# A Generic Framework for Visualizing the News Article Domain and its Application to Real-World Data

Elisabeth Lex, Christin Seifert, Wolfgang Kienreich and Michael Granitzer

Know-Center, Competence Centre for Knowledge-Based Applications and Systems
Inffeldgasse 21a
8010 Graz, Austria
{elex|ceifert|wkien|mgrani}@know-center.at

***ABSTRACT:*** *In this work we present APA Labs, a generic framework for visualizing the news article domain. APA Labs is a web-based platform enabling retrieval and analysis of news repositories provided by the Austrian Press Agency. APA Labs is designed as a rich internet application combined with a modular system of interactive visualizations. News articles are analyzed using domain specific named entity extraction methods combined with language specific heuristics. The proposed methods were subject to an evaluation procedure outlined in this contribution. This article illustrates the domain, the underlying concepts and implementation details. Several visualization modules are presented and an outlook on planned modules is given. Being online for around six months the community feedback as well as the easy integration of new modules shows the success of the underlying concept and the platform itself.*

***Keywords:*** H.3.3 [Information Search and Retrieval]: Search process, H.2.4 [Systems]: Textual Databases, [H.3.5] Online Information Services: Web-based services

## 1. Introduction

With the advent of the Internet and the evolution of Web 2.0 innovative scenarios for producing, providing and consuming information emerged. Before the rise of the Web, press and news agencies have been in charge for gathering and distributing news items to the wide public; thus granting access to content was the main task.
However, with the grown popularity oft the web the absolute monopoly of the news agencies disappeared and news items became increasingly available to the public. Naturally, this conflicts with the traditional business concept of news and press agencies. Due to the changing consumer needs resulting from ongoing developments in Web 2.0, solely providing content is no longer sufficient for attracting paying customers. News agencies have to extend their services to not only provide high quality content but also to develop novel intelligent services combined with high quality and noise-free content. Many recent advances in information retrieval and information visualization have been applied to news article repositories. Unfortunately, resulting applications have been accessible mostly to experts and closed communities. For example, systems like Galaxy of News [Rennison, 1994], Lighthouse [Leuski and Allen, 2002] and InfoSky [Granitzer et al., 2004] proposed novel visual metaphors to facilitate explorative analysis of large news article repositories. Also, the APA Online Manager [Kienreich, 2005] has provided a number of interactive visualizations to support analysis of search results obtained from news repositories. While incorporating new visualizations and access metaphors, the APA Online Manager is only available as a rich-client application to paying subscribers.
We present the generic framework APA Labs [APALabs], first introduced in [Kienreich et al., 2008], an experimental, web-based platform supporting retrieval and analysis of news articles obtained from the archives of the Austrian Press Agency [APA]. APA Labs utilizes many concepts usually summarized under the term Web 2.0 [OReilly, 2005] and invites users to participate in developing, testing and evaluating novel ways to access news agency repositories. Especially the perpetual beta paradigm and the involvement of users for judging usability of visualization are important steps towards a living laboratory for analyzing news data. The evolvement of the Web has shown that when developing innovative methods participation of users, usability and user acceptance are crucial.
In this work we refer to the version of APA Labs as of September 2008. Due to ongoing development the appearance of the application may change over time.

The remainder of this contribution is organized as follows: We briefly introduce the application domain in section 2 and then outline the concept underlying APA Labs in section 3. We describe implementation details in section 4 and report on the modules currently available in APA Labs in section 5. Section 6 gives an evaluation of the accuracy of our pre-processing techniques while a future outlook and a conclusion is provided in Section 7 and Section 8.

## 2. Domain

News articles are published by a wide variety of sources. News agencies collect, store and organize news articles and distribute them to paying subscribers. The work outlined in this contribution is based on services provided by the Austrian news agency APA. The news article archive of APA contains 100 Million articles gathered from over 200 sources. Approximately 10.000 articles are added every day. The majority of articles stored and distributed by APA are in German language.

News articles exhibit distinct structures. For example, the first paragraph of an article often forms a summary of the article as a whole. The first occurrence of a person in an article is usually composed of title and full name, while subsequent occurrences mention family name, title or a combination of both. Further, structures can be identified and leveraged for domain-dependent enhancement of retrieval and visualization results.

