A webarchiválás válogatott bibliográfiája összefoglalókkal

Szerkeszti: Németh Márton <nemeth.marton@oszk.hu>

Frissítés dátuma: 2018.09.11.

Beinert, T. (2017). Webarchivierung an der Bayerischen Staatsbibliothek. (German). *Web Archiving at the Bayerische Staatsbibliothek. (English)*, *51*(6), 490. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The Bayerische Staatsbibliothek has been collecting and archiving websites dealing with regional studies and science since the year 2010. The article provides a survey of the collection and archiving profiles of the Bayerische Staatsbibliothek concerning websites, the legal basis, the workflow which has been developed as well as the registration and making available of websites in the archives. Finally, further perspectives for the future are presented. (English) [ABSTRACT FROM AUTHOR]

Boruna, A. E., & Rahme, N. (2011). Arhivarea paginilor Web – inițiative relevante de păstrare a patrimoniului digital european. *Biblioteca Nationala a Romaniei. Informare si Documentare*, *4*, 39–52,. Retrieved from https://search.proquest.com/docview/1443688144?accountid=27464

Brakker, N. V., & Kujbyshev, L. A. (2013). The Experience of the National Libraries Abroad of the Collection and Longterm Preservation of Internet Resources. *Bibliotekovedenie [Library and Information Science (Russia)]*, (2), 88–96. https://doi.org/10.25281/0869-608X-2013-0-2-88-96

A review of National Libraries experience of WEB harvesting, archiving technologies and legal issues. The paper suggests an overlook of experience and experiments of National Libraries of Austria, Germany, China, Lithuania, the Netherlands, New Zeeland, Northway, Portugal, United Kingdom, USA, Finland, France, Czech Republic and Sweden.

Buel, J. W. (2018). Assembling the Living Archive: A Media-Archaeological Excavation of Occupy Wall Street. *Public Culture*, *30*(2), 283–303. https://doi.org/10.1215/08992363-4310930

The article discusses the issues behind the social protest called Occupy Wall Street (OWS) that was staged in Zuccotti Park, Manhattan, New York in September 2011. Also cited are the efforts to archive the movement to preserve its history in a decentralized online archive, as well as the efforts by the OWS Archives Working Group in the archival process.

Careless, J. (2013). Archiving Web Content: An Online Searcher Roundtable. *Online Searcher*, *37*(2), 44–46. Retrieved from https://search.proquest.com/docview/1417518328?accountid=27464

In a roundtable discussion, several executives shared their views about archiving web content. Library of Congress' Office of Strategic Initiatives leader Abbie Grotke said the Library's web archiving project preserves web content around events, such as the US National Elections or September 11, or related themes such as public policy topics or the US Congress. They also archive their own Web site at loc.gov. Las Vegas-Clark County Library District virtual library manager Lauren Stokes said they archive their video and audio content in a variety of media. They use local server storage, portable hard drive backups as well as CD backups. Server storage is also backed up on tapes rotated into cold storage. Boston Public Library's director of administration and technology David Leonard said their digitization efforts are focused on accessibility. Web portal accessibility -- whether as part of their own web presence or the positing of materials to other Internet sites as well as some social media sites -- all help with accessibility. Adapted from the source document.

Duke, J. (2013). Internet Archive, Reed Tech Agree. *Advanced Technology Libraries*, *42*(12), 6–7. Retrieved from https://search.proquest.com/docview/1622279345?accountid=27464

Internet Archive and Reed Technology and Information Services Inc., part of the LexisNexis family, have agreed to jointly market and sell Internet Archives Archive-It service and continue to support the growing community of organizations currently using the service. First launched at Internet Archive in early 2006, