Towards Integrative Enterprise Knowledge Portals

Torsten Priebe
Department of Information Systems
University of Regensburg
D-93040 Regensburg, Germany
torsten.priebe@wiwi.uni-regensburg.de

Günther Pernul
Department of Information Systems
University of Regensburg
D-93040 Regensburg, Germany
guenther.pernul@wiwi.uni-regensburg.de

ABSTRACT

Knowledge portals make an important contribution to enabling enterprise knowledge management by providing users with a consolidated, personalized user interface that allows efficient access to various types of (structured and unstructured) information. Today's portal systems allow combining access modules to different information sources side by side on a single portal webpage. However, there is no interaction between those so called portlets. When a user navigates within one portlet, the others remain unchanged, which means that each source has to be searched individually for relevant information.

This paper discusses integration aspects within enterprise knowledge portals and presents an approach for communicating the user context (revealing the user's information need) among portlets, utilizing Semantic Web technologies. For example, the query context of an OLAP portlet, which provides access to structured data stored in a data warehouse, can be used by an information retrieval portlet in order to automatically provide the user with related documents found in the organization's document management system. The paper shortly presents a prototype that we are building to evaluate our approach, demonstrating such an OLAP and information retrieval integration.

Categories & Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H5.m [**Information Interfaces and Presentation**]: Miscellaneous

General Terms: Design, Management

Keywords: Knowledge Management, Integration, Portals, OLAP, Information Retrieval, Semantic Web.

1. INTRODUCTION

A major challenge of today's information systems is to provide the user with the right information at the right time. Using Web-based technologies, knowledge portals are an emerging approach for providing a single point of access to various types of information, making an important contribution to enabling enterprise knowledge management. This paper discusses integration aspects in the context of enterprise knowledge portals. In particular, the integration of structured information (like OLAP data stored in a data warehouse) and unstructured information (e.g. in form of documents) is a key issue of this paper.

We base our approach on integrated metadata, using an ontology for concept mapping, together with an approach for context integration. Today's portal systems allow combining different portal components side by side on a single portal webpage. However, there is no interaction between those so called portlets. When a user navigates within one portlet, the others remain unchanged, which means that each source has to be searched individually for relevant information.

This paper presents an approach for global searching and for communicating the user context among portlets. This approach is, to our knowledge, unique. In the concrete case of integrating structured data warehouse data and unstructured documents, the query context of an OLAP portlet (i.e. the information shown within a certain OLAP report) can be used by an information retrieval portlet to automatically provide the user with related documents found in the organization's document management system.

The rest of this paper is organized as follows: In section 2 enterprise knowledge portals and organizational memory systems are introduced. Section 3 discusses approaches for global searching over multiple information sources. The main contribution of this paper is, however, the context integration approach presented in section 4. Section 5 shortly introduces the prototype portal system we are currently building to evaluate our ideas. Section 6 presents related work and compares our approach to others in the literature. Finally, section 7 concludes the paper and discusses remaining open issues and possible future work.

2. ENTERPRISE KNOWLEDGE PORTALS AND ORGANIZATIONAL MEMORY

In Latin the term "portal" means something like door or gate. Accordingly it is used for webpages which provide an entry point to the Internet or an intranet. In contrast to Web portals, community portals, etc., enterprise (also B2E, business-to-employee) portals focus on corporate information and services which should be provided to the employees of an enterprise. The terms enterprise portal and enterprise information portal are used

CIKM'03, November 3-8, 2003, New Orleans, Louisiana, USA.

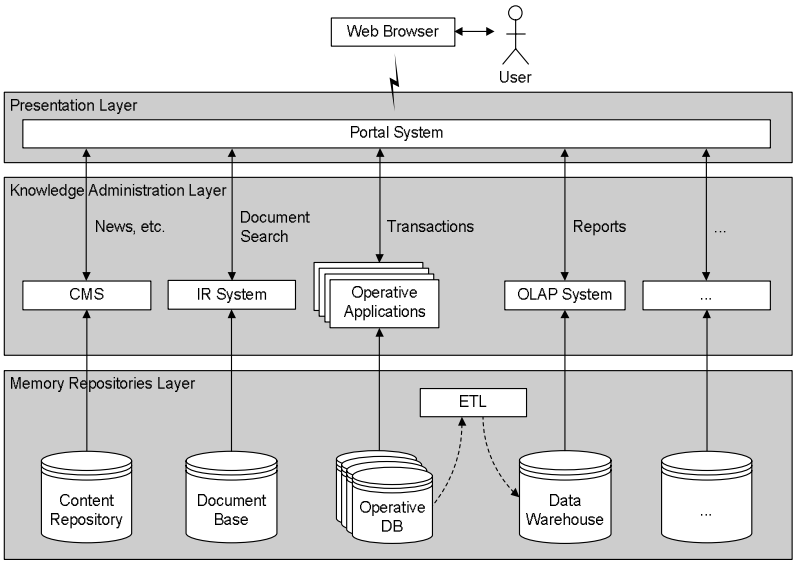

Figure 1. Organizational memory system with knowledge portal

interchangeably. The goal is to provide the user with a consolidated, personalized user interface to all information he needs for his daily tasks.

