From Full Text Search to Semantic Web: The Infofox Project

Barbara Sigrist and Petra Schubert
University of Applied Sciences, Basel
Firstname.Lastname@fhbb.ch

Abstract

Infofox is a project launched by the Ecademy, the National Network of Excellence of the Swiss Universities of Applied Sciences for e-business and e-government. The paper discusses an envisioned improvement of the service level of the Ecademy Web site, an information portal which is set to integrate valuable information from the different member schools. The goal of the Infofox project is the development of an intelligent search engine. In order to fully exploit the potential of the information provided by the Ecademy members, the Infofox project team will need to implement aspects of a Semantic Web – namely an ontology built upon a joint classification scheme and a set of inference rules. The RIP paper presents the basic concept and the results of the first phase of the project.

1 Introduction

The management of information is a pivotal aspect of success or failure of any undertaking, be it of private, academic, or commercial nature. There are many terms related to information management: content management, enterprise content management, knowledge management, workflow management, etc., all of which try to frame a specific problem and match it with a specific solution. As is the case of many Internet related issues, the World Wide Web Consortium tried to tackle the problems of adding meta-information to Web content with a concept introduced back in 2001 known as Semantic Web.

Berners-Lee et al. [2001] describe the *Semantic Web* as a new form of Web content that is meaningful to computers. The authors stress the fact that we will only be able to fully unleash the potential of information published on the Net once electronic agents will be able to interpret the data automatically and independently enabling them to fulfil information needs defined by the user. Today, most Web pages are only meaningful to human readers. HTML is a means to describe the format of displaying information – the meaning being unreadable for computers. Adding markup-information to Web pages is a basic prerequisite for the fulfillment of complex user tasks by electronic agents. The Semantic Web will bring structure to the meaningful content of Web pages. But the idea of the Semantic Web does not stop at adding some tags and putting formerly HTML-written pages into XML pages. It also entails a language for defining rules

– thus adding logic for the processing of data. A basic component of the Semantic Web is the concept of ontologies, containing taxonomies (with defined classes and relations among the classes) and inference rules (which describe semantic relationships between content classes).

The Infofox project which will be described in this paper is targeted at the creation of a smart search engine for information discovery. It does *not* aim at making the Web content computer-readable but it uses the underlying paradigms of the Semantic Web in order to build an interactive search interface for the user. A possible user question for which Infofox will offer an answer could be the following: “Find all people who are willing to teach a class on data mining, who are experts in this field (at least have personal project experience), and who would be willing to travel to Basel”. Today, the information needed to answer this request is available in the Infofox database but there is no way of connecting three different kinds of classes (area of interest, expert skill level and mobility preferences) into one joint search query. An information seeker would have to search the database first for the area of interest and then drill down on expert skill and mobility preferences for every single person – a cumbersome process.

The authors believe that the concept of the Semantic Web and its underpinnings such as the Resource Description Framework (RDF), XML (syntax), and URIs (naming) is a promising solution for some of the problems related to the digital realm as described by distinguished experts (e.g. Berners-Lee, Staab, Matthews).

2 Background and Players

Information availability is literally boundless. Data and information become progressively more inaccessible for a variety of reasons. This fact has gained increasing attention and importance throughout the digital world. Commerce has suffered from it ever since the Internet became business critical, and academia has spurred efforts in recent years to meet the shortcomings of the technical implementations of what is generally addressed as “knowledge management”. In knowledge management, there are three players inherent: the information *seeker*, the information *provider*, and the *intermediary*.

Everyone is an information *seeker* in that we all collect data, transform it or have it transformed into information, and apply it once it becomes knowledge. In conventional digital search settings, information seekers have only very limited influence on the search process – they have to live with the possibilities offered by the search interface (generally a Web form with one field where the user can enter the search phrases).

An information *provider* can be anyone from individuals to institutions and businesses. Information providing itself has turned into a veritable business with its own turfs. Within this paper, the authors attribute this role to any entity offering information free of charge or for a fee.

