Intelligent News Publishing and Distribution Through Web Services, Peer-to-Peer Technologies and Vector Space Content Relations

Markus W. Schranz
Pressetext Nachrichtenagentur GmbH
Josefstaedter Strasse 44, 1080 Vienna, Austria
schranz@pressetext.at

Christian Platzer
Vienna University of Technology
Argentinierstr. 8/184-1, 1040 Vienna, Austria
c.platzer@infosys.tuwien.ac.at

Abstract

Besides a clear research and archive focus electronic publishing technologies become increasingly important to business application domains to improve the service values for their end users and customers. The specific domains news publishing and news distribution are facing significant performance requirements in information handling while transferring hundreds of thousands of articles daily to their auditorium. Additionally, the business application of electronic publishing technologies in the specific news domain is confronted with the fact that European business today is highly segmented and widely unrecognized beyond national borders, mainly due to language differences and economical gaps. This paper discusses a project that integrates news agency services from existing European organizations supported by University research in the are of electronic publishing, distributed information management and AI in order to form an intelligent multinational and multilingual business news publishing and distribution network, based on Web Services and a peer-to-peer inter-agency communication network. Highly relevant business contents are related to a specific article automatically using AI methods from the research area of information retrieval. Contents are distributed in any language the provider chooses and the vector space model has been utilized to provide easy access to related and most relevant business news articles within a multilingual and multinational context.

1 Introduction

Electronic publishing has introduced a wide application range within the information management field, including the digitisation of ancient archives, the manageability and availability of enormous amounts of data or the visualization and provision of contents to research and open public. Modern approaches have enriched the field of electronic publishing with the research results of distributed computing, engineering of complex web services (Schranz 2000), information retrieval methodologies, eLearning, theories of infocracy, security, privacy, semantic relations and metadata towards the Semantic Web.

Beside the large variety of scientific approaches, serious business impacts have been provided by electronic publishing. Online information such as news have been spread with a new class of speed and range all over the world, making news available to virtually everyone within a very short time.

Within the overwhelming amount of available information, specific branches of news management have developed and integrate modern concepts of the electronic publishing research area (Bueno 2002). Business news mostly bear national relevance but hold the potential to spread cooperation opportunities and business changes towards an economically and socially integrated global information society.

European business as of today is highly segmented and widely unrecognised beyond national and language borders. This holds especially true for the small and medium enterprises which are constituting the majority of the European business. News publishing and distribution is thus facing both national borders and language obstacles, hindering global cooperation and business success. Technically, these obstacles can be overcome by introducing up-to-date distributed computing standards like Web Services (W3C 2001; Gudgin 2003), multilingual integration and international networks of technical and business cooperation.

This paper focuses on the creation and the conception of an intelligent news publishing and distribution network consisting of existing local news agencies that use modern electronic publishing and distributed computing technologies to build up a new kind of multinational and multilingual service. Research fields such as network communication have been involved to create a scaleable peer-to-peer architecture, artificial intelligence is utilized to identify the most relevant related articles within the entire multinational network that can add to the quality of the business news currently in focus and modern synchronous and asynchronous electronic publishing

mechanisms (such as web-based information access frontends and business news archives or email-based mass news distribution or content provision for next generation user devices) are used to publish and distribute the current business news.

2 Application Domain

The network discussed within this paper is targeting at a multinational and multilingual integration of such business cases, thus allowing the news agencies of different European countries to share their contents and exchange their business news towards an integrated network for news aggregation, creation and dissemination. The idea of integrating several European markets is most attractive to the lately identified class of customers, the private, commercial and institutional content providers.
Aside from the obvious business benefits of such a service integration, there are necessary steps to be taken to technically and organizationally bring the services and the systems of the existing news agencies to an integrated network. Beyond this, the project consortium is currently developing a demonstrator and initial business service that shall attract additional agency partners throughout Europe to join the network in order to

- Have access to relevant business news at an international level
- offer a distribution and dissemination interface for their customers that provide news to the network