### 3. Concept

The emergence of Web 2.0 has been both beneficial and challenging for news agencies. The technological advances associated with Web 2.0 have enabled news agencies to provide services on a higher level of quality and to a wider audience. Rich internet applications have proven capable of replacing the specialized clients used to access large news repositories in the past.
However, the commercial concepts associated with Web 2.0 have been met with less enthusiasm. The traditional business model of news agencies offers services to registered subscribers paying a per-article charge. The content of an article is considered the primary commodity which generates value. This business model limits the application of Web 2.0 concepts like the Long Tail or Mash-ups because once the content of an article has become available in the public domain its value is greatly reduced.
The basic idea behind APA Labs is to give the general public access to novel retrieval and visualization services applied to news article repositories in the framework of a rich internet application. Visitors are invited to evaluate the services and to provide feedback. The number of news articles available for evaluation purposes has been limited to avoid conflicts with the business model outlined. However, the imposed limitations have been carefully balanced to retain added value for visitors. Within the range of available articles, no further restrictions are applied and the full article content can be accessed free of charge.
The expected benefits are manifold: APA Labs generates public awareness for the provided services and documents technological leadership. The rich internet application enables the Austrian Press Agency to field-test new services early in the development cycle, in accordance with the concept of the perpetual beta proposed by the Web 2.0 paradigm. Services which have been proven as useful by visitors can rapidly be integrated into the business model, enabling the Austrian Press Agency to adequately respond to new trends in today's highly volatile markets.

### 4. Implementation

APA Labs has been implemented in Java as a web application based on J2EE technology [J2EE]. A client-side rich internet application utilizes JavaScript and AJAX technology to communicate with an Apache Tomcat Web Server Version 5.5 [Tomcat] providing content through Java Servlets and Java Server Pages. The APA Labs server is built around a central request handler servlet which accepts and forwards requests issued by registered sessions. A session is registered on creation and assigned light-weight, volatile state data through session attributes. Heavy-weight, persistent state data is stored in a separate repository and referenced by session identifiers to reduce synchronization cost in clustered environments. Figure 1 illustrates the architecture of the APA Labs server and client elements. Functional components connected via well-defined interfaces are described in more detail in this section.

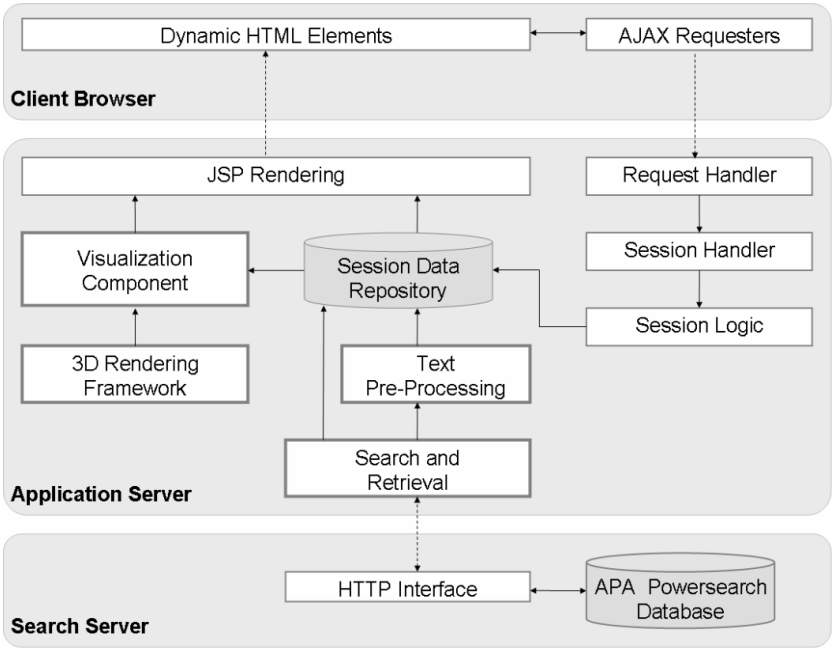


Figure 1: Architecture of the APA Labs Framework.

### 4.1 Functional Components

All components have been implemented following a singleton design pattern. Requests requiring access to such components are queued and processed in sequence.

#### 4.1.1 Search and Retrieval

The underlying repository of news articles is accessed using a HTTP-based query interface to a generic search engine, the APA PowerSearch engine [APAPowerSearch]. This search engine supports Boolean queries with a wide range of operators and returns a relevance-ranked list of news articles. For each news article the title, the medium and the publication date of the document is provided. Search results are stored in the session-specific repository of heavy-weight data. In order to avoid page-loading delays on the client-side result details like article content are loaded only when needed. For instance, displaying the search result list does not require the article content. Instead it will be loaded only on user demand.

#### 4.1.2 Pre-processing

Pre-processing works on the content of the retrieved news articles. First, relevant noun sequences are identified using stop word lists, stemming and language-specific heuristics. Second, an entity extraction procedure uses these noun sequences to extract Named Entities (NE) by applying statistics, heuristics and gazetteer lists. The entity types recognized by the framework currently include people, geographic locations, web addresses, date and time expressions and predefined topics of interest. Details on pre-processing and estimates for the quality achievable is given in section 6 in conjunction with an evaluation of the implemented search and retrieval component.

#### 4.1.3 Rendering Framework

The server-side generation of visualizations has been implemented using a rendering framework based on the Java bindings for OpenGL [Segal and Akeley, 2006] [JavaOpenGL]. Most modern server machines feature rudimentary graphical capabilities which are usually not employed in a web application. The rendering framework exploits these capabilities to generate complex visualizations at a rate of several frames per second while placing minimal load on the central processing unit. The computed visualizations are delivered to clients as a combination of compressed images and structured image maps designating interaction areas.