Recently the term enterprise knowledge portal is more and more used instead of enterprise information portal. Advanced techniques (like discussed in this paper) try to help the user with accessing the right information at the right time. This implies the support of organizational learning and corporate knowledge processes. Therefore enterprise knowledge portals are the ideal user interface to a knowledge management or organizational memory system (OMS) [Lehn02].

Typical elements of an OMS are operative applications, an OLAP system [ChDa97] to access data warehouse data, and an information retrieval (IR) system [BaRi99] to search for documents in a corporate document base (e.g. managed by a document management system). Additional components might include geographical information systems (GIS) or other decision support systems like expert systems. Being integrated into a corporate intranet through the portal, typical intranet content such as news articles is obviously also an integral part of an OMS. The overall architecture of an organizational memory system using a knowledge portal as a user interface is shown in figure 1.

Similar to authors like [Lehn02] we divide the architecture of an OMS into three layers. The memory repositories layer includes all the data stores that together build the organization's knowledge base. The knowledge administration layer contains the software components that are used to access and interpret the different data sources. Finally, the presentation layer is responsible for transporting the information to the end user. Our proposal is to use a web-based portal for this purpose.

There is a number of commercial portal platforms available today (e.g. IBM, BEA, Plumtree, etc.). The individual portal components (representing different information sources) which are rendered together to a portal webpage are called portlets [Wege02]. The screen design of our prototype in figure 8 and 9 at the end of this paper shows a typical portal page with three portlets. Many (e.g. OLAP) software vendors already provide specific portlet implementations of their systems for common portal platforms. Portal systems provide an integration platform on user interface level (i.e. within the presentation layer). The problem of available standard solutions is, however, the lack of interaction between the individual portlets. We will address this issue in section 4.

Before, we will however discuss some integration issues in the two other OMS layers. Integration in the memory repositories layer obviously means data integration. One of the most prominent examples is the ETL (extract, transform, load) process which extracts data from different source databases, and feeds it into a data warehouse. This process is depicted by the dashed arrows and the ETL component in figure 1. Additional integration initiatives on this layer would, among others, involve metadata integration which is discussed in more detail in the next section.

Integrating different information sources within the knowledge administration layer involves coupling the software components that represent them. Technologies like CORBA [http://www.omg.org/ corba/] can be utilized for this purpose. Another example for integration within this layer is the meta search approach presented in the next section.

3. GLOBAL SEARCHING

A major requirement for an enterprise knowledge portal is to be able to globally search for information, no matter where this information is stored or which piece of software manages it. The system should find documents from the DMS, news articles and other content, as well as predefined OLAP reports. We call all these resources throughout this paper. Within the next subsections we will discuss two possible approaches for providing a global search functionality.

3.1 Meta Searching

Meta searching assumes that each individual system managing data from a certain source has its own search capabilities. The problem of providing global searching is to integrate these individual search mechanisms. A portlet with a global search interface (like the one in figure 8) would communicate with a meta search engine rather than a search engine for a particular data source. This meta search engine forwards the user's search request to the individual search engines (e.g. an information retrieval system) and consolidates the search results.

Figure 2 shows the architecture of such a meta search system. Coming back to the three layers in figure 1, the approach can be seen as integration within the knowledge administration layer as it is realized by coupling different (search) software components.

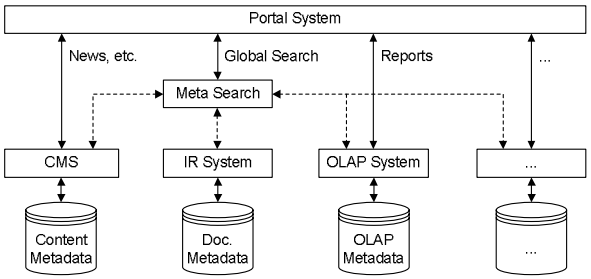

Figure 2. Global searching through meta searching

3.2 Metadata Integration

As you could already see in figure 2 we assume that the individual systems use metadata for their searches rather than searching the data itself. For example, for the IR system this means searching for certain document metadata (author, title, topic, etc.) rather than keywords that occur in the full-text. We argue that this search approach also allows semantic searching, e.g. for resources dealing with a certain product. This is particularly true for a corporate setting where the existence of rich metadata is easier to assure. We shortly present a possible approach for information retrieval on metadata in the next subsection.

