, character strings (author, title, abstract) are stored in the relational database, but other such as bag of words representation or binary contents of the SRD document are stored either in a specialized database or directly in the file system. Such abstraction allows the users of the framework to concentrate on analysis tasks and allows flexibility in addition of new framework components (in this case at the data level).

Apart from the objects' attributes there are relations between SRDs and other objects, e.g. *similar* is a many to many relation that links documents with similar content. The *linked_to* relation describes the existence of links between found documents: Contents of documents represented by SRDs in AMI-SME database may contain links to documents not yet retrieved, while new document (SRD) additions would require full analysis of all stored SRD content to create *linked to* relationship to existing SRDs. In order to avoid it, content analysis includes link extraction and indexing, which allows quick updates of the *linked_to* relationship by looking links up in the index.

5. Usage of the Ontology

The software application distinguishes system ontologies and project ontologies. System ontologies structure the domain of several projects, e.g. internationalisation. Project ontologies are specific for one project, e.g. selling simulation software in France.

System ontologies

One system ontology is already designed for and integrated in AMI-SME. It describes important concepts in internationalisation and is used to test the system. That ontology has four main concepts, which are crucial to understand a given situation: product, company, target market and regional area. All of them are further specified with sub-concepts, relations and attributes, but also other concepts such as events and associations exist.

In order to obtain valuable search results it could be wise to augment such an industry independent ontology with industry or company specific concepts attributes or relations. This follows the approach in other domains, where generic ontologies exist, e.g. for organisational knowledge that can be instantiated to a specific situation, e.g. to a specific company [10]. In addition it is possible to connect the system ontology with project specific ontologies; for example to specify the concepts “product” or “company”. Moreover, it is possible to define other system ontologies related to a different domain, e.g. innovation management, and thus use the system for this other purpose.

All changes to a system ontology have to be done outside the AMI-SME system, using ontology modelling tools. The sample internationalisation ontology was modelled with the graphically oriented software tool Sementalk [11]. You can also import existing industry or region specific ontologies and use them as system ontologies, as far as they follow some formal restrictions regarding the supported relations and constraints.

Project ontologies

Within the AMI-SME system, it is possible to work on several independent search projects in parallel or successively, e.g. in case of internationalisation for different products or different countries. If you start your project, you select the appropriate system ontology, and all changes will be stored to its project specific copy. Thus, each project will have its own project ontology, extended and instantiated by modifications during the usage of the system. As summarised in figure 1, the project ontology ...

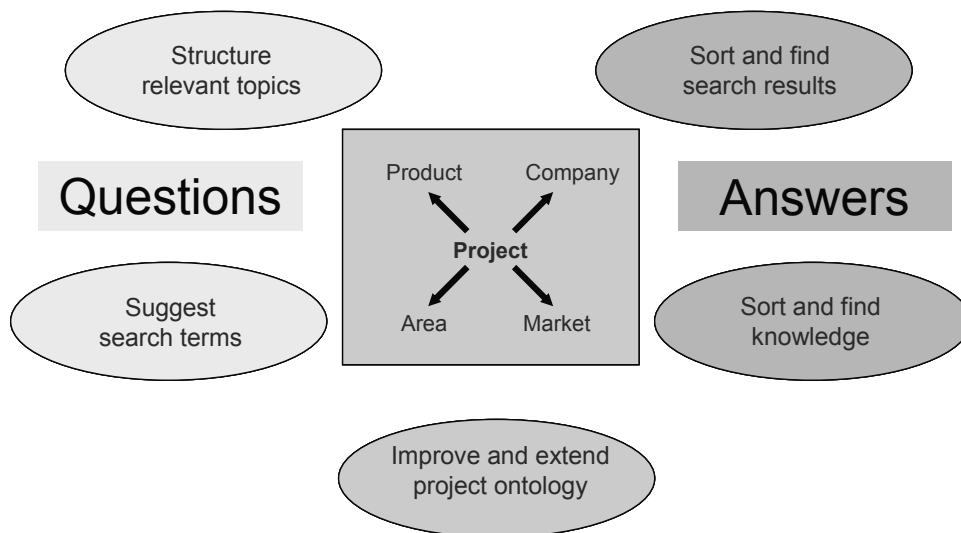


Figure 1: Ontology usage overview

- is used to label search results manually, which allows to find them within the ontology structure easily; as soon as enough results are labelled manually it is also used to suggest labels for search results automatically;
- allows saving and retrieving the personally extracted information about relevant knowledge items, e.g. about specific competitors or relevant regions;

- assists in the definition of queries by topic suggestions related to the concepts and instances of the ontology, e.g. legal issues or distinct products;
- suggests keywords and synonyms as well as relevant Internet pages for concepts and instances.