#### 4.1.4 Visualization

The visualization component is an abstract container enabling the definition of different user interfaces to visually analyze search results. A concrete implementation determines the types of extracted entities used and their visual representation. As depicted in Figure 1 the required data is loaded from the session data repository where the results of the search and retrieval component and the pre-processing component are stored. The visualization component uses the 3D rendering framework to generate the final visualization and delivers the result to the client. When creating new visualizations only this component needs to be implemented. Visualization modules currently available in APA Labs are described in more detail in section 5.

#### 4.1.5 Feedback Component

The feedback component provides functionality to collect and evaluate user generated feedback. User feedback is commonly used in Web 2.0 applications to gather the opinion of the community and enables the application providers to react on usage trends and user approval. The feedback component provided in APA Labs is added to each visualization module and enables the user to rate the visualization according to functionality, design and usability on a 5-point Likert scale [Likert, 1932]. Further, users can comment on the visualizations via email or a feedback form. A module manager at the Austrian Press Agency collects the user feedback which serves as a basis for further decision processes like improving the visualization or adding the originally experimental visualization to the product range of APA.

### 4.2 Interfaces

All components of APA Labs are connected via well-defined interfaces. There are two main interfaces available in APA Labs. The first interface links the APA Labs framework and the APA search engine. The search engine provides a Representational State Transfer Architecture (REST) [Fielding and Taylor, 2002] interface and is accessed via HTTP using a clear syntax.
The second interface connects the framework itself and all visualization modules. Modules following the interface definition can be easily integrated in the system. On the developer side new visualization modules can be implemented without knowing the underlying logic of the system. Also data exchange and session handling is accomplished by framework components.

## 5 Modules

The user interface of APA Labs provides conventional means to search for news articles, to navigate search result sets through relevance-ranked lists and to display article content. In addition, the user interface integrates a set of custom modules which feature consistent design and interactivity. The general layout of the platform is illustrated in Figure 2.

There are two different types of modules available: (i) modules operating on a set of documents and (ii) modules analyzing a single document. All modules share the ability to provide an alternative way to formulate a search query, to navigate a result

set or to analyze article content. The modules are implemented as classes within the application framework and, like described earlier, access result sets, extracted entities and rendering facilities through unified interfaces. This section describes the modules currently available in APA Labs.

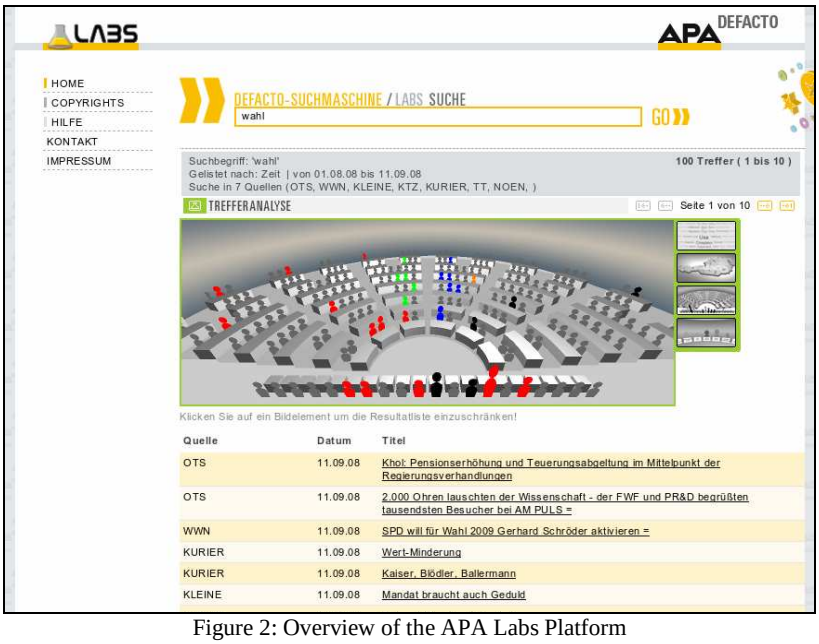

Figure 2: Overview of the APA Labs Platform

### 5.1 Geospatial Visualization

Generally, geospatial visualizations display information entities referencing geographical locations on appropriate maps [Scharl and Tochtermann, 2007]. This area of research has recently attracted much attention. Geospatial visualization is a natural extension for systems presenting news articles because most news articles reference one or more geographic locations. More than 85% of all articles available in the archives of the Austrian Press Agency contain at least one geographical reference. The geospatial visualization present in APA Labs extracts geographic locations from a set of documents resulting from a preceding search query.

