Web Archive Retrieval Tools (WebART)

University of Amsterdam
National Research Center for Mathematics and Computer Science (CWI)
National Library of the Netherlands (KB)

NWO/Catch Program 2010

Abstract

The advent of the Web has had a revolutionary impact on how we acquire, share and publish information. In fact, it has a fundamental impact on our daily lives that increasingly take place "on the Web." However, this increasing dependence on the Web comes at a price: the ease of publishing on the Web also results in the easy loss of information—Web content tends to be ephemeral. This project addresses the problem of our *future* cultural heritage. Globally this has been addressed head on by the Internet Archive, now supplemented by many national initiatives. Web archiving has so far concentrated primarily at preservation, and much less on the use of the archived Web material. However, the chosen selection and archiving strategies of Web material will have a crucial impact on their future value as cultural heritage. Hence there is an urgent need to understand how the Web archive will be used, not only in superficial exploration, but especially as the object of study of scientific researchers.

The crucial issue, and the main research problem of this proposal, is to critically assess the value of Web archives for realistic research scenarios, and develop information access tools and methods that maximize the archive's utility for research. Our approach is to conduct actual Web archive research hand-in-hand with the development of Web archive access tools tailored to the realistic research scenarios. Within the project, our focus is on the use-case of a humanities researcher, although tools will no doubt be useful for other use-cases as well.

1 Description of the Proposed Research

1.a) Scientific Aspects

Scientific Problem The advent of the Web has had a revolutionary impact on how we acquire, share and publish information. In fact, it has a fundamental impact on our daily lives that increasingly take place "on the Web." However, this increasing dependence on the Web comes at a price: the ease of publishing on the Web also results in the easy loss of information—Web content tends to be ephemeral. Hence, as the UNESCO [31, art.3,4] recognized:

> *The world's digital heritage is at risk of being lost to posterity. ... Attitudinal change has fallen behind technological change. ... The threat to the economic, social, intellectual and cultural potential of the heritage – the building blocks of the future – has not been fully grasped. Unless the prevailing threats are addressed, the loss of the digital heritage will be rapid and inevitable.*

This project addresses the problem of our *future* cultural heritage. Globally this has been addressed head on by the Internet Archive, now supplemented by many national initiatives such as the archiving of the Dutch Web space at the *Koninklijke Bibliotheek*, the National Library of the Netherlands. Web archiving has so far concentrated primarily at preservation, and much less on the use of the archived Web material. This is entirely understandable, since the preservation of Web pages is a necessary requirement for their later use. However, the chosen selection and archiving strategies of Web material will have a crucial impact on their future value as cultural heritage. Hence there is an urgent need to understand how the Web Archive will be used, not only in superficial exploration, but especially as the object of study of scientific researchers.

The crucial issue, and the main research problem of this proposal, is

- *to critically assess the value of Web archives for realistic research scenarios, and develop novel information access tools and digital research methods that maximize the archive's utility for research.*

Our approach is to conduct actual Web archive research hand-in-hand with the development of Web archive access tools tailored to the realistic research scenarios. Concretely, we will address the following research questions:

1. Which methods of research are privileged by current Web archives, and which are precluded? What research assumptions are embedded into current Web archiving practices and access tools?

2. Which access tools may be build on top of the archive, supporting humanities scholars to use emerging digital research methods?

3. How to engage researchers in Web Archive enrichment, and harness their knowledge of local collections?

The focus on future cultural heritage is irrevocably tied to the problem of appraisal—what is the heritage value of objects to preserve—similar to the selection criteria of archives, libraries, and contemporary art galleries. We regard cultural heritage in a broad sense, e.g., both covering the high arts as well as popular culture. The current selection criteria of the Dutch Web archive focus on content that has clear heritage value, such as Web sites of cultural institutions, science, politics and government.

Research Method The focus on novel methodology for the humanities makes the current CATCH round by far the most ambitious round so far. The main complexity is that the change toward using digital methods is also a change of research paradigm [see also 23]. Traditional humanities research is not based on the classical truth-finding paradigm of the natural sciences and the empirical sciences. It is highly individualistic and interpretative paradigm in which parallel truths or "realities" can be constructed and are judged on the insight they provide. The most common critique, as voiced by traditional scholars in the humanities, is that they don't recognize their traditional paradigm in the current work on digital methods in the humanities.