If searching is done on metadata, another approach to provide a global search facility would be to integrate the individual metadata sets. This way an information retrieval system could not only search for documents but also other resources. As shown in figure 3, this could either be achieved by making sure the different systems all use a centralized metadata repository or by replicating the metadata from a proprietary local repository (like the OLAP repository shown in the figure) to a global one.

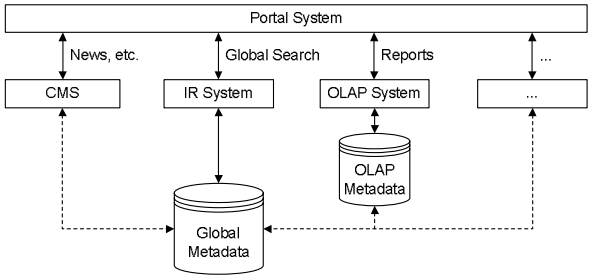

Figure 3. Global searching on integrated metadata

In general, the ideal situation would obviously be to have a single enterprise-wide metadata repository that is used by all system components. However, due to the independence and heterogeneity of (standard) software components that are being used, this is a difficult issue. Standardization efforts like the Common Warehouse Metamodel (CWM) [http://www.omg.org/cwm/] try to develop a standard metamodel that should be supported by participating software vendors.

On the other hand, in the Semantic Web [BeHL01] environment only the representation of the metadata (RDF) is standardized. The semantic integration is done by means of ontological mapping (synonyms, subclassing, etc.). This way, the individual systems can store their metadata using "their own language". This is also what we propose for our approach and use within our prototype. Figure 4 shows how a global ontology can help with mapping constructs of different metadata models.

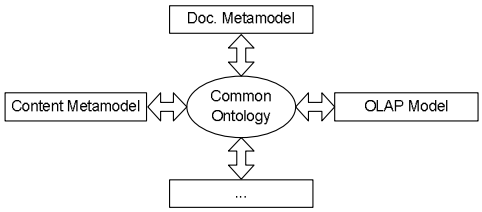

Figure 4. Ontology for model construct mapping

In addition to the resource metadata we propose to include business objects in the repository. By business objects we mean entities that are of enterprise wide importance, e.g. the product portfolio, customers, etc. These can be used for describing the topic of resources. For example a document dealing with a certain product can be explicitly linked to a business object representing that product.

Similarly, a predefined OLAP report can be described by the elements shown on its rows and columns. For example, the OLAP report shown in the OLAP portlet in figure 9 shows the sales of certain audio electronics products within the four quarters of 1998. The metadata of this report can be represented in RDF with a description like the one in figure 5 (see next page). For now, just note the "about" properties pointing to the different product business objects and the date ranges representing the quarters. The scenario used is a mail-order retail company that sells various consumer goods via calling centers. For details refer to [PrPe03].

3.3 Information Retrieval on Metadata

The main point about searching metadata rather than full-text keywords is the semantics that can be used, e.g. by utilizing the above mentioned business objects. For example, take a look at the query shown in figure 8. The user wishes to search for resources about a specific product (the Freeplay Solar Radio) that have the word "sales" in the title. Note that rather than just being a string literal, "Freeplay Solar Radio" was selected by the user from the metadata such that the system can identify it as a business object.

Now, the most straight forward search approach would be to simply perform an exact query on the metadata repository and return such (and only such) resources that fulfill the criteria specified as search constraints by the user. However, there are two problems. First, the metadata quality depends on the users' tagging, which is a voluntary process (that creates extra work). Second, we expect an enterprise metadata model to become quite complex, it is thus hard for users to build search queries that "perfectly" represent their information need.

For these reasons we propose an information retrieval approach similar to classical retrieval models like the vector room model (VRM) [BaRi99]. Usual IR queries which represent keyword-based full-text searches are of a fuzzy nature. A query returns a

ranked list of documents to the user showing most relevant ones first. In particular, the VRM is based on the similarity of document and query representations. Every document is represented as a vector of term frequencies (i.e. how frequently certain keywords occur in the documents). Another vector is created from the query keywords. The matching (and ranking) of the documents is done on the basis of the similarity of the individual document vectors and the query vector. Distance measures such as the cosine similarity are used.