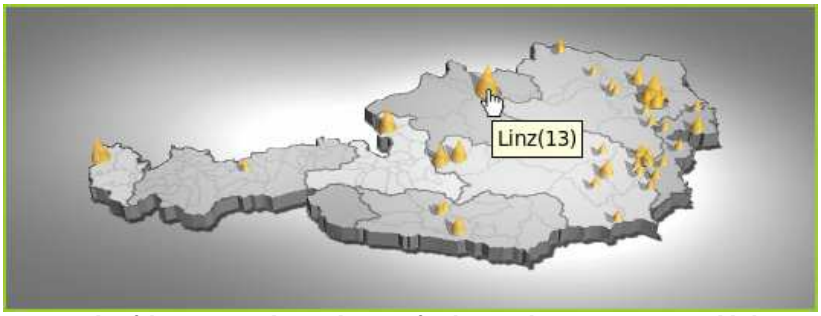

Figure 3: Example of the Geospatial Visualization for the search query "WM" (world championship)

Figure 3 displays the Geospatial Visualization of a search result set obtained in APA Labs. Cones have been positioned on a map of Austria to denote locations mentioned in one or more articles. The size of each cone encodes the number of occurrences identified for its location. The cones are rendered using a semi-transparent material to alleviate occlusion effects. Moving the mouse pointer over a cone displays the name of the location and the number of references identified for it in form of a tool tip window. Clicking on a location instantly filters the search result set to contain only articles referencing the selected location.
One benefit of the Geospatial Visualization is the ability to identify geographical hot spots for a particular topic at a glance. Another benefit is the ability to quickly refine the search results by region.

### 5.2 Tag Cloud Visualization

Tag clouds are text-based visual representations of a set of words (tags) usually depicting tag importance by font size. The popularity of this type of visualization has steadily grown due to recent trends in social and collaborative software. In contrast to many other types of visualizations, tag clouds do not use real-world models or metaphors. A tag cloud is a visual

abstraction and thus suitable for visualizing information entities of arbitrary types. This fact makes tag clouds uniquely suited for topical browsing [Kuo et al., 2007] and for browsing of news articles.

Tag clouds have become very popular in Web 2.0 applications, such as del.icio.us and flickr. Most state-of-the-art tag cloud algorithms layout tags inside rectangular boundaries. This constrains the general layout of web sites. In contrast, proposed non-rectangular layouts suffer from huge white-spaces between tags or tag overlap. Recently we presented an algorithm capable of dealing with polygonal boundaries and therefore allowing more flexible website designs [Seifert et al., 2008].

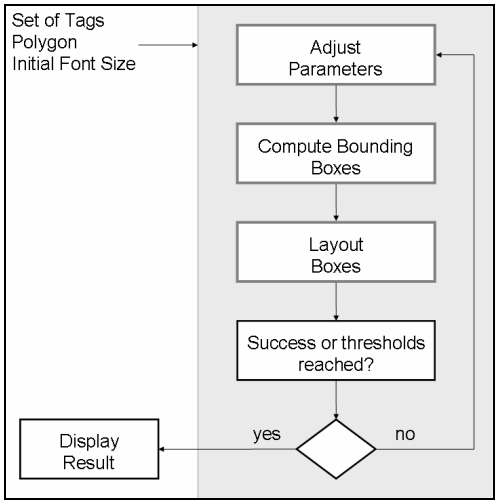

Figure 4: Process overview of the tag layout algorithm

Figure 4 shows an overview of the algorithm for tag layout. The algorithm takes a set of tags and a convex polygon as input. Each tag is assigned a relevance value. Besides initial values for the minimal font size, the maximal allowed and the minimal allowed font sizes are given. The resulting font size for a particular tag is calculated from its relevance value and the current font size interval. From the tag's font size and the string itself a bounding box for each tag is computed. The bounding boxes serve as input for the core tag layout algorithm. The core algorithm tries to place the bounding boxes circularly starting from the centre of mass of the bounding polygon. In order to account for westernized reading direction the algorithm uses a heuristic to preferably place tags to the left and the right of existing tags. The box layout has to start again if not all bounding boxes fit inside the polygon. In this case a parameter adjustment takes place and the box layout starts again. The whole layout trial stops if either the thresholds for string truncation and font size are reached or all tags were successfully laid out. For details on the algorithm, a technical evaluation and a user study refer to [Seifert et al., 2008].

Figure 5 shows two example tag clouds inscribed in a regular polygon and a circle, which in the discrete 2D space can be approximated by a polygon.

Figure 5: Examples tag layouts in arbitrary polygons. Left: 30 tags in a regular octagon. Right: 10 tags in a circle

Figure 6 displays the tag cloud visualization of a search result set obtained in APA Labs. Despite the algorithm can lay out tags in arbitrary convex polygons, for the APA Labs visualization a rectangular border is used to guarantee a common visual appearance of all visualizations within APA Labs.

In the context of APA Labs tags have been derived from extracted entities. The entity extraction procedure is applied to a set of documents. Each tag denotes either the name of a person, the label of a geographical location or a general term. Font size and color are measures for tag weight: Highly relevant tags are rendered using a larger font and have less transparency applied. Tag weights are determined based on the number of occurrences in the result set.

