ARCHITECTURE DESIGN AND APPLICATION FOR AN INTELLIGENT DISTRIBUTED NEWS ARCHIVE

MARKUS W. SCHRANZ[1]; BERT PAEPEN[2]

[1] Distributed Systems Group, Institute of Information Systems,
Technical University of Vienna & pressetext.austria
Argentinierstrasse 8/184-1, Vienna, Austria
e-mail: schranz@infosys.tuwien.ac.at; schranz@pressetext.at

[2]Research Group on Document Architectures, K.U.Leuven
Kasteelpark Arenberg 10, 3001 Leuven
e-mail: bert.paepen@esat.kuleuven.ac.be

Throughout an entire decade, the Internet has brought unmanageable amounts of information to the average user's fingertips. Since this growth will only continue, it is vital that users are supported in converting this universe of information into improved productivity and opportunity instead of being swamped and paralyzed. Failing to address information overload will cost enterprises and individuals money, often in ways that are not easily measured: Costs that result from lowered productivity and from mislead business decisions. To really satisfy user needs and restricted budgets, the myriads of information need to be structured and organized in an intelligent and user-oriented way. Technically, appropriate architectures to integrate existing archives with an intelligent news retrieval engine are to be developed. The research approach in the discussed OmniPaper project is investigating ways for drastically enhancing access to many different types of distributed information resources. The key objective is the creation of a multilingual navigation and linking layer on top of distributed information resources in a self-learning environment, thus providing a sophisticated approach to manage multinational news archives with strong semantic coupling, delivering to the user more than the sum of the individual service features.

Keywords: News archive architecture, multilingual information retrieval

INTRODUCTION

Recent developments for the information society set the Internet into a leading medium for information transport and submission. Multiple content application areas allowed the Internet to become an information transport medium equally important to TV broadcast, radio or printed news, in particular fields even with unique quality and support for the interested and professional consumers.

Electronic information and access to it are scattered throughout the Internet. The data is geographically spread throughout the modern world. It has numerous access methods, storage formats and information structures[1]. Countries in which information is physically stored all have their own legislation, bringing along different approaches how to handle information. And last but not least, information can be stored in many different languages. To satisfy an information need, relevant libraries have to be selected, the information need has to be reformulated for every library with respect to its schema and query syntax, and the results have to be semantically joined[2]. Up to now, these are inefficient manual tasks for which accurate tools are desirable.

Promising research activities in the area of digital libraries[3] provide end-to-end solutions for federated digital libraries which cover most of the problematic issues. Information retrieval techniques, retrieval quality and the integration of non-cooperating libraries are the research focus. Especially in digital news archives, the integration of various existing non-standardized services is a demanding challenge to system architects. The OmniPaper project[4] follows the research work in exploiting the special application area of digitally available news libraries to provide professional access via the Internet.

This paper focuses on the results of architectural research for digital libraries and the integration of multilingual news archives into an intelligent international news search and retrieval engine for the WWW. It explains the research work conducted and is structured as follows. Section 2 defines the basic ideas, concepts, and requirements of the OmniPaper project. Section 3 focuses on the architectural design and reasoning and provides an overview on system components and news archive processes. Section 4 explains the integration of the proposed architecture for the Web prototype developed during the project and gives a critical evaluation of its utility and usability. The current status of the project, scheduled activities and results, as well as the planned business use round up the paper in the conclusion.

ARCHITECTURAL REQUIREMENTS

Since the emerging boom of the Internet all different kinds of information has been published and various formats and media types are transferred to the fastest growing information distribution service on earth, the World Wide Web. Particularly also a lot of newspapers are being published on the World Wide Web. This increasing amount of news items remains scattered throughout various archives, countries and languages. Searching for news is still mostly done using full-text search robots which lead to a search result quality that highly depends on the sophistication of the user's search input. In fact, finding semantically connected news from various international newspapers is still easier in an airport news kiosk than on the Internet.

The OmniPaper project is investigating techniques to create a novel online news experience, using up-to-date XML- and AI-related technologies. The OmniPaper architecture starts from distributed news archives, all within different operating environments, database formats and indexing mechanisms. Heterogeneity, performance and usability are challenges to the responsible system architects. In a standardization effort, SOAP (Simple Object Access Protocol[5]) has been selected to create a uniform access method to the existing archives. In addition to the simple access requirements, the intelligent news archive is required to extract specific contents and create relations between information units. Rich indexing and meta-data structures, such as Topic Maps[6] and RDF are utilized to make intelligent search possible. A cross-archive intelligent index built from the news and metadata processing, contains concepts, relationships between them and occurrences of the news items in different languages.

