What Do You Want to Collect from the Web?*

Thomas Risse, Elena Demidova, and Gerhard Gossen

L3S Research Center and Leibniz University of Hanover, Germany
{risse, demidova, gossen}@L3S.de

Abstract. Today an increasing interest in collecting, analyzing and preserving Web content can be observed in many digital humanities. Especially the Social Web is attractive for many humanities disciplines as it provides a direct access to statements of many people about politics, popular topics or events. In this paper we present an exemplary study that we have conducted with the aim to understand the needs of scientists in social sciences, historical sciences and law with respect to the creation of Web archives.

Keywords: User Study, Web Crawling, Research Data

1 Introduction

Since 1996 when the Internet Archive started collecting the Web, Web harvesting and archiving became an increasingly interesting topic. In the beginning Web archiving was mainly driven by libraries and archives with the aim to preserve the Web out of cultural necessity. Due to the sheer size and the rapid growth of the Web, capturing was at that time focused on crawling and archiving rather than on the access and usage of the content.

Today an increasing scientific interest in using the Web content can be observed. Digital humanities, at the intersection of humanities and computer science, is moving more and more towards the Web not only for publishing but also as a source of information. With Web Science a new scientific discipline emerged where the Web itself became the object to be studied. Documenting the user activities on the Social Web in Web archives allows getting a better understanding for example of the perception of topics, evolution of crises or the propagation of information. It should be noted that even though Web archives have great potential for research, they also raise legal and ethical issues about privacy and copyright that need to be respected during crawling and Web archive access.

For the creation of Web archives for research purposes some common requirements across scientific disciplines following good scientific practice as well as discipline dependent requirements can be observed. Common requirements are authenticity, citability and provenance. The way how this is implemented in different scientific disciplines differs depending on how they work with the Web. For example, on the one hand, Law scholars want to cite in their own work

* This work is partly funded by the European Research Council under ALEXANDRIA (ERC 339233).

different comments about regulations found in blogs. But this requires that blog entries remain unchanged and stay accessible. On the other hand, historical sciences and social sciences are more event driven and more interested in the social context. In this case the entire data within the Web archive builds the context for each specific document and the provenance of the archive as a whole needs to be documented.

In this paper we present an exemplary study that we have conducted with the aim to better understand the needs of scientists in social sciences, historical sciences and law with respect to Web crawling. The study has been conducted as part of the project "iCrawl: The Integrated Focused Crawling ToolBox"[1] which develops a novel Web archive crawler that specifically addresses the needs of researchers in various disciplines. The study has been performed by interviewing six members of the scientific disciplines at the Leibniz University of Hanover, Germany, and GESIS (Leibniz-Institute for the Social Sciences) and by developing exemplary use cases. This study gives important first insights into the needs but cannot be comprehensive as many sub-disciplines and specializations exist. During the interviews it also turned out that there is often no complete vision on how to work with Web content. Therefore the study can only be a starting point for the discussion and the development of first crawler implementations. We expect that needs will increase with the availability of the technology and new ideas will come up.

2 Use Cases

In the following we present the summary of the interviews that we conducted with researchers from the social sciences, historical sciences and jurisprudence.

2.1 Historical Sciences

Historical sciences study the past and how it relates to humans. Traditionally they work with limited information or information fragments and try to reconstruct from these the whole picture. It turned out that the acceptance of the Web and Web archives as a source of information differs among countries. While these sources are more commonly accepted e.g. in North America and Britain [10], they are rarely used in Germany as pointed out by the participants of our study.

The reasons why Web sources are rarely cited by German historians are diverse. First of all, Web sources are mostly anonymous, non-permanent and lack unique identifiers for specific, unalterable versions. Given frequent changes within the page content coupled with the lack of proper archival mechanisms, citation of Web sources does not appear feasible and is normally avoided, raising the requirement of citation-enabling archiving of Web resources. Furthermore, a lack of trust in the data quality on the Web largely prevents usage of Web data by historians in Germany.

[1] http://www.l3s.de/projekte/intelligenter-zugang-zu-informationen/projekt/article/icrawl/

The necessity of preservation of historically relevant Web content is already well-recognized in other countries [10]. In this context, official pages, such as government information, online media and other journalistic sources appear to be good starting points for archiving the Web for historians because they have a known provenance.