Translating this into the world of semantic metadata means that we represent both the resources and queries as metadata descriptions. As a user defines his query by constraining certain properties, this set of properties can be represented as a (virtual) resource. In RDF this leads to an anonymous description as shown in figure 6.

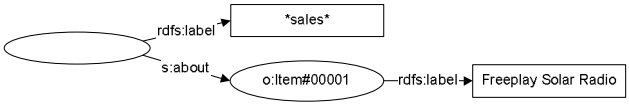

Figure 6. Sample query as RDF description

The search can then be performed by calculating a distance measure (or match percentage) between the query and each individual resource description. A good similarity measure for such metadata descriptions still has to be determined (and tested in a practical setting). One issue is to deal with distant relations. A resource about the Audio product group, to which the Freeplay Solar Radio belongs, should also be found by the above query, but with a lower match value than a resource that is directly related to the Freeplay Solar Radio product. Secondly, a weighting of property search constraints is required.

A possible search result of the query shown in figure 8 is found in figure 9. Again, for more details on our metadata search approach and the metadata model used in the example, see [PrPe03].

4. CONTEXT INTEGRATION

As mentioned earlier, many portal platforms exist as commercial systems. Also, many vendors, e.g. of OLAP systems, provide portlet implementations for such platforms. However, the problem of these standard solutions is the lack of interaction between the individual portlets. When a user navigates within a certain portlet (representing a certain knowledge source), the other portlets remain static.

In order to provide an efficient knowledge access within the organizational memory system it would be desirable that the user's information need, revealed by his navigation within one portlet, could be provided to the other portlets enabling them to automatically find related information. This leads to integration on user interface level, or, using the terms of our above architecture in figure 1, within the presentation layer.

A classic example is a finance portal providing access to stock quotes and news feeds, e.g. from Reuters. A user querying stock quotes for a particular ticker symbol is automatically provided with recent news concerning this company. This is, however, a special "hard-wired" solution. The approach presented in this paper develops a framework for communicating the user context between portlets to provide such integration in a generic way.