In between, there are the *intermediaries*. These intermediaries provide portals, search engines, and indexes in an attempt to alleviate the problem arising with an abundance of information. Information providers and intermediaries push information to the seeker in predefined ways instead of offering the seeker possibilities to adapt the query and possibly the results. The following figure illustrates the players’ triangle, the situation with regards to information distribution, and the information flow amongst the constituencies:

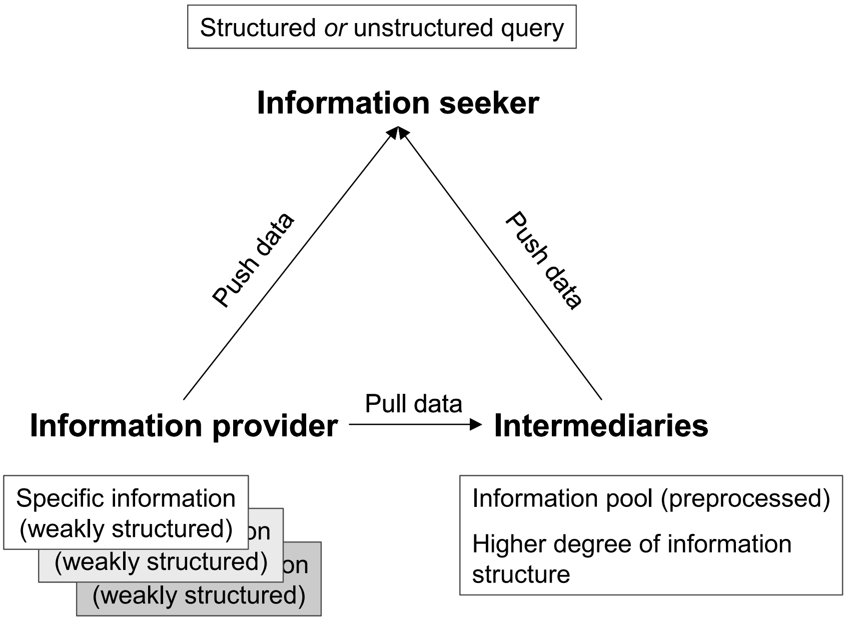

Structured *or* unstructured query
Information seeker
Push data
Push data
Information provider
Pull data
Intermediaries
Specific information (weakly structured)
(weakly structured)
(weakly structured)
Information pool (preprocessed)
Higher degree of information structure

Fig. 1: Players' triangle and information flow

Information providers and intermediaries provide either topic centered solutions (e.g. indexes for medical topics) or topic neutral approaches (e.g. google, search.ch). A combination of these two is rarely found, and the one commercial application set out to conquer this specific digital challenge, Northern Light, closed operations effective January 1st 2003.

According to a Gartner research conducted in 2001, the time needed for document management alone will have increased from 20 % in 1997 to roughly 40 % in 2003 [ZyLAB 2002]. This assumption is contradictory to the hopes and hails of information management gurus who predicted a decrease in time and paper used to handle information in the digital era. Furthermore, information is to be collected and digested in order to be turned into personal and ultimately corporate or organizational knowledge. This leads back to the triangle "information seeker", "information provider", and the "intermediaries". In order for data or information to be useful, (1) the information seeker *and* the intermediaries both have to know the specific search context and the implicit situational context of the query, and (2) this information has to be translated and transferred into appropriate forms in order to yield the relevant query results. To date, XML has evolved as primer choice for describing the structure of information. RDF and XML schemes describe the metadata semantics of information, XSL helps to convert class instances into XML documents, and XSLT supports the transformation of XML documents into further XML documents [Fensel 2001].

These standardization efforts are supported by the fact that almost 80 % of all responses of an online survey conducted by the Universität Karlsruhe reported that information retrieval is the key activity when using the Internet [Leibold/Stroborn 2003].

On the grounds of a specific project owned by the Ecademy, the authors will describe in this research-in-progress paper (1) the limitations of current information management systems; (2) how the Ecademy approaches the existing limitations in order to successfully integrate distributed knowledge from different network sources in the future; (3) a proposal for an approach using ontologies; (4) next steps to a successful ontology implementation and maintenance.

