

Supporting the Collaborative Editing of Documents with Real-Time Content Recommendations

Arno Scharl, Carmina Reyes, Alexander Hubmann-Haidvogel
MODUL University Vienna, Department of New Media Technology
Am Kahlenberg 1, A-1190 Vienna, Austria
{arno.scharl, carmina.reyes, alexander.hubmann}@modul.ac.at

ABSTRACT

A new culture of participation and social innovation is driven by advances in Web technology and the proliferation of social media platforms. The document editing environment introduced in this paper reflects this trend, enabling users to collaboratively create and edit documents. In order to support this process, the system provides tailored information services in the form of content recommendations based on the evolving text of the co-authored documents. The context-sensitive editor is part of the real-time synchronization framework of the *Media Watch on Climate Change*, a public Web intelligence portal that provides a rich repository of environmental knowledge and a portfolio of visual components including tag clouds, keyword graphs and geographic projections.

Author Keywords

Knowledge co-creation; document editor; automated annotation; content recommendations, climate change.

ACM Classification Keywords

H.3.3 [Information Search and Retrieval]: Information Filtering; H.5.2 [Information Interfaces and Presentation]: User Interfaces and Interaction styles; H.5.3 [Group and Organization Interfaces]: Web-based Interaction.

INTRODUCTION

The DecarboNet project (www.decarbonet.eu) develops a *Collective Awareness Platform* to empower citizens, help translate awareness into behavioral change, and provide analytical and visual methods to understand the processes that underlie this behavioral change. The platform aims to engage various environmental stakeholders with a focus on carbon footprint reductions. Among the central components of DecarboNet is a document editor for managing environmental information in a cooperative manner, increasing

awareness and providing citizens with a tool to manage evolving community knowledge. Observations and best practices can be shared, resulting in an iterative process of social innovation that is driven by active user participation. This process can help resolve conflicts among competing interests, build consensus, and improve the quality of decisions. To promote this type of community interaction and translate environmental knowledge into collective action, the platform will:

- aggregate information from multiple online sources,
- extract relevant knowledge and organize it into a contextualized information space,
- provide information search and visualization services,
- support knowledge co-creation through a collaborative document editor, and
- maintain user profiles to support the customization of the platform's content and services.

KNOWLEDGE REPOSITORY

The *Media Watch on Climate Change* (MWCC) is a Web intelligence and online collaboration platform available at www.ecoresearch.net/climate. It compiles large archives of digital content from multiple sources, and provides a variety of knowledge co-creation and visualization services [4]. Within DecarboNet, MWCC serves as the central knowledge repository. It integrates multilingual content from English, French and German source: social media including *Twitter*, *Facebook*, *Google+* and *YouTube*, and the Web sites of news organizations, companies, municipalities, and environmental NGOs. To classify and annotate this content, MWCC utilizes the *webLyzard* media monitoring and Web intelligence platform (www.weblyzard.com). The platforms' data acquisition and information extraction services have been optimized for Web-scale applications in terms of throughput and scalability. The result is a comprehensive information space spanning geospatial, temporal and thematic dimensions.

A visual dashboard shows the dominant issues that are discussed in conjunction with a selected topic and displays them through interactive trend charts (see Figure 1). The results of a search are mapped on geographic and semantic maps to show the regional distribution of the coverage (e.g., places most talked about), as well as its semantic context (e.g., number of documents about related issues).