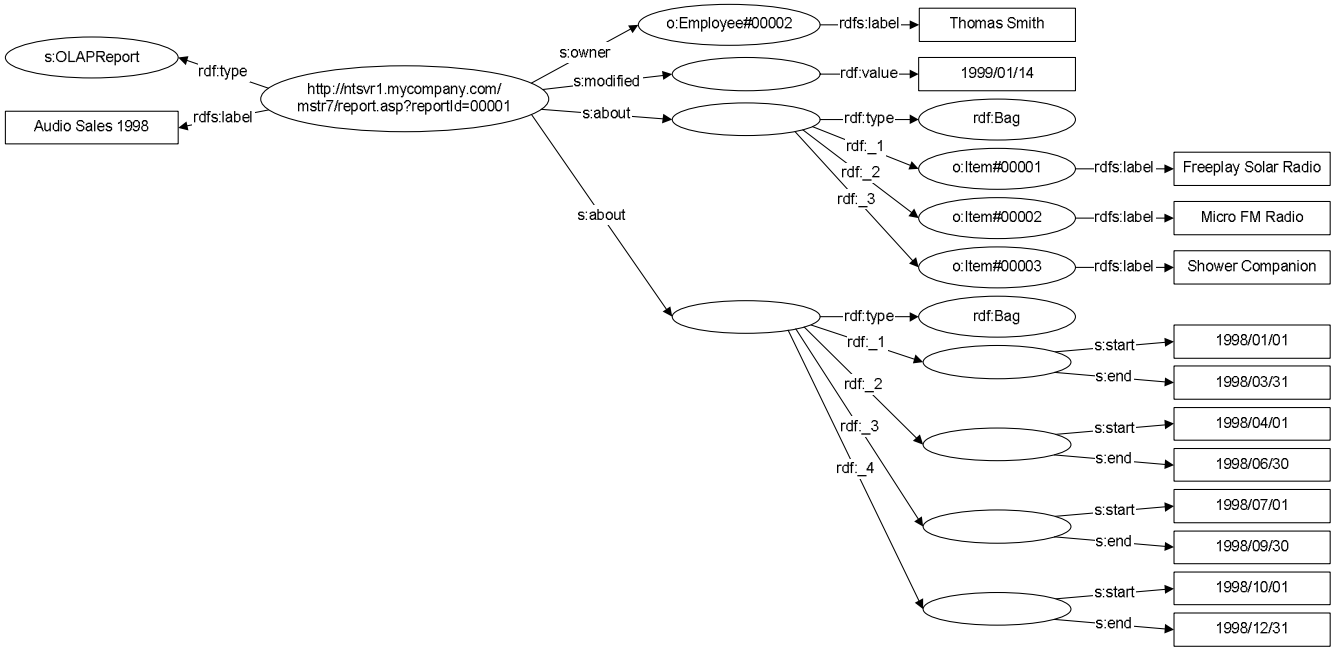

Figure 5. Metadata for sample OLAP report

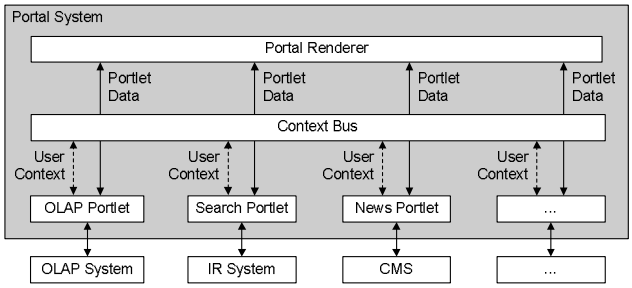

Figure 7. Architecture for context integration

In regular portal systems portlets only provide their portlet data for rendering the user interface. In addition, we introduce a communication bus where portlets can publish their current user context. Other portlets can pick up that context and use it to also show related information. Figure 7 shows the overall architecture of our context-based portlet integration.

Obviously, the problem is the heterogeneity of the portlets and the underlying systems that manage the information displayed by them. An OLAP portlet will use its OLAP data model to formulate the user context, while a portlet responsible for accessing a legacy application component will rather use an underlying operational data model or an application object model. Finally, a (metadata-based) information retrieval portlet will use the document metamodel for this purpose. To solve this problem, we again propose to use ontological concept mapping, just like the one used for global searching over integrated metadata in the previous section.

The main idea is to use the metadata description of a (possibly virtual) resource to represent the user context. For example, if a user displays the OLAP report shown in the OLAP portlet in figure 9, the user context can be represented as an RDF description just like the one in figure 5 (the same one that was used for describing the predefined OLAP report to provide global searching). However, this also works for ad-hoc reports that are dynamically created through drilling and slicing/dicing. Such reports might have no title and URI associated with them, but the principle remains the same.

After all, the approach is not only valid for an OLAP portlet. For example, a portlet representing a CRM system displaying information about a certain customer can use a similar metadata description pointing to a customer business object to represent its user context.

What remains is to show that such a user context can actually be used by other portlets to show related information. It turns out that the integrated metadata and the information retrieval approach described in the previous section can be used quite well to utilize such context descriptions within a search portlet. As the context is represented in the same way as resource metadata descriptions (which in turn are represented in the same way as search queries), the similarity-based information retrieval approach mentioned above can simply use this context description as a query.

5. PROTOTYPE

In order to evaluate our approach we are building a prototypical knowledge portal system based on the open source Apache Jetspeed portal platform [http://jakarta.apache.org/jetspeed/]. At this point we provide three portlets. One is responsible for displaying portal content (in particular news articles) and a second one provides access to an OLAP system (we use the MicroStrategy 7i [http://www.microstrategy.com] software as an OLAP engine). We also use the VMall demo data set provided by MicroStrategy as a case study for our prototype. As mentioned before, the scenario is a mail-order retail company.

A third portlet is responsible for metadata-based information retrieval. To provide global searching we use an integrated RDF (and RDF schema) [W3C99, W3C03a] based metadata set with concept mapping to translate different terms from different systems (e.g. the OLAP systems uses “owner” for what is called “author” within the document metadata). The expressiveness of RDF(S) is, however, somehow limited. At some point it might thus be advisable to switch to RDF(S) extensions like DAML+OIL or OWL [W3C03b], or even languages that support axiom formalisms. Nevertheless plain RDF(S) has the advantage of being a well accepted W3C standard with a still emerging but promising tool support. We use the open source framework Sesame [http://sesame.aidministrator.nl] for managing the RDF repository. Sesame supports the RQL query language [KCPA01, BrKa01] which can be used for metadata querying.

We assume that the content and document management systems have appropriate interfaces to directly access our RDF repository, while we integrate the OLAP metadata (stored in a proprietary repository managed by the MicroStrategy software) in an ETL-like process. The business objects (products, customers, etc.) are generated within the same process from OLAP dimension elements.