The emerging digital humanities put distinctly more emphasis on empirical evidence, and transcends into neighboring disciplines such as the (interpretative) social sciences [e.g., 4]. This resembles the empirical focus of the natural sciences, yet the success criterion didn't change—it's about finding many interpretations that provide insights, rather than uncovering the single underlying truth. It is clear that new and unprecedented opportunities arise due to the large-scale availability of heritage material in digital form in our memory institutions. It is also clear that this makes great potential contributions to the interpretations and insights we have from the traditional methods.

What then, in essence, is needed to tackle the problem of digital methods, and information access tools that support these digital methods?

- Systems that allow a humanities scholar, without particular knowledge of ICT methods or data formats, to execute these methods by her- or himself. This is the major barrier to success, although the current tech-savvy generation of young humanities students are hopeful candidates to engage with digital methods.

- Systems allowing them to start with a simple question or query and iteratively refine it into a complex question or search strategy, while at every stage being able to explore the consequences—it is about the interpretation or insight it provides. So rather than a rigid query to answer retrieval system, the system should be a highly interactive, exploratory search system [34].

- The used digital methods, which when stored as search strategies, can be used, reused and shared among different scholars. In addition, the stored search strategies and their process of construction are explicitly available in logs and a valuable resource for further analysis.

The proposed project aims to deliver exactly that. That is, we work with the use-case of a national Web archive using the massive data collection crawled and maintained by the National Library of the Netherlands. To give a feel of its magnitude: in 2010 (up to December 9th), we crawled 66 million files (48 million text files) covering 3.5 Terabytes of data uncompressed. This is more than twice the volume that we crawled in 2009, and we expect this exponential growth to continue over the coming years.

The driving force of the project will be the actual use of the archived Web for research purposes. One of the most obvious candidates are New Media researchers that take the Web as an object of study [28, 33]. We focus here on the Web as unique medium, and investigate a range of questions in Web epistemology—an area of study where the main claim is that the Web is a knowledge culture distinct from other media.[1] The project will essentially set up a Living Lab for Web Archive research [20].

How should we support them? We are currently developing systems that support scholars to formulate complex search strategies [6]—a complex "query" covering a whole search episode (i.e., a whole task) rather than a single navigational query (i.e., Google style 2.4 words)—and to store and share these strategies. With this system, a new media scholar interested in national Webs can do so by formulating an appropriate search strategy (say, a selection of sites; with a projection of a particular field, value, or relation; and how it evolved over time, or compared to another selection). Currently, individual scholars have to spend days or weeks to script and screen-scrape one particular slice of this data. Our proposed DB/IR approach uses a formalism based on a probabilistic relational model that encapsulates both information retrieval and traditional database queries [see also 5] and supported by a rich user interface to empower end-users to carry out complex search tasks [following 6]. In order to deal with the volume of data, we need to deal with terabytes of highly structured data by developing a distributed, Hadoop-based implementation.

This has high potential gain, but what about the risk? Clearly even relatively simple search strategies may look highly complex to the uninitiated. Still we believe there are enough securities put into place. First of all, also traditional methods require substantial learning effort, and provided there is sufficient payoff, we would expect a scholar to be willing to invest time and effort in learning the new methods, step-by-step. Second, the system will support iterative strategy development, e.g., starting with a simple keyword query, the result may be suggesting an interesting selection criterion based on a facet, which can be added as a condition to the search strategy, etc. Third, it is key to build on the wisdom of the crowd by reusing previous search strategies, when stored and shared, to educate other scholars to conceive the potential of particular digital methods, apply them to different use-cases, or extend and refine them to offer different interpretations. Some excellent examples of use-cases are at [7, 32].

As Howell [11] has it "gathering evidence of prior versions of Web sites should be performed in a careful forensic manner with cognizance of the underlying technology used in the archiving process." In this light, it is interesting to

[1] Although our focus is on the use-case of a humanities scholar, in particular in new media studies, the resulting tools will no doubt be useful for other use-cases within and beyond the humanities.

observe that the core of the exact same system we will apply in the humanities area has found a practical application in digital forensics [1].

