Future Internet **2014**, *6*, 433-456; doi:10.3390/fi6030433

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Analysing and Enriching Focused Semantic Web Archives for Parliament Applications

Elena Demidova [1,*], Nicola Barbieri [2], Stefan Dietze [1], Adam Funk [3], Helge Holzmann [1], Diana Maynard [3], Nikolaos Papailiou [4], Wim Peters [3], Thomas Risse [1] and Dimitris Spiliotopoulos [5]

[1] L3S Research Center, Leibniz Universität Hannover, 30167 Hannover, Germany; E-Mails: dietze@L3S.de (S.D.); holzmann@L3S.de (H.H.); risse@L3S.de (T.R.)

[2] Yahoo Labs Barcelona, 08018 Catalunya, Spain; E-Mail: barbieri@yahoo-inc.com

[3] NLP Group, Department of Computer Science, University of Sheffield, S1 4DP Sheffield, UK; E-Mails: a.funk@dcs.shef.ac.uk (A.F.); d.maynard@dcs.shef.ac.uk (D.M.); w.peters@dcs.shef.ac.uk (W.P.)

[4] ATHENA - Research and Innovation Center in Information, Communication and Knowledge Technologies, 15125 Maroussi, Athens, Greece; E-Mail: npapa@cslab.ntua.gr

[5] Athens Technology Center (ATC), 15233 Halandri Athens, Greece; E-Mail: d.spiliotopoulos@atc.gr

* Author to whom correspondence should be addressed; E-Mail: demidova@L3S.de; Tel.: +49-511-762-17732; Fax: +49-511-762-17779.

Received: 16 April 2014; in revised form: 19 June 2014 / Accepted: 11 July 2014 / Published: 30 July 2014

Abstract: The web and the social web play an increasingly important role as an information source for Members of Parliament and their assistants, journalists, political analysts and researchers. It provides important and crucial background information, like reactions to political events and comments made by the general public. The case study presented in this paper is driven by two European parliaments (the Greek and the Austrian parliament) and targets an effective exploration of political web archives. In this paper, we describe semantic technologies deployed to ease the exploration of the archived web and social web content and present evaluation results.

Keywords: web archiving; semantic content analysis; entity and event extraction; enrichment; topic detection; parliament libraries

---

1. Introduction

Parliament libraries provide Members of Parliament (MP) and their assistants, journalists, political analysts and researchers information and documentation for parliamentary issues. Besides traditional publications, the web and the social web play an increasingly important role as information sources, which provide important and crucial background information, like reactions to political events and comments made by the general public. It is in the interest of the parliaments to create a platform for preserving, managing, mining and analyzing all of the information provided on the web and in social media.

The ARCOMEM application (the official release of the ARCOMEM system is available online [1]) presented in this paper is driven by two European parliaments (the Greek and the Austrian parliament) and targets the effective creation of political archives based on the web and social media. Through ARCOMEM, the Greek and Austrian parliaments aspire to transform their flat digital content archives to historical and community memories.

Community memories largely revolve around events, as well as entities, topics and opinions related to these events. These may be unique events, such as the first landing on the moon or a natural disaster, or regularly occurring events, such as elections or TV serials. In this context, the main logical concepts considered in ARCOMEM extraction and enrichment activities are entities, topics, opinions and events (ETOEs). To create incrementally-enriched web archives that allow access to all sorts of web content in a structured and semantically meaningful way, extraction, enrichment and consolidation of ETOEs are of crucial importance [2]. To this extent, the main challenges we face are related to the extraction, detection and correlation of ETOEs and related information in a vast number of heterogeneous web resources. These processes face issues arising from the diversity of the nature and quality of web content, in particular when considering social media and user-generated content, where further issues are posed by the informal use of language.

While entities and events resulting from the automatic extraction processes provide an initial classification and structure for the crawled web documents, they can be heterogeneous, ambiguous and provide only very limited information. Therefore, data enrichment and consolidation in ARCOMEM follows two aims: (1) enrich and disambiguate extracted entities with related publicly available knowledge; and (2) identify data correlations by aligning ARCOMEM entities with reference datasets. To achieve these aims, we enrich ARCOMEM entities using Linked Open Data (LOD) sources, such as DBpedia [3] and Freebase [4], and correlate the entities using their direct and indirect relationships within the LOD graph (see, e.g., [5]).

In addition to entity and event extraction, topic discovery and analysis techniques provide an effective way of analyzing and browsing large collections of textual data, such as the ones collected during the ARCOMEM's crawling campaigns, as they are able to uncover the hidden thematic and semantic structure of the data. By exploiting the potentiality of this kind of analysis, the ARCOMEM framework provides state-of-the-art tools for probabilistic topic detection and to analyze the trendiness and dynamical change of topics over time. These tools can provide a high level perspective of the textual data retrieved during a crawling campaign and help the user to understand the semantic concepts behind each document. In addition, we can investigate relationships between topics and their social influence of

users. A proposed topic-aware social influence framework jointly learns topics and the users' influence in those topics [6].

In this paper, we first provide an overview of the ARCOMEM offline processing chain that supports the extraction of semantic information from web archives and then present the search and retrieval application (SARA) that enables users to navigate through the archived content using semantic information. Finally, we present evaluation results.

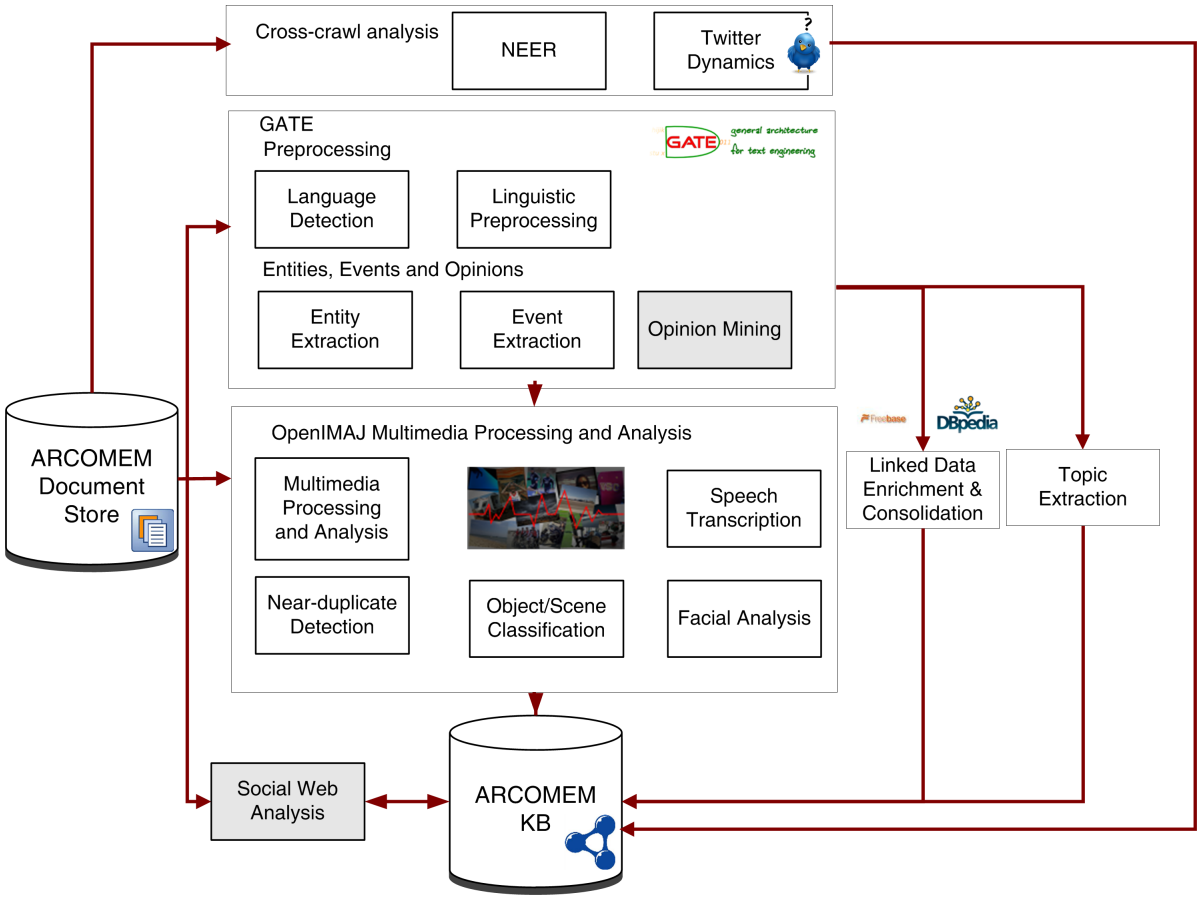
2. The Offline Processing Chain and Cross Crawl Analysis

The overall ARCOMEM architecture includes several levels [7]. First, in the crawler level, the system decides and fetches the relevant web objects as initially defined by the archivists. In parallel to the crawler level, the online processing level performs time-efficient analysis of the fetched resources to support prioritization of the crawler processing queue. Following that, in the offline processing level, fully-featured versions of the entity, topic, opinion and event (ETOE) analysis and the analysis of the social content operate over the cleansed data from the crawl that are stored in the ARCOMEM Knowledge Base (ARCOMEM KB). These processing tools make use of linguistic, machine learning and Natural Language Processing (NLP) methods in order to provide a rich set of metadata annotations that are interlinked with the original data. The respective annotations are stored back in the ARCOMEM database and are available for further processing and information mining. After all of the relevant processing has taken place, the web pages to be archived and preserved are selected in a semi-automatic way. The selected original pages are transferred to the web archive (in the form of Web ARChive file format (WARC) files). Finally, the cross-crawl analysis level operates on collections of web objects that have been collected over time in order to register the evolution of various aspects identified by the ETOE and web analysis components. As such, it produces aggregate results that pertain to a set of web objects crawled at different points in time, rather than to a single crawl. Both the web archive and the ARCOMEM KB are accessible for the SARA application to support its functionality.