Figures 8 and 9 show preliminary screen designs of our prototype. While the search portlet in figure 8 depicts the mentioned user search query, figure 9 shows possible search results. The OLAP and news portlets apply a best fit strategy and display the report resp. news article most relevant to the query.

The context integration is represented by the “Find Related” link within the OLAP portlet. When a user displays a (predefined or dynamically generated) OLAP report, clicking on this link will cause the system to generate an RDF description of the query context (i.e. the elements shown on the report). The search portlet picks up this context description and performs a similarity based search to find related resources, like documents, news articles, etc. in the metadata repository. This is, as mentioned in the introduction, a unique feature of our approach.

Within the scope of this paper it was only possible to very roughly present the general idea of our implementation, which in addition is still under development. For more details see [PrPe03].

6. RELATED WORK

The integration of different data sources as a means of enabling enterprise knowledge management has been studied by various authors like [Lehn00]. They propose architectures for organizational memory systems using knowledge portals as an integrated user interface. However there are usually no (technical) details given how true integration can be achieved.

In the EU funded project GOAL (Geographic Information Online Analysis, INCO COPERNICUS project no. 977071) [KMM00] in which the authors were involved, the integration of data warehouse (or more precisely OLAP) technology and geographical information systems (GIS) was analyzed. The basic idea behind such integration is that a geographical OLAP dimension can be mapped to GIS objects such that maps can be used to navigate through OLAP data. The context integration approach presented in this paper can be seen as a generalization of this idea.

Like us, [RiKM00] address the integration of unstructured (or semi-structured) documents with structured OLAP data. However, their approach is quite different from ours. They treat OLAP cubes as "documents" stored in a digital library like repository and use manually created metadata to link them to related documents. Our approach goes a step further, the OLAP cubes are not treated as a black box, but the navigation inside the cubes (the above mentioned query context) is also considered for retrieving related documents.

Such context based information retrieval has been studied by [HeMo02] who propose a high-level architecture for finding documents relevant to the context the user is in. They propose to include plug-ins in client applications that would communicate the user's working context to an information retrieval engine. However, most application programs will not easily allow the integration of plug-ins. In addition, the identification of a user context and its translation to an IR query in a totally generic way seems problematic, which is why in more recent publications like [HeMo03] the authors have concentrated on the particular use within the software development environment.

For the same reason we focus on the (controllable) environment within a knowledge portal system. The use of an MIS (or OLAP) query context is explicitly mentioned as promising by [HeMo02] but not elaborated in more detail. Our retrieval approach, based on the similarity of metadata sets, is a first move towards how context-based IR could actually work in practice.

Information retrieval on the Semantic Web is also discussed by [ShFJ02]. They propose a hybrid approach of metadata- and full-text-based searching. Actually, combining our metadata-based approach with regular key-word based IR would be quite interesting and will be part of our future work.

[Aude03] also combines OLAP and information retrieval functionality. In technical terms his approach is quite similar to ours (he also uses RDF-based metadata). However, as for [RiKM00] OLAP cubes are treated as monolithic elements. The navigation within cubes and the query contexts of specific OLAP reports are not considered for the retrieval.

7. CONCLUSIONS

Nowadays, efficient access to information of all kinds is becoming more an more important. Organizational memory systems and enterprise knowledge portals provide a means of addressing this issue. In this paper we discussed integration aspects in enterprise knowledge portals. In particular we presented approaches for global searching over different information sources. We proposed to use an integrated metadata set and an ontology for concept mapping for this purpose.

In addition we introduced an approach for context integration to communicate the user context among different portlets representing different information sources. Using this approach an information retrieval system can, for example, automatically provide the user with documents from the organization's document management system that are related to what he is currently viewing in an OLAP report.

As it turned out that existing search mechanisms for metadata are not suitable for information retrieval we also sketched a similarity based retrieval approach. Finalizing this approach (especially the matching algorithm resp. the distance measure) is what we are currently working on. In parallel we are working on the development of a prototype portal system as a "proof of concept".

Future work will involve the integration of other additional information sources, for example operative (legacy) systems. In addition, combining our metadata-base retrieval approach with full-text searching (like proposed by [ShFJ02]) seems promising. Finally, the concept of personalization which is typical for portal systems should be merged with our context integration. After all, the identity (or other attributes) of a user is also part of the user context.

This leads to another field of future work: security. In earlier research we worked on access control in data warehouse and OLAP systems. Commercial OLAP systems nowadays provide quite mature security mechanisms to control what data can be accessed by whom. On the other side there are still quite a few open issues with controlling access to unstructured (textual) data. Another research project of our group deals with access control based on document metadata and user credentials. We are planning to integrate this security mechanism into our portal system. As mentioned above, the use of the user credentials not only for security purposes, but also for adaptive (context-based) searching seems promising, too.

8. REFERENCES

[Aude03] Audersch, S. XML-basiertes Content-Management mit OLAP-Funktionalität auf der Basis von RDF. Workshop of the GI working group "Konzepte des Data Warehousing", Munich, Germany, March 2002.

[BaRi99] Baeza-Yates, R. and Ribeiro-Neto, B. (Eds.). Modern Information Retrieval. Addison Wesley Longman Limited, 1999.

[BeHL01] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. In: Scientific American, May 2001.

[BrKa01] Broekstra, J. and Kampman, A. Query Language Definition. Technical Report, On-To-Knowledge (EU-IST-1999-10132) deliverable 9, Aidministrator Nederland bv, February 2001.

[ChDa97] Chaudhuri, S. and Dayal, U. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record, Volume 26, Issue 1, March 1997.

[HeMo02] Henrich, A. and Morgenroth, K. Integration von kontextunterstütztem Information Retrieval in Portalsysteme. Teilkonferenz Management der Mitarbeiter-Expertise in IT-Beratungsunternehmen, MKWI 2002, Nürnberg, Germany, 2002.

[HeMo03] Henrich, A. and Morgenroth, K. Supporting Collaborative Software Development by Context-Aware Information Retrieval Facilities. Proc. of the DEXA 2003 Workshop on Web Based Collaboration (WBC 2003), Prague, Czech Republic, September 2003.

[KCPA01] G. Karvounarakis, V. Christophides, D. Plexousakis, and S. Alexaki, “Querying RDF Descriptions for Community Web Portals”, Proc. French National Conference on Databases (BDA'01), Agadir, Maroc, November, 2001.

[KMM00] Kouba, Z., Matousek, K., and Miksovsky, P. On Data Warehouse and GIS Integration. Proc. 11th International Conference on Database and Expert Systems Applications (DEXA 2000), Greenwich, UK, September 2000.

[Lehn00] Lehner, F. Organizational Memory Systems. Carl Hanser Verlag, Munich, Germany, 2000.

[MiSR02] Miller, L., Seaborne, A. and Reggiori, A. Three Implementations of SquishQL, a Simple RDF Query Language. First International Semantic Web Conference (ISWC2002), Sardinia, June, 2002.

[PrPe03] Priebe, T. and Pernul, G. Ontology-based Integration of OLAP and Information Retrieval. Proc. of the DEXA 2003 Workshop on Web Semantics (WebS 2003), Prague, Czech Republic, September 2003.

[RiKM00] Rieger, B., Kleber, A., and von Maur, E. Metadata-based Integration of Qualitative and Quantitative Information Resources Approaching Knowledge Management. Proc. 8th European Conference of Information Systems, Vienna, Austria, July 2000.

[ShFJ02] Shah, U., Finin, T., and Joshi, A. Information Retrieval on the Semantic Web. Proc. of the Eleventh International Conference on Information and Knowledge Management, McLean, VA, USA, 2002.

[W3C99] Lassila, O. and Swick, R.R. (Eds.). Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, February 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[W3C03a] Brickley, D. and Guha, R.V. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Working Draft, January 2003. http://www.w3.org/TR/2003/WD-rdf-schema-20030123/

[W3C03b] McGuinnes, D.L. and van Harmelen, F. OWL Web Ontology Language Overview. W3C Working Draft, March 2003. http://www.w3.org/TR/2003/WD-owl-features-20030331/

[Wege02] Wege, C. Portal Server Technology. IEEE Internet Computing, May/June 2002.