3 Limitations of Present Information Management Systems

Seen from a business perspective, the ultimate goal of technical advancement is to facilitate competitive advantage. The strategy of every company, organization, and even individual already does or soon will comprise knowledge creation, processing, use, and reuse. The specific strategy for the underlying knowledge management will vary depending on the viewpoint and the business profile. The focus, though, is in the process of shifting from pure information providing and seeking to knowledge building and (re)using.

Intermediaries such as Google and other non-topical and unspecific search engines as well as information providers still heavily deploy full-text search as their main vehicle to information discovery placed at the disposal of the general public and professionals alike. However, the shortcomings of full text search are pressing for more elaborate approaches:

The search methods are only targeted at information retrieval using keywords as matching mechanism;

Web pages contain vast amounts of information that are accessible to humans only. Content management systems provide simple keyword supplements but do not support the management of markup information which could enable software agents to automatically process the information provided in the HTML code.

Data and information are available and searchable. Knowledge as an extension of information may be available, but full-text search largely ignores the added value provided through experience and logical deduction (e.g. with the help of inference rules) which extends information to knowledge;

Thesauri add structure to the largely non-hierarchical information vastness, masking the underlying mismatch of information provided versus information sought;

The definition of “valuable information” is not addressed, thus effectively leaving the information seeker, the information provider, and the intermediaries with little more than information chunks to do business with that are out of most context;

The trustfulness of information sources is difficult to evaluate;

Wrappers used to translate between systems are static and therefore difficult to maintain; they are difficult to write in order to yield meaningful translations. Inherently stand-alone applications are forced to “talk to each other” or draw information from one another on cumbersome grounds, leaving much of the work to educated guessing and error handling.

4 Case Study: Ecademy and the Infofox Project

Infofox is a project run by the Institute for Business Economics at the University of Applied Sciences of Basel on behalf of the Ecademy. The Ecademy is the National Network of Excellence of the Swiss Universities of Applied Sciences for the thematic clusters “E-Business” and “E-Government”. It currently has 13 member institutions from all regions of Switzerland (comprising three different language regions). The project aims at improving information access and knowledge management in the networked environment of the Ecademy. Infofox is therefore about extending data beyond information and focuses on funneling knowledge by means of the semantic web as defined by Berners-Lee et. al. [2001]

The Ecademy member schools use a broad range of Internet applications. The central services (see Fig. 2) maintained by the Ecademy head office in Basel are Lotus Notes based. The Ecademy and its content services serve as the pilot case for the Infofox project. The authors believe that the Ecademy Web services are an excellent environment for a Semantic Web based on a common ontology (comprising a taxonomy as well as inference rules).

Communication among members (individuals and organizations) and inter-operability of applications and systems have proven rather towing because of the distributed character of the network and the manifold resources available at the member institutions. The network has a single thematic domain: e-business and e-government. Throughout the last years the member have develop and adopted a joint framework for a common terminology (of basic E-Business/E-Government concepts) along with a common E-Business/E-Government glossary. The distributed knowledge and information could be made accessible more effectively using a set of common ontologies across the network (for a detailed description of ontology uses see [Uschold/Gruninger 1996]). The envisioned goal of Infofox is to bundle existing information containers into a *common virtual database* using common ontologies accessible to at least one query engine. On the one hand, this will be achieved by providing an efficient framework for information providers allowing them to add the required meta-information to their content. Information seekers, on the other hand, will be offered easy to use query interfaces with extended functionality powered by additional markup information and inference rules. The Ecademy Web team is information provider and intermediary at the same time.

The Ecademy and its ontological approach in information and knowledge sharing may serve as an example, guide, and maybe as a master copy for other Swiss Networks of Excellence. To the knowledge of the authors, no other network is actively engaging in knowledge management research for the benefit of any of the Swiss networks.