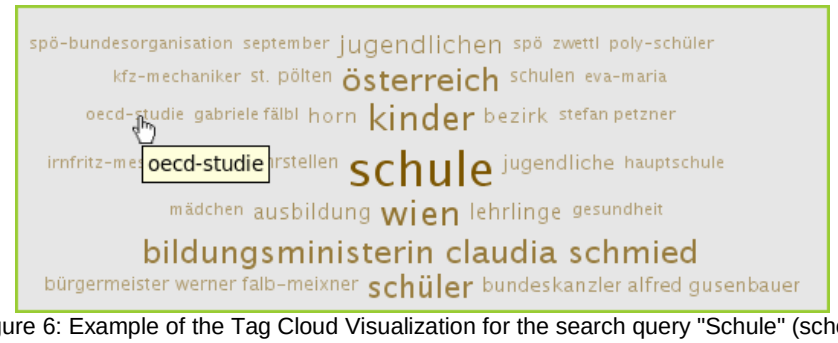

Figure 6: Example of the Tag Cloud Visualization for the search query "Schule" (school)

Due to the fact that the bounding boxes of the tags are not visible clicking a small tag may be difficult because the user can not immediately grasp whether the mouse pointer is exactly above the particular tag. Therefore the tool tip provides a quick feedback which tag is currently selected. Clicking on a tag instantly filters the search result set to contain only articles referencing the selected tag.
The most prominent benefit offered by the tag cloud visualization is the ability to identify the major subtopics present in a result set at a glance. Besides the user can immediately skim through the most important words in the retrieved news articles and therefore gets a general idea about the contents provided within the resulting document set.

### 5.3 Parliament Visualization

One of the most common forms of media observation carried out by experts is the analysis of the impact statements made by public figures in news articles. This type of observation is also referred to as media diffusion analysis. For the general public, the most interesting public figures are probably the voted representatives and leading politicians of a country. The Parliament Visualization module integrated in APA Labs enables users to instantly determine which members of the Austrian government and parliament have been mentioned in the context of a search result set.

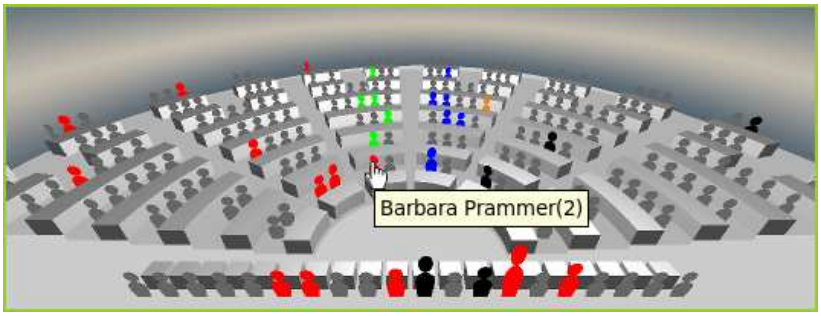

Figure 7: Example of the Parliament Visualization for the search query "Wahl" (election)

Figure 7 displays the Parliament Visualization of a search result set obtained in APA Labs. The basis of the visualization is formed by a three-dimensional, stylized model of the Austrian parliament. The curved rows of seats visible in the background of the image belong to the Members of Parliament. In the foreground, a straight row of seats is reserved for the ministers forming the government. A three-dimensional icon in the shape of a stylized human is attached to each set and oriented to face the observer's point of view. The icon represents the person holding the seat in parliament. For each seat, the name and the party association of the person holding it is known.
Initially, all icons are colored grey and have the same size. For each person mentioned at least once in the current result set, the according icon is colored in the colors of that person's political party. Icons are scaled along the vertical axis based on the number of references present in the current result set.
Moving the mouse pointer over an icon displays the name of the Member of Parliament or minister and the number of references identified for him or her in the form of a tool tip window. Clicking on an icon instantly filters the search result set to contain only articles referencing the selected person.
The major benefit provided by the Parliament Visualization is the ability to instantly identify which politicians are associated with a specified topic.

### 5.4 Round Table Visualization

The Round Table Visualization shows the current seven top candidates for the Austrian parliament elections taking place in autumn 2008. This visualization allows users to identify who of the top candidates is associated with a specific subject in the media or is mentioned in context of a certain issue discussed in the news. For instance, users can immediately learn what opinion a particular politician states when e.g. the reformation of the Austrian health system is discussed in public.