Technical Architecture Overview

Electronic publishing in the area of business news distribution involves technical features to manage scalability and performance in mass information provision (millions of page impressions) and mass distribution (millions of electronic mails sent daily). Since scalability is managed by integrating local strength into a powerful network, modern networking features and capabilities are within the focus of the technical architecture.
The final approach taken in the project is based on a peer-to-peer network (Akavipat 2004), which is a unique way to create a network for news exchange between European news agencies. The developed state machines and interaction models define the entry point for the implementation that is starting to create a first prototype. New technologies like NewsML(IPTC 2005), Web Services via SOAP(Gudgin 2003), WSDL (W3C 2001) or XML in general, are used to create this decentralized system and connect it to all participants. With the methods outlined, we have established a reliable network, which is easy to use and easy to integrate into the target systems.
The core architectural concepts handle data management, the publishing interface and the interface connections. The interfaces between the communicating peers and the interfaces between a peer and an existing local system are defined in detail by Web Service descriptions. Existing services use modern web services to retrieve remote business news information to be electronically published for the local service and/or distributed via local channels and the network for remote distribution and electronic publishing.

3 Multinational and Multilingual News Network Approaches

Nedine composes a network to integrate multiple European national information sources consisting of participating news agencies, PR agencies and independent journalists into an international information service for news professionals and decision makers. The project provides a network of news exchange and distribution that supports mutual awareness of relevant topics and information areas within multiple European countries. With its main focus on widespread availability and affordability for all partners it addresses news providers to transport national and international information to the relevant target group, regardless of the origin, nationality and financial capability of the information provider.

3.1 Multinational Service Integration

One of the initial objectives of the Nedine project has been the development of a sophisticated news platform and a high performing distribution network based on modern digital news exchange technologies, not only targeted at newsrooms but aimed at reaching as many business leaders and decision makers as possible directly and personalized at their desktop.
The Technical University of Vienna and the Universidad Polytecnica de Madrid have contributed their knowledge in distributed computing and have researched for feasible network architectures to integrate existing news publishing and distribution services in Europe. Based on the existing news services of local agencies in Austria, Czech Republic, Germany, Slovakia, and Switzerland local adaptation and network integration requirements have been summarized and reported in a System Architecture and Development Plan. The main achievements within the multinational integration of national services include format standards, communication protocols and news publishing and distribution management processes, which are described in the following subsections.

3.1.1 Standardized News Exchange Formats

Although most significant information in the news area is stored in traditional text files, the information management in the news context has been modernized and adopted towards a unique set of contents within international cooperation work. The international Press and Telecommunication Council IPTC, has been developing news formats and standards to capture data and meta-information on news, following the specific needs and requirements of the multimedia news industry. Lately IPTC's activities have primarily focused on developing and publishing Industry Standards for the interchange of news data, namely NITF (News Industry Text Format, current Version 3.2), and NewsML.
NewsML can be applied at all stages in the (electronic) news lifecycle. It would be used in and between editorial systems, between news agencies and their customers, between publishers and news aggregators, and between news service providers and end users. Because it is intended for use in electronic production, delivery and archiving it does not include specific provision for traditional paper-based publishing, though formats intended for this purpose - such as the News Industry Text Format (NITF) can be accommodated. Multimedia content types such as image formats, audio- and video files are integrated with appropriate markup and description elements in the NewsML language.
An early achievement of the current project has been the selection of NewsML as the appropriate content format for the international integration following a thorough research and evaluation period. All participating news agencies have already adopted their local services and provide interfaces that support NewsML contents for the Nedine exchange protocols.

3.1.2 Local Service to Peer Communication Protocol

In order to integrate existing local services into a multinational network, communication protocols had to be defined and established, ideally following state-of-the-art models and research standards. We have chosen Web Services as the most appropriate and modern exchange mechanism, built on top of a Peer-to-Peer architecture to integrate the national services into a multinational network as shown in Figure 1.