4.1 Existing Information Containers

The Ecademy mainly deploys Lotus Domino applications in order to make information available to the interested audience. The content is managed via the Internet/Extranet using standard Web browsers. The content management system (CMS) makes use of the extended functionality of the Lotus Notes client. The following figure illustrates the five main information containers and the primary graphical user interface (Web browser and Lotus Notes client, respectively):

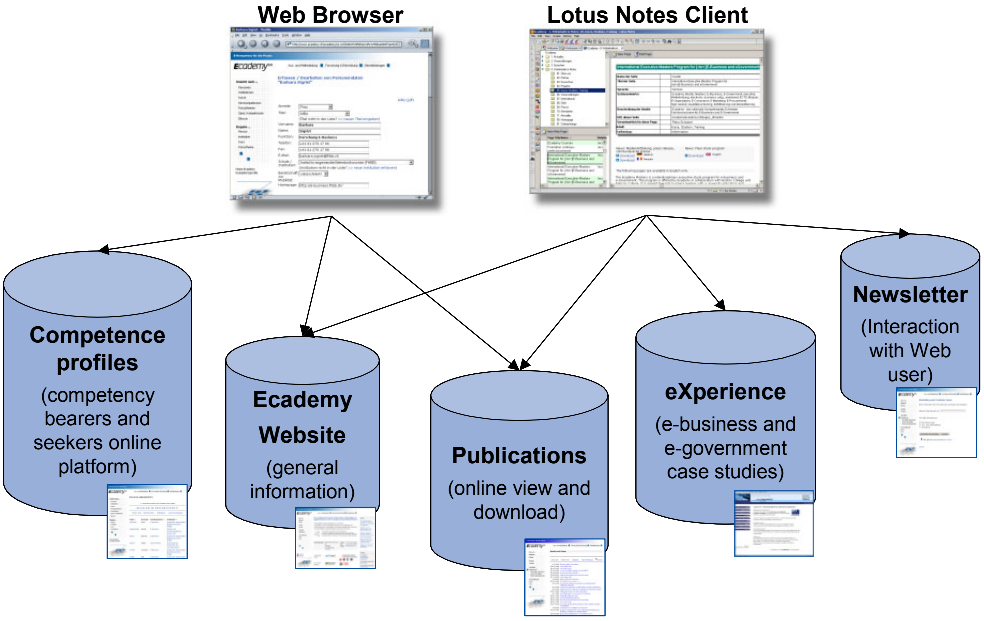

Web Browser
Lotus Notes Client
Competence profiles
(competency bearers and seekers online platform)
Ecademy Website
(general information)
Publications
(online view and download)
eXperience
(e-business and e-government case studies)
Newsletter
(Interaction with Web user)

Fig. 2: Ecademy information containers (Lotus Notes databases)

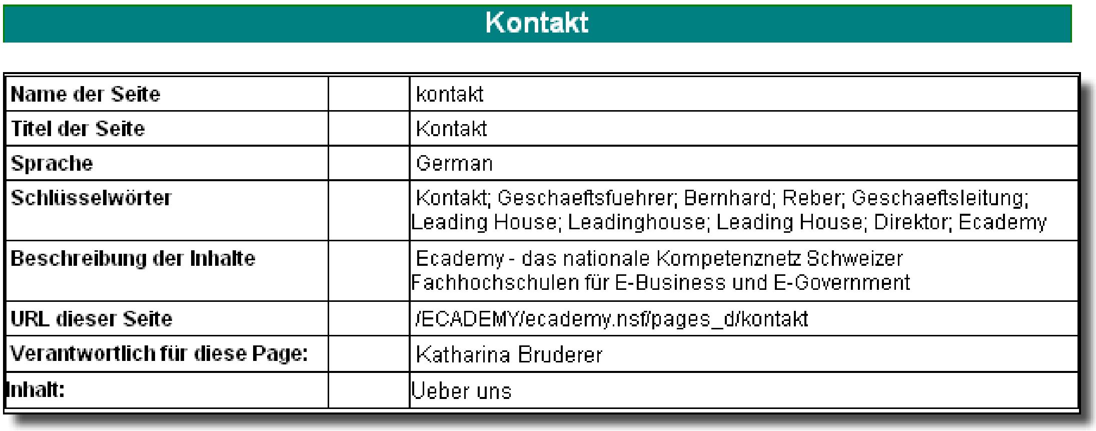
The "competence profiles" database contains a first approximation of a common Ecademy ontology. In this database, experts (people), classes and projects have been stored according to a common classification scheme [Schubert/Sigrist 2002].