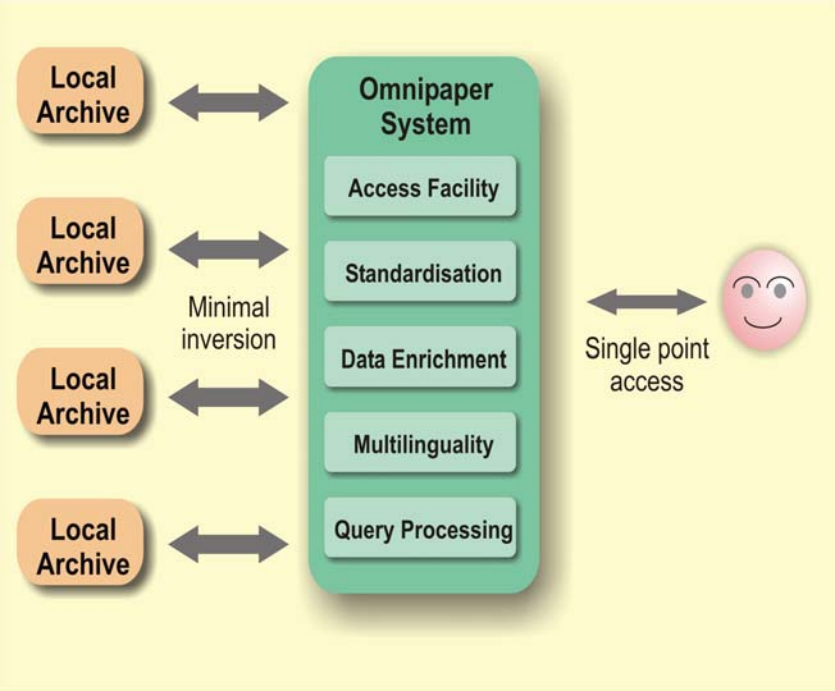

FIGURE 1 – ABSTRACT ARCHITECTURE OF OMNIPAPER

Figure 1 shows an abstract architecture of the OmniPaper system. Distributed source data exists in various standard formats, different languages and with varying depth of available information. The existing local archives are connected to the OmniPaper system using minimal inversion i.e. existing access methods are reused and wrapped to standardized SOAP requests with consideration of moderate change and cost effort with the information providing partners. Search and browse results are retrieved from all local archives.

ARCHITECTURAL DESIGN AND IMPLEMENTATION

In order to meet the goals and fulfil the requirements of the multilingual news archive, research on and evaluation of related international projects[2], retrieval methodologies[7] and semantic relation approaches[8] has been applied in the field of digital libraries and news archives. Integrating the experience from research and practical applications in complex technical environments, especially in the area of Internet Services[9], we created a system architecture for web-based distributed heterogeneous news archives like the OmniPaper service.

The architecture is discussed from three different perspectives: we defined views on system components and system processes as well as on system interfaces.

The top level system architecture contains a multi-layer view on the technical architecture of a distributed news archive. The architecture is based on existing digital news archives as the bottom layer. The distributed information retrieval or “local knowledge” layer contains components and control processes that access the existing archives via standardized interfaces for news retrieval and metadata management. The “overall knowledge layer” combines the features of integrating distributed information with the capability of creating semantic coupling of the corresponding content.

The multilingual aspect is supported by extracting existing keywords and metadata from the heterogeneous archive information and by associating them with existing domain-specific thesauri for the relevant language. The overall knowledge layer contains a network of thesauri, coupling corresponding standardized terms and enabling the intelligent news engine

to find corresponding articles in news archives over different countries and languages. Based on these layers, the topmost user interface layer allows journalists and researchers to investigate material on specific topics in a multilingual environment, relying on high result quality and content relevance.

The top level system architecture is constructed mainly from the point of view of a news provider. Existing archives and news repositories act as data sources for the distributed news archive and provide their contents via standardized SOAP interfaces, using the World Wide Web as widely available transport medium. The system itself offers simply a data feed interface, acting as black box for the existing archives. To provide openness and facilitate easy archive extension with considerations in cost-effectiveness the set of SOAP queries is limited to four messages: FullTextSearch, IdentifiedSearch, NewsUpdate and MetaDataFeedback (see item *access processes*). The relations and attributes for each article created within the intelligent multilingual distributed news archive are managed and updated regularly.