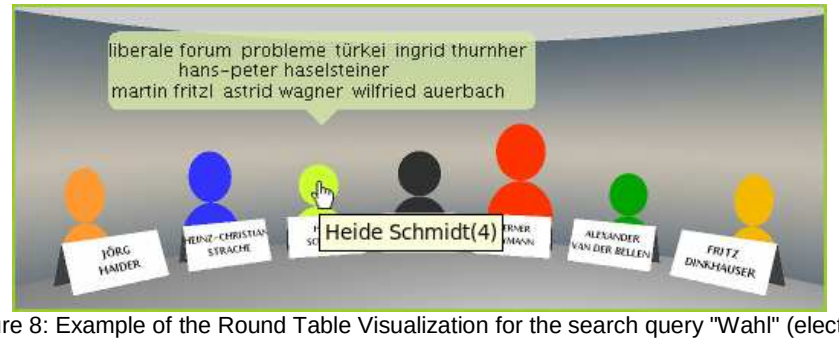


Figure 8: Example of the Round Table Visualization for the search query "Wahl" (election)

Figure 8 gives an example of the Round Table Visualization. The three-dimensional visualization provides a semicircular arrangement of seven figures. The figures correspond to the top candidates of the parties which are eligible for election. The figures are colored according to the particular party affiliation. In front of each figure a label providing the name of the politician is placed. The size of the figures corresponds to the number of occurrences of the particular person in the search result set. Politicians occurring more often in relevant thematic context to the original search query are displayed larger than others. The names of the politicians and the exact number of hits are also available as tool tip. In addition to that a speech bubble represented by a small tag cloud is provided over each figure. The speech bubble contains the most important names, locations and terms extracted from the set of documents related to the particular politician. The document set related to a particular person can be obtained by clicking the figure corresponding to the person. This results in a refinement of the original search query.
The Round Table Visualization provides a profound overview of the current political discussion taking place in the present election campaign in Austria. Because the Round Table Visualization also shows the most important keywords each top candidate states to an individual topic it supports users in finding out which issues are covered by each politician. The persons shown in the visualization can, of course, be changed to reflect the actual political situation in Austria.

### 5.5 Brockhaus Look-Up

Mashups are one of the core concepts commonly associated with the term Web 2.0. A mashup combines data and services to create new functionality. This agile concept has proven to be beneficial for both, non-commercial projects and business-to-business applications. APA Labs features a Brockhaus Look-Up module which integrates services provided by the Brockhaus Encyclopaedia [Brockhaus] with services provided by the APA to present users with a lexical context for articles or selected text ranges.


Figure 9: Example of the Brockhaus Look-Up, displaying content of a selected article after search for "Mehrwertssteuer" (engl. VAT)

In Figure 9 the full text of a news article is shown as text to the left and the Brockhaus Look-Up is shown in the sidebar to the right. The Brockhaus Look-Up identifies relevant terms within a text using the entity extraction methods described in section 4.1.2. It then looks up these terms in the Brockhaus Encyclopedia and returns a relevance-ranked list of found encyclopedia articles which constitute a topical context of the analyzed text.
The Brockhaus Encyclopedia is considered the prevalent multimedia encyclopedia in the German-speaking domain. It contains approximately 240.000 articles and 350.000 keywords and thus ranks among the world's largest encyclopedias.
The generation of lexical context is started automatically for the full text of an article whenever article content is being displayed. Clicking the button "from selection" in the sidebar starts generation of lexical context for the currently selected text range. Found encyclopedia articles are displayed in the sidebar with title and text preview.
The major benefit of the Brockhaus module is the availability of a lexical context explaining unfamiliar domains to users. Another benefit is that users can directly look up unknown terms in the encyclopedia.

**6. Evaluation of Pre-processing Techniques**

All proposed modules strongly depend on the quality of the pre-processing components. Especially the Brockhaus Look-up module relies on the capability of the components identifying relevant terms and their context. Therefore, in the following section the accuracy of the pre-processing methods is estimated. Hereby the focus lies on the identification of key terms and named entities as well as on resolving context information; essential criteria for automatically linking encyclopedic resources to news articles.
Because no specialized test dataset was available for our domain and creating such a dataset is time consuming and costly, our evaluation is performed on a standard test dataset.
Since the online encyclopedia Wikipedia exhibits similar structures as the Brockhaus encyclopedia, a Wikipedia-based standard dataset, the INEX 2007 Link-the-Wiki dataset [Huang et al., 2007], was used for evaluation. Wikipedia pages contain links manually annotated by users. These links can be regarded as important key phrases as well as links pointing to relevant encyclopedic resources. These aspects lead to the conclusion that the Wikipedia serves as a suitable test dataset for the intended evaluation.
The INEX 2007 Link-the-Wiki dataset consists of 659,413 Wikipedia pages taken from the Wikipedia XML Corpus described in [Denoyer and Gallinari, 2005]. Like in the 2007 INEX track, 90 topics from the dataset are used as test set whereas one topic corresponds to a Wikipedia page. The challenge is to identify the 8,392 annotated links available in the test set using the pre-processing methods. Each test page contains 94.29 links on average, with a minimum of 6 and a maximum of 521 links. In total, the test collection links to 5590 unique Wikipedia pages.
In general, the pre-processing is based on standard tasks like tokenization, sentence boundary detection and part-of-speech tagging enabling the elimination of non-nouns. In order to identify named entities finite state machines for gazetteer matching are used - similarly to those used in GATE [Cunningham et. al., 2002] - as well as simple grammatical rules. For instance, person detection exploits lists of known names (e.g. names of politicians) as well as forename lists combined with grammars based on noun phrases for detecting possible surnames.
Encyclopedic resources exhibit structures appropriate for being used as gazetteer lists because nearly every page in either the Brockhaus or the Wikipedia has a precise title describing the actual topic, e.g. the name of a person or a location. However, such a topic matching process is purely syntactical in nature and requires disambiguation strategies as post processing steps.
After detecting all possible entities using the outlined matching approach, the words in the neighborhood denote the entity's context. The number of words to consider depends on the segmentation level, which is defined as either the whole document, an automatically detected part of a document, the so-called topic [Choi, 2000], or a single sentence. Considering the segmentation level all identified nouns are used as input for a Boolean OR query performed on the indexed training data sets. The Java based Open Source Search Engine Lucene [Lucene] has been used as an underlying search backend. If the entity of the context is contained in the result set, the score for this entity is used as confidence in the correctness of the annotation.
For each document the pre-processing procedure returns a ranked list of entities and their position within the document. These results can be evaluated using standard Information Retrieval measures.