Desired Results The main desired results of the proposed project are:

- to put the problem of the loss of digital heritage, or the cultural heritage of the future, on the map within CATCH;
- to critically evaluate and provide concrete recommendations on curatorial practices and selection decisions of current Web archives—which are now characterized by ad hoc choices;
- to review curatorial practices from the viewpoint of archival appraisal and selection (rather than from the viewpoint of library collection development), and especially in the light of their cultural and historical value for (future) use in humanities research;
- to gain insight in the use of Web Archives by humanities scholars, in particular those working on new media and digital culture, resulting in new digital methods;
- to build a system that supports scholars to formulate complex search strategies [6]—a complex "query" covering the whole task rather than a single, 2.4 word, navigational query—and to store and share these strategies;
- to reuse previous search strategies when stored and shared, to educate other scholars to conceive the potential particular digital methods, apply them to different use-cases, or extend and refine them to offer different interpretations;
- to let the search strategies, or particular digital methods, evolve in parallel with their use and user population in a living lab for Web archive research [20]; and
- to scale our proposed DB/IR approach to deal with terabytes of highly structured data—powerful enough to handle any cultural heritage collection.

Related Research An encyclopedic overview of related research is beyond the scope of the proposal. We focus on Web archiving here, and briefly discuss some of the main institutions, Web archive access tools, and Web archive use cases. We also discuss of relevant related techniques.

An excellent overview of all aspect of current Web archiving can be found in Masanès [22]. Web archiving started over a decade ago at the Internet Archive [16]. This was followed by many other institutions, often national libraries focusing on national Webs, or institutions focusing on specific topics. An example is the archiving of the Dutch Web by the National Library of the

Netherlands [19]. Since 2003, these institutions are collaborating in the International Internet Preservation Consortium [12]. In 2004, the European Archive [8] was founded. Since 2001, the annual *International Web Archiving Workshop* (IWAW) has been held, where research on all aspects of Web archiving is published. Although the main emphasis over the last decade has been on Web preservation, there has been also important progress in access tools [9]. Users are being addressed explicitly in the IIPC Access Working Group (with Marcel Ras and René Voorburg as active members). Clear use cases have been defined in [15], some of which deal with research use of the archive, as well as prototypes of access tools [14]. Currently, the IIPC is working on requirements for Web archive access [13].

In 2001 the Internet Archive introduced the Wayback Machine, allowing users to look up specific URLs and browse through archived versions of Web pages across time. Hence, access is limited to known URLs, and no keyword based search is available. Recently the NutchWAX full-text search engine is developed, based on a standard Web search engine adapted to Web archives data formats. Allowing full-text search greatly enhances access to archives. In terms of the retrieval functionality, NutchWAX and Nutch rely on Lucene, which deploys a standard vector-space model. Tools for managing collections of Archived Web pages are also emerging, the Internet Archive offers the facility for registered users to bookmark pages, and Archive-it offers more facilities to partner institutions to create collections based on seed pages and by adding metadata to the seeds and the overall collection. The proposed project will significantly extend such tools and will tailor it to the needs of researchers. The results are based on open-source software, and will be made available through the IIPC.

The 2011 IIPC General Assembly will be organized at the National Library of the Netherlands, with already substantial involvement of the applicants in the opening event "Out of the Box: Building and Using Web Archives" where, inspired by the current proposal, Web curators meet with the researchers using Web archives. At this meeting, we will present one of the case studies that prompted the current collaboration and research proposal: a study of the Dutch Blogosphere based on an early compilation of Dutch blogs in 2001 [10]. This study required dedicated crawling and screen-scraping of data from the Internet Archive, and development of various specific post-processing tools. The proposed project will offer comprehensive solution that allow for complex selections on various criteria (beyond an explicit list of sites) where each selection is also a layer of annotation feeding back into the archive. The results contain many sites (rather than the Wayback Machine's single site) which can be explored on various dimensions and facets, including the time and sites based on the link structure (without the need for ad hoc tools and methods). Also new facets can be introduced (e.g., specific blog features like blogrolls or RSS feeds) based on formulating appropriate selection criteria that demonstrate the appearance and popularity, or demise, of such features over time or other dimensions. In short, the proposed tools allow such case studies to be conducted and shared as complex search strategies within a single Web archive research system.