Not included on this figure are the numerous information containers of member institutions and their affiliates.

4.2 Semi-Ontological Approach

Early on in the development history of the information containers, the application designers and programmers recognized the need to structure the content further in order to provide a value adding information service. The Ecademy Web development team is able to react swiftly to requests from users. They have taken up the task of classifying the input of each piece of information following the Ecademy classification scheme. This reflects a semi-ontological approach which was born out of situational needs rather than strategic planning and coordinated execution. The metadata used by the Web team to describe documents has not yet been widely communicated across the network, let alone agreed upon. As for now, it is rather imposed upon the content managers and users who rarely recognize the additional information which is added after they enter their information in the content management system. Thus, a crucial precondition of a working ontology (see definition below in section "4.3 Ontologies and Meta Data") is not yet met and will consequently be part of the ongoing Infofox project.

The following screenshot of an Ecademy Web page (displayed in the CMS) reveals the meta-information either added explicitly by the content manager or automatically generated by the CMS (built-in in the template) in order to describe an unspecific Web page:



Kontakt

Name der Seite		kontakt
Titel der Seite		Kontakt
Sprache		German
Schlüsselwörter		Kontakt; Geschaeftsfuehrer; Bernhard; Reber; Geschaeftsleitung; Leading House; Leadinghouse; Leading House; Direktor; Ecademy
Beschreibung der Inhalte		Ecademy - das nationale Kompetenznetz Schweizer Fachhochschulen für E-Business und E-Government
URL dieser Seite		/ECADEMY/ecademy.nsf/pages_d/kontakt
Verantwortlich für diese Page:		Katharina Bruderer
Inhalt:		Ueber uns

Fig. 3: Sample page with meta-information and body text

External content managers are generally also asked to provide some meta- information. The following table explicates the meta-information displayed in Fig. 3:

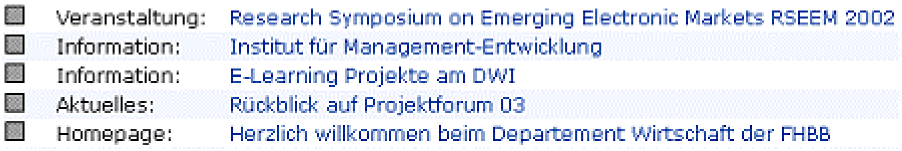
Meta Data	Content
Page name:	Mainly used for programming purposes, i.e. composing the URL.
Page title:	Displayed in the Window title. Must be descriptive and concise.
Language:	Language of content. The value controls the language of the display menu.
Key words:	HTML meta tag “keywords”.
Description:	HTML meta tag “description”.
URL:	Computed. This is of informative character and can be used to add links between different pages.
Content responsible:	Every page is owned by a person.
Content type:	The Ecademy distinguishes several content types: “About Us”, “Partners”, “Know-How”, “Projects”, “Courses, Study, Training”, “Events”, “International”, “Orbit”, “Media”, “Newsletter”, “News”, “Extranet”. This value controls the indicator (dot) in the navigation bar.
Page type:	Computed. Page types describe the character of the page itself rather than the content. The Ecademy distinguishes the following basic types: “information”, “news”, “events”, “homepage”. Search results display the page type:  ▨ Veranstaltung: Research Symposium on Emerging Electronic Markets RSEEM 2002 ▨ Information: Institut für Management-Entwicklung ▨ Information: E-Learning Projekte am DWI ▨ Aktuelles: Rückblick auf Projektforum 03 ▨ Homepage: Herzlich willkommen beim Departement Wirtschaft der FHBB Fig. 4: Search results displaying the page type and the page title

Table 1: Metadata scheme

This semi-ontological approach has evolved naturally over the last years. Following the distinction in [Fensel 2001], the aspired ontology type is best matched by the metadata ontology as described in the Dublin Core [Dublin Core 2003]. Following Heflin [2003] the exemplified use cases and the Web portal combined with web site management tasks could be used in the Infofox project.