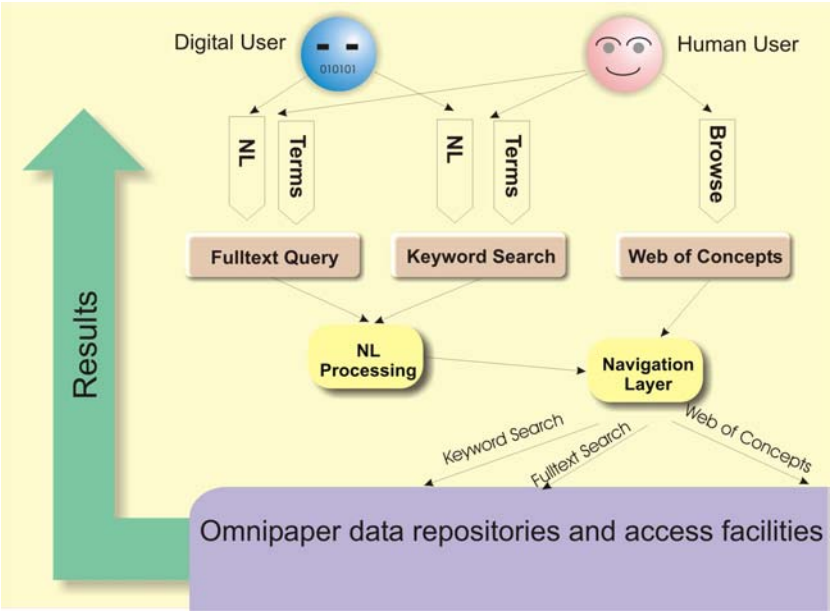

FIGURE 2: USER INTERFACE TO OMNIPAPER

From the user's perspective the distributed news archive offers a news professionalist interface. Either digital users (applications) or human users may access the archive using natural language queries, sets of terms or browse through a web of related terms (concepts) as depicted in Figure 2. The user interface is designed in the system architecture as a user-centred Web application, providing convenient ways to access multilingual and widely distributed news archives. For digital users both simple access to Web forms and XML Web Services are designed in the system architecture.

DISTRIBUTED NEWS ARCHIVE COMPONENTS WITHIN THE ARCHITECTURE

Several information repositories with different types of content handling and interrelation mechanisms are utilized to compose an intelligent distributed news archive. The following table denominates and explains the major system components.

TABLE 1. OMNIPAPER COMPONENTS

Component	Intention and usage explanation
Existing Archive	News archive(s) that manage significant amounts of information and provide proprietary (Web or legacy) interfaces for search and retrieval. SOAP query handling for three mandatory requests types (query 4 is optional) is required
TNDB	Temporary News Database, an auxiliary database for unprocessed news articles. This repository within the distributed news archive system is accessed within the feed news process. The content is processed within regular intervals.
ODB	Ontology Database, listing keys & concepts. Attributes are stored for each news article based on AI technology. Vectors of keywords and weights are managed to be recognized as related to abstract or concrete concepts
Linguistic Resources	The Linguistic Resources DB holds a set of language specific thesauri and relations between terms and keywords across multiple languages.
Extracted Links DB	Database that holds automatically extracted links from a specific article to other information within the distributed news archive.
MDDB	The Metadata-Database manages a set of data for each article within the distributed news archive. User queries are targeted to this database and forwarded to specific repositories if necessary.

Table 1 focuses on data repositories as part of the architecture. Several processes are managed to maintain the news information and create, store, maintain and evolve knowledge within this components. The overall system architecture is described in Figure 3, providing a conceptual view on the system components and their interaction following specific OmniPaper processes.