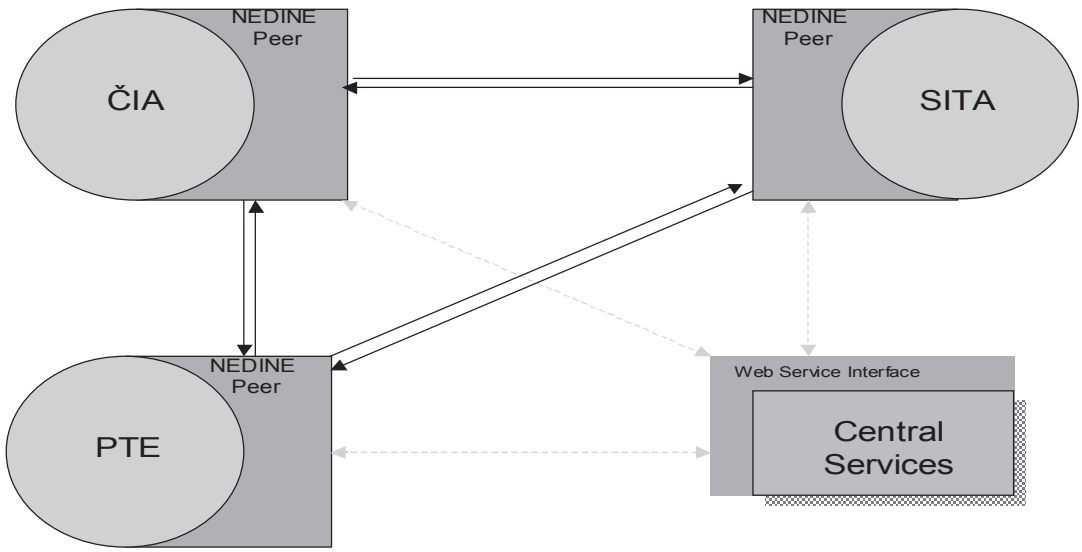


Figure 1: Nedine Peer-to-Peer Architecture

The existing news services are shown as circles in Figure 1, Nedine Peers are depicted as attachments to the local services. The decision of using a P2P network has been based on the following reasoning:

1. local service providers hold business critical information
   Specific data like customer addresses and subscriber databases are strictly limited to internal use of a local service and must be kept secret and isolated from network access. Thus a centralized storage would not be accepted by any of the current or future partners.
2. The installation of a local peer with well-known (open) source increases trust of the participating organizations and underlines the local character of the relevant business data.

The communication between the symmetric peer software and the existing services follows the Web Service definitions in the Nedine WSDL documentation. The communication includes the upload of local news data, the registration of local services to the network, the distribution initiation and parameters as well as the regular polling of news data to be distributed on behalf of the network partners. WSDL details are out of scope of this paper but have been documented within the project's deliverables.

3.1.3 Peer-to-Peer Communication Protocol

According to the P2P architecture outlined in Figure 1 the news distribution network provides business news presentation and distribution beyond national and language borders by transporting data between peers as network nodes. The multinational news distribution is restricted to P2P communication, so local services cannot access each other's services directly.
This restriction guarantees standardized communication and service scalability following normalized communication interfaces. Peers only communicate on events triggered by any local service. Such events include the distribution initiation of a news item, the presentation request of an item abroad or the registration or status requests issued by a local service.
P2P communication is again standardized by using Web Services. Additionally, peers negotiate and calculate semantic relations between multinationally distributed news items and forward the resulting ranking lists to each partner node involved in the distribution of the news item in focus.

3.1.4 Multinational News Presentation and Distribution Management

News presentation and distribution has been based on local news presentation and distribution features. This approach starts with the local strength of each participating partner and integrates the local powers to a forceful international service, reaching as many business news subscribers as possible in an international context.
Each partner benefits from the Nedine network since in addition to its local services he now is able to provide his news distribution customers with the highest international range of coverage using the P2P communication network. Local services request from the local peer the domain of all available business news categories, languages and subscriber countries and consecutively offer their customers news presentation and distribution services according to their choice. On selection of specific parameters the Nedine network initiated news distribution by utilizing each involved partners local distribution capabilities.

3.2 Multilingual News Publishing and Distribution

Beyond the technical details of news presentation and distribution, Nedine provides a multinational and multilingual news network throughout Europe. Identifying the special needs in the multicultural context of Europe and following the increased economical cooperation on the Old Continent since the establishment and the recent extension of the European Union, Nedine is targeting towards a business network capable of crossing both national and language borders.
Multilinguality is a heavily discussed topic in at least linguistic and technical research. Depending on the application domain, multilinguality can bring enormous challenges to technical and business oriented projects, including

- Automated translation
- Multilingual service presentation
- Semantic relations between multilingual presentations
- Multilingual information retrieval.