| Document Level | Segmentation | 11-pt. AP | R-prec | Micro Precision | Micro Recall |
|---|---|---|---|---|---|
| Topic | | 0.2833 | 0.3681 | 0.2215 | 0.4317 |
| Sentence | | 0.2842 | 0.3679 | 0.2244 | 0.4318 |
| Document | | 0.2854 | 0.3539 | 0.1449 | 0.5810 |

Table 1: Evaluation of disambiguation quality of different segmentation levels

In Table 1 the result for different segmentation levels in terms of 11-point average precision (11-pt. AP), R-precision as well as micro precision and recall are given. While precision is around 20 percent, half of the entities can be correctly identified.
However, the precision achieved is determined by the values of R-prec. R-prec is defined as the precision at R where R denotes the number of relevant documents in the test set .In this case 1/3 of the links can be considered as relevant.
Employing this evaluation strategy it has to be considered that Wikipedia pages are annotated by humans, following not only logical rules but also aesthetic criteria. As outlined in [Wu and Weld, 2007] entities are most often annotated only on their first occurrence, therefore it is assumed that accuracy can be increased significantly.

However, user feedback indicates that the quality of the proposed pre-processing methods is sufficient for the APA Labs community's needs. Also, evaluations done by domain experts revealed an acceptable quality. Nevertheless, further improvement will focus on the use of machine learning techniques and supporting a broader range of entities as well as a higher semantic richness by utilizing ontologies.

## 7. Future Work

The design of APA Labs encourages a continuous process of gathering evaluation results and developing and integrating novel modules. User feedback for the modules outlined in this paper will be collected and analyzed in detail. The Austrian Press Agency will decide which modules are candidates for being integrated into its commercial services based on the collected user feedback. A preliminary analysis of the feedback gathered so far indicates that the Parliament Visualization and the Geospatial Visualization are likely candidates for such a step.
New modules will be developed based on input from the community and on research findings in the field of information and knowledge visualization. For instance, a new visualization module is planned to illustrate the different branches of Austrian industries. In this visualization, Austrian industry companies will be extracted from a collection of APA news articles and will be displayed on a map of Austria. Each individual industry branch will be represented by a characteristic icon. If e.g. the name of a factory is mentioned in conjunction with a search query a small icon in shape of a factory will be placed on the correct geographic location on the map of Austria. The aim of this visualization is to provide an instant overview of the Austrian industrial development.
We also plan to extend our applications to trend visualizations [Piche, 1995] where a document set is analyzed over time. These visualizations are very promising to identify trends in the fields of tourism, industry or the financial world. Investigation of news article collections using trend visualizations could lead to a better identification of trends emerging within a specific time period.

## 8. Conclusions

This contribution presented the generic framework APA Labs. The experimental web-based platform provides novel methods to access the repository of the news archive of the Austrian Press Agency APA. Traditional technologies and business models usually favored by news agencies are modified and augmented by several Web 2.0 concepts. APA Labs is designed as a combination of a rich internet application with a modular system of interactive visualizations using server-side entity extraction and three-dimensional rendering capability. The labs platform enables APA to test the acceptance of new services and get early user feedback in the development cycle. Due to its online availability APA Labs generates public awareness to the products of the Austrian Press Agency and hopefully initiates innovative developments beneficial for both company and customers.

## 9. Acknowledgement

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

## References

[APA] Austrian Press Agency Homepage, http://www.apa.at, last accessed Sept 2008.

[APALabs] APA Labs Homepage, http://www.apa.at/labs, last accessed Sept 2008.

[APAPowerSearch] APA-IT Informations Technologie GmbH Powersearch Database Homepage, http://www.apa-it.at/cms/it/DE/loesungen.html?channel=CH0265&doc=CMS1147877187557, last accessed Sept 2008.