1.b) Multidisciplinary cooperation

Concrete application The project's main scientific questions are inseparably connected to the concrete application, Web Archiving. That is, rather than doing technology-driven research that requires substantial operationalization and embedding to be deployed in practice, the insights and tools developed will be directly integrated with the application context of the Web Archive. Moreover, since we will build on the emerging international standards for Web Archives (as developed within the IIPC), they will also be directly applicable at numerous other Web archives.

The proposal's topic is remarkable in that the different members in the project team study the same research object and share essentially the same research questions. Just to mention some obvious examples: The cultural heritage partner is involved in the large-scale crawling of Web content—an activity commonly associated with computer science. Likewise, the humanities partner is actively gathering Web data from search engines and Web archives, and developing tools to visualize the network structure of the Web—again activities more commonly associated with computer science. Of course, the technology partner is also building tools, but is at the same time actively involved in user studies and evaluation efforts. Hence the project will result in in-depth cross-disciplinary collaboration, where scholars from different research traditions make important contributions.

Concrete Role of the Cultural Heritage Institution The heritage partner provides the optimal stage for the proposed research. In terms of data, a huge collection of archived Web sites is available and the collection is growing at an increasing rate. There is a mirror of the data available at separate servers that allow the development of Web archive access tools without endangering the operational Web preservation systems. In fact, the current state-of-the-art access tools—such as the Heretrix crawler and the Wayback Machine—are already available, and we have experimented with NutchWAX (Nutch with Web archive extensions) and WERA interface (combining Nutch-based search and Wayback Machine like browsing) [27]. This will greatly facilitate the proposed research by allowing it to focus, from the very first day, on the research questions of researcher-use of Web archives.

1.c) Relevance

Scientific Significance and Impact The proposed project makes a number of contributions in the *humanities*, mainly in terms of novel digital methods and in terms of the new media studies based on the Web archive, and to *computer science*, mainly the innovative approach to interactively construct complex queries and explore the results using various facets or dimensions, as well as the resulting back-end. More generally, it will contribute to cultural heritage and Web archiving. First, and foremost, it may lead to reconsideration and refinement of current Web archiving methods, that will lead to more effective selection strategies that may preserve future heritage that would otherwise inevitably be lost.

Web Archives form by themselves a novel form of cultural heritage institutions—the name "archive" signals the relations with traditional archives, but the most Web Archives refer to themselves as "digital libraries" signaling the relation with traditional libraries. Since the traditional boundaries between libraries and archives are eroding in the digital age, studying Web Archives provides insight into the heritage institutions of the future. Web Archives are digital libraries, making the insights and tools of the project are applicable in digital libraries, especially when they scale to the level that exceeds traditional organization through explicit metadata, and are also applicable in the cataloguing systems of traditional research libraries. Web Archives are archives, making the insights and tools of the project are applicable in digital archives. The main research questions have a direct bearing on theory and practices of selection and appraisal, and the project's approach focusing on the future heritage value of "Web records" is a showcase of the archival records continuum [24]. The proposed researcher-support environment significantly extends related efforts in archival science, such as [35].

Even if Web material is preserved its sheer quantity will create substantial barriers to use, and makes it impossible for Web Archives to provide the substantial metadata that is the main means of organization in digital libraries. Actively involving researchers themselves to provide such organization and descriptive metadata can provide valuable additional retrieval cues, and help other researchers to explore the Web Archive. To draw the parallel with the archival world: historical researchers rely on peers and citations as primary means of identifying collections [2, 30]. Hence, such contributed descriptions may turn out to be the crucial factor in bringing out all the hidden treasures of the Web Archive.

1.d) Research Utilization

Our research utilization partners are a public institution, the National Library of the Netherlands, and an industrial partner, Spinque, `http://spinque.com/`.

A recent start-up company Spinque is the ideal partner to make the results of the project available to a wider audience, and a broader range of heritage data sets. Spinque is a spin-off company specializing in the DB/IR technology we apply in the project. Prof. De Vries is co-founder of Spinque, ensuring seamless communication and transfer of knowledge between the proposed project and the industrial party.

The National Library of the Netherlands is committed to make the archived Web available to the general public by 2013 latest, and access from within the institution starting in 2011. By then a monumental amount of data will be available, and we are working on ways to deal with the legal barriers to publishing the archived Web on a public server [3].