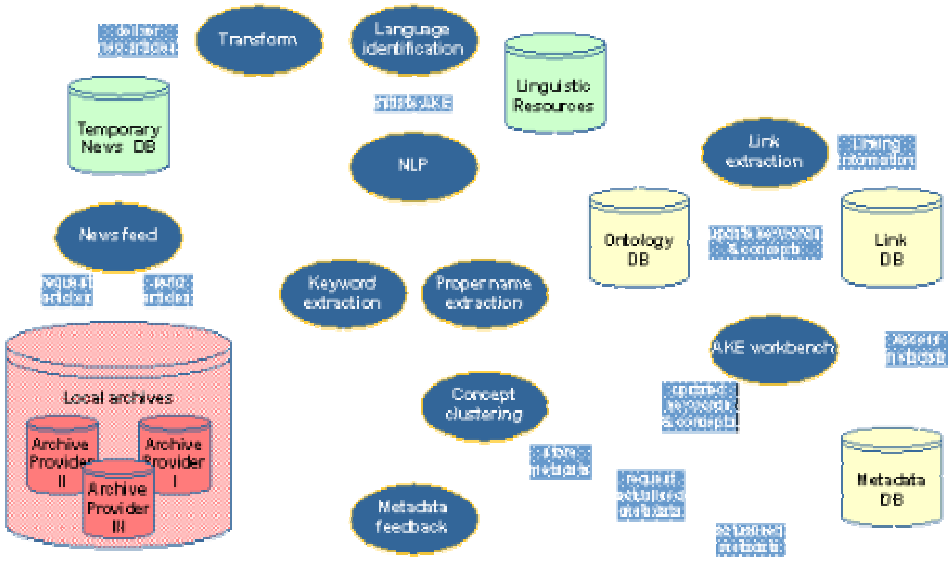

FIGURE 3: OMNIPAPER ARCHITECTURE AND COMPONENT INTERACTION THROUGH PROCESSES
OMNIPAPER PROCESS DESCRIPTIONS

This section describes the processes that enrich the system architecture depicted in Figure 3 with real news retireval activities. They explain the utilization of the intelligent components such as multilingual thesauri and content relation management in order to build this modern application in the area of Semantic Web. OmniPaper components and processes are either close to the news providers, thus providing small and manageable functionality to access and retrieve news information or they are used to build and maintain knowledge and create an intelligent distributed news archive.

In order to distinguish the utilization of the processes and describe a structured picture of the process architecture, the subprocesses are organized in groups of "acquisition processes", "knowledge management processes" and "user access processes".

Acquisition Processes

These processes manage activities that involve the news providers in sending/retrieving news articles to the distributed news archive.

The process *feed news* informs the system about the existence of new articles in the participating archives. The archive provides unique IDs to access the articles within and from the distributed news archive. The existing archive triggers this process by calling the SOAP query „NewsUpdate". The process feed news accepts the SOAP call and enters the parameter IDs into the auxiliary database TNDB. The TNDB stores article IDs and all provided information for the articles that are not analyzed yet within the distributed news archive. It logically separates IDs from different news providers to maintain source information and provide hints for specific access methods/restrictions.

Through the process *metadata feedback* the news provider is able to take advantage of the metadata extraction and knowledge generation processes of the overall system, to be used within the distributed news archive. Extracted metadata can be requested by the existing archives to eventually integrate them for their local services. The process *metadata feedback* accepts requests for metadata feedback on an unique article id from the existing archives. The feedback on metadata process requests the relevant meta data from the MD Database and delivers it to the requesting archive. It has to be considered that Meta data may change over time, since the knowledge management components regularly calculate relations between articles based on metadata, keywords and concepts. Newly identified concepts or updated weights in keyword vectors can influence relations and links between articles.

Knowledge Management Processes

The metadata information within the knowledge databases is used to create relations between articles to form a semantic network of news. Relevant activities are triggered either by receiving a new article or regularly based on predefined or system-dependant time intervals.

The process *transform* starts the complex set of tasks to extract, generate, manage and store metadata information from newly provided articles by the existing archives. The process fetches (and removes) the entries of non-analyzed articles (serially for all providers) from the TNDB. The periodic fetching is time triggered within the distributed news archive system (e.g. run every 10 minutes). All metadata extraction processes are repeated for each fetched article. The transform process converts the information presented by the existing archives into a standardized format within the distributed news archive (i.e. NITF 3.0 as the appropriate XML application for online news).

The *natural language process* is the first task within the keyword extraction activities. The standardized content (not the existing metadata information) of each article is provided to the natural language processing by the preceding transform to standardized format process. The *natural language process* compares contained words with linguistic resources in order to normalize the content for keys & concepts extraction. During normalization stopwords are

removed, wordstemming and grammatics are considered. The remaining terms set the base for frequency calculations.

The extracted normalized words and the original metadata information is provided to the *concept clustering process* and results are stored both in the ontology DB (ODB) and in the metadata DB (MDDB).