4.3 Ontologies and Meta Data

As described above, the Ecademy originally pursued the traditional way of sharing and using information deploying a central content management system. However, over time internal discussions started about if Ecademy could integrate information directly from member source (e.g. the Web sites run by the member schools)? Other questions were: Why Ecademy started its own glossary in the face of existing glossaries provided by members? Who are the typical users of the competence profiles and in what way do they use them? Who decides what kind of information should be included and how it should be displayed? The natural way to solve these issues was to take the best from each question and adopt the current content management system to offer the best answers. The result was the semi-ontological approach adopted by the designer team in compliance with the Web project owner. The most visible, though ultimately insufficient result was the introduction of meta-information in addition to the different types of content pages.

The emergence of concepts and methods such as the Semantic Web, ontologies, and the connected technical specifications (XML, RDF) set the direction for further development. The project team realized that a common set of ontologies would ultimately alleviate the unresolved discussion and in some areas even the existing dissent among the Ecademy members. Furthermore, promising application areas and helper programs with reasonable readiness for market had emerged in recent years: KAON [KAON 2003] and OntoEdit [OntoEdit 2003] to name two examples.

Many definitions evolved during the past years as to what defines the term “ontologies” a term which emerged from a philosophical context. This paper concurs with the definition in [Fensel 2001, 11]: “An ontology is a formal, explicit specification of a shared conceptualization.” Taking into consideration the structure and logical architecture of the existing information container, concentrating on ontologies bears the strategic advantage of not abandoning any digital asset. On the contrary, the Ecademy will have a set of processes at hand with which a conceptualisation (e-business, e-government) can be shared in a planned and controlled manner. The specification most probably will be semi-formal, taking into account the Notes legacy. Above all, the ontology schemes must be kept as open as possible in order to allow future ontologies to dock in.

4.3.1 Finding Matching Ontology Schemes

As discussed in detail in [Staab 2002], the development process should start with a feasibility study, continue with an in-depth analysis, proceed with the development of specifications, then introduce a semi-formal description of one or more ontologies, continue with an evaluation of the concepts and rules, and end with a full-grown application. The Ecademy is still in the first phase of the development process. From what we know today about the existing technology and resources which could be put into the Infofox project, we assume that the project is feasible. The first basic specifications are currently being developed. The project team is primarily concerned with defining an ontology which will best suit all member needs.

Ontologies will play an important role at two levels:

1. Documents containing unspecific data and general information (information containers "Ecademy Website" and "Newsletter");
2. Documents with specific and hierarchical data (information containers "Competence Profiles" and "Publications")

Existing metadata will be taken and transformed into one ontology for unspecific documents. Metadata generation usually involves either automatic, manual, or semi-automatic methods [Staab 2002]. Automatic generation will play a marginal to nonexistent role in this project and will only happen through the application of content templates. Manual processing of metadata has been used until now and will play a decisive role in the new approach. The prospect is to move on to semi-automatic methods using XML schemes within RDF without abandoning the existing infrastructure based on Lotus Notes.

