Successful 2nd International Alexandria Workshop

The second Alexandria Workshop took place in L3S Research Center on 2-3rd November 2015. The workshop was aimed at bringing together communities involved in web archiving, digital preservation, digital humanities and information retrieval to encourage a closer dialogue between researchers from computer science, digital humanities and cultural heritage institutions. It was widely attended from participants from national libraries, humanities to computer scientists from varying disciplines like Information retrieval, natural language processing, database systems and distributed systems. The workshop, spanning two days, included two keynotes, several research talks, system demonstrations and a panel discussion on shortcomings, research infrastructures, and future directions.

The first keynote, given by Prof. Wolfgang Nejdl from the L3S Research Center focused on the challenges and solutions surrounding searching and exploring Web archives. Putting in context some of the findings in the BUDDAH project, which underlined the need for combining qualitative and quantitative analysis of archives he surmised the need for novel and better access methods for Web archives. He argued that better access methods are not only useful for searching archived collections, but are potentially useful for corpus creation, which is a fundamental task for historians and researchers in humanities. Keyword search, widely studied in the area of information retrieval, is a natural and easier access method but is fraught with uncertainty of keyword generation especially for advanced search tasks. Also, the intent of the user searching archives is markedly different from the traditional search behavior of end users. To this extent, the keynote talk touched on the recent contributions made in the Alexandria Project, which he is a principal investigator in, towards devising novel retrieval models for such specialized search and exploratory behaviors.

Complementing the need for improved access methods, other talks were targeted towards application specific usability of Web archives. Ivana Marenzi discussed the possibility of improving usability of improve the user experience when working with web archive collections. Focus was also on exploiting news archives for a variety of tasks ranging from Wikipedia enrichment, event-based ranking, to timeline summarization. Prof. Maarten De Rijk, also acknowledged the potential of Web archives was acknowledged in promising results from evolving text collections used in vocabulary shifts over time. Elisabeth Niigemamm, from German National Library, gave an insight about how the library is into Web harvesting for focused events and collections. Websites and their content become relevant objects when cited by researchers and thus citations need permanency to ensure validity and reproducibility of research results. She however expressed concerns about “link rot” which makes links unreliable and transient.

Prof. Niels Brügger from the Digital Media Lab (Aarhus University) gave the second keynote about Web History, Web archives, and Web Research Infrastructure. He highlighted the challenges faced in setting up such infrastructure and shared experiences about RESAW -- A Research Infrastructure for the Study of Archived Web Materials, and NETLAB -- An internet research infrastructure within the Danish research infrastructure for the humanities Digital Humanities Lab. Similarly, Thomas Risse also talked about the SoBigData project which creates a research
infrastructure providing an integrated ecosystem for ethic-sensitive social data mining.

In addition to these wide ranging set of talks, there was a demo session showcasing the proof-of-concept systems developed in L3S relating to the topics of interest to the workshop. Finally, the workshop concluded by a panel discussion on the issues that affects archives in general. A key challenge that was identified was legality of openness and data sharing. In the second year of the workshop we saw tangible research results and prototype systems emerging validating the real potential of Web archives and temporal collections. Although some concerns regarding data sharing, web infrastructures and data persistence still remain; the fact that an interdisciplinary set of researchers could come together and discuss about the possibilities, challenges and opportunities made it a success.

A full overview about the Alexandria project can be found on http://alexandria-project.eu/