The project will deliver operational prototypes, with a functional UI. Spinque could turn the project's prototype into a fully-fledged operational service, using their experience in optimization of high volume Web-sites, and in user interface (UI) and user experience (UX) design.

This is a two-way street: a public portal for the archived Dutch Web will greatly benefit the proposed project. Recall, that the used digital methods are stored as search strategies to be used, reused and shared among different scholars. In addition, the stored search strategies and their process of construction are explicitly available in logs and a valuable resource for further analysis.

In addition, we can apply the project results to other heritage data sets, such as the catalogue descriptions of National Library (with millions of book records in the field of humanities), with only limited changes to the server backend, and to the interface front-end—demonstrating their generic applicability. This will benefit the project substantially by allowing for a comparison of search strategies of scholars working on literature search, or on Web archives.

2 Description of the Proposed Plan of Work

The project is greatly facilitated by the availability of basic versions of all the needed components. These need to be adapted to the scale and demands of Web archives—both with respect to data, and with respect to the types and needs for digital methods in web archiving and new media research. This leads to three work-packages:

WP1: Distributed Structured Querying In this work-package, we scale our infrastructure to the level needed for Web-scale retrieval. WP1 will be the main responsibility of the *Research Programmer*.

WP2: Complex Search Systems In this work-package, we set up a living lab for Web Archive research and implement access tools aiming for incremental query or search strategy construction and interactive result exploration. WP2 will be the main responsibility of the *PhD Student*.

WP3: Applied Digital Methods In this work-package we conduct a range of case studies of Web archive based research in media studies, study novel digital research methods, and critically reflect upon the chosen selection strategies. WP3 will be the main responsibility of the *Postdoctoral Researcher*

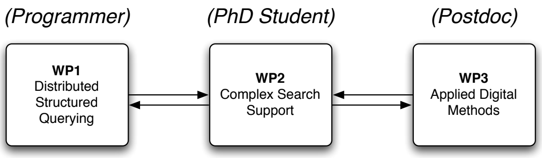

Figure 1: Work-packages and relations.

Figure 1 shows the dependencies between the work-packages: WP2 builds on the infrastructure of WP1, and WP3 uses the tools made available by WP2. Conversely, feedback and needs from WP3 will factor in WP2 resulting in new or improved access methods, and the demands from WP2 in terms of scale and efficiency will feed into WP1.

The global planning of the project (Gantt-style) is as follows:

Work-package	Year 1	Year 2	Year 3	Year 4
WP1: *Distributed Structured Querying*	T1.1			
		T1.2		
WP2: *Complex Search Systems*	T2.1			
		T2.2		
WP3: *Applied Digital Methods*	T3.1			
		T3.2		

The project will essentially set up a Living Lab for Web Archive research, and follow a spiral development cycle with an initial, operational system available within the first year which is enhanced and refined in the rest of the project.

WP1: Distributed Structured Querying

WP1 starts with Task 1.1 in which a selection of a few 100 million Web pages is indexed using the currently available MonetDB/PFtijah infrastructure. The main work will be in Task 1.2 in which a distributed version of MonetDB/PFtijah is implemented using the Hadoop framework.

Task 1.1: Initial Storage Structure with Powerful Querying

Goals: Set up an initial search system allowing for powerful structured search (XQuery with various text retrieval extensions) using the off-the-shelf, open source MonetDB [25] storage structure with relational SQL, full XQuery, and PFtijah [26] text retrieval extensions. This allows for defining particular retrieval strategies (WP2) in a declarative manner.

Data: Current system can deal with millions of pages, depending on CPU/memory available, and hence a selection of available Web Data can be indexed. We focus on one of the key selections of the Dutch Web Archive: several thousands of dutch Web sites in the cultural domain, including popular culture.

Evaluation: Main evaluation will be in terms of its use to support novel user interfaces and to support researchers in action using the project's Living Lab (WP2). In addition, the TREC Web track uses a Web crawl in WAR-format, with an appropriate evaluation suite, that can be used for quantitative evaluation and comparisons to other state-of-the-art methods.

Dependencies: Task 1.1 will feed into Task 2.1 (WP2) and Task 1.2 (WP1).

Risks: Risk is minimal since all required data is available, and only existing software is used.

Task 1.2: Distributed Probabilistic Relational Database

Goals: Port the existing MonetDB relational database backend to the open-source Hadoop framework for data-intensive distributed applications. This will allow for almost unlimited scaling over a cluster of relatively cheap hardware, while retaining the powerful querying using relational SQL, full XQuery, and PFTijah text retrieval extensions.

Data: The resulting infrastructure will allow for indexing all data, provided a sufficiently large cluster is available.

Evaluation: Again, the main evaluation will be in terms of its use to support novel user interfaces and to support researchers in action using the project's Living Lab (WP2).

Dependencies: Task 1.2 will get input from Task 1.2 (WP1) and Task 2.2 (WP2), and feed into Task 2.2 (WP2).

Risks: There is no special budget to install a large cluster of storage and computing nodes. While this does not affect the development of the infrastructure, given the expected exponential growth of the Dutch Web Archive it may necessitate to limit experiments to a selection of the available data. The additional budget for research utilization may be used, in part, for hardware costs whether on-site or in the clouds.

WP2: Complex Search Systems

WP2 starts with Task 2.1 in which a Living Lab for Web Archive research is set up, providing an experimental environment in which the access tools can be implemented and iteratively refined. It continues with Task 2.2 in which support for complex search strategy construction and result exploration are developed.

Task 2.1: Living Lab for Web Archive Research

Goals: Set up a Living Lab for Web archive research. Enriching Web page and aggregated Web site representations based on the Web archives selection criteria and derived metadata (esp. dates, links and anchor text). Use the initial search system (WP1) to test-drive state-of-the-art Web retrieval methods for pages, sites and entities [e.g., 17, 18, 21] and integrate them with non-topical relevance (temporal and site dimensions). Develop initial building blocks for complex search strategies (such as site and temporal selections both for the whole collection, or for specific queries).

Data: The initial system and selection of Dutch Web Archive from WP1. Standard retrieval models are available from PFtijah [26], which can be adapted to the data and case at hand. Initial case studies (WP3) will suggest building blocks for complex queries or search strategies for Web archive research, which can be stored and shared amongst researchers.

Evaluation: Retrieval effectiveness can be evaluated against bench-marks from the TREC Web Track. Main evaluation of resulting systems is through the case studies of WP3.

Dependencies: Task 2.1 will get input from Task 1.1 (WP1) and Task 3.1 (WP3), and will feed into Task 3.1 (WP3) and Task 2.2 (WP2).

Risks: Risk is minimal since all required data and tools are available.

Task 2.2: Complex Querying and Result Exploration

Goals: Rather than a rigid query to answer retrieval system, we will build a highly interactive, exploratory search system, allowing searchers to explore the results by various facets or dimensions, such as the relevance, times, sites, (aggregated) link-structure, notable entities, and any other contributed metadata. The system supports searchers to construct a complex search query or search strategy by starting with a simple question or query and iteratively refining it by exploring the results through various facets suggesting the which semantic structures are useful for their particular search problems.

Data: The basic system allows for highly powerful querying which we will turn into modular building blocks, and develop user interfaces hiding low-level details [6, 18]. Available metadata can be used for faceted searching and browsing.

Evaluation: Methods will be developed in parallel with use-case and experiments (WP3) in the project's Living Lab.

Dependencies: Task 2.2 will get input from Task 2.1 (WP2), Task 1.2 (WP1), and Task 3.2 (WP3). Task 2.2 will feed into Task 1.2 (WP1) and Task 3.2 (WP3).

Risks: Complex search strategies have an unavoidable inherent complexity that may hamper adoption. We address this by the iterative construction of search strategies, and the sharing of effective strategies, as well as the development of user interfaces that hide much of the complexity.

WP3: Applied Digital Methods

WP3 starts with Task 3.1 in which initial case-studies of Web archive research are done based on ongoing work in Rogers' `http://www.digitalmethods.net/`.

It continues in Task 3.2 in which the new opportunities offered by the powerful search tools are explored in advanced case-studies.

Task 3.1: Initial Case Studies

Goals: Identify typical research questions for Web archive research, and conduct initial case studies formulating experiments in whole or in part as appropriate search strategies, say, a selection of sites; with a projection of a particular field, value, or relation; and how it evolved over time, or compared to another selection.

Data: Current research [7] focuses on seven areas: the link structure (social network analysis), the Web site (single-site/page histories), the search engine/Web archive (ranking and selection determinants), the spheres (Web sphere, Blogosphere, Open sphere), the national Web (geographical analysis), social networking sites (demographics and profiles), and Wikipedia (evolving control).

Evaluation: Through experiments on the project's Living Lab (WP2) that demonstrate the utility of the tools, and insight in the corresponding best practices. The main external success criterion is that it leads to insights into emerging digital research methods the humanities.

Dependencies: Task 3.1 will receive input from Task 2.1 (WP2) and feed into Task 2.1 (WP2).

Risks: Risk is minimal since it is based on straightforward extensions of ongoing research [7].

Task 3.2: Advanced Digital Methods

Goals: Further substantive studies of increasing complexity, exploring the diachronic evolution of sub-collections (politics, media, culture, etc.) or pairs of sub-collections (high arts versus pop-culture, traditional versus social media, left and right wing politics, etc), or the broken Web (what links/content is blocked from the Archive and why) and the lost Web (accumulating indirect evidence from links and textual references); and so on. Explore novel Web archive research methods in parallel with evolving support tools in the project's Living Lab (WP2). Critically assess the selection strategies and its underlying assumptions, similar to [29].

Data: The tools emerging from WP2 will open up new opportunities for research in Task 3.2, which will in turn raise new questions suggesting novel search strategies, changes in the user interface, or ways of annotation/sharing/reusing data.

Evaluation: Again, through experiments on the project's Living Lab (WP2) leading further tools and established digital research methods. This should lead to insights and relevant publications in the humanities, both in terms of the Web and Web archives and in terms of methodology.

Dependencies: Task 3.2 will get input from Task 3.1 (WP3) and Task 2.2 (WP2), and feed into Task 2.2 (WP2).

Risks: Some use-cases may require access tools unique to the case at hand. The intent is to develop only generic access tools in WP2, but allow for exporting data in various formats to allow for dedicated visualisations or post-processing methods.

Training and education are aimed at developing scientific expertise, and acquiring professional competencies. With respect to the scientific expertise, the PhD student should become able to fully understand, critically analyze, and contribute to research at the frontiers of science. The PhD student has two supervisors, who together draw up a training and education plan, which is revised (if necessary) annually; this is a highly individual plan, taking into account the particular background of the PhD candidate. Towards the end of their first year, the PhD student will write a detailed proposal for their thesis research.

As to professional competencies, we actively work to equip our students to function well in the professional environment of a university or research institute. This includes a range of academic skills, such as the ability to communicate research ideas clearly and effectively, the ability to pursue research individually, as well as being able to function as a team member. Practically, as part of the day-to-day activities, the applicants are an avid proponents and practitioners of an "open door" environment in which informal meetings occur naturally and frequently.

Finally, we will organize a range of related workshops and conferences in the coming years, including the SIGIR 2011 Workshop on Supporting Complex Search Tasks in Beijing; the CIKM 2011 Workshop on Exploiting Semantic Annotations in Information Retrieval in Glasgow; the INEX 2011 Workshop on Focused Retrieval from Structured Documents in Saarbrücken; the Fifth Digital Methods Summer School 2011 in Amsterdam (on Data-rich Media); and the fourth Information Interaction in Context (IIiX) Conference in 2012 in the Netherlands. These events will provide the project employees with unique opportunities for looking-under-the-hood of the field, and for extending their network of peers.

3 Literature

[1] W. Alink, R. A. F. Bhoedjang, P. A. Boncz, and A. P. de Vries. XIRAF – XML-based indexing and querying for digital forensics. *Digital Investigations*, 3:50–58, 2006.

[2] I. Anderson. Are you being served? historians and the search for primary sources. *Archivaria*, 58:81–129, 2004.

[3] A. Bleicher. A memory of Webs past. *IEEE Spectrum*, March 2011. http://spectrum.ieee.org/telecom/internet/a-memory-of-webs-past.

[4] P. Cohen. Digital keys for unlocking the humanities riches. *The New York Times*, 2010. http://www.nytimes.com/2010/11/17/arts/17digital.html.

[5] R. Cornacchio, S. Héman, M. Zukowski, A. P. de Vries, and P. Boncz. Flexible and efficient IR using array databases. *VLDB Journal*, 17(1):151–168, 2008.

[6] A. P. de Vries, W. Alink, and R. Cornacchio. Search by strategy. In *Proceedings of the Third Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2010)*, pages 27–28. ACM Press, New York USA, 2010.

[7] DMI. Digital Methods Initiative, 2011. http://digitalmethods.net/.

[8] European Archive. Web preservation and digital cultural access, 2011. http://europarchive.org/.

[9] T. Hallgrimsson. *Access and Finding Aids*, chapter 6, pages 131–151. In , Masanès [22], 2006.

[10] A. Helmond and E. Weltevrede. Mapping the Dutch Blogosphere with the Internet Archive. In *Out of the Box: Building and Using Web Archives*, 2011. Open Session of the 2011 IIPC General Assembly.

[11] B. Howell. Proving web history: How to use the Internet archive. *Journal of Internet Law*, pages 3–9, February 2006.

[12] IIPC. International internet preservation consortium, 2011. http://netpreserve.org/.

[13] IIPC Access Working Group. Web archive access requirements. Technical report, International Internet Preservation Consortium, 2008. Draft.

[14] IIPC Access Working Group. Prototypes related to IIPC access working group use cases. Technical report, International Internet Preservation Consortium, 2006. Version 1.

[15] IIPC Access Working Group. Use cases for access to internet archives. Technical report, International Internet Preservation Consortium, 2006. Version 1.

[16] Internet Archive. Universal access to human knowledge, 2011. http://archive.org/.

[17] J. Kamps. Web-centric language models. In *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 307–308. ACM Press, New York NY, USA, 2005.

[18] R. Kaptein, P. Serdyukov, A. P. de Vries, and J. Kamps. Entity ranking using Wikipedia as a pivot. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010)*, pages 69–78. ACM Press, New York USA, 2010.

[19] KB. Web archiving at the national library of the netherlands, 2011. http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html.

[20] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *Computer*, 42:60–66, 2009.

[21] M. Koolen and J. Kamps. The importance of anchor-text for ad hoc search revisited. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129. ACM Press, New York NY, USA, 2010.

[22] J. Masanès, editor. *Web Archiving*. Springer, 2006.

[23] W. McCarty. Beyond retrieval? computer science and the humanities. *CATCH Midterm Event, Den Haag*, 2007.

[24] S. McKemmish and M. Piggott, editors. *The Records Continuum: Ian Maclean and the Australian Archives First Fifty Years*. Ancora Press, 1994.

[25] MonetDB. Database system with XQuery front-end, 2011. http://www.monetdb-xquery.org/.

[26] PFtijah. Flexible open source text search system, 2011. http://dbappl.cs.utwente.nl/pftijah/.

[27] M. Ras and S. van Bussel. Web archiving: User survey. Technical report, National

Libary of the Netherlands, 2007.

[28] S. Schneider and K. Foot. The web as an object of study. *New Media & Society*, 6:114–122, 2004.

[29] M. Thelwall and L. Vaughan. A fair history of the web? examining country balance in the internet archive. *Library & Information Science Research*, 26:162–176, 2004.

[30] H. R. Tibbo. Primary history in America: How U.S. historians search for primary materials at the dawn of the digital age. *American Archivist*, 66:9–50, 2003.

[31] UNESCO. Charter on the preservation of the digital heritage. United Nations Educational, Scientific and Cultural Organization, 2003.

[32] WebArchivist. University of washington and the suny institute of technology, 2011. http://webarchivist.org/.

[33] E. Weltevrede. Thinking nationally with the web: A medium-specific approach to the national turn in web archiving. Master's thesis, Department of Media Studies, University of Amsterdam, 2009. http://wiki.digitalmethods.net/pub/Dmi/DmiSummer09/weltevrede_national_webs.pdf.

[34] M. L. Wilson, B. Kules, m. c. schraefel, and B. Shneiderman. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010.

[35] E. Yakel, S. Shaw, and P. Reynolds. Creating the next generation of archival finding aids. *D-Lib Magazine*, May/June 2005. http://www.dlib.org/dlib/may07/yakel/05yakel.html.