# The Recent Challenge in Web Archiving: Archiving the Social Web

Günther Schefbeck[1], Dimitris Spiliotopoulos[2], Thomas Risse[3]

[1]Austrian Parliament, Austria, guenther.schefbeck@parlament.gv.at
[2]Athens Technology Centre, Greece, d.spiliotopoulos@atc.gr
[3]L3S Research Centre, Germany, risse@l3s.de

**Abstract.** Setting the public as the driving force for formulating the future: Social Web is exploited as the main democratic channel to express the public's opinion on crucial social events. Leveraging the wisdom of the crowds, socially aware digital preservation is the key to the future of archiving. Future internet services will be offering social web analysis and opinion mining integration to the events, people, countries, the everyday events that will be shaping our lives in the near future. One main question surfaces: how can we make this work? Web archives can evolve to exploit the information from the social networks, analyse and expand the potential of the vast amount of social input from the people in a number of social networks (blogs, microblogs, wikis, Twitter, Instagram, Youtube, Facebook, Flickr, etc). Future internet is also about the future of internet content and socially aware digital preservation is the key to ensuring that the content lives on.

**Keywords:** Social Networks, Digital Preservation.

## 1 Introduction

Future internet for the Web is already here: the Social Web plays a crucial role for information and services for all domains, allows contributions by every citizen, giving room for the articulation for a multitude of stakeholders and reflects all types of events, opinions, developments within society, science, politics, environment, business. Future Internet for memory institutions like archives, museums, and libraries is the age of the Social Web. Memory institutions are more important now than ever: as we face greater economic and environmental challenges we need our understanding of the past to help us navigate to a sustainable future. This is a core function of democracies, but this function faces stiff new challenges in face of the Social Web, and of the radical changes in information creation, communication and citizen involvement that currently characterise our information society (e.g., there are now more social network hits than Google searches). Social media are becoming more and more pervasive in all areas of life. In the UK, for example, it is now not unknown for a government minister to answer a parliamentary question using Twitter, and this material is both ephemeral and highly contextualised, making it increasingly difficult for a political archivist to decide what to preserve.

This work reports on the EU-ICT Archive Communities Memories (ARCOMEM) project [1] progress, a project that aims to develop methods and tools for transforming digital archives into community memories based on novel socially-aware and socially-driven preservation models. This will be done (a) by leveraging the Wisdom of the Crowds reflected in the rich context and reflective information in the Social Web for driving innovative, concise and socially-aware content appraisal and selection processes for preservation, taking events, entities and topics as seeds, and by encapsulating this functionality into an adaptive decision support tool for the archivist, and (b) by using Social Web contextualization as well as extracted information on events, topics, and entities for creating richer and socially contextualized digital archives.

The vision of the project approach is to leverage the Wisdom of the Crowds for rapid and automated content appraisal, selection and preservation of digital-born content, in order to create and maintain archives that reflect collective memory and social content perception, and are, thus, closer to their community of current and future users. For this purpose, an innovative socially-aware and socially-driven preservation model will be developed and investigated along three dimensions, building upon earlier works [11,12]:

- Investigation on the leveraging of Social Web information for socially-aware processes of content appraisal and selection as well as for contextualizing content in archives with information about their perception within society and for preserving this context. [7,9]
- Reflecting primary methods of cognitive information perception, structuring and memorization in collective memory by considering events and related entities and topics as a crystallization point in content selection and content organization in the archive as well as for considering evolution and long-term interpretation of the archives (semantic preservation). [3,4,5]
- Exploring an explicit Social Web style for creation of archives by involving communities, sharing effort, and strong interlinking. [2,6,8,10]

In the following paragraphs, the requirements, methodologies and challenges towards the opinion-driven web archiving of the future are described as well as the architecture derived from the user requirements.

## 2 Web archiving and web archivists

The proposed methods effectively introduce the transformation of content from digital archives to communities memories. The clear benefits of this approach need to be made clear to both researchers (and technology providers) and the actual users.

### 2.1 The social network angle

Harnessing the power and potential of social networks is not only about extracting the information but also making the best use for it. In order to do that, the advantages of the social web to digital preservation must be identified.

Social web analysis is all about the users that are actively engaged and generate content. This content is dynamic, rapidly changing to reflect the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. Making use of the social web means the inclusion of the vast sources of user-generated content to the candidate archived content. The social networks are pools of a wide range of articulation methods, from simple "I like it" buttons to complete articles, their content representing the diversity of opinions of the public. The user activities on the social networking are often triggered by specific events and related entities (e.g. sport events, celebrations, crises, news articles, persons, locations) and topics (e.g. global warming, financial crisis, swine flu). In order to include this information, a semantic-aware and socially-driven preservation model is a natural way to go.