The ETOE processing chain within the offline processing and cross-crawl analysis levels of the ARCOMEM architecture is presented in Figure 1. In the following sections, we briefly introduce components responsible for processing text resources.

The ARCOMEM KB and Data Model: The ARCOMEM Knowledge Base (ARCOMEM KB) serves as a central integration point of the chain, enabling its components to store and exchange processing results. The ARCOMEM KB is implemented as H2RDF [8], a fully-distributed RDF store that combines the MapReduce processing framework with a NoSQL distributed data store. H2RDF is able to provide a fully-distributed SPARQL query functionality on a virtually unlimited number of RDF triples. The information within the ARCOMEM KB is stored according to the ARCOMEM data model that describes concepts used in ARCOMEM and the relationships among them.

Figure 1. Processing chain for entities, topics, opinions and events (ETOEs) extraction, analysis and enrichment.



3. Entity Extraction from Text

The work performed in this task aims to advance the state-of-the-art in the development and adaptation of language processing resources to new domains and languages, especially in the domain of social media, where such tools are not widespread and suffer from a variety of problems. Our main aim is to investigate new methodologies for language processing on social media and particularly on degraded texts (especially tweets [9]).

The entity recognition components of ARCOMEM are all developed in GATE—an open source solution for text processing [10]. The application for entity recognition consists of a set of processing resources executed sequentially over a corpus of documents. The application can be divided into the following components:

(1) Document pre-processing, which separates the body of the content from the rest;
(2) Linguistic pre-processing, such as tokenization, part of speech tagging, lemmatization and language identification for documents or even document parts in different languages;
(3) Named entity recognition to detect crawl-related entities that occur in the document, such as persons (e.g., artists, politicians, Web 2.0 users), organizations (e.g., companies, music bands, political parties), locations (e.g., cities, countries), dates and times;
(4) Term extraction;

(5) RDF generation.

In ARCOMEM, we have created a new branch of the GATE application, which deals specifically with tweets, including specialized techniques to language identification, tokenization, normalization and POS-tagging in tweets [11].

4. Event Extraction from Text

Along with entities, event recognition is one of the major tasks within information extraction [12]. In ARCOMEM, we refer to an event as a situation within the domain (states, actions, processes, properties) expressed by one or more relations. This perspective on events as relational constructs between participants is in line with foundational ontological definitions [13] and the conceptual organization of existing *de facto* event models (e.g., linked open descriptions of events (LODE) [14], event-model-F [15], and the event ontology [16]. Events can express these relational constructs at any level of semantic granularity: coarse-grained events, such as the first landing on the moon or a natural disaster, regularly occurring events, such as elections or TV serials, or fine-grained events, such as the resignation of a politician.

There are various strategies to recognize events that can be broadly categorized into top-down and bottom-up approaches, both of which were implemented in the ARCOMEM event extraction module. The top-down method involves predefined sets of relations, possibly structured into an ontology, which makes such tasks domain-dependent, but feasible. This approach gained great popularity within the ACE context [17] and has been adopted in ARCOMEM for semantically well-defined events, such as strikes and music performances. The top-down approach involves a form of template filling, by selecting a number of known events in advance, and then identifying relevant verbs and their subjects and objects to match the slots. For example, a "performance" event might consist of a band name, a verb denoting some kind of "performing" and, optionally, a date and location. This kind of approach tends to produce high precision, but relatively low recall. Domain-independent techniques involve a more bottom-up identification of entities and relations, which can incrementally be generalized into events [18]. Exploiting natural language technology enables the discovery of n-tuples of relevant items (entities) belonging to an n-ary relation in natural language documents in an unsupervised way [19]. In ARCOMEM, we follow a bottom-up approach, which consists of identifying verbal relations in the text and classifying them into semantic categories, from which new events can be suggested. This kind of approach produces higher recall, but lower precision.

The key NLP components for event extraction are named entity recognition, term extraction and relation extraction. Once extracted, acquired event knowledge is formalized and stored according to (*de facto*) standard representation models. The GATE application we have developed for event recognition is similar in structure to the one for entity recognition and is designed to be run after the entity recognition application has first been run on the corpus.

5. ETOE Enrichment and Correlation

ETOEs extracted by the GATE component of the ARCOMEM offline analysis depicted in Figure 1 provide an initial classification and structure for the crawled web documents, for instance, the association

of terms with entity types defined in the ARCOMEM data model. However, as the content analysis extracts structured data from unstructured resources, such as text and images, the generated data is: (i) heterogeneous, *i.e.*, not well interlinked; (ii) ambiguous; and (iii) provides only very limited information. This is due to the data being generated by different components and during independent processing cycles.

Data enrichment and consolidation (achieved by the linked data enrichment and consolidation component in Figure 1) follows three aims: (1) enrich existing entities with related publicly-available knowledge; (2) disambiguation; and (3) identify data correlations, such as the ones above, by aligning ARCOMEM entities with reference datasets. Both (1) and (2) exploit publicly available data from the Linked Open Data Cloud [20], which offers a vast amount of data of both a domain-specific and domain-independent nature (the current release of the LOD Cloud consists of more than 31 billion distinct triples, *i.e.*, RDF statements). The ETOE enrichment approach [2] first identifies correlating enrichments from reference datasets, which are associated with the respective entities and, secondly, uses these shared enrichments to identify correlating entities in the ARCOMEM KB. In particular, the enrichment approach of ARCOMEM uses DBpedia [3] and Freebase [4] as reference datasets.

Named Entity Disambiguation: An important aspect of ETOE enrichment and correlation is the disambiguation of named entities in the face of multiple enrichment candidates within a knowledge base. Whereas the DBpedia Spotlight service [21] used in our enrichment process ranks enrichment candidates based on their popularity within the DBpedia graph, such a ranking is independent of the particular entity context and does not resolve the problem of entity ambiguity. In the context of Freebase, structured queries used for enrichment do not solve this problem either as they return a set of all potential enrichment candidates. The disambiguation problem occurs whenever a named entity corresponds to more than one resource within the knowledge base. For example, the most popular entity corresponding to "Gary Johnson" in DBpedia is an English football manager and former player: "http://dbpedia.org/page/Gary_Johnson_(footballer)". However, in the context of the U.S. Elections 2012 dataset, the entity "Gary Johnson" is most likely referring to an American politician: "http://dbpedia.org/page/Gary_Johnson". Another example is the location "Athens", which typically corresponds to the capital of Greece. However, in the context of the U.S. Elections 2012, it most likely referring to one of the locations in U.S., such as Athens, Georgia, or Athens, Ohio.

In order to decide which knowledge base resource from a list of candidates should be used for semantic enrichment of a particular named entity, we take the similarity of the entity contexts in both sources, the web resource and the knowledge base, into account. Other work on this topic has been done using Wikipedia as a resource, such as the famous learning to link approach by [22]. However, this method requires interlinked description of related entities, which are currently not available in linked data knowledge bases.

In the ARCOMEM project, we implemented a context similarity computation for DBpedia. To this extent, we represent the contexts of entities in web resources and enrichment candidates in DBpedia as context vectors. A context vector contains frequencies of entities occurring in this context (*i.e.*, in the text of the corresponding paragraph in the web resource or in the property values of the DBpedia resource). Then, we compute the context similarity score using the cosine similarity of both context vectors. Finally, the candidate with the highest context similarity score is used for enrichment. We

envision that this disambiguation technique is also applicable for enrichment in knowledge bases with limited textual representation of resources (e.g., Freebase). In this case, the vector representing entity context can be built using closely-connected entities within the knowledge base graph. Such graph-based approaches have already been proposed in different ways [23,24], for instance by performing a random walk from co-occurring entities.