Since all these activities relate to the same project specific ontology, wherever you add a new knowledge item (instance), a knowledge type (concept) or a relation, they are immediately available for the other purposes as well.

6. User Interface

The user interface screens are optimised to offer much intelligent functionality while hiding the complexity of language processing and ontology manipulation for the user.

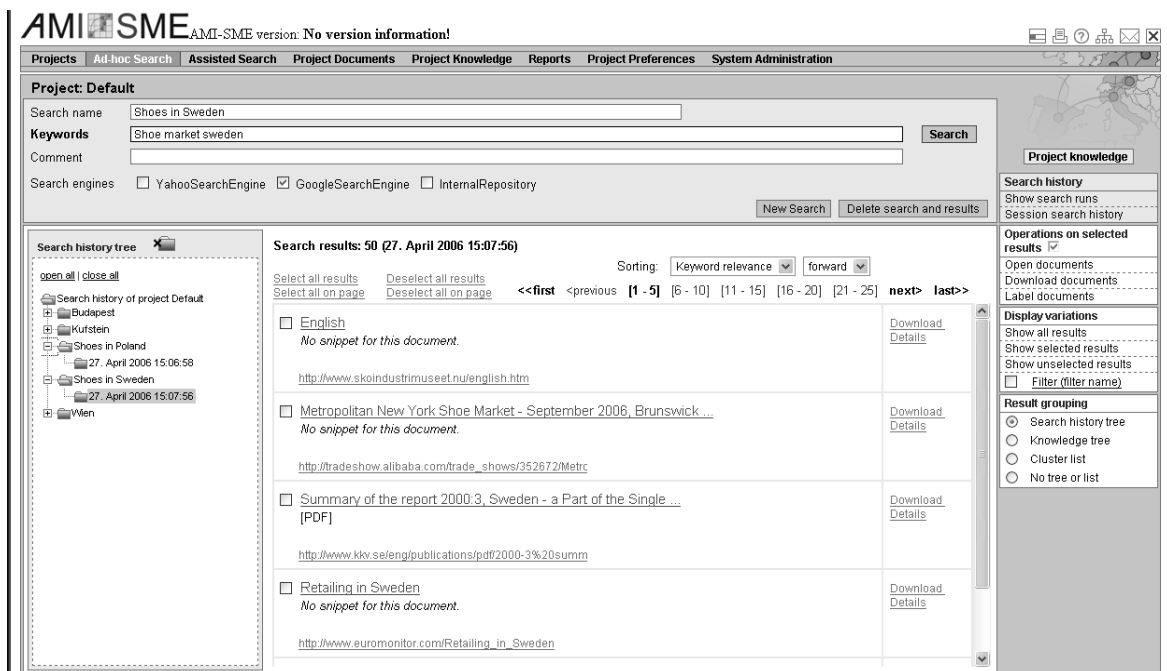


Figure 2: Search screen

- There are three main project related views:
1. Search screen (see figure 2): An ad-hoc search screen allows to run a query as simple as in common meta search engines, with additional options to comment and name the search, which can be stored as well. In an assisted search mode, also topics, keywords and information sources are suggested by the ontology. For the results of both search modes, clustering structures the results, and filtering searches in the results or its metadata. For each result, a details pop-up (see figure 3) for annotation is available, partly with system suggested values, e.g. by a summariser, language detector or metadata extracted from the search engine); additionally the result document can be labelled with concepts or instances of the project ontology.
 2. Document screen: Here you can see all results either for selected search executions, or all results with a specific ontology label. This allows the user to concentration on the evaluation of the results of all project related searches.
 3. Knowledge screen (see figure 4): This view allows editing values to specific instances of the project ontology, e.g. the number of a company's employees, or adding relations, e.g. the products of a company. From search or document screen, it is possible to open a pop-up window to directly view and edit information about identified knowledge items.