The *keyword extraction process* generates a vector-based metadata description for each article. The normalized content and the original metadata information of the examined article is compared to the set of normalized words in the linguistic DB to determine weights of keywords within a larger context (distinguished on language level). Natural language techniques are used to identify compound keywords or syntactic patterns in order to update co-occurences of terms. The calculated weights for the identified keywords are stored as a vector in the ODB. The resulting vectors are compared to existing concepts, which are also available as vectors. Vectors are linked to concepts on similarity thresholds. Concepts and related vectors are stored a-priori, or newly created in the ODB. They are either represented by predefined example vectors (e.g. "sports"-vector) or dynamically generated as anonymous concepts. Anonymous concepts are identified using AI methods to determine significant similarity.

The *workbench for AKE process* is initiated to verify the validity of automatically extracted keys and concepts, and their relations to the articles. The workbench process fetches single articles from the MD database and displays it to the evaluator (ideally enriched by the fulltext of the article). The evaluator accepts or rejects the provided set of keys and concepts to the discussed article. The corrected values are stored in the MDDB and relevant actions are initiated to update the ODB accordingly.

The *link article process* enriches the relations between articles based on automatic keyword extraction and multilinguality through several subprocesses. The extract links between articles subprocess uses keyword similarities to generate links to other articles (same or other language, translation of keywords by use of the linguistic ressources). The extract proper names subprocess creates links to further descriptions (e.g. company infos) in a predefined Name database. The topic linking subprocess fetches the corresponding concepts from the KCDB/KCDB-CR. It generates a set of links to related articles within the same concepts. For performance and scalability reasons, the link article process stores the previously extracted links into the MD Database for each article. As a result, the MDDB is able to provide links to variously related news for each article.

Access Processes

Access activities in the distributed news archive are triggered by human or digital user interaction. The access processes are not shown in Figure 3, since they do not influence the metadata archive as such. This type of processes take requests from the user interface and make use of the results of the previously described processes.

The *fulltext search process* accepts a set of terms from the user interface. A preceding natural language processing may transform a provided NL-query into the appropriate set of terms. The set of terms is passed to the SOAP FulltextSearch query that is targeted to all involved existing archives. Results are collected and fused by the distributed news archive. Specific cost measures, e.g. timeout, system availability, preferences determine the quantity of the result. As a result, the user interface is provided with an XML listing of matched news articles.

The *keyword search process* accepts a similar input as the fulltext search process: a set of terms is passed to it and handed over to a query on the MDDB. The caching nature of the MDDB within the distributed news archive system provides fast access and immediate response for the user interface compared to the fulltext search process. The information stored

in the MDDB is extracted according to the keyword search and the user interface is provided with an XML listing and weight information of matched articles within the distributed news archive.

The process *browse web of concepts* provides an hypermedia approach to information retrieval within the distributed news archive. At the user interface, named concepts are visualized with their relations between them and an appropriate browser (e.g. visualization in SVG) provides the navigation through this web of concepts according to the user's input. After selecting a specific concept, the process browse web of concepts is provided with the set of terms corresponding to the chosen concept. The set of terms is forwarded to the MDDB for information retrieval. As a result, the user interface is provided with an XML listing of news articles corresponding to the currently selected concept. The user interface is responsible for visualizing the returned results.

The process *show article details* focuses on the visualization of a single news item within the OmniPaper system. All previously described access processes produce as results XML listings of matching news articles. Every time a user interface requests a specific article, the process show article details is triggered. Depending on the requested details, the process requests the information solely from the MD database or initiates the SOAP IdentifiedSearch request to the source archive, for which the ID is also retrieved from the MD database. The article details are provided together with metadata information on related keywords, concepts and links to further related articles within the distributed news archives. As a result, the article details are provided to the user interface, which is responsible for visualization and enhancement of content and metainformation presentation.

ARCHITECTURE EVALUATION

OmniPaper users can use a sophisticated approach to access multinational news archives with strong semantic coupling of the retrieved contents. As a mature result of the project consortium, the discussed system architecture has been implemented and verified with different system prototypes.

Architecture Evaluation in OmniPaper

The design and implementation of the OmniPaper technical prototype follows the abstract concept defined in the architecture description above. As defined in the abstract top level system architecture in Figure 2, the concrete implementation follows a logical structuring: the logical layering of architecture components.