As entities in linked data sources are often interlinked among each other, incorporating data from other sources can further improve the disambiguation results, especially on entities with very limited information available. Possibly useful resources in this matter could be GeoNames [25] for geographic entities or the Virtual International Authority File (VIAF) [26], given that links to the target knowledge base are available or can be established. The VIAF is maintained by different libraries and covers local versions of names, as well as name variations for several kinds of entities. Therefore, this would be especially useful for linking and disambiguating entities in multilingual crawls. The investigation of incorporating further linked knowledge bases into the disambiguation process remains for future work.

Correlation and Clustering: With respect to data correlation, we distinguish direct and indirect correlation. To give an example, during one particular cycle, the text analysis component might detect an entity from the term "Ireland", while during later cycles, entities based on the term "Republic of Ireland" or the German term "Irland" might be extracted. Each of the three entities, all referencing the same real-world entity, is associated with the same enrichments to the respective Freebase ("http://www.freebase.com/view/en/ireland") and DBpedia ("http://dbpedia.org/resource/Ireland") entries. Therefore, correlated ARCOMEM entities (and hence, web objects) can be clustered directly by identifying joint enrichments between individual entities. In addition, the retrieved enrichments associate (interlink) the ARCOMEM data and web objects with the vast knowledge, *i.e.*, data graph, available in the LOD cloud, thus allowing one to retrieve additional related information for particular ARCOMEM entities. For instance, the DBpedia RDF description of Ireland ("http://dbpedia.org/resource/Ireland") provides additional facts and knowledge (for instance, a classification as an island or a country, geodata, the capital or population, a list of famous Irish people and similar information) in a structured and, therefore, machine-processable form. That knowledge is used to further enrich ARCOMEM entities and create a rich and well-interlinked (RDF) graph of web objects and related information. Thus, we can perform additional clustering and correlation of entities (and hence, crawled web resources) to uncover indirect relationships between web resources related in one way or another.

6. Topic Detection Based on Probabilistic Topic Models

Exploring and retrieving meaningful information from large collections of textual documents is a challenging task. The hidden thematic structure in such collections can be discovered by applying recently proposed statistical analysis tools, such as probabilistic topic models [27]. At a high level, the idea is to study the co-occurrence of words, assuming that words that tend to co-occur frequently belong to or express the same semantic concept. The purpose of the topic detection module is to uncover the hidden thematic structure of a document collection related to an ARCOMEM campaign and to identify semantic topics of interest for future data analysis and navigation. The detected topics provide a low-dimensional representation, in terms of the abstract co-occurrence patterns, of the textual

data analyzed. After detecting topics, the projection of each document into the topic space can be exploited to classify documents and named entities into semantic categories, allowing a more effective browsing of the content crawled during each campaign.

The simple instantiation of probabilistic topic model is known as latent Dirichlet allocation [28]. Its underlying generative process can be formalized as follows:

(1) For each latent topic $k = 1, \ldots, K$ sample a multinomial distribution over words $\boldsymbol{\phi}_k \sim \mathit{Dir}(\boldsymbol{\beta})$;
(2) For each document d in the corpus,
   (a) Sample the number n_d of words to be generated;
   (b) Choose $\boldsymbol{\theta}_d \sim \mathit{Dir}(\boldsymbol{\alpha})$;
   (c) For each of the n_d words to be generated,
      i. Sample a topic $z \sim \mathit{Discrete}(\boldsymbol{\theta}_d)$
      ii. Sample $w \sim \mathit{Discrete}(\boldsymbol{\phi}_z)$

Exact inference in Latent Dirichlet Allocation (LDA) is intractable, due to the exponential number of possible word-factor assignments. The solution to this is to use approximate inference algorithms, such as mean-field variational approximation [28], expectation propagation [29] and Gibbs sampling [30]. The topic detection module in ARCOMEM is based on the open source implementation of a collapsed variational Bayes inference procedure provided by Mahout [31].

The nature of the documents crawled within a campaign could be intrinsically multilingual, since we are often interested in tracking events with importance at the worldwide dimension. As a result, running the topic module on the whole textual collection corresponding to a crawl could produce low quality results: (i) the low-dimensional decomposition of documents could simply try to identify language-specific topics (English, German, *etc.*); (ii) from the user point of view, topics should be directly interpretable by analyzing their keywords, and the language of users should hence be considered. A preliminary evaluation by users on the topics detected on the U.S. Elections 2012 data highlighted the importance of addressing these issues. In particular, users were reported to engage with topics, by using them as a tool for browsing the collection of documents, but also to have problems in understanding some of the keywords associated with a topic, since they belonged to a different language. This valuable feedback motivated the integration of a language selection phase into the data pre-processing module. An updated version of the topic detection module takes as input a list of languages that should be considered for the analysis. Language detection is accomplished by integrating and running a detector based on a naive Bayes open source implementation [32].

To enhance the capabilities of topics as a tool for organizing and browsing a collection of large textual data, it is interesting to detect relationships between topics. The connection between topics and relevant documents allows users to navigate documents, focus on the ones they are interested in and keep exploring data by browsing documents having the same topics. At the same time, users may also be interested in exploring documents having different, yet similar topics, with respect to the one which is relevant for the current document. Hence, in this user scenario, discovering relationships between topics boils down to detecting their similarities. As topics are formally defined as a multinomial distribution

over words of a dictionary, we can compute the distance between two topics by employing the Kullback-Leibler divergence:

$$KL(\phi_k, \phi_l) = \sum_{w=1}^{V} \phi_{k,w} \cdot \log \left(\frac{\phi_{k,w}}{\phi_{l,w}} \right)$$

that computes the expectation of the difference between two distribution ϕ_k and ϕ_l over the first distribution. KL is asymmetric; for the sake of simplicity and for consistency reason (a user would naturally expect that if k is similar to l then these relationships should hold also in the opposite direction), we measure the distance between two topics as the Jensen–Shannon divergence:

$$d(\phi_k, \phi_l) = 0.5 \cdot KL(\phi_k, \phi_l) + 0.5 \cdot KL(\phi_l, \phi_k)$$

which is symmetric. Distance weights between pairs of topics are stored in the ARCOMEM KB and can be used by SARA to detect the most similar topics to the one currently explored by the user.

The topic module provides also a compact description of each topic by considering named entities rather than textual tokens. Essentially, this corresponds to a multinomial distribution over named entities (retrieved from the ARCOMEM KB) for each topic. This description is complementary to the one provided by keywords and provides further insights into the semantic meaning of each topic.

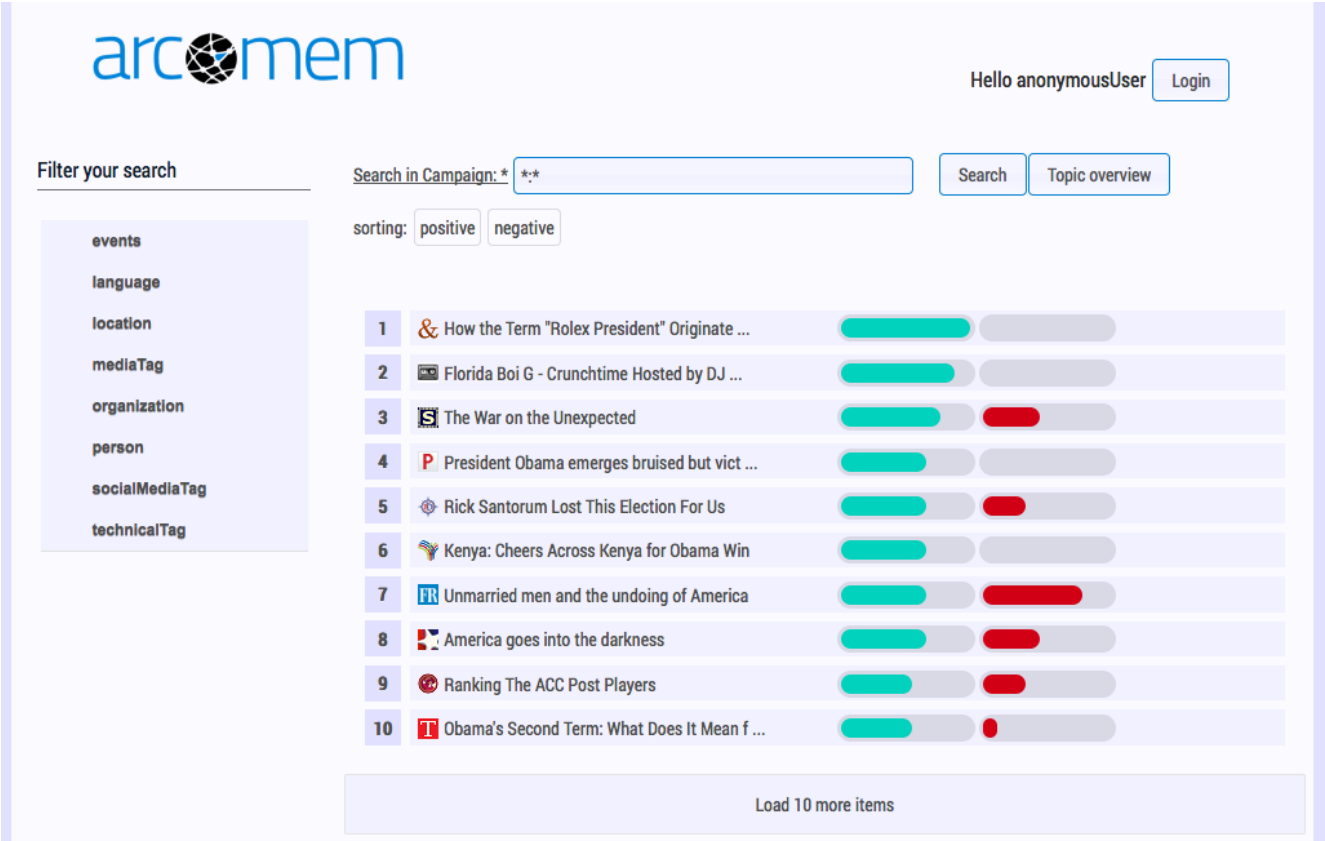
Topics also constitute an important dimension for identifying experts and influential users in social media. To this aim, we proposed a machine learning framework that jointly models social influence and topics and is able to effectively detect users' authoritativeness and interests for retrieved topics [6].

