Scalable Annotation Mechanisms for Digital Content Aggregation and Context-Aware Authoring

Arno Scharl, Alexander Hubmann-Haidvogel
MODUL University Vienna, New Media Department
Am Kahlenberg 1,
1190 Vienna, Austria
+43 320 3555 301
{arno.scharl, alexander.hubmann}@modul.ac.at

Gerhard Wohlgenannt, Albert Weichselbraun
Vienna Univ. of Economics and Business, Research Institute for Computational Methods
Augasse 2, 1090 Vienna, Austria
+43 1 31336 5228
{gerhard.wohlgenannt, albert.weichselbraun}@wu.ac.at

Astrid Dickinger
MODUL University Vienna, Department of Tourism and Hospitality Management
Am Kahlenberg 1,
1190 Vienna, Austria
+43 320 3555 412
astrid.dickinger@modul.ac.at

ABSTRACT
This paper discusses the role of context information in building the next generation of human-centered information systems, and classifies the various aspects of contextualization with a special emphasis on the production and consumption of digital content. The real-time annotation of resources is a crucial element when moving from content aggregators (which process third-party digital content) to context-aware visual authoring environments (which allow users to create and edit their own documents). We present a publicly available prototype of such an environment, which required a major redesign of an existing Web intelligence and media monitoring framework to provide real-time data services and synchronize the text editor with the frontend’s visual components. The paper concludes with a summary of achieved results and an outlook on possible future research avenues including multi-user support and the visualization of document evolution.

Keywords
Content production and consumption, context-awareness, classification, real-time annotation, collaborative authoring.

INTRODUCTION
The next generation of human-centered information systems will transcend individual disciplines such as adaptive systems, natural language processing, and human-computer interaction. The challenge is to gather and organize multimodal data from disparate sources, find the best mix to communicate a message or experience, customize content to maximize perceived quality, advance human knowledge, create shared meaning and facilitate collaboration in virtual communities [11; 16]. The embedding of fixed and portable communication devices into our local physical spaces will further increase this challenge and amplify observable trends towards social computing and adaptive real-time services. Studies clearly show the potential of such services, but the extent to which user behavior is influenced by contextual variables has not been sufficiently investigated. With mobile devices becoming multi-purpose entertainment, transaction and communication tools [6; 17], a deeper understanding of context is imperative – including its geospatial, temporal, semantic and social aspects (the latter category includes both the users’ identities as well as their social networks).

Context has a significant impact on the way humans act and on how they interpret things. The term does not simply refer to a profile, but to an active process dealing with the way humans weave their experience [3]. Information systems research conceptualized context as a hybrid construct determined by the specifics of the computing environment, user environment, and physical environment [4; 5]. Based on an extensive review of the literature, Hong et al. [9] present an abstract reference architecture for context-aware systems comprising four layers: (i) network, (ii) middleware, (iii) application, and (iv) interface. Disregarding the network layer and placing special emphasis upon the application and interface layer, this paper distinguishes two main types of context:

- **User and Application Context.** This includes situational aspects of a person in a particular usage situation – i.e., physical surroundings, social surroundings, temporal perspective, task definition, and antecedent states [2]. All of those influence the person in the stage of content production as well as content consumption, which are both addressed by this paper.
- **Context of Digital Content.** The context information related to authors and readers needs to be distinguished from the context of the virtual environment itself and its embedded information objects (= digital content).

Advanced Web intelligence solutions automatically enrich information objects with context information along multiple dimensions to improve the accuracy and relevance of information exploration and retrieval services. This approach goes beyond automated metadata extraction, which focuses on the thematic aspect and ignores the specific conditions of authoring and using documents.
The importance of time and location [21] as contextual variables has been acknowledged – but to truly understand users and harness their collective intelligence, further dimensions of context such as tasks, antecedent states and the structural characteristics of their social networks [7] need to be incorporated into a comprehensive framework of real-time contextualization. Such a framework is an important step in extending the Web with multidimensional annotation, querying and reasoning capabilities – not only at the infrastructure and middleware level [20], but also at the application level when developing adaptive user interfaces for collaborative authoring environments.