3.2.1 Multilingual Information Retrieval

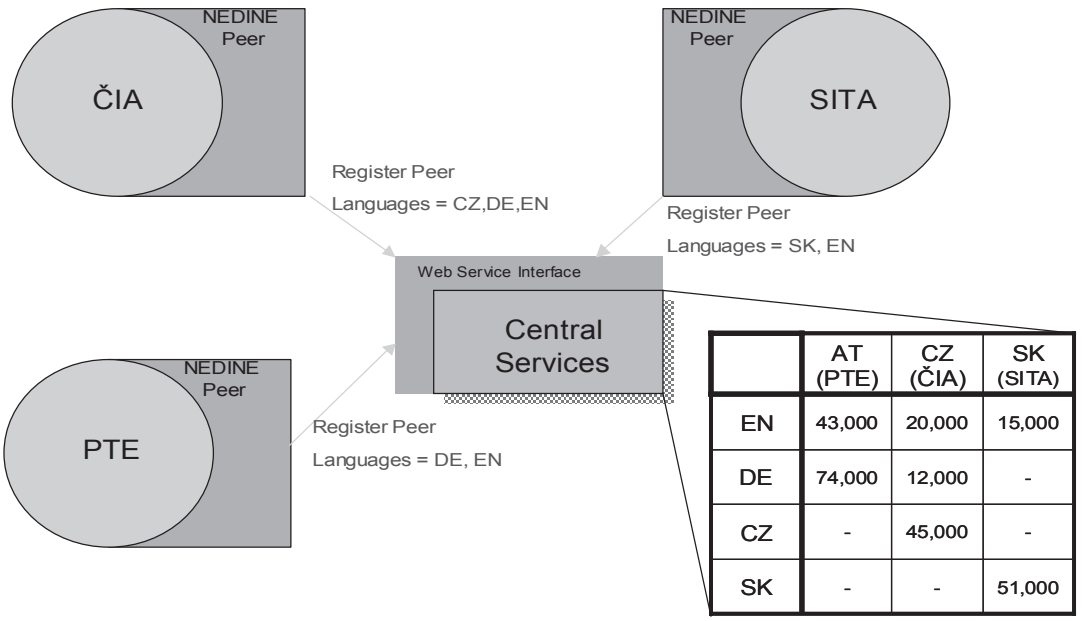
Since business news are most relevant for both news professionals like journalists and opinion leaders to establish and assess economic decisions, multinational and multilingual services are required to fulfil highest expectations on terms of recall and precision in the multilingual information retrieval area. Recent research has shown feasible approaches to assign news contents to multilingual thesauri such as EuroWordNet (EWN, Vossen 2001) and identify relevant related contents according to same-language synonyms or based on the Interlingual Index offered by tools like EWN.
Based on the business model of the news presentation and distribution network of Nedine, cross-lingual relations between single business news items are less relevant than the possibility to spread business news in more than one language to the clearly identified target group of subscribers. Multilingual news subscribers thus

may receive contents in different languages describing the same business facts, but both the number of multiply addressed readers is relatively small (currently below 8%) and the effect of reaching subscribers more than once is considered positively.

3.2.2 Automated Content Translation

One of the most relevant business scenarios considered in the Nedine project is a small enterprise intending to distribute a public relations news item in more than one language and to subscribers in more than one country to best present his business services to his potential international customers.

Multilinguality in the form of content translation thus is an essential feature of the Nedine network. Automatic translation has been researched and developed recently. Despite the increasing quality of such services (cf. Babel Fish at http://babelfish.altavista.digital.com/) business news items and PR contents heavily depend on top quality and contain most sensitive wording which are definitely beyond the quality levels available by automatic translation. Since Nedine is powered by the local strengths of the existing news agency services, the news presentation and distribution network offers content translation as a specific service. Technically, the business news items are distributed to the editorial service centers of the participating news agencies and professional journalists and editors translate the sensible business news into the optimal local versions for the targeted subscribers and markets. Nedine thus offers a multilingual service with bilateral interpretation and translation of news contents according to the language capabilities offered through the Peer systems as shown in Figure 2.



	AT (PTE)	CZ (ČIA)	SK (SITA)
EN	43,000	20,000	15,000
DE	74,000	12,000	-
CZ	-	45,000	-
SK	-	-	51,000

Figure 2: Language translation registration within the P2P network

A virtually centralized service in the Nedine network registers language translation capabilities of existing news agencies and third party translation services obeying the Nedine network communication protocol. Language translation services include both quality and quantity of services in the offered language, thus including the source and target language and the number of registered subscribers for the target language within the reach of the local provider.