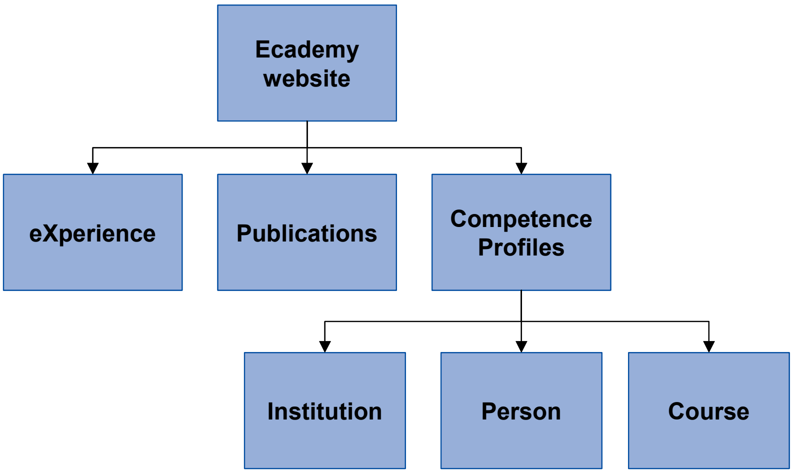


Fig. 5: Ecademy ontology scheme

4.3.2 Ecademy Ontologies: a Proposal

A first proposal of an ontology scheme can be obtained by extending and reordering Table 1. The Ecademy Web team distinguishes between document information and subject matter information. Document information is what is colloquially labeled metadata and describes the ambit information of a document. Subject matter information describes the content of each page and is unique for every page. Table 2 represents the future element set. The "New Element Name" is in accordance with the Dublin Core Metadata Element Set

Current Element Name	**New Element Name**	**Reference**	**Description**
Content responsible	Creator	Document information	Every page is owned by a person. Orphans are not allowed.
Language	Language	Document information	

Current Element Name	New Element Name	Reference	Description
Page name	*-Obsolete-*	Document information	The source and the URI uniquely identify each document, the page name is therefore obsolete.
Page type	Coverage	Document information	Coverage describes the character of the page itself rather than the content. The Ecademy currently distinguishes the following: "information", "news", "events", "homepage". This classification scheme will be revised.
URI	Identifier	Document information	Reference to the source within Ecademy.
Content type	Type	Subject matter	The Ecademy distinguishes several content types: "About Us", "Partners", "Know-How", "Projects", "Courses, Study, Training", "Events", "International", "Orbit", "Media", "Newsletter", "News", "Extranet". This classification scheme will be revised, e.g. mapped with the DCMI Type Vocabulary scheme .
Description	Description	Subject matter	
Key words	Subject	Subject matter	
Page title	Title	Subject matter	

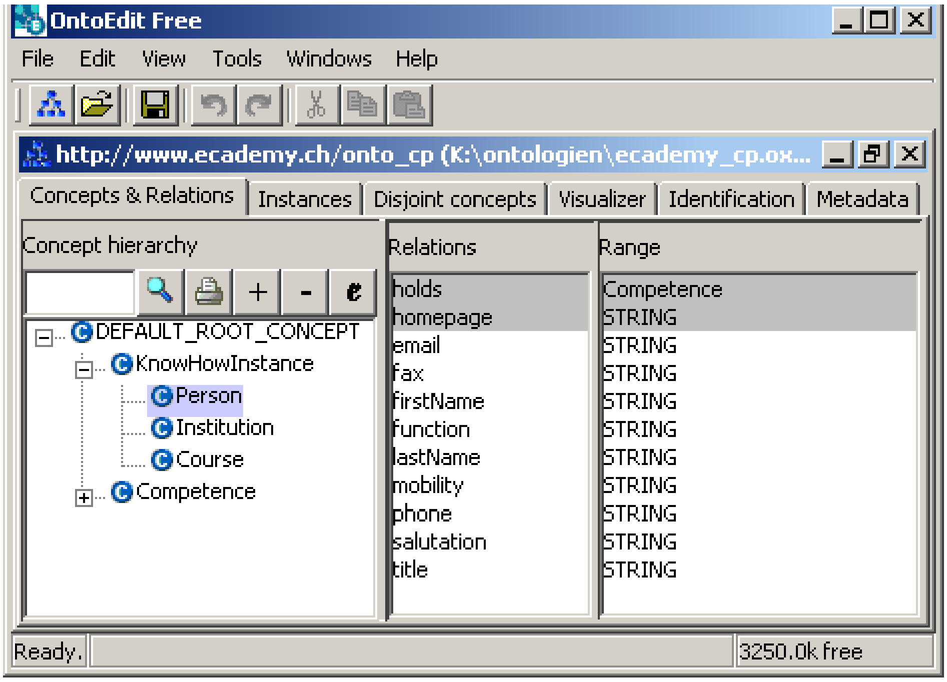
Table 2: Future set of metadata (ontology)

The list will be extended by new elements.