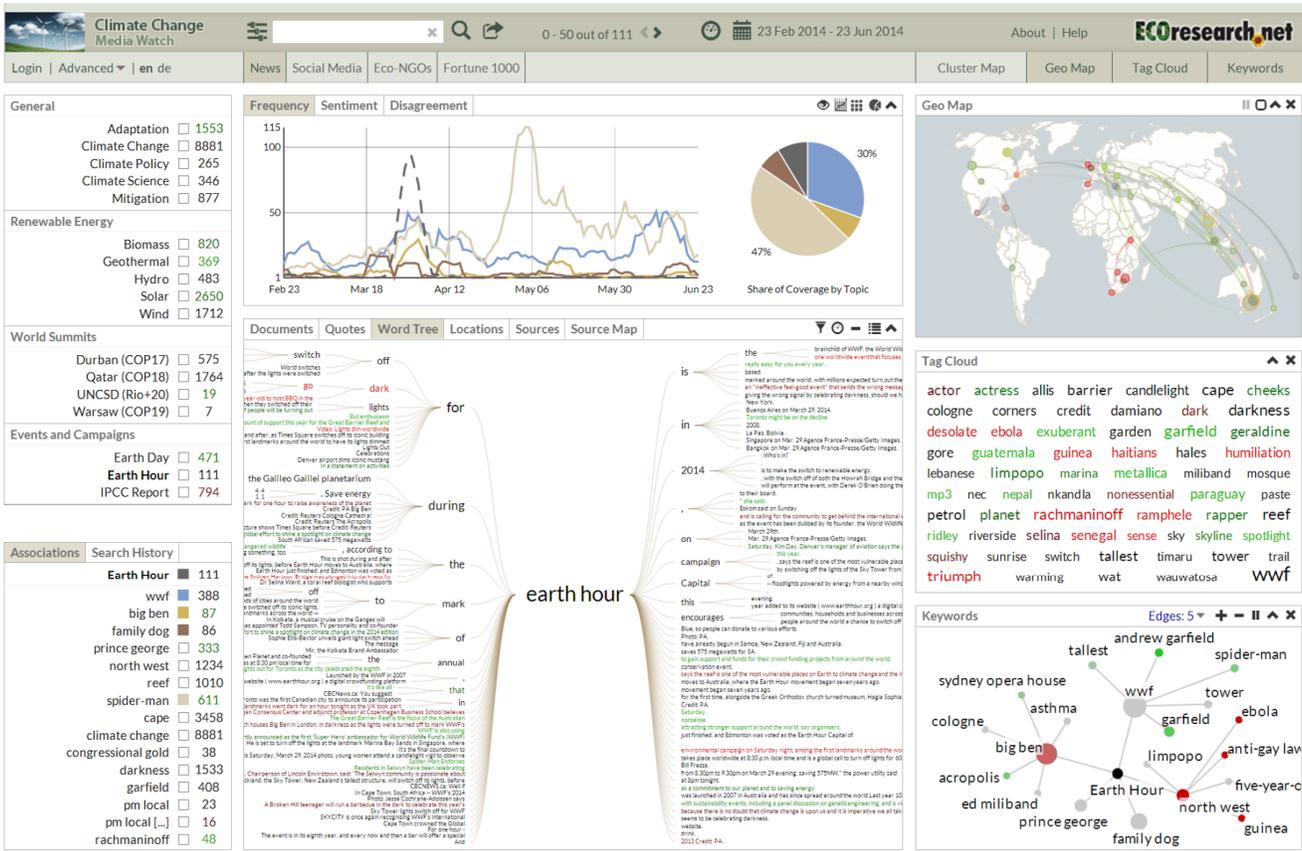


Figure 1. Screenshot of the Media Watch on Climate Change, showing Anglo-American news coverage about the *WWF Earth Hour* between February and June 2014 (www.ecoresearch.net/climate)

The search query for “Earth Hour” depicted in Figure 1, for example, lists *WWF* (initiator and organizer of the event) and *Big Ben* (landmark in London, which had switched off its lights) as the two top news media associations in the context of climate change between February and June 2014. *Earth Hour* as a large environmental campaign [1; 5] represents an ideal use case of DecarboNet. Given its global impact, the event is a good opportunity to test the project’s knowledge co-creation framework and apply its analytical methods – e.g., using advanced geospatial MWCC projections to create dynamic displays of user actions and geo-referenced information flows.

COLLABORATIVE AUTHORING MODELS

DecarboNet not only aims to collect data from external sources, but also to enable its users to collaboratively create documents and participate in social innovation processes that promote sustainability and increase community resilience [3] – e.g. by sharing opinions or best practices that they have identified to reduce their carbon footprint.

Traditional authoring models assume that users first investigate a topic (review the literature, scan the latest news media coverage, etc.) and communicate with their co-authors, and then start drafting and revising the manuscript in either an asynchronous or synchronous manner [2]. The

sequential character of these phases and fragmented workflow is not supportive of collective awareness. DecarboNet will thus merge these distinct phases through automated content recommendations. While a user is editing a document – e.g., a press release or a scientific article – the system automatically recommends related content from a desired set of sources (news media, social media, archives of environmental reports, etc.), as well as appropriate tags to classify the document (geographic location, sentiment, keywords, etc.). These recommendations will be made on-the-fly, while the user is typing. Edited documents become part of the knowledge repository and are processed together with content from the other sources outlined above.