[Brockhaus] Bibliographisches Institut & F. A. Brockhaus AG, Brockhaus Enzyklopädie, Germany, http://www.brockhaus.de, last accessed Sept 2008.

[Choi, 2000] Choi, F.Y. Y. (2000). Advances in domain-independent linear text segmentation, In: *Proceedings of the First Conference of the Association for Computational Linguistics.* ACM. Seattle, Washington, USA, pp 26-33.

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002), GATE: A framework and graphical development environment for robust NLP tools and applications, In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL2002),* Philadelphia, USA.

[Denoyer and Gallinari, 2005] Denoyer, L. and Gallinari, P. (2005). The Wikipedia XML Corpus. *SIGIR Forum, pp. 64-69*.

[Fielding and Taylor, 2002] Fielding, R. T. and Taylor, R. N. (2002). Principled design of the modern Web architecture. *ACM Transactions of Internet Technology* 2, 2, pp. 115-150.

[Granitzer et al., 2004] Granitzer, M., Kienreich, W., Sabol, V, Andrews, K. and Klieber, W. (2004). Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories, In: *Proceedings of the IEEE Symposium on Information Visualization 2004 (InfoVis2004),* IEEE, Austin, Texas, USA.

[Huang et al., 2007] Huang, D.W.C., Xu Y., Trotman, A. and Geva, S. (2007). Overview of INEX 2007 Link the Wiki Track, In: *Proceedings of the Initiative for the Evaluation of XML Retrieval (INEX 2007*).

[J2EE] Java™ 2 Platform Enterprise Edition Specification, v1.4, http://java.sun.com/j2ee/j2ee-1_4-fr-spec.pdf, last accessed Sept 2008.

[JavaOpenGl] Java Specification Request (JSR) 231: Java Bindings for the OpenGL API, http://jcp.org/en/jsr/detail?id=231, last accessed Sept 2008.

[Kienreich, 2005] Kienreich, W. (2005). A Visual Query Interface for a Very Large Newspaper Article Repository, In: *Proceedings of the Sixth International Workshop on Theory and Applications of Knowledge Management (TAKMA2005),* IEEE, Copenhagen, Denmark.

[Kienreich et al., 2008] Kienreich, W., Lex, E. and Seifert, C.(2008). APA Labs: An Experimental Web-Based Platform for the Retrieval and Analysis of News Articles, In: *Proceedings of the first International Conference on the Applications and Digital Information and Web Technologies (ICADIWT08)*, Ostrava, Czech Republic, pp. 58-62.

[Kuo et al., 2007] Kuo, B.Y., Hentrich T., Good, B.M. and Wilkinson M.D. (2007). Tag clouds for summarizing web search results, In: *Proceedings of the 16th international Conference on World Wide Web (WWW2007)*, Banff, Alberta, Canada.

[Leuski and Allen, 2002] Leuski, A. and Allan, J. (2002). Lighthouse: Showing the way to relevant information, In: *Proceedings of the IEEE Information Visualization (IV2002),* IEEE, Boston, Massachusetts, USA.

[Likert, 1932] Likert, R. (1932). A Technique for the Measurement of Attitudes, *Archives of Psychology* 140: pp. 1-55.

[Lucene] Apache Lucene Search Engine Library. http://lucene.apache.org/, last accessed Sept 2008.

[Rennison, 1994] Rennison, E. (1994). Galaxy of news: An approach to visualizing and understanding expansive news landscapes, In: *Proceedings of Seventeenth Annual ACM Symposium on User Interface Software and Technology* (*UIST1994)*, ACM, Marina del Rey, California, USA.

[OReilly, 2005] O'Reilly, T. (2005). What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software, http://www.oreillynet.com, last accessed Sept 2008.

[Piche, 1995] Piche, S.W. (1995). Trend visualization, In: *Computational Intelligence for Financial Engineering, Proceedings of the IEEE/IAFE 1995*, pp. 146-150.

[Scharl and Tochtermann, 2007] Scharl, A. and Tochtermann, K. (Eds.). The Geospatial Web, *Advanced Information and Knowledge Processing Series*, Springer, London, 2007.

[Segal and Akeley, 2006] Segal, M. and Akeley, K. (2006). The OpenGL Graphics System: A Specification, (Version 2.1), http://www.opengl.org/documentation/specs/version2.1/glspec21.pdf, Silicon Graphics, Inc.

[Seifert et al., 2008] Seifert, C., Kump, B., Kienreich W., Granitzer, G. and Granitzer, M. (2008). On the beauty and usability of tag clouds, In: *Proceedings of the 12th International Conference Information Visualisation (IV2008)*, IEEE, London, UK, pp. ref17-25.

[Tomcat] The Apache Software Foundation. Apache Tomcat, http://tomcat.apache.org, last accessed Sept 2008.

[Wu and Weld, 2007] Wu, F. and Weld, D. S. (2007), Autonomously semantifying wikipedia, In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM2007)*, ACM, New York, NY, USA, pp. 41—50.