7. The Search and Retrieval Application (SARA)

The SARA interface is the consolidated means for accessing the archived metadata from the ARCOMEM KB. In this application, the user may select a campaign and receive an overview page containing metadata and statistics, including the owner of the campaign, duration, number of documents, a word cloud of the most frequent entities, web and social media document distribution and cultural statistics. In the next step, the user may search the campaign archive using text queries. The full text search of SARA returns lists of results at the web resource level, dynamic facets and similarity-based topic suggestions (see Figure 2).

The dynamic facets of SARA interface presented in Figure 2 are generated from the metadata in the result set and are mainly comprised of named entities and events. Events are returned as a top-level facet item for any search for which the results contain web resources that have events associated with them at the document level. The user can employ these facets to filter the results. In addition, the topics in the topic overview are triggered as suggestions by the interface based on the interaction with the main entities that describe each topic. The list of results includes the web resource title, an indication for the volume of positive and negative opinions expressed in the text of each web resource, and the actual source for each web resource, such as blog, Twitter [33], YouTube [34], *etc.*

Figure 2. Search results page in the search and retrieval application (SARA): A list of results at the web resource level sorted by positive opinions expressed in the text, dynamic facets (on the left side) and access to the topic overview (through a button at the top right).



Topics, events and entities can also be explored from the individual web resource pages, as shown in Figure 3. In the web resource view, the user may examine the following:

- List of entities, grouped by named entity type *i.e.*, person, organization, location and sorted alphabetically. For each entity, the following information can be accessed:
  - Opinion
  - Enrichment
  - Entity clusters;
- Lists of events;
- Topics.

The detailed view of the available metadata, such as ETOEs, enrichments and clusters in the SARA interface, is shown in Figure 4. The SARA interface adopted the continuous forward going search model from the early stages of the design [35]. This means that the search and refinement may always continue by clicking the metadata items.

Figure 3. An individual web resource page in SARA.

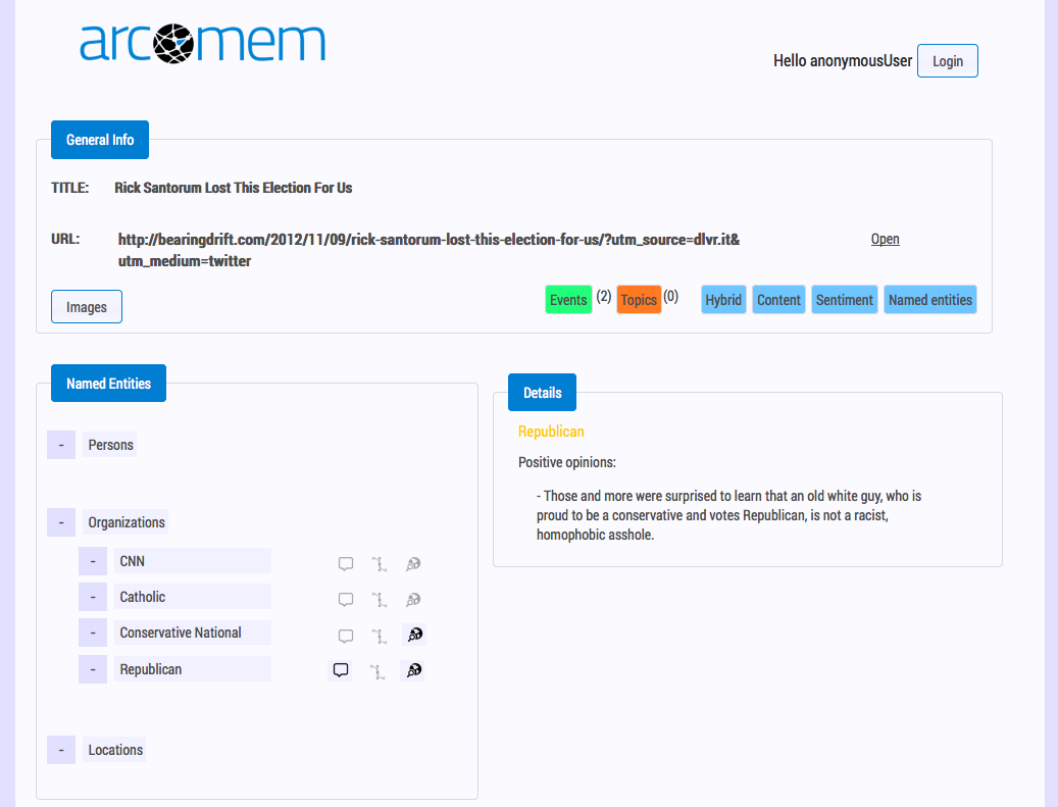
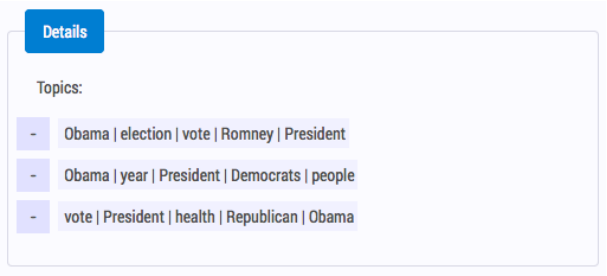
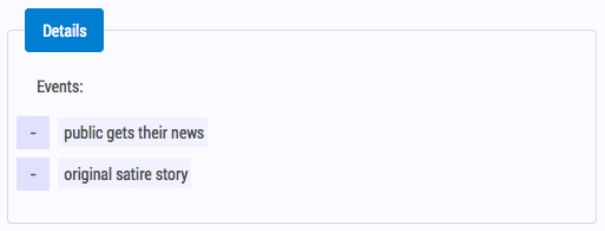


Figure 4. The detailed view on the topics, events, clusters and enrichments in SARA. (**a**) Topics; (**b**) Events; (**c**) Cluster; (**d**) Enrichment.



(**a**)



(**b**)

(**c**)

(**d**)

8. Datasets

Two crawl campaigns were used for the evaluation, a sample of the U.S. Elections 2012 and the investment for growth. The U.S. Elections 2012 campaign was gathered by an earlier version of the ARCOMEM system and included social media resources from Facebook, YouTube and Twitter. Based on the users' feedback at the early development stages, we identified Twitter to be the most relevant source. Later on, Facebook has been excluded from the ARCOMEM system, due to the privacy considerations. The investment for growth data was collected at the later stage of the system development. At this stage, the social media resources were restricted to Twitter—the most relevant social media source according to the users' feedback.

The U.S. Elections 2012 web crawl was run for a period of 25 days. The total number of resources gathered in this crawl was 2,912,413 containing about 33% multimedia resources. For the testing and fine tuning of the framework components, the ARCOMEM system worked with a subset of the U.S. Elections 2012 crawl, containing 7521 resources. The API crawl for U.S. Elections 2012 ran for a period of 10 days. It gathered 3,832,338 social media resources, including 35,011 Facebook, 65,189 YouTube and 3,732,138 Twitter posts. All ARCOMEM modules were run on this data set, albeit at different development stages.

The investment for growth web crawl was run for a period of five days in August, 2013. In total, the crawlers collected more than 344,616 resources for this campaign, including about 13% multimedia resources. The API crawler was run for six days in June, 2013, on Twitter and collected 14,775 unique posts. At the time of the evaluation, the final versions of the ARCOMEM modules were run for entities, opinions, events, enrichment and clustering.

9. Evaluation

The aims of the evaluation were the assessment of the overall usability of the SARA system, user perception of quality and importance of metadata generated by different modules (ETOEs, enrichments and clusters) and the contributions of the particular modules to an overall user satisfaction. The evaluation was set up for all modules of SARA.