The *local layer* provides standardized interfaces to lately three news (re)distributors and enables to access about 8.7 million documents. The SOAP based interface provides unified access to the local databases (existing as Oracle® and Mysql systems), hosted on Windows and Linux systems. The local layer contains XML data structures to retrieve newspaper information from distributed archives. The queries focus on article selection by search criteria and keyword extraction.

The results from the local layer queries constitute the input for data management on the *overall layer* which is developed in distributed units. According to the architecture the required components and repositories are used and new techniques are analyzed and compared. XTM and RDF based auxiliary databases are core units of this layer, which enhances the initial available data by the use of AI (automated keyword extraction, ontological structuring, similarity measures of concepts, etc.), Multilingualism and Knowledge Management.

A *layer for user-friendly presentation* of the system, based on current HCI understandings, is set up on top of the overall knowledge layer. Efforts in the corresponding project work packages focus on the news search engine, the display of the results from processes in the overall knowledge layer, and the visualization of newspaper articles and

cross-links. Obeying current technological developments, the user interface has been planned and is currently developed based on modern web service and web browser technology[9].

The described architecture has been applied to the OmniPaper prototype system between January and October 2003. The real architecture is too complex to be expressed and outlined in detail within this article. Nevertheless major components are implemented as described in the designed model from Figure 3. Existing archives have been integrated with SOAP implementations on both sides, the existing archives and the OmniPaper prototype. The abstract system architecture has been applied for the specific application area of newspaper articles and online news distribution. The participating three news distributors have implemented the SOAP definitions and deliver online news in NITF3.0 format.

The prototype includes implementations of all data repository components, distributed on servers within the EC project consortium partners. The components are logically structured and centered around process definitions as described above. The grouped modules were developed as parts of individual prototypes and have been integrated recently into a overall OmniPaper prototype. The modules interact on the same technical basis as the integration with the existing archives has been realized: using SOAP queries. This design decision has been jointly developed to allow service distribution and load balancing within the available project resources.

The acquisition processes were implemented as defined in the abstract architecture. The TNDB fetches the entire extracted metadata information for new and unhandled news articles at once to avoid multiple archive accesses. The knowledge management processes are realized as described, the ODB holds vectors of weights of keywords for all articles to be used to maintain semantic relations between news. No article fulltext is stored at any time of the processing and metadata management for document rights reasons. The metadata sets are stored in standardized format (RDF) in the MDDB, controlled by a current implementation of an RDF management engine. For each article ID one metadata set is stored in the MD Database. On a regular basis, triggered by predefined time events the metadata is updated within the OmniPaper prototype.

The user interface is targeted towards news professionals and news delivery customers. Besides fulltext and NL querying, the interface provides a highly interactive approach with the possibility to browse a web of concepts. Based on the ontology of the structured multilingual thesaurus EuroWordNet[10] a web of individual concepts can be browsed semantically by the end user, thus identifying related news articles at any position of the web of concepts for further investigation. Additionally, the search and retrieval access can be widened or narrowed by accessing a broader resp. narrower term within the EuroWordNet ontology. On top of the comfortable news retrieval, the multilingual nature of the EuroWordNet integration provides the access to articles in multiple languages, all related to the currently active concept and thus semantically related to increase the quality and multilinguality of the OmniPaper retrieval results.

Results and Experiences gathered from OmniPaper

Although the project is scheduled to be finished in December 2004, the technical design and development phases have been already finished to a wide extent and the experiences of the design and implementation phases can be used to report on results in positive and negative qualities.

The project consortium identified general positive results in project management, especially in time discipline and the application of software engineering methodologies, based on highly experienced project partners.

Concerning the defined architecture for an intelligent distributed news archive, the authors could identify the following positive results:

1. Satisfactory application to involved news archives
   All involved partners were able to implement the SOAP services within the estimated resource effort, which was scheduled low to allow easy system enlargement by attracting further archive participants. The utilization of SOAP as access protocol proved simple and reliable and allows a quick and easy expansion of the prototype to an European news network.
2. Positive User-Feedback for semantic web of concept
   The browsing of the semantic web of concepts, based on the multilingual ontology of EuroWordNet has earned very positive user feedbacks during system tests. Especially this feature is considered a strong business benefit and USP of the OmniPaper prototype.
3. Applicability to business cases
   Two of the participating news archive providers have identified the features of the OmniPaper prototype as promising enrichments of their current business solutions. The consortium is currently defining business cases to utilize the research prototypes in future business cases.

Besides the satisfactory results some issues remain open to transfer the research results to eventually successful business applications:

1. Optimization of component distribution
   As proof of concept the implementation of the defined architecture within the OmniPaper prototype is satisfactory. Depending on logically inevitable clusterings, the components have been allocated to hardware close to the mostly experienced partners within the project consortium, thus creating a widely scattered distribution of the logically centralized service.
2. Performance tuning
   Due to performance reasons, a final business implementation following the prototype to build an economically successful service requires an optimal definition of the distribution topology and adequate employment of preferably homogenous hardware and software resources, which is yet undefined.
   Furthermore, the scalability of the research prototype is yet to be investigated, demanding optimization in process structuring and replication and multiple instantiation of architectural components regarding improved performance to provide the users with satisfactory results, both in time delay and content quality.
3. Definition of a legal entity to provide services
   Although the proved architecture and the working prototype are sufficient results for the research project OmniPaper and despite all promising technical and research results, the consortium is interested in the economic fortune of the project results after the predefined project deadline. In order to build a successful market product out of the research prototype, also legal issues need to be set. This includes the digital rights management of the contents of all involved news archive providers and the definition of a legal entity that is entitled to bring the service to market.

CONCLUSION

Developed for very large-scale distributed collections, the proposed architecture is targeted to serve systems that improve access to cultural and scientific knowledge sources. Access to digital news services will not only be improved quantitatively by combining a large number of digital newspapers in one system. The architecture also improves the quality of

access by supporting the building of personalized, cross-lingual and self-learning interfaces to the distributed collections.

The identified components and processes provide a stable and well-performing basis for an intelligent distributed information archive. Basic processes for content acquisition and knowledge management are defined and implemented by the project consortium, which is constituted by news providers, system architecture experts, modern information standard experts and usability engineers. A concrete implementation of the defined architecture has been engineered and evaluated within the EU-funded project OmniPaper. The prototype provides access to personalized and context-specific content, and organizes heterogeneous information sources using ontology and semantic cross-lingual search. The self-learning aspect of the prototype system includes the analysis of user behavior and the enriching of the knowledge layer with lessons learned from this behavior. Thus, people can share knowledge without even realizing it.

The project has been including cross-testings of several approaches and election of the most successful techniques in order to build a framework and a blueprint as a guideline for future intelligent and multilingual news archives on the World Wide Web. The discussed architecture, applied in the implemented and herein evaluated prototype, is part of a final blueprint documentation.

ACKNOWLEDGMENTS

This work was partially funded by the EU 5th Framework project OmniPaper (Smart Access to European Newspapers, IST-2001-32174).

NOTES AND REFERENCES

1 Rosenfeld, L. and Morville, P. *Information Architecture for the World Wide Web*, O'Reilly & Associates, Aug 2002.

2 Nottelmann, H. and Fuhr, N. MIND: An architecture for multimedia information retrieval in federated digital libraries, Proceedings of the DELOS-Workshop on Interoperability in Digital Libraries. DELOS-Network of Excellence on Digital Libraries., 2001

3 Fuhr, N. Towards data abstraction in networked information retrieval systems. Information Processing and Management, 35(2), 1999, p. 101-119.

4 Bueno, F. et. al. *OmniPaper – Smart Access to European Newspapers*, EU project IST 2001-32174, http://www.omnipaper.org/, Jan 2002.

5 Gudgin M. et. al. *SOAP Version 1.2 Part 1: Messaging Framework*, http://www.w3.org/TR/SOAP/, 24 June 2003

6 Pepper S. and Moore G. *XML Topic Maps (XTM) 1.0*, http://www.topicmaps.org/xtm/1.0/, August 2001.

7 Mantzaris S.L. et al. Integrated search tools for newspaper digital libraries, in proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR 2000, July 24-28 Athens, Greece, 2000.

8 Kramer R. et al. Thesaurus federations : loosely integrated thesauri for document retrieval in networks based on Internet technologies, International Journal on Digital Libraries 1(2) pp 122-131, June 1997.

9 Schranz, M. et. al. Engineering Complex World Wide Web Services with JESSICA and UML. In proceedings (ISBN 0-7695-0493-0) of the 'Hawaii International Conference On System Sciences HICSS-33', Maui, Hawaii, USA, Jan 4-7, Jan 2000, p. 167.

10 Vossen, P., *EuroWordNet*, EU-funded project, LE-4 8328, http://www.illc.uva.nl/EuroWordNet/ September 2001.