The context-sensitive editor built into the MWCC dashboard supports concurrent editing by multiple users – each minor change triggers knowledge extraction services that analyze the changed document in order to suggest annotations, retrieve related documents from external sources, and visualize related trends and content patterns. While previous versions of the editing environment relied on the *Aloha Editor* (www.aloha-editor.org), the latest release developed within DecarboNet is based *Etherpad Lite* (www.github.com/ether/etherpad-lite), a Web-based collaborative word processor written in JavaScript and made available under an Apache open source license. The plugin

framework and modular structure of *Etherpad Lite* facilitate its integration into third-party solutions, with an Application Programming Interface (API) to manage documents and (groups of) users.

Recent development work has focused on the management of document-level metadata elements and the ranking of content recommendations:

- **Metadata Management.** The Aloha-based version of the editor provided access to the textual content of a document [4], but did not support the management of document-level metadata. The new version continuously analyses the evolving content and parses the text for occurrences of terms with positive or negative polarity from a sentiment lexicon [6]. These terms are highlighted through color-coding, using green for positive and red for negative terms. Planned extensions within DecarboNet will add mechanisms to highlight named entities, suggest keyword labels, and modify semantic annotations on both the term and document levels.
- **Ranking Strategy.** Content recommendations based on vector space similarity only consider the content of documents, but do not reflect the interests of the user, or the specific context of the authoring process. A new cross-platform authentication and user modelling component provides the means to store such context information and will enhance the MWCC ranking algorithm to select not only the most relevant *documents* from a given source, but also the most relevant *sources* in a given context (e.g., news media versus social media).

USING THE EDITOR

Activating the *Document Management* mode of MWCC replaces the *Topic Management* section on the left side of the dashboard. This mode is only accessible to registered users. It activates the editor and lists documents created by the user (“My Documents”) as well as all documents shared with the user (“Documents Shared with Me”).

Users can group documents into specific content categories, providing custom labels and creating their own personal view on the content space. A ‘plus’ button can be used to add new categories, before assigning documents to one of these categories using drag-and-drop operations.

Creating and Editing Documents. Clicking either on the *New Document* icon or an existing item in the *Document List* activates the editor. It provides users with the appropriate editing rights access to common formatting tools. Multiple users can contribute and edit concurrently. As shown in the screenshot of the prototype (Figure 2), the system generates metadata about the document automatically, while the users are typing:

- **Positive and Negative Sentiment Terms.** The color coding of sentiment terms varies in saturation, depending on the degree of polarity – the more vivid the color, the stronger the polarity.

- **Named Entities.** The system also identifies and highlights persons, organizations, and locations in a given document. If the identification process returns multiple matches (e.g., the name *Jane Seymour* could pertain to the English actress or to the third wife of King Henry VIII), the entity with the highest degree of confidence is listed first. Users can access further information on an entity via the portal sidebar, and modify the ranking if the suggested entity with the highest degree of confidence is not the correct match.

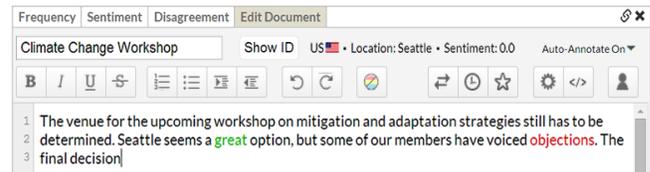


Figure 2. Screenshot of the MWCC editor prototype

Sharing Documents. New documents are automatically added to the document management section. By default, they are not assigned to any category. The initial author is annotated as the *Creator* of the document. The *Creator* of a document can share it with other users by providing their email addresses and specifying them as either *Viewers* who only have read access, *Editors* who can modify the document, or *Owners* who can also share it with other users. If an email address is not connected to an existing user profile, the system sends out an email with an invitation and link to the shared document. An icon next to the sharing button retrieves the list of contributors and viewers – i.e., all users with appropriate rights to access the active document (the *Creator* of the document is always listed first).

Version History. The full history of a document is shown via the time slider of *Etherpad Lite*, which allows accessing and retrieving previous versions of the document. Each snapshot contains the document’s content, the specific date and time, and the co-author who modified the document. Users with appropriate access rights can start editing from any stored instance of the document.

EVALUATION

Following an evolutionary systems development approach, rapid feedback cycles and agile software development have been instrumental in the conceptualization and implementation of the editor. The authors incorporated feedback from the *NOAA Climate Program Office* (www.climate.gov), an early adopter of the platform, and several other partners.