9.1. Objectives

Usability testing is an essential element of the quality assurance and represents a true test of how people actually use the user interface. The evaluation aim was also to reveal the importance of analytics results produced by different ARCOMEM modules. Finally, the presentation, discoverability, ease of use and user understanding of the generated metadata (ETOEs, enrichments and clusters) were also evaluated. The latter is an important objective, since it is directly related to the digital preservation approach of ARCOMEM. The combined feedback of perceived accuracy, importance and acceptance provides a strong indication as the level of success of the approach, which is focused preservation and retrieval of web and social web resources. Finally, the evaluation of the actual quality of generated metadata was another key objective. For the end user, the quality of detected entities, topics and events

can be measured by their subjective feedback. The user expectations were that entities are correctly detected and classified according to their type, while topics and events are descriptive.

9.2. Methodology

Overall, our evaluation methodology included focus groups, summative usability testing, the think-aloud protocol and the cognitive walkthrough method. Focus groups were used at the earlier stages in order to identify the usability requirements [35]. Led by experienced moderators, the focus groups sessions took place at the Hellenic Parliament, where the archivists and the end users collectively discussed their expectation of the functionality and content of the SARA application.

The focus groups included few people at a time who discussed preservation methods and the types of semantic data required for retrieval of archived documents as a part of the design and the formative evaluation during the development iterations. Major findings included the need for metadata-driven semantic search, the use of the semantic information, such as entities, topics and opinions and provision of information that would help end users to better understand the content of the archived documents. The latter was exemplified as an overview of the semantic metadata, as well as links interconnecting metadata items.

The main usability evaluation method that was used in the evaluation presented in this paper was the summative usability testing, that is the summative evaluation of the SARA interface. The summative evaluation is used at the end of the design and development life-cycle for finished user interfaces. In this evaluation methodology, real users perform concrete tasks that are designed to measure effectiveness, efficiency and satisfaction. The feedback from the users is evaluated against the usability requirements. The summative usability testing is performed in a controlled environment and aims to collect both subjective and objective feedback. The Likert scale is a commonly used representation for graded user feedback. The subjective feedback from the participants was collected using the complete online questionnaires using the Likert scale of 1 (very low) to 7 (very high).

In our study, the think-aloud protocol was used to collect partial subjective feedback from the participants. The participants were encouraged by the facilitator to vocalize the rationale for their actions, as well as the intended interactions while they were performing the tasks of the usability testing. This enables deeper understanding of the user actions, providing insights in the user attitude. It complements with the user feedback in traditional questionnaire web forms and the interaction logs.

In addition, the cognitive walkthrough method was used before the main evaluation sessions. Few experienced user interface evaluators and the designers interacted with the SARA interface looking for the best ways to perform certain tasks in search and retrieval. The aim was to identify potential shortcomings or issues and also help the evaluation facilitators verify the task-based scenarios that would be used for the evaluation.

Discoverability and efficiency were major performance indicators for the SARA interface; however, the feedback from the users distinctly mentioned that the quality of the key semantic metadata of the web archives could be the most crucial content-specific reason for their acceptance. That statement was put to the test during the final evaluation and was particularly articulated by specific questions for the subjective feedback of the users at the end of each session.

The importance of a metadata type (e.g., entity, event or topic) is the perceived usefulness of this type in the context of the user query. The quality is the direct qualitative aspect of the actual metadata values. For example, for topics the quality translates to easiness to understand, descriptiveness and logical connection to encompassed entities. Normally, the importance, quality and an overall acceptance of metadata are not affected by means of presentation as long as these means do not favor or hinder particular metadata types. In order to facilitate evaluation of metadata importance and quality, we made all metadata types directly accessible through the SARA user interface. To this extent, we developed specific visualizations that present metadata to the user and enable metadata-based search and navigation. For example, SARA interface provides direct access to the entities of different types, as well as events and opinions through dynamic facets. In addition, the topic visualization was a design approach to present a data collection through an interactive graph between all topics (in the collection), their top-ranked associated entities and the opinions on those entities. Through this visualization, the user could get an immediate overview of the topics in the whole collection and either start a topic-driven search or drill down to the documents containing selected topics. This way, SARA enables end users to freely select available metadata types to explore at any time during the interaction process.

9.3. Participants

Thirty-two users participated in the evaluation. They were computer science students familiar with HCI and evaluation methods. They were experienced evaluators, and about half of them had been involved in earlier usability studies of SARA. Their average age was 22. They were divided into two groups. To make the student evaluators aware of the requirements expert users pose on the system, the students were briefed with the feedback from the professional parliament library users that have used preliminary versions of SARA and the two campaign datasets. Two facilitators were available for all evaluation sessions.

9.4. Training

Before the main evaluation, all participants were given 20 minutes to access the online training material. There were three types of training material available. In SARA, an interactive visual guide to the interface elements was provided using pageguide.js [36]. The users could trigger it anytime in order to access information about specific sections of the interface. An overview presentation of the ARCOMEM digital preservation approach, as well as two courses, introductory and advanced, about the output of all modules were also available online. Most of the participants familiarised themselves with the SARA interface using the in-app guide. More than a half of the participants also spent the majority of the training time reading the advanced course material.

9.5. Search and Retrieval Scenarios

This evaluation included two scenarios: a guided scenario on the U.S. Elections 2012 campaign and a free stroll scenario for the investment for growth.

The first scenario was obligatory and directed to guide the users to specific actions in order to obtain their feedback. It ensured that the users searched using all the available ways and looked into enough web resources in order to familiarise themselves with the wealth of multi-perspective data available through SARA. The second was a free-stroll scenario advising the users to search items of interest in SARA using their intuition. The goal was to allow the users to explore their own search paths, become experts in specific tasks and reveal high-impact or potentially interesting types of data analysis.

The former was an explicit step-by-step task asking the users to perform specific actions to accomplish pre-determined search and retrieval tasks. The scenario was a severely revised, combined version of scenarios used in earlier usability testing sessions. Each step was designed to access specific type of metadata and triggered specific content to show on the interface. The users were required to voice if the presented metadata met their expectations, which metadata items were the most useful in their search so far, and if they agreed that the proposed next step of action was optimal.

Similar scenarios, but more targeted to the evaluation of the metadata quality, were constructed for the interaction with the investment for growth archive. The participants were asked to verify acceptance and assess potential usefulness of each specific metadata type. In this respect, the scenarios for both campaigns were complementary. That effectively means that specific key questions were included in both scenarios. That lead towards identifying the qualitative differences between the versions of the metadata generation modules.

For the investment for growth campaign, the users were asked to retrieve specific content. They were free to use any search path they deemed optimal. The participants could try several paths and report on the optimal ones. The discoverability of information, as well as assessment of the importance and usefulness of the content itself and the way it was made available and presented were measured both qualitatively and quantitatively.

All tasks were expected to be completed within 40 min. For the guided scenario, the tasks effectively guided the users into accessing metadata. The free-stroll scenario tasks assessed the efficiency and true potential of the interface and the overall acceptance of each specific metadata type.

9.6. Methods for Feedback Collection

Four ways of collecting feedback from the formal evaluation sessions were used:

- The records from the facilitators, who thoroughly observed the participants during their interaction;
- The subjective feedback from the participants using the complete online questionnaires;
- The recorded participant information on the intentions, expectations and actions during the interaction (think aloud);
- The interaction logs from SARA that recorded every interaction per user, providing information about direct feedback times and choices regarding search approaches.

The later was used to verify the feedback from the participants and facilitators. All feedback was collected during the evaluation sessions. The feedback from the participants and the facilitators was immediately available. The interaction logs were also retrieved and all feedback was then processed.

9.7. Sessions and Execution

Two sessions took place on the same day for the two groups of participants. They shared the same lab room and the same facilitators. Both session participants were given 20 min to optionally familiarise themselves more with the system using any of the three types of training material available. The facilitators also provided introductory presentations about the overview of the ARCOMEM approach, the available modules that analysed the collected data and the available tools for creating and accessing campaigns.

The first session was the guided scenario on the U.S. Elections 2012 dataset, and the second was the free task-oriented evaluation using the investment for growth campaign data. Both sessions lasted less than 40 min. The think-aloud protocol was used in both scenarios to collect the real-time user feedback on intentions and expectations from the data and the interface. The formal feedback from the users was collected via web forms right after each session. The system-level interaction logs were collected for both sessions. The maximum allocated time for each participant group was 60 min. The average completion time was 49 min.

10. Evaluation Results

Entity-based and topic-based search for the U.S. Elections 2012 campaign archive revealed the impact that well designed data visualization has over simple long lists of metadata. In fact, the combination of topics, entities and opinions in an interactive visual search resulted in completion of search tasks multiple times faster than the entity-based ones. The average completion times of the topic-based searches were, on average, more than eight times lower than the entity-based ones, as calculated by the system. This denotes the advantage of advanced visual analytics to information understanding, especially if complex semantic links are present in the data.