An important challenge in the context of Web archiving for historians is the identification of the topics of interest for future historians in today's Web. As historical importance of the topics and documents can only be judged after at least 25 years, some estimates are needed. One possible estimate of topic importance is an increased media attention. The assumption is that the topics attracting a lot of attention today (e.g. financial crisis) are likely to be relevant for historical research in the future. Further targets of interest can be multicultural, political or controversial topics. Optimally, the topics selected for archiving in the context of history-oriented use cases should be continuously observed.

Authenticity of content is also an important requirement. For historians it is important to see the content as it was presented at crawl time. This includes the layout, pictures and videos but also embedded advertisements.

2.2 Social Sciences

The aim of the social sciences is to understand society and the relationships among individuals within the society. The importance of digital data analysis for social research has been recognized especially in the sub-field of digital sociology [8]. In contrast to the historians' use case where Web data collection is performed well in advance as the topics of interest are not completely known yet, sociologists can define topics of their interest in today's Web and collect and analyze the data on these topics directly.

Starting points could be regular shallow observation crawls of major news sites for identifying interesting topics for research. Once the topic has been identified, focused crawls around the topic in official, journalistic sources as well as Social Media should be conducted. Especially with respect to controversial topics, such as immigration, education and social inequality, the collected resources can represent the diversity of opinions on the topic in various online sources thus facilitating sociological research in this area. In addition, topic changes over time should be recognized and actively followed during the data collection process.

Similar to the historian's use case the context of collected information plays a very important role for sociologists. To enable sociological research, the context and metadata need to be carefully documented. For example, it is important to know who is authoring the information provided and which interests this organization or person has. With respect to user generated content, personal data such as profession, gender, location, or political affiliation of the users play an important role to enable data interpretation by sociologists. For the interpretation and verification of the results obtained from social networks it is necessary to have also demographic information like the size of a social network, the age distribution of the participants, etc.

Furthermore, an important factor to achieve acceptance of Web-based research in digital sociology is transparency and detailed documentation of the data collection process. It has to be clear when a page has been collected and which strategy has been used to guide the crawler.

2.3 Law

Research on law is mainly conducted based on official sources like law publications and official protocols from parliaments or comments released by publishers. However, many lawyers and interested laymen blog their comments and discuss law related issues in the Social Media. Even though many interesting information can be found, Web sources are mainly used to get background information but rarely cited in publications, studies, etc. The reasons are the same as in the historical sciences, namely the citability and authenticity of the source.

Another interesting aspect for law research is the documentation of the genesis of laws as well as law changes. The genesis can be used to understand the original intention of the law at a later point in time and why certain decisions and formulations have been taken. The degree of genesis documentation is different among countries. While there is for example a good documentation on the European level, there is no comprehensive documentation available on the German national level even though individual minutes of official parliament and committee meetings are published and are freely accessible. Part of a good genesis of law is also the documentation of the public discussion. The general public plays an important role in the development, evolution and instantiation of law. Therefore in democratic systems a complete, transparent and sustainable documentation of the genesis of law is important to understand the diversity in arguments from different political parties and individuals.

3 Derived Requirements

From the use cases presented in Section 2 we derive in the following the requirements for the crawl specification and the user workflow.

Topical Dimension: Current Web archives are typically created using *Web crawlers*, i.e. programs that follow links between Web pages and download all the pages they traverse (e.g. Heritrix [11]). In the discussion with the researchers the need for focused crawls [3], i.e. crawls that only contain pages related to a specific topic, was regularly highlighted. The discussed crawl intentions mostly focused around events, sometimes also around entities like important persons. Both social sciences and historical sciences have the need to observe official and major journalistic sites for events or to identify trends.

From the technical perspective, the focusing of the crawl can be done by keywords or using more complex semantic descriptions. The challenge is to find the right terms that cover most of the interesting content without knowing everything about a topic and its representation on the Web in advance. Semi-automatic tools that analyze a sample of relevant pages could help researchers to identify important terms to create good crawl specifications. The focus of the

crawl may also need to change as an event unfolds or a topic evolves. Therefore the researcher must be able to constantly inspect and monitor the crawled pages and if necessary limit, expand or shift the focus of the crawl.

Time Dimension: The time dimension of crawls covers the aspects of the start time, the duration and the termination of a crawl. The start time in the use cases were mostly event driven but a concrete time point is not always available. Whereas for sudden events (e.g. a terrorist attack) the time of the event is the best start time, for planned events the time before the event is also important. For example in the case of elections it is interesting to see how the public opinion is influenced by the political parties. Therefore the start time depends a lot on the research plans and the crawl intention.