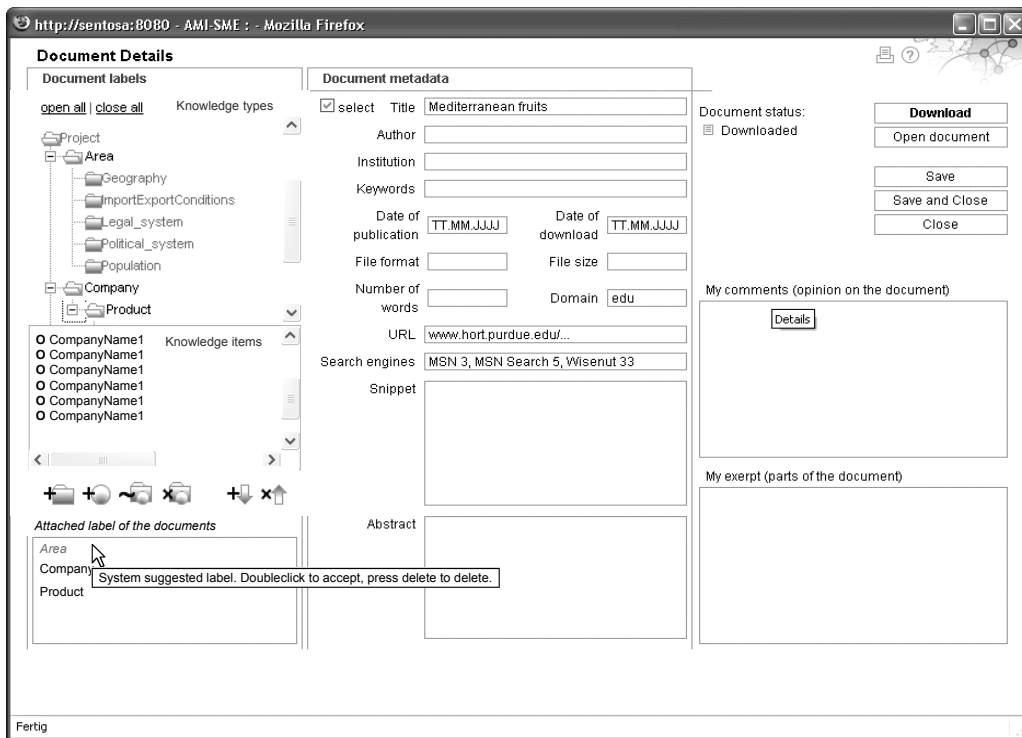


Figure 3: Document details pop-up screen

The concepts of the specific project ontology are displayed in the user interface comparable to a folder structure in form of a tree in the left column, with the instances of the selected concept in a list below. In the knowledge screen (figure 4) also relations to other knowledge items and to labelled documents are presented in two separated lists in the right column, and also the related documents are presented. This allows the inexperienced user a simple orientation and a well-known navigation, despite of the complex structure of the ontology itself.