The synchronization of the editor with the other dashboard components is fundamental to its ability to provide context-specific recommendations. To gain insight into the user experience of the dashboard’s synchronization mechanism, usability evaluations were conducted in regular intervals. The aim of these evaluations was to determine strengths and weaknesses of the interaction design, distinguishing two types of assessment: (i) *heuristic evaluation*, where

experts examined the interface and judged the extent to which it is compliant with established usability principles; (ii) *formative usability tests*, where users were observed while working on predefined tasks in realistic settings. Their gaze data was recorded with a *Tobii X60 Eye Tracker* to generate heat maps that show which elements of the dashboard were used to complete a given task, and in which sequence (see Figure 3).



Figure 3. Heatmap of the Media Watch on Climate Change for a search query on “solar energy”.

The feedback from test users showed that analysts have no difficulty using the system after receiving proper training. They appreciate the synchronized views to keep track of the semantic and geospatial context of their current tasks. For untrained first-time users, the complexity of the dashboard can be overwhelming. An upcoming mobile version of the user interface will address the needs of such users.

Further evaluations to be conducted prior to the public release of the MWCC editing platform will integrate new task types into the assessment process, with the aim of revealing common usage patterns when authors are presented with the real-time content recommendations of MWCC, as compared to traditional authoring models.

OUTLOOK AND CONCLUSION

This paper presents a knowledge co-creation environment as part of DecarboNet, a research project to build a “Decarbonisation Platform for Citizen Empowerment and Translating Collective Awareness into Behavioral Change”. The system considers documents open, evolvable resources as part of a *knowledge commons* [3], rather than finished products. It builds upon the *Media Watch on Climate Change* (www.ecoresearch.net/climate), an interactive Web portal that aggregates environmental knowledge in order to investigate stakeholder perceptions and common framing strategies in the public discourse about climate change.

The presented system implements and promotes an alternative to traditional authoring models. It aims to integrate the typically fragmented workflow of (i) investigating a topic,

(ii) communicating with co-authors, and (iii) drafting and revising the actual manuscript. Real-time content recommendations are at the heart of this integration. Such recommendations require advanced knowledge extraction algorithms, scalable enough to analyze evolving documents in real time, while users type.

Due to the complexity of the climate change domain and the diversity of unstructured content, optimizing such knowledge extraction algorithms represents a challenging task. Ambiguity, varied language patterns, the colloquial and contextual nature of social media content, and the lack of co-reference information (unlike longer news articles, tweets often do not follow a conversation thread and appear in isolation from other tweets) all complicate the task.

Addressing these challenges, future research will extend the analytical portfolio of MWCC to capture and visualize the complex phenomenon of collective awareness, processing information at several layers of abstraction – from inter-individual communication and threaded dialogs to large-scale community structures and information diffusion processes that serve as catalysts of collective awareness. Once applied to the collective knowledge gathered from various online sources, the analytical portfolio will not only provide customized information services for citizens, but also important insights for policy makers, NGOs and other environmental stakeholders.

Acknowledgement. The presented research has been conducted as part of DecarboNet (www.decarbonet.eu), which has received funding by the European Union’s 7th Framework Program for research, technology development and demonstration under the Grant Agreement No. 610829.

REFERENCES

1. Aldred, J. (2014). Earth Hour: Millions to Switch off Lights around the World. *The Guardian*. 28-03-2014. <http://www.theguardian.com/environment/2014/mar/28/earth-hour-lights-out-wwf>.
2. Golovchinsky, G., Qvarfordt, P. and Pickens, J. (2009). “Collaborative Information Seeking”, *IEEE Computer*, 42(3): 47-51.
3. Pitt, J., Bourazeri, A., et al. (2013). “Transforming Big Data into Collective Awareness”, *IEEE Computer*, 46(6): 40-45.
4. Scharl, A., Hubmann-Haidvogel, A., et al. (2013). “From Web Intelligence to Knowledge Co-Creation – A Platform to Analyze and Support Stakeholder Communication”, *IEEE Internet Computing*, 17(5): 21-29.
5. Sison, M.D. (2013). “Creative Strategic Communications: A Case Study of Earth Hour”, *International Journal of Strategic Communication*, 7(4): 227-240.
6. Weichselbraun, A., Gindl, S. and Scharl, A. (2013). “Extracting and Grounding Contextualized Sentiment Lexicons”, *IEEE Intelligent Systems*, 28(2): 39-46.