New Element Name	Reference	Description
Source	Document information	A formal identification scheme – either organizational or thematic – will be used to describe the different sources of the Ecademy.
Format	Document information	Each document has a codification scheme applied. Format information describes the possible codification schemata of a document (MIME types).

Table 3: Metadata ontology extensions

Ontologies for the competence profiles are created on top of them. They are of hierarchical structure with moderate relations [Schubert/Sigrist 2002]. The concept for the competence profiles describes a hierarchical approach cross relating certain entries, though without specific inferencing and querying handling. A visualization done with OntoEdit shows the hierarchy and a possible set of relations for the competence profiles:



DEFAULT_ROOT_CONCEPT
- KnowHowInstance
  - Person
  - Institution
  - Course
- Competence

Relations	Range
holds	Competence
homepage	STRING
email	STRING
fax	STRING
firstName	STRING
function	STRING
lastName	STRING
mobility	STRING
phone	STRING
salutation	STRING
title	STRING

Fig. 6: proposed structure of the competence profiles

In this example, a person inherits two relations from "Bearer". The remaining relations are person inherent and therefore not handed down. Similarly, institutions and courses carry competences and have homepages but have nothing else in common with a person. The three subconcepts do not share many superconcepts.

5 Final Remarks

The project owner, i.e. the Ecademy board, will install a team in order to discuss and develop authoritative ontology schemes for the different information containers within the Ecademy network. The most difficult task will be the definition of a common structure which is suited for all information objects contained in the multiple Ecademy databases. It is planned that the team will seek for consulting advise (coaching) by experienced institutions in the field of ontologies and search engines. Once the ontologies are defined and agreed upon, the Ecademy portal will be equipped with the Infofox search engine.

The idea of the Semantic Web is a promising approach for the improvement of knowledge discovery on the World Wide Web. Nevertheless, the great challenge will always be the definition of domain-specific ontologies among partners with (partly) differing interests and opinions. And finally, we should not forget that adding meta-information to Web pages requires discipline, sophistication and time from every single content provider worldwide.

6 Bibliography

Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001): The Semantic Web, in: Scientific American, May 17, 2001.

Dublin Core (2003): [http://dublincore.org/documents/dcmi-terms/]. [Zugriff: 04.09.2003].

Fensel, Dieter (2001): Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Springer.

Gilbert, M.; Drakos, N.; Latham, L. (2001): The Web Content Management Magic Quadrant for 2001, Research Note. Accessed: August 4, 2003 <http://www.oit.duke.edu/CMSsub/docs/Fatwire/Gartner_WCM.pdf>.

Heflin, Jeff (Ed.) (2003): Web Ontology Language (OWL) Use Cases and Requirements. Accessed: August 4, 2003. <http://www.w3.org/TR/webont-req>

KAON (2003): [http://kaon.semanticweb.org/]. [Accessed: 04.09.2003].

Leibold, Kay; Stroborn, Karsten: Internet-Zahlungssysteme aus Sicht der Verbraucher. Ergebnisse der Online-Umfrage IZV6. Accessed: August 4, 2003 <http://www.iww.uni-karlsruhe.de/izv6/>.

OntoEdit (2003): [http://www.ontoprise.de/]. [Accessed: 04.09.2003].

Schubert, Petra; Sigrist, Barbara (2002): Ecademy Kompetenzprofile: Eine Internet-basierte Applikation zur Erfassung und Abfrage von verteiltem Wissen, Basel: Fachhochschule beider Basel (FHBB), Institut für angewandte Betriebsökonomie (IAB), Arbeitsbericht E-Business No. 10, 2002.

Staab, Steffen (2002): Wissensmanagement mit Ontologien und Metadaten. Habilitation, Accessed: August 4, 2003 <http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/habil.zip>.

Uschold, Mike; Gruninger, Michael (1996): Ontologies: Principles, Methods and Applications, in: Knowledge Engineering Review 11(2).

ZyLAB (2002): Know the cost of filing your paper documents. Accessed: August 4, 2003. <http://www.zylab.nl/zylab2002/US/Downloads/whitepapers/PDF/21%20-%20Know%20the%20cost%20of%20filing%20your%20paper%20documents.pdf>.