The crawl duration and termination play an important role with state of the art crawlers since there is no formal stop criterion except that the crawler queue (i.e. the list of the pages to be visited) is empty. Therefore crawls are typically stopped after a certain number of pages has been crawled or the crawl ran for some time. Semantic information adds a possible new stop criterion: the content relevance. The average relevance of a crawled page wrt. the crawl specification decreases over time when most of the relevant pages have been crawled. So a threshold can be given at which the crawler automatically stops. Alternatively a fast quality analysis could also give the researchers insights into the data quality and allow them to stop the crawl explicitly at any point.

Flexible Crawling Strategies: In the use cases, two types of crawls have been highlighted. On the one hand, social and political scientists expressed their interest in regular shallow observation crawls of news sites for identifying interesting topics for research. On the other hand, whenever the researcher identified a topic of interest, the need for focused crawls gathering information relevant for this topic was regularly highlighted by all disciplines. Both scenarios require different strategies for focusing, guidance and prioritization of the crawler.

For regular observation crawls that aim to identify potential topics of interest, a breadth-first strategy traversing all outgoing links of the pages seems appropriate to get diverse content. Other strategy to prioritize important Web pages in this context is the ordering based on PageRank [4], an interlinking-based measure of page importance in the Web graph. After identifying an interesting topic or event, focused crawling strategies that take the semantics into account for prioritization of the pages are the better strategy.

Social Web Crawling: In the interviews special interest has been shown in collecting content from Social Media sites. Most of the Social Web sites support the access to content by using application programming interfaces (API). APIs focus on the core content (e.g. a tweet, a Facebook message) and allow to filter content regarding keywords, users and other application specific properties. As the functionalities offered by each site are different and are a subject to change, application specific API crawlers need to be implemented and maintained for each site of interest.

Authenticity: Authenticity was described in the interviews as being able to see a Web page or a Social Media message in the same way as the targeted reader

at the crawl time. Authenticity raises a number of issues in the dynamic Web environment. One important issue for the authenticity is that a Web page has to be crawled within a short time interval [16]. This means that all information that are necessary to construct a page e.g. images, style sheets or embedded objects are crawled possibly together with the main page.

Context and Provenance: For better interpretation of analysis results it is crucial to also have demographic information such as the size and age distribution of each community in a Social Web site. These numbers are often found in press releases or official reporting e.g. in conjunction with financial reports. Once the pages or site which provide the necessary information are identified they can be archived together with the main content. For access purposes they should also be annotated as additional information that does not belong to the core research content. Furthermore, researchers from different disciplines underlined that documentation of the crawl specification and crawl process as well as unique identification and verification methods for collected Web documents are required to ensure re-usability and citability of the collected resources.

4 User Workflow

In order to turn a research idea into a Web archive the researcher has to perform a number of steps that can be categorized into the *Crawl Preparation*, *Crawl Execution* and *Crawl Finalization* phases.

Crawl Preparation: The quality of the final Web archive depends to a large extent on the quality of the crawl preparation. For the user, the crawl preparation phase includes the following steps: (i) Selection of crawling strategy: focused or observation crawl; (ii) Definition of data sources: media sites, Social Web, video sites, etc.; and (iii) Definition of the time frame of the crawl including start and end time as well as frequency (e.g. once, repeating, continuous).

In addition, for focused crawls: (iv) Description of the crawl intention by specifying relevant keywords and entities; (v) Identification of a number of relevant Web pages that are close to the intention of the crawl; (vi) Validation of the crawl intention model learned by the crawler.

Crawl Execution: During this phase the user can perform regular monitoring of the crawl results and an incremental refinement of the crawl specification if necessary. The monitoring of the crawl quality can be facilitated using aggregated views of the content and intuitive visualization techniques (e.g. a tag cloud). From the user point of view, the crawl execution phase includes the following steps: (i) Start the crawl; (ii) Monitor the quality of the crawl; and (iii) If necessary, refine the crawl specification.

Crawl Finalization: For the user, after the crawl termination, the crawl finalization includes the following steps: (i) Final validation of the Web archive quality; (ii) Export of crawl results to WARC files for archiving and preservation; and (iii) Cataloging and registration of archives for long term preservation (if required by the use case).