### 2.2 Challenges

There are certain areas that require careful planning since they present challenging issues, either research or technology-wise:

- Thorough analysis of crawled Web objects
  - Extraction of Entities, Topics, Opinions, Events (ETOE henceforth)
  - Analysis of interaction on the Social networks Supporting the crawler
  - Learn more about the crawl intention
  - Identification of video duplicates by leveraging the Social Web
- Archive Enrichment
  - Semantic Information about Topics, Entities and Events
  - Sentiments of user content in the Social Web
  - Social and cultural contextualization of the content
- Understanding the dynamics of the Web content
  - Evolution of languages especially in Social Web
  - Evolution of entities over time e.g. Mario Monti: Professor → EU Commissioner → Prime Minister
  - Evolution of opinions
  - Better understanding of the public perception

### 2.3 Target group: web archivists

The first step is to collect the use cases relevant to the users, the web archivists. Understanding user requirements is crucial for designing a correct software system [13]. Incorrect requirements lead to systems that answer to the wrong questions. However, requirements analysis is a difficult task: there are well-documented reasons in the bibliography explaining why it is impossible to be definitive about the specification of a software problem. Factors influencing the detailed architecture of our system can be grouped, according to their characteristics, into categories:

- Functional requirements (What the users want the system to do)
  The goals that users want to reach and the tasks they intend to perform with the new software must be determined. By recognizing the Functional Requirements, we understand the tasks that involve the abstraction of why the user performs certain activities, what his constraints and preferences are, and how the user would make

trade-offs between different products- software applications. The important point to note is that *what* is wanted is specified, and not *how* it will be delivered.

- Data organization
  The logical organization of the data used by the system, and its interrelationships fall into this category.
- Infrastructure requirements
  Special hardware or already existing hardware / software systems that must be used in the project fall into this category.
- Non-functional requirements (The restrictions on the types of solutions that will meet the functional requirements)
  Specification of non-functional requirements includes the categorization of the users (professionals and personal users), the description of user characteristics such as prior knowledge and experiences, the special needs of professional (journalists, editors, etc) and personal users (news audience), subjective preferences, and the description of the users' environment, in which the product or service will be used. Legal issues, intellectual property rights, security and privacy requirements are also an issue.

## 3 Discussion and future work

The extensive use cases have lead to a list of very specific requirements that quantify and qualify the impact of events/topics/entities in the social web. Investigation on how people talk about events, how are they recommended, how the sentiment of comments, reactions and appraisal can be utilized and how is the content spread out on different social platforms results in a list of required analyses:

1. Timeline of feedback
2. Sentiment analysis
3. Social and demographic information
4. Mentions and backlinks
5. Loyalty of users
6. Context of feedback
7. Discover of key influencers
8. Monitor topics, events and persons
9. Access to social multimedia content
10. Track the dissemination of information
11. Identify reliable content sources
12. Support content verification
13. Differentiate facts and opinions
14. Information Retrieval
15. Geographical information
16. User types

Content analysis and archive enrichment with detailed semantic and social meta-information holds the key to the future of digital preservation. It is in the scope of this work to enable the web archivists with an integrated approach on socially-driven preservation.

## References


1. Archive Communities Memories, EU-FP7, www.arcomem.eu
2. G. Zenz, N. Tahmasebi, T. Risse (2012): Towards mobile language evolution exploitation In: Multimedia Tools and Applications – Special Issue on Semantic Ambient Media Experience. Springer, Netherlands; DOI: 10.1007/s11042-011-0973-0
3. E. Demidova, X. Zhou, W. Nejdl (2011): A Probabilistic Scheme for Keyword-Based Incremental Query Construction IEEE Transactions on Knowledge and Data Engineering, 07 Feb. 2011. IEEE computer Society Digital Library. IEEE Computer Society.
4. G. Papadakis, E. Ioannou, C. Niederée, P. Fankhauser (2011): Efficient Entity Resolution for Large Heterogeneous Information Spaces. Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM), Feb. 2011, Hong Kong.
5. D. Maynard, A. Funk (2011): Automatic detection of political opinions in tweets. In Proceedings of MSM 2011: Making Sense of Microposts. Workshop at 8th Extended Semantic Web Conference (ESWC 2011). Heraklion, Greece. June 2011.
6. S. Maniu, B. Cautis, T. Abdessalem (2011): Building a Signed Network from Interactions in Wikipedia In First ACM SIGMOD Workshop on Databases and Social Networks (DBSocial), June 2011, Athens, Greece.
7. O. Edelstein, M. Factor, R. King, T. Risse, E. Salant, P. Taylor (2011): Evolving Domains, Problems and Solutions for Long Term Digital Preservation. In Proc. of 8th International Conference on Preservation of Digital Objects, Singapore, November 1-4, 2011.
8. N. Tahmasebi, T. Risse, S. Dietze (2011): Towards automatic language evolution tracking, A study on word sense tracking. In Proc. of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011), Bonn, Germany, October 24, 2011.
9. T. Risse, S. Dietze, D. Maynard, N. Tahmasebi, W. Peters (2011): Using Events for Content Appraisal and Selection in Web Archives. In Proc. of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), Bonn, Germany, October 23, 2011.
10. K. Doka, D. Tsoumakos, N. Koziris (2011): KANIS: Preserving k-Anonymity Over Distributed Data. Proc. 5th International Workshop on Personalized Access, Profile Management and Context Awareness in Databases, Seatle, WA, September 2, 2011.
11. Living Web Archives (LiWA), www.liwa-project.eu
12. LivingKnowledge, www.livingknowledge-project.eu/
13. D. Spiliotopoulos, E. Tzoannos, P. Stavropoulou, G. Kouroupetroglou, A. Pino (2012): Designing User Interfaces for Social Media Driven Digital Preservation and Information Retrieval. In Proc. 13th International Conference on Computers Helping People with Special Needs (ICCHP 2012), Lecture Notes in Computer Science, vol 7382, pp 581-584, Springer, Berlin, Heidelberg.