Entities and opinions were found to be the most useful metadata that the end users required. Entities were primarily used for querying the archives and refining the search within the SARA interface. Opinions were naturally bound to the entities. The requirements of that time stressed the need for exposing the opinions and their target entities, optimally extracting the opinionated sentences from the web documents and making them available for search.

Topics were individually evaluated during exploratory pluralistic walkthroughs for the U.S. Elections 2012 session. The participant list included all the evaluation users, as well as the members of the design team. Earlier evaluation session had taken place around the middle of the design life-cycle when the topic detection module became available. The goal was to measure the importance of topics within the scope of web archive analysis and subsequently explore novel ways for representing topics to an optimal value within the context of SARA. As a result, topics, just like entities, were present on the main result page, on the web resource page, as well as exclusive visual overview, for the purpose of topic-based searching. The results of this evaluation verified that the topic-based search was a significant asset of the SARA interface. The accuracy and descriptive nature of the results of the visual representation for the relation between topics, main entities per topic and opinions for those entities were very successful.

Topics were rated quite high in terms of importance in both datasets (Figure 5). The advanced topic visualisation was an experimental approach available for evaluation only for the U.S. Elections 2012

dataset. In the investment for growth evaluation, the topics were suggested by the SARA application only after the queries and subsequent user actions retrieved enough entities to partially describe a topic. Effectively, the users could not reveal the topics until they searched for a sufficient number of entities to identify at least one topic. On this dataset, the quality of the topics was perceived as worse (Figure 6), mostly due to the generic nature of the investment for growth dataset. Importance of topics was reported identical for both data set, while acceptance in the investment for growth was deemed much worse than on the U.S. Election 2012, due to both quality and the lack of the advanced topic visualisation feature connecting the topics with their context (*i.e.*, entities and opinions).

Figure 5. Perceived metadata importance.

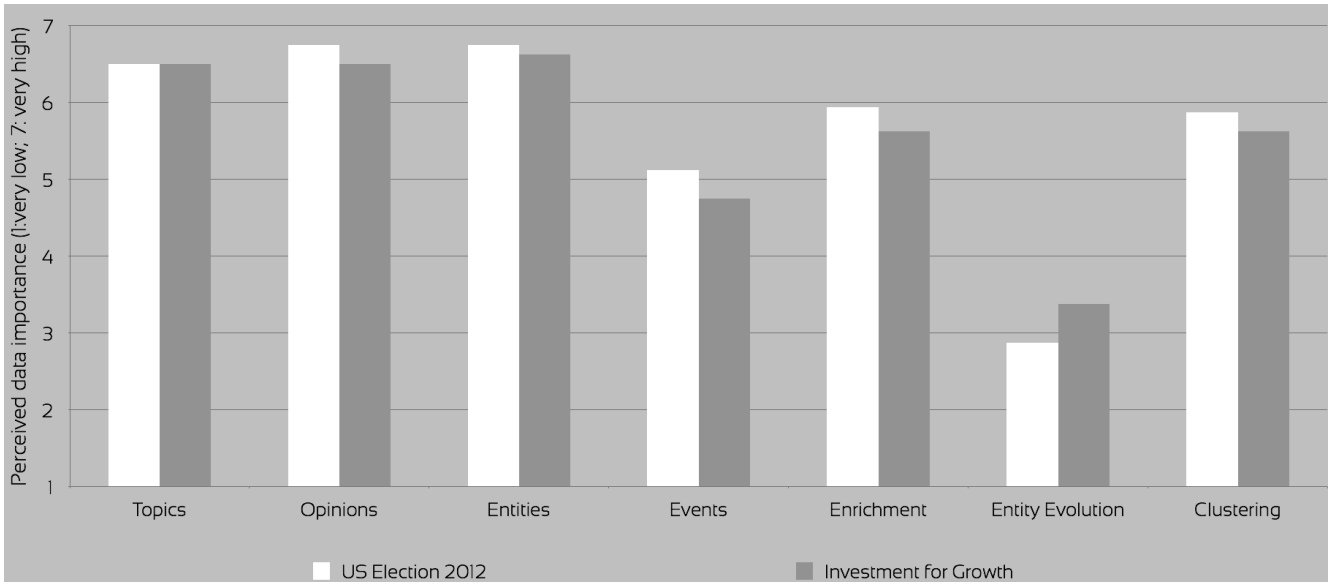
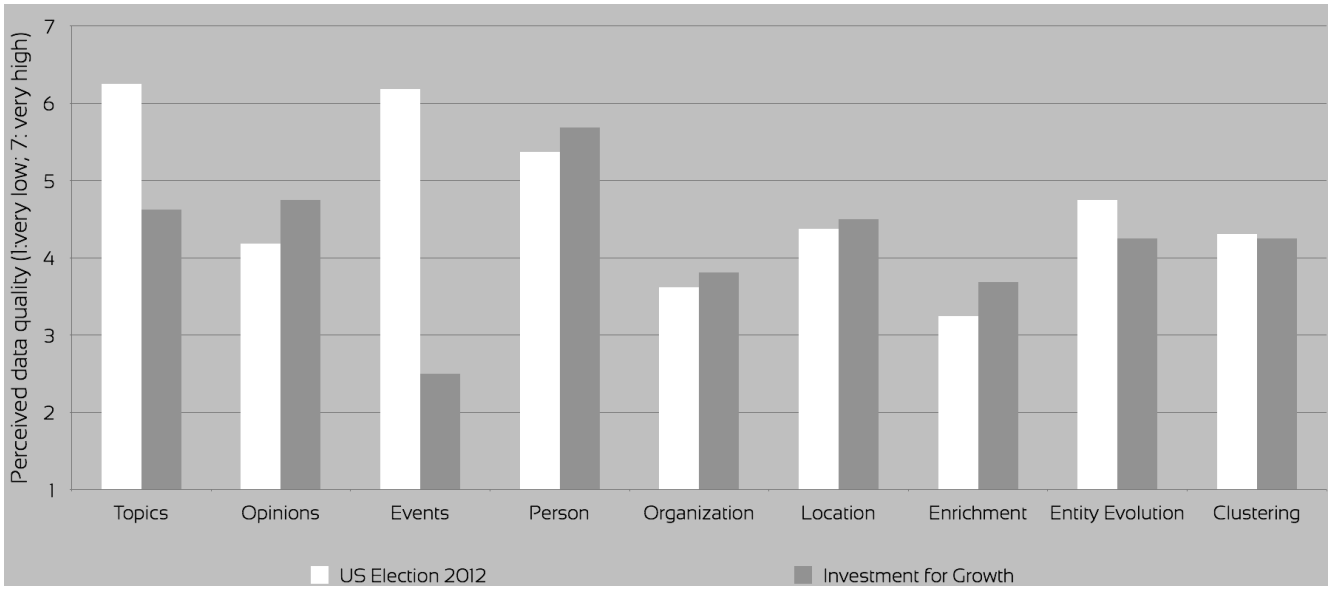


Figure 6. Perceived metadata quality.



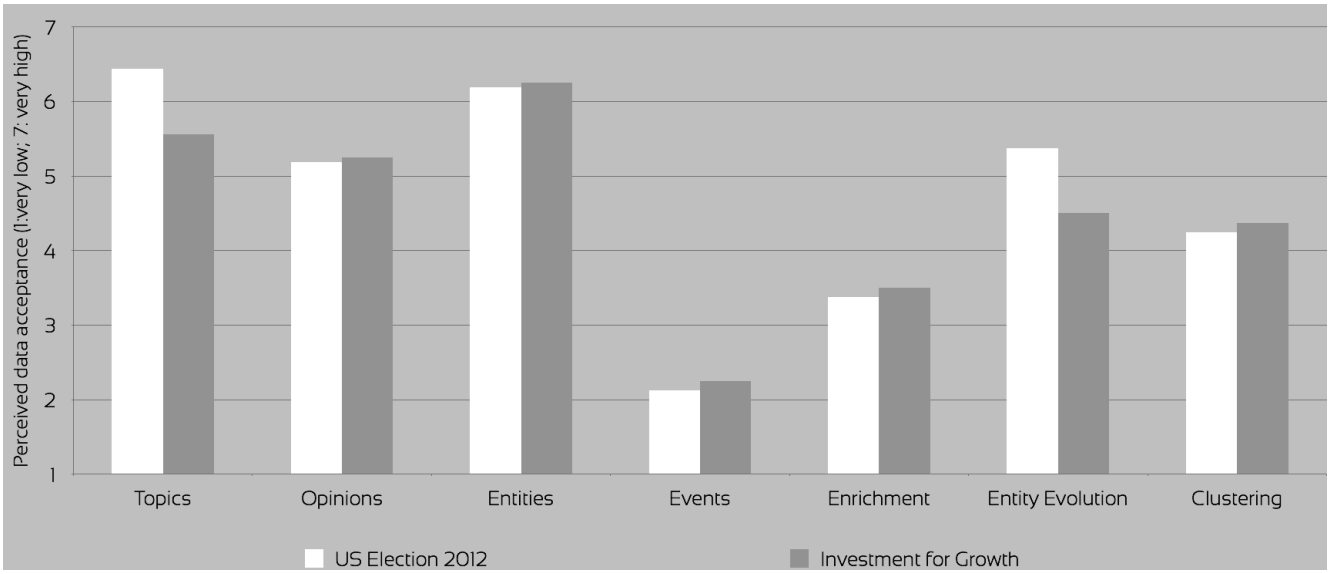
The ETOEs, enrichments and clusters were also fully evaluated in the investment for growth sessions. That dataset used the metadata generated by the final version of the ARCOMEM modules. Entities,

opinions and enrichments were found to be more accurate, while events much less so. A critical claim was that event titles were incomprehensible sometimes. Enrichment and clustering provided additional value for the entities that were closely examined for specific web resources (Figure 6). A major comment about enrichment was a suggestion for improvement. The users suggested that the enrichment could suggest the best resource at the top of the list, instead of providing the full list of the, sometimes, identical or supplementary resource links. Additionally, a few lines preview text from that resource could be included, with a link to a "read more" expanded panel. Regarding clustering, the users also commented that they felt that it could be more on the semantic level.

The feedback on the perceived importance of the data clearly indicated that topics, entities and opinions were the most important for the aforementioned tasks, followed by enrichment and clustering (Figure 5). The entity evolution—a query expansion technique identifying historical names of the entities [37]—although deemed quite accurate, was much less important, only used very infrequently during the evaluation. The reason for that was its limited use, which was confined to the initial text search.

Finally, the overall acceptance of the data for both data sets reflected either poor accuracy or intelligibility as a major factor for not accepting event detection (Figure 7). In our experiments, we found our event extraction mechanisms to perform around 65% F1, which partly accounts for this judgement. In addition, our events express fine-grained relations between event participants as occurring within sentences. Combined with the high number of extracted events of this nature, this hampers interpretability for the user. Because of the fine-grainedness, diversity and size of the extracted event information, the semantic cohesion of the metadata is easily lost. In order to remedy this, the number of events should be reduced in order only to cover participants that are important for the domain, *i.e.*, frequently occurring named entities and terms. This reduced set will be easier for the users to display and understand. Another way to reduce complexity is to group the events into more coarse-grained semantic categories. This will provide a semantically more intuitive and informative set of metadata for users while browsing, and more general event classes provide a more coherent and usable set for querying.

Figure 7. Perceived metadata acceptance.



Furthermore, in order to further increase the overall metadata acceptance, the users made suggestions to enhance visibility of the enrichment and clustering in the system (Figure 7)

11. Related Work

Over the last decade, researchers from different disciplines, such as sociology, politics and history, studied the importance of the web in the context of several application domains and discussed the necessity of web archiving to enable retrospective analysis of web content in these domains (e.g., [38–42]). At the same time, several projects have pursued web archiving (e.g., [43,44]). A large number of national libraries and national archives are now actively archiving the web as part of their heritage preservation mission. The Heritrix crawler [45], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet Preservation Consortium (IIPC) [46], is a mature and efficient tool for large-scale, archival-quality crawling.

Standard crawling methods aim to capture as much of the web as possible. In contrast, focused crawling [47] aims to only crawl pages that are related to a specific topic. Focused crawlers (e.g., [48,49]) learn a representation of the topic from the set of initial pages (seed URLs) and follow links only if the containing page matches that representation. Extensions of this model use ontologies to incorporate semantic knowledge into the matching process (e.g., [50]), "tunnel" between disjoint page clusters (e.g., [51]) or learn navigation structures necessary to find relevant pages (e.g., [52]). ARCOMEM complements existing approaches to long-term preservation by developing new ways into intelligent and adaptive decision support for content appraisal and selection for web archiving by leveraging the social web and taking events, entities and topics into account. Furthermore ARCOMEM preserves the semantic context of web objects by enriching web archives with the extracted information.

The social web provides an additional source of data for parliament applications. Many services, such as Twitter, YouTube or Flickr [53], provide through their APIs access to structured information about users, user networks and created content and are therefore attractive to researchers. Data collection from these services is not supported by standard web crawlers. Usually it is conducted in an *ad hoc* manner, although some structured approaches exist [54–56]. However, these platforms focus on API access and do not archive web pages linked from social web platforms. In the ARCOMEM project, the first approaches were investigated to implement a social- and semantic-driven appraisal and selection model for web and social web content.

A limited set of tools exists for accessing web archives, like NutchWAX [57] and Wayback [58]. NutchWAX provides URL and keyword-based access based on Apache Nutch [59], an open source web search project. Wayback is an open source implementation of the Internet Archive Wayback Machine. It allows browsing the history of a page or domain over time. Overall, the possibilities to explore web archives are limited to basic functionalities. In ARCOMEM, we developed and evaluated SARA—a search and retrieval application that enables users to navigate through the archived content using semantic information.

12. Conclusions

The ARCOMEM application presented in this paper targets an effective creation and exploration of political archives based on the web and social media. The metadata extraction and enrichment pipeline presented in this paper plays a key role for this application, providing a focused ETOE-centric view on the archived information. In this paper, we introduced the overall processing chain of the offline analysis in the ARCOMEM architecture and presented components for entity, event and topic extraction, as well as enrichment and correlation involved in this chain. Furthermore, we presented SARA—a search and retrieval application that uses generated metadata to facilitate archive exploration.

The user evaluation results revealed the most important types of semantic data that the evaluators deemed useful for the search and retrieval in digital archives. In summary, topics and associated entities were identified as the most important types, even more so when combined to provide an overview of the entire dataset. Topic-driven search was very well received and was the preferred starting point for most of the tasks that included exploration of data. Overall, the evaluation verified the ARCOMEM digital preservation approach regarding the provision of many types of semantic data that enables the successful exploration of political web and social media archives.

Acknowledgments

This work was partially funded by the European Commission under Grant Agreement No. 270239 (ARCOMEM), the European Research Council under ALEXANDRIA (ERC 339233) and the COST Action IC1302 (KEYSTONE). Nikolaos Papailiou has received funding from IKY fellowships of excellence for postgraduate studies in Greece, SIEMENS program.

Author Contributions

All authors contributed equally to the paper.

Conflicts of Interest

Thomas Risse and Wim Peters are co-editors of the special issue on Archiving Community Memories.

References

1. The ARCOMEM Consortium. ARCOMEM system release. Available online: http://sourceforge.net/projects/arcomem/ (accessed on 25 July 2014).
2. Dietze, S.; Maynard, D.; Demidova, E.; Risse, T.; Peters, W.; Doka, K.; Stavrakas, Y. Preservation of Social Web Content based on Entity Extraction and Consolidation. In Proceedings of the 2nd International Workshop on Semantic Digital Archives (SDA), Paphos, Cyprus, 27 September 2012; pp. 18–29.
3. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; Bizer, C. DBpedia—A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semant. Web J.* **2014**, in press.

4. Bollacker, K.D.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the ACM SIGMOD Conference (SIGMOD '08), Vancouver, Canada, 9–12 June 2008; pp. 1247–1250.
5. Nunes, B.P.; Dietze, S.; Casanova, M.; Kawase, R.; Fetahu, B.; Nejdl, W. Combining a co-occurrence-based and a semantic measure for entity linking. In Proceedings of 10th European Semantic Web Symposium (ESWS 2013), Semantics and Big Data, Montpellier, France, 26–30 May 2013.
6. Barbieri, N.; Bonchi, F.; Manco, G. Topic-aware social influence propagation models. *Knowl. inf. syst.* **2013**, *37*, 555–584.
7. Risse, T.; Dietze, S.; Peters, W.; Doka, K.; Stavrakas, Y.; Senellart, P. Exploiting the Social and Semantic Web for guided Web Archiving. In Proceedings of 2nd International Conference on Theory and Practice of Digital Libraries (TPDL 2012), Paphos, Cyprus, 23–27 September 2012.
8. Papailiou, N. H2RDF: adaptive query processing on RDF data in the cloud. In Proceedings of the 21st World Wide Web Conference (WWW 2012), Lyon, France, 16–20 April 2012.
9. Bontcheva, K.; Derczynski, L.; Funk, A.; Greenwood, M.A.; Maynard, D.; Aswani, N. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In Proceedings of the Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, 9–11 September 2013; pp. 83–90.
10. Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V.; Aswani, N.; Roberts, I.; Gorrell, G.; Funk, A.; Roberts, A.; Damljanovic, D.; *et al. Text Processing with GATE (Version 6)*; The GATE team: Sheffield, UK, 2011.
11. Derczynski, L.; Maynard, D.; Aswani, N.; Bontcheva, K. Microblog-genre noise and impact on semantic annotation accuracy. In Preceedings of 24th ACM Conference on Hypertext and Social Media (part of European Computing Research Congress), Paris, France, 2–4 May 2013.
12. Piskorski, J.; Yangarber, R. Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization. Volume in the Series: Theory and Applications of Natural Language Processing*; Poibeau, T.; Saggion, H.; Piskorski, J.; Yangarber, R., Eds.; Springer-Verlag: Berlin, Germany, 2013.
13. Masolo, C.; Borgo, S.; Gangemi, A.; Guarino, N.; Oltramari, A. WonderWeb Deliverable D18: Ontology Library. Available online: http://wonderweb.man.ac.uk/deliverables/documents/D18.pdf (accessed on 25 July 2014).
14. Shaw, R.; Troncy, R.; Hardman, L. LODE: Linking Open Descriptions of Events. In Proceedings of 4th Annual Asian Semantic Web Conference (ASWC 2009), Shanghai, China, 6–9 December 2009; pp. 153–167.
15. Scherp, A.; Franz, T.; Saathoff, C.; Staab, S. F-A Model of Events based on the Foundational Ontology DOLCE+DnS Ultralight. In Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009), Redondo Beach, CA, USA, 1–4 September 2009.
16. Raimond, Y.; Abdallah, S. The Event Ontology. Available online: http://motools.sourceforge.net/event/event.html (accessed on 25 July 2014).
17. NIST Multimodal Information Group. Automatic Content Extraction 2008 Evaluation (ACE08). Available online: http://itl.nist.gov/iad/mig/tests/ace/2008/, (accessed on 25 July 2014).

18. Akbik, A.; Visengeriyeva, L.; Herger, P.; Hemsen, H.; Löser, A. Unsupervised Discovery of Relations and Discriminative Extraction Patterns. In Proceedings of International Conference on Computational Linguistics (COLING 2012), Mumbai, India, 8–15 December 2012; pp. 17–32.
19. Xu, F., U.H.; Li, H. A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 584–591.
20. Heath, T.; Bizer, C. *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1*; Morgan & Claypool: City, Country, 2011; pp. 1–136.
21. Mendes, P.N.; Jakob, M.; García-Silva, A.; Bizer, C. DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, Graz, Austria, 7–9 September 2011; pp. 1–8.
22. Milne, D.N.; Witten, I.H. Learning to link with wikipedia. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), Napa Valley, CA, USA, 26–30 October 2008; pp. 509–518.
23. Hachey, B.; Radford, W.; Curran, J.R. Graph-Based Named Entity Linking with Wikipedia. In Proceedings of the 12th International Conference on Web Information System Engineering, Sydney, Australia, 13–14 October 2011; pp. 213–226.
24. Han, X.; Sun, L.; Zhao, J. Collective entity linking in web text: A graph-based method. In Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011), Beijing, China, 25–29 July 2011; pp. 765–774.
25. Marc Wick, Unxos GmbH. The GeoNames geographical database. Available online: http://www.geonames.org (accessed on 25 July 2014).
26. OCLC. The Virtual International Authority File (VIAF). Available online: http://www.VIAF.org (accessed on 25 July 2014).
27. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84.
28. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. mach. Learn. Res.* **2003**, *3*, 993–1022.
29. Minka, T.; Lafferty, J. Expectation-propagation for the Generative Aspect Model. In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, Alberta, Canada, 1–4 August 2002; pp. 352–359.
30. Steyvers, M.; Griffiths, T. *Latent Semantic Analysis: A Road to Meaning*; Laurence Erlbaum: Mahwah, NJ, USA, 2007.
31. The Apache Software Foundation. Apache Mahout. Available online: https://mahout.apache.org/ (accessed on 25 July 2014).
32. Cybozu Labs Inc.. Language detection library for Java. Available online: https://code.google.com/p/language-detection/ (accessed on 25 July 2014).
33. Twitter Inc. (US). Twitter. Available online: https://dev.twitter.com/ (accessed on 25 July 2014).
34. Google Developers. YouTube. Available online: https://developers.google.com/youtube/ (accessed on 25 July 2014).

35. Spiliotopoulos, D.; Tzoannos, E.; Cabulea, C.; Frey, D. Digital Archives: Semantic Search and Retrieval. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*; Holzinger, A., Pasi, G., Eds.; Springer-Verlag: Berlin, Germany, 2013; pp. 173–182.
36. AppNeta. Pageguide, an interactive guide for web elements using jQuery and CSS3. Available online: http://tracelytics.github.io/pageguide/ (accessed on 25 July 2014).
37. Tahmasebi, N.; Gossen, G.; Kanhabua, N.; Holzmann, H.; Risse, T. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In Proceedings of International Conference on Computational Linguistics (COLING 2012), Mumbai, India, 8–15 December 2012.
38. Jankowski, N.W.; Foot, K.; Kluver, R.; Schneider, S. The Web and the 2004 EP Election: Comparing Political Actor Web Sites in 11 EU Member States. *Info. Pol.* **2005**, *10*, 165–176.
39. Foot, K.A.; Schneider, S.M. *Web Campaigning*; The MIT Press: Cambridge, MA, USA, 2006.
40. *The Internet and National Elections. A Comparative Study of Web Campaigning*; Kluver, R., Jankowski, N., Foot, K., Schneider, S.M., Eds.; Routledge: New York, NY, USA, 2007.
41. Brügger, N. Web historiography and Internet Studies: Challenges and perspectives. *New Media Soc.* **2013**, *15*, 752–764.
42. Rogers, R. *Digital Methods*; MIT Press: Cambridge, MA, USA, 2013.
43. Arvidson, A.; Lettenström, F. The Kulturarw Project—The Swedish Royal Web Archive. *Electron. libr.* **1998**, *16*, 105–108.
44. Abiteboul, S.; Cobena, G.; Masanes, J.; Sedrati, G. A First Experience in Archiving the French Web. In Proceedings of 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002), Rome, Italy, 16–18 September 2002; pp. 1–15.
45. Mohr, G.; Kimpton, M.; Stack, M.; Ranitovic, I. Introduction to Heritrix, an archival quality web crawler. In Proceedings of 4th International Web Archiving Workshop (IWAW04), Bath, UK, 16 September 2004.
46. International Internet Preservation Consortium. Available online: http://netpreserve.org/ (accessed on 25 July 2014).
47. Chakrabarti, S.; van den Berg, M.; Dom, B. Focused crawling: A new approach to topic-specific Web resource discovery. *Comput. Netw.* **1999**, *31*, 1623–1640.
48. Diligenti, M.; Coetzee, F.; Lawrence, S.; Giles, C.L.; Gori, M. Focused Crawling Using Context Graphs. In Proceedings of 26th International Conference on Very Large Data Bases (VLDB 2000), Cairo, Egypt, 10–14 September 2000; pp. 527–534.
49. Aggarwal, C.; Al-Garawi, F.; Yu, P.S. Intelligent crawling on the World Wide Web with arbitrary predicates. In Proceedings of the Tenth International World Wide Web Conference, Hong Kong, China, 1–5 May 2001; pp. 96–105.
50. Dong, H.; Hussain, F.K. SOF: A semi-supervised ontology-learning-based focused crawler. *Concurr. Comput. Pract. Exp.* **2013**, *25*, 1755–1770.
51. Qin, J.; Zhou, Y.; Chau, M. Building domain-specific Web collections for scientific digital libraries. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2004), Tucson, AZ, USA, 7–11 June 2004; pp. 135–141.
52. Jiang, J.; Song, X.; Yu, N.; Lin, C.Y. FoCUS: Learning to Crawl Web Forums. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 1293–1306.

53. Flickr API. Available online: http://www.flickr.com/services/api/ (accessed on 25 July 2014).
54. Boanjak, M.; Oliveira, E.; Martins, J.; Mendes Rodrigues, E.; Sarmento, L. TwitterEcho: A Distributed Focused Crawler to Support Open Research with Twitter Data. In Proceedings of the 21st World Wide Web Conference, Lyon, France, 16–20 April 2012; pp. 1233–1240.
55. Psallidas, F.; Ntoulas, A.; Delis, A. Soc Web: Efficient Monitoring of Social Network Activities. In *Web Information Systems Engineering 2013*; Springer: Berlin, Germany, 2013; pp. 118–136.
56. Blackburn, J.; Iamnitchi, A. An architecture for collecting longitudinal social data. In Proceedings of IEEE Workshop on Beyond Social Networks: Collective Awareness, Budapest, Hungary, 9–13 June 2013; pp. 184–188.
57. Internet Archive. NutchWAX. Available online: http://archive-access.sourceforge.net/projects/nutch/, (accessed on 25 July 2014).
58. Internet Archive. Wayback. Available online: http://archive-access.sourceforge.net/projects/wayback/, (accessed on 25 July 2014).
59. The Apache Software Foundation. Apache Nutch. Available online: http://nutch.apache.org/, (accessed on 25 July 2014).