5 State of the Art

Web archives are typically created using *Web crawlers*, i.e. programs that follow links between Web pages and download all the pages they traverse. This method is derived from search engine crawling [9]. Several projects have pursued Web archiving (e.g., [1]). The Heritrix crawler [11], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet Preservation Consortium (IIPC)[2], is a mature and efficient tool for large-scale, archival-quality crawling. However, such crawlers are seldom feasible for the use of individual researchers, as they require extensive domain knowledge in how to set up and conduct a crawl.

Standard crawling methods aim to capture as much of the Web as possible. In contrast, *focused crawling* [3] aims to only crawl pages that are related to a specific topic. Focused crawlers [e.g. 5, 12] 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 [6], 'tunnel' between disjoint page clusters [14] or learn navigation structures necessary to find relevant pages [5, 7].

The Social Web provides an additional source of data for Web Science researchers. Many services such as Twitter, Youtube or Flickr 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 [2, 13]. However, these platforms focus on API access and do not archive Web pages linked from Social Web platforms.

The ARCOMEM project [15] implemented first approaches for a social and semantic driven appraisal and selection model for Web and Social Web content. However, special requirements for science and research have not been taken into account.

6 Conclusions

In this paper we presented an exemplary study that we have conducted with the aim to better understand the needs of scientists in social sciences, historical sciences and law with respect to Web crawling. Although the study cannot be comprehensive, it already uncovers many common requirements from these disciplines, but also highlights issues specific for each of them. The results also unveil special needs and usage scenarios for Web archives. Especially the reliable citability of Web resources has been mentioned quite often. In the historian use case also the need for active preservation for content became obvious since content from today will be used in 25 years or later. Based on the requirements identified in this study we are currently implementing the iCrawl system which will be a flexible crawler framework that is easily extensible and enables an integrated way of crawling Web and Social Web content.

[2] `http://netpreserve.org/`

7 Acknowledgments

We would like to thank Tina Krügel and her team at the Institut für Rechtsinformatik, Phillip Nordmeyer from the Historisches Seminar, Axel Philipps from the Social Sciences (all at Universität Hannover) and Katrin Weller from GESIS for their kind support.

References

1. A. Arvidson and F. Lettenström. The Kulturarw Project – The Swedish Royal Web Archive. *Electronic library*, 16(2), 1998.
2. J. Blackburn and A. Iamnitchi. An architecture for collecting longitudinal social data. In *IEEE ICC'13 Workshop on Beyond Social Networks: Collective Awareness*, pages 184–188, June 2013.
3. S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.
4. J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *World Wide Web Conference*, pages 161–172, 1998.
5. M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused crawling using context graphs. In *Conference on Very Large Data Bases*, pages 527–534, 2000.
6. H. Dong and F. K. Hussain. SOF: a semi-supervised ontology-learning-based focused crawler. *Concurrency and Computation: Practice and Experience*, 25(12):1755–1770, 2013.
7. J. Jiang, X. Song, N. Yu, and C.-Y. Lin. Focus: Learning to crawl web forums. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1293–1306, June 2013.
8. D. Lupton. *Public Sociology: An Introduction to Australian Society*, chapter Digital Sociology: An Introduction. Allen & Unwin, 2012.
9. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
10. I. Milligan. Rethinking the archival box; Herrenhausen Lightning Talk on Historians and Web Archives. `http://ianmilligan.ca/2013/12/09/lightning_talk/`, 2013. [Online; accessed 2014-02-10].
11. G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to Heritrix, an archival quality web crawler. In *4th Int. Web Archiving Workshop (IWAW04)*, 2004.
12. G. Pant and P. Srinivasan. Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems*, 23(4):430–462, Oct. 2005.
13. F. Psallidas, A. Ntoulas, and A. Delis. Soc web: Efficient monitoring of social network activities. In *Web Information Systems Engineering 2013*, pages 118–136. Springer, 2013.
14. J. Qin, Y. Zhou, and M. Chau. Building domain-specific web collections for scientific digital libraries. In *Joint ACM/IEEE Conference on Digital Libraries, 2004*, pages 135– 141. IEEE, June 2004.
15. T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic web for guided web archiving. In *Theory and Practice of Digital Libraries*, volume 7489 of *LNCS*, pages 426–432. Springer, 2012.
16. M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. "Catch me if you can": Visual analysis of coherence defects in web archiving. In *Int. Web Archiving Workshop (IWAW 2009)*, Sept./Oct. 2009.