3.2.3 Multilingual service presentation

Based on the business model of Nedine future providers may dominantly present their business news contents in a single or a few languages, usually strictly separated on language-specific news portals. Nevertheless, multilingual news presentations are technically feasible and are shown in the prototype described in section 4.

3.3 Semantic Content Coupling

Business news reach a specific audience: researchers that observe markets for specific branches or analyse modern trends as well as competitors and commercial organizations observing trends to optimise their business. Publishers hold the power to bias the market by filtering or targeting the business news towards their own needs. With sophisticated methods of the area of artificial intelligence, automatic content relation management has

been applied to the business news publishing and distribution service in order to provide a fair and informative enhancement to existing solutions. The most relevant semantically related business news applied to the current news item in focus are retrieved and provided via the electronic publishing service.
For this purpose we utilize a search mechanism common in modern information retrieval systems: The Vector Space Model (Salton 1983). This approach is mainly used for search engines such as Google, based on natural language. The underlying concept is quite simple. Any document is split up in keywords. Each of these keywords constitutes a dimension in a n-dimensional vector space. Therefore a document can be seen as a vector within this ”term space” (Wong 1985). The position of this vector to other vectors within the same vector space describes their similarity to each other. The mathematical method to evaluate how similar two documents are, is to calculate a cosine value for them and express the result as a percentage rating. This method produces very good results for natural language but it is not limited to this field alone (Wong 1985). Virtually any document collection can be mapped to a vector space to create an efficient search environment. In the case of business news publishing and distribution, specific weighting can be applied according to the data formats used within the information network. Important keywords like the business news category or the document title could be used to categorize news items according to the assumed behaviour.

4 News Network Prototypes

Technical developments at the TU Vienna and the R&D departments of the participating news agencies have created several prototypes, including P2P communication components, multinational and multilingual content aggregation servers, news presentation portals, and vector space model components.
The multilingual news presentation prototype www.nedine.org provides a proof of concept implementation of the content aggregation services and multilingual news presentation features developed in the project Nedine. Business news contents from Austria, Czech Republic, Germany, Slovakia, and Switzerland are fetched via XML Web Services in NewsML format from the respective local services and presented in a customisable news portal interface, providing news in the language of the web user’s choice.
Content relations according to vector space modelling have been calculated in Peer prototypes and news distribution protocol details are exchanged on prototype peer servers. According to currently evaluated test results and optimisations, the service is planned for going public at the end of 2005.

5 Conclusion

The main goal in our research is to create a intelligent business news publishing and distribution network, engaging modern Web Services, peer-to-peer networking and artificial intelligence for service optimisation. Electronic publishing techniques have been discussed and experimented with and the most feasible approaches are followed to prove the power of modern research results and to use them in the application of building a successful multilingual and multinational business news platform in Europe. Modern Peer-to-Peer architectures, sophisticated content translation service platforms and semantic content coupling are utilized to allow international and multilingual business integration, especially targeted at the small and medium enterprise market by utilizing efficient and effective network services.

Aknowledgements

This research work has been partially funded by the European Commission’s eContent Project Nedine (EDC-22225).

References

Akavipat F. M. R. and Wu L. (2004), Small world peer networks in distributed web search. In Proceedings of the 14th WWW conference, Budapest, Hungary, May 17-22 2004.

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

Salton M. J. M. G. (1983), Introduction to Modern Information Retrieval, volume 1. McGraw-Hill, Inc., 1983.

Wong, W. Z. and Wong P. (1985), Generalized vector space model in information retrieval. ACM, 1985.

World Wide Web Consortium (W3C 2001), Web Services Description Language (WSDL) 1.1, http://www.w3.org/TR/wsdl/, April 2005.

IPTC Internet description, Standard NewsML, http://www.iptc.org, status of April 2005.

Gudgin M. et. al. (2003), SOAP Version 1.2 Part 1: Messaging Framework, http://www.w3.org/TR/SOAP/, 24 April 2005

Schranz, M. et. al. (2000), 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.

Vossen, P., EuroWordNet, (2001), EU-funded project, LE-4 8328, http://www.illc.uva.nl/EuroWordNet/, status of April 2005.