INTEGRATING CONTENT PRODUCTION AND CONTENT CONSUMPTION PROCESSES

Reliable information on user context not only allows the provision of adaptive services, but also enhances the automated annotation of resources. The webLyzard suite of Web mining tools [28] provides a comprehensive framework for the real-time contextualization of digital content from heterogeneous sources.¹ This contextualization process requires automated methods to generate, manage and apply context information in dynamic Web and social media environments. In such environments, users serve as both producer and consumer of information, often switching between these roles within a single session. Technological advances support such user-generated content, and encourage ad-hoc collaboration and joint authorship of continuously evolving information resources.

The webLyzard framework helps to investigate the impact of these advances on content production, distribution and consumption. webLyzard not only extracts Microformats, RDFa and social data from Web resources, but also harvest social ‘lifstreams’ of services like Twitter and Facebook, which provide aggregated information about a user’s online activities [8]. The goal is not to create another lifestream aggregator, but to act upon the already syndicated and machine-readable data of multiple sources, build a repository of documents and micro-content assets, and develop appropriate mediation strategies to identify redundancies and solve conflicts.

Evaluations of information and communication technology have typically focused on system performance, usability and acceptance to predict consumer trends. The users’ contextual variables as triggers of consumption have received less attention. With the convergence of devices, the choice of the medium will increasingly match individual usage needs. Targeted information provision is possible and necessary, but traditional approaches require users to explicitly specify their informational needs via adaptive choice prompts, which tend to be time-consuming and cumbersome. The webLyzard platform, by contrast, seamlessly integrates the usage context without the need for lengthy dialogs. The results of geotagging published articles [16] or tapping into third-party services such as foursquare [26], for example, can be used as a proxy for information requirements [13] in specific domains like travel and tourism [18], political campaigns [19], and climate change communication [10].

¹ Under development since the late 1990s, the webLyzard platform was significantly extended through two FIT-IT Semantic Systems Projects: Information Diffusion across Interactive Online Media (www.idiom.at), and Relation Analysis and Visualization for Evolving Networks (www.modul.ac.at/nmt/raven).
This requires strategies for efficiently aggregating and mediating context information from multiple sources, and providing contextualized content repositories. Interoperability issues involved in combining multiple context dimensions into a holistic view pose a serious hindrance for understanding and applying context. Therefore, many projects do not consider the contextual attributes of content production and consumption summarized in Table 1.

**COLLABORATIVE AUTHORING**

Context-aware applications pave the way for a close coupling of content production and consumption, which can significantly improve the customization and annotation of digital content – e.g., by using associative retrieval techniques to serve knowledge items that optimally match a given context, and annotate those resources in real-time.

The collaborative authoring environment presented in this paper is based on the Media Watch on Climate Change [10; 27], a public Web portal that provides a comprehensive and continuously updated account of online media coverage on climate change and related issues. The portal aggregates, filters and visualizes environmental content from the Web sites of various stakeholders: 150 Anglo-American news media sites, Web logs (blogs), environmental non-profit organizations, and the Fortune 1000 companies – the largest U.S. companies in terms of revenues [25].

Moving from a content aggregator to a real-time authoring environment entailed a complete redesign of many of the underlying services, since real-time context extraction and adaptation require a highly scalable approach.

**Scalability and Real-Time Data Services**

The scalability of the annotation and document enrichment processing pipeline had to be radically improved. While a content aggregator usually works in batch mode and faces hourly updates at most, users of real-time editing environments expect rapid response times and no noticeable delays during the authoring process. In order to extend the Media Watch on Climate Change [10; 27] into a Climate Change Collaboratory [24], the throughput of four services required significant methodological changes:

![Figure 1. Screenshot of the Media Watch on Climate Change (www.ecoressearch.net/climate)](image)

Technical Session on Viewing, interacting, annotating
the geotagging component to identify geographic references in segments of plain text, rank these references according to their relevance within the text segment, match the most relevant references to unique locations, and annotate the text segment with the geographic coordinates (longitude, latitude) of these locations;

(ii) the computation of semantic similarity to determine a document’s precise position in the information landscape (the visualization shown in the upper right corner of Figure 1). Information landscapes are visual representations based on a geographic map metaphor, used for analyzing relationships in large, high-dimensional data sets (e.g., document archives). The height of a hill serves as an indicator for the amount of documents belonging to the topic, while its compactness is an indicator of topical cohesion. [14; 15].

(iii) the identification of keywords to label document clusters in the information landscape, and to explain observable fluctuations and peaks in the trend charts when moving the mouse pointer over a particular data point. The trend chart in the upper left corner of Figure 1 exemplifies this functionality, showing the top five keywords as of 23 May 2011, their frequency in April and May 2011, and the specific topics that Anglo-American news media (= the chosen source) associated with the term “direct action” on 28 March 2011.

(iv) the serving and caching of image tiles for the various visualizations. More aggressive caching had an immediate positive impact on the system, allowing users to hover above the information landscape and geographic map to preview documents in real time (a subsequent click then activates the previewed document).

Implementation of the Prototype Editor

Once the required scalability and throughput were achieved, the development efforts focused on the actual document editing environment. The initial prototype stores the edited document as a simple HTTP Cookie [1], allowing users to continuously edit a document while performing other activities within the portal.

The required annotations to determine the geographic and semantic context are now being computed in less than 100 milliseconds, ensuring a seamless user experience and making sure that the most relevant content (e.g., news media publications and scientific articles on climate change adaptation) is immediately identified and displayed across all multiple coordinated views:

- The target geography of the edited document is determined by a geotagging service, which holds a gazetteer database in memory to speed up the identification of referenced geographic references. A geographic IP lookup database is used to identify the source geography of the active user on a country level.
- The (temporary) position of the document in the information landscape is calculated by interpolating the positions of the most similar documents using document similarities calculated on a Lucene [22] full-text index. Both lists of geographically (‘Nearby Locations’) and semantically (‘Similar Topics’) similar documents are updated in real time as well (see Figure 1).

The final public release of the Climate Change Collaboratory will provide (i) concurrent editing by multiple users, (ii) advanced formatting options, (iii) embedded images and other multimedia content, (iv) temporal controls to track incremental changes, and (iv) interactive visualizations to show a document’s evolution.

CONCLUSION AND OUTLOOK

Situational variables describe collaborative editing environments as momentary interactions with specific content items in the time-space continuum. To support the editing process and better understand user decision making, it is therefore essential to not only consider basic service characteristics, but also the surroundings and circumstances of content production. Authors and users alike feel comfortable in situations with static or gradually changing context [11]. Yet research on information and communication technology usage patterns has often neglected the importance of context for service provision, focusing instead on acceptance, usability, ease of use, and formal system tests.

We will fill this gap by conducting large-scale empirical studies among the users of the Climate Change Collaboratory to improve the quality of user experiences. The set of methods will include multivariate analysis of questionnaire data (experimental and non-experimental settings), conjoint analysis to test for contextual variables, and structural equation modeling to capture causal effects. Building upon the lessons learnt from these studies, the development of specific solutions for different domains (politics, tourism, financial markets) will underscore the generic and flexible nature of this platform.

Even before the advent of user-generated content, users produced implicit feedback in the form of clickstreams. Web 2.0 technology further blurs the boundary between producers and consumers of information. The webLyzard platform processes social lifestreams to build personal contextualized repositories. Analyzing these repositories in real time, enriched by the user’s social graph, supports the customized just-in-time provision of information services. In conjunction with standards such as the Atom Syndication Format [12] and the Attention Profile Markup Language [23], the webLyzard contextualization backend and the context-aware editor outlined in this paper will provide a sound platform for the creation of adaptive real-time applications.

ACKNOWLEDGMENTS

The collaborative authoring environment presented in this paper is being developed as part of the Climate Change Collaboratory (www.modul.ac.at/nmt/triple-e), a research project funded by the Austrian Climate and Energy Fund within the program line ACRP (Austrian Climate Research Program; www.klimafonds.gv.at), and the DIVINE Project (www.weblyzard.com/divine), which is supported by the Austrian Research Promotion Agency within the strategic objective FIT-IT Semantic Systems (www.ffg.at/fit-it).
REFERENCES

ONLINE RESOURCES
23. Attention Profile Markup Language (APML) http://www.apml.org/
24. Climate Change Collaboratory http://www.modul.ac.at/nmt/triple-c
25. Fortune Magazine http://www.fortune.com/
27. Media Watch on Climate Change http://www.ecoresearch.net/climate/