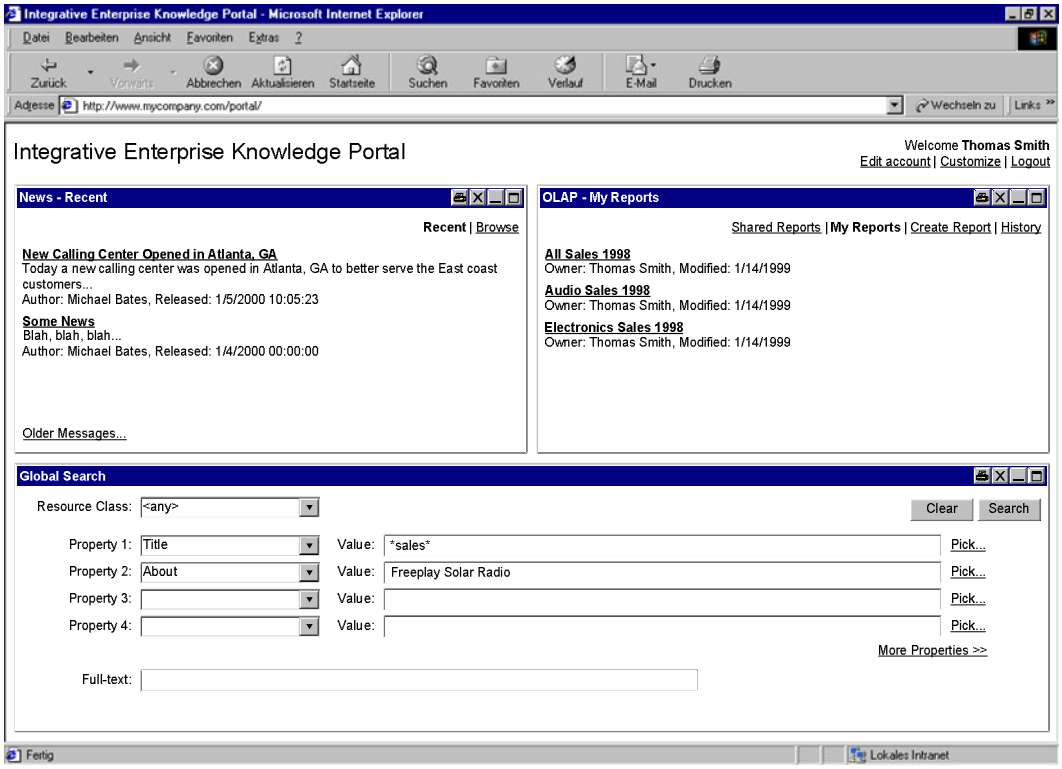

Figure 8. Portal prototype

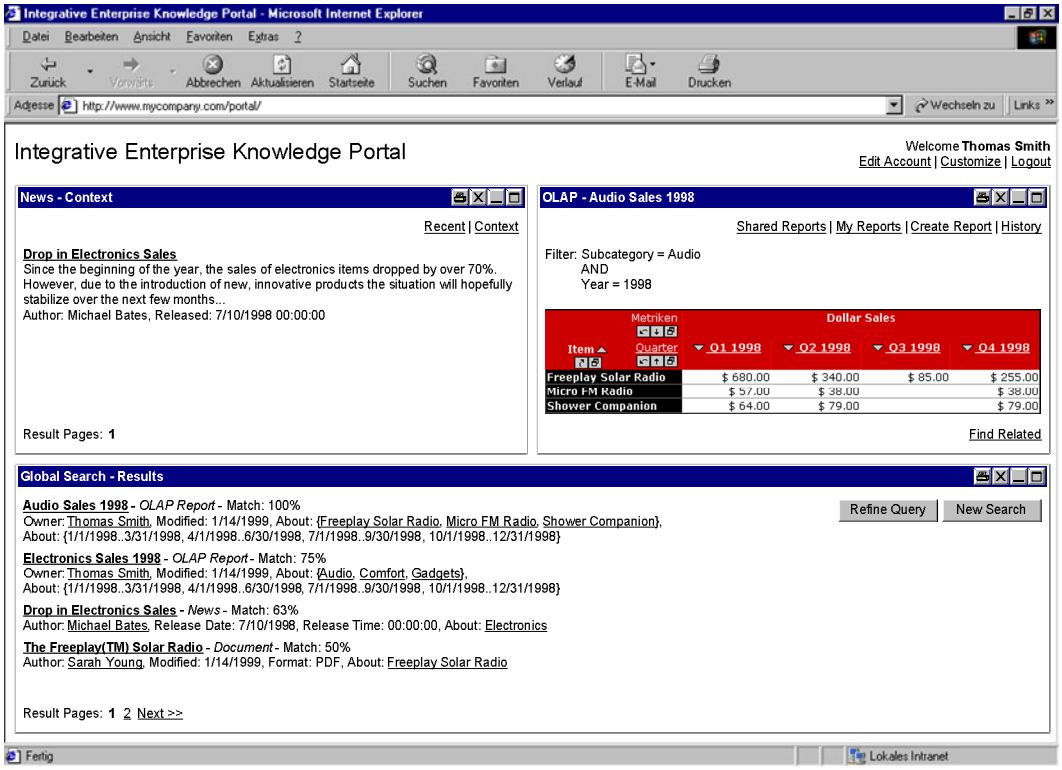

Figure 9. Portal after performing search