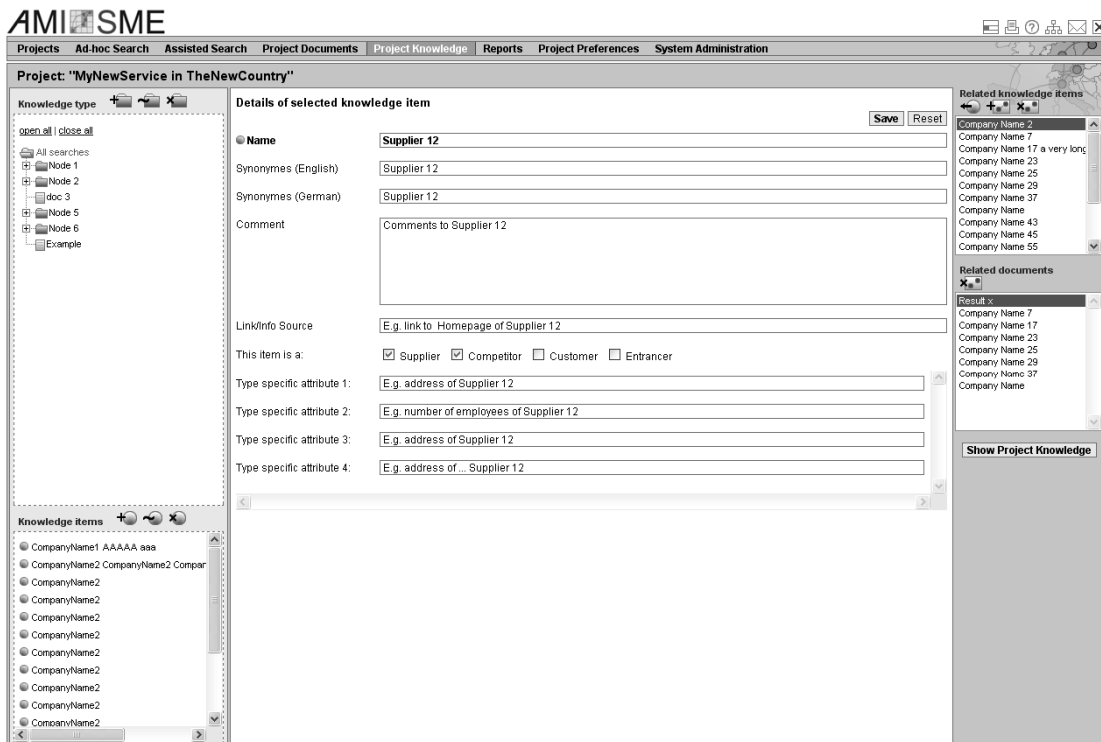


Figure 4: Knowledge screen

7. Conclusions

The realised concept of ontology based search and storage improves the interface between search engines and knowledge management: it makes Internet search more intelligent and integrates it closely to the users' project context. Especially for tasks that companies seldom conduct, like internationalising, such a guiding structure helps not to forget any important issue. In case of weakly structured information, as for individual industrial niches, that is even more important since the reader has the effort to extract relevant content on his own and is now supported to store results in a proven way.

The general idea can be transferred to other topics than internationalisation, e.g. product development or innovation management. Also the automatic extraction of available content from the Internet to the knowledge base is a research activity that would additionally extend the usage of the software, either by RDF source integration [12] or with specific wrappers [13]. During the exploitation in other domains, the project partners will offer additional services to the users of the software, such as ontology modelling, system integration, design adaptation, training as well as support in the interpretation of the search results.

References

- [1] R. Hackathorn, "Web Farming For The Data Warehouse", Morgan Kaufman, 1999
- [2] M. Q. Wang Baldonado, T. Winograd: "SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests", Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, 1997
- [3] H. Reiterer, G. Müller, T. M. Mann, S. Handschuh, "INSYDER - An Information Assistant for Business Intelligence", Proc. of the 23rd Annual Int'l ACM SIGIR Conf., 2000
- [4] A. Hotho, A. Maedche, S. Staab, V. Zacharias: "On Knowledgeable Unsupervised Text Mining". *Text Mining*, 2003: 131-152
- [5] H. Rybiński *et al.* "Experiments with TM platform and basic DM algorithms adopted to TM", Report 2, WUT, 2006 (unpublished)
- [6] J. Gersh, B. Lewis, J. Montemayor, Ch. Piatko, R. Turner, "Supporting exploratory search: Supporting insight-based information exploration in intelligence analysis", *Communications of the ACM*, Vol. 49(4), 2006
- [7] *Communications of the ACM*, Vol. 49(4), April 2006
- [8] The software is being developed within the EU-supported CRAFT project AMI-SME: "Analysis Of Marketing Information For Small And Medium-Sized Enterprises". For more details see <http://www.ami-sme.org>
- [9] <http://objectledge.org>
- [10] A. Gualteri and M. Ruffolo, "An Ontology-Based Framework for Representing Organizational Knowledge", In: Proceedings of I-Know 05, Graz, Austria, June 29-July 1, 2005, pages 71-78.
- [11] <http://www.semtalk.com>
- [12] e.g. <http://www.cia.gov/cia/publications/factbook>
- [13] e.g. <http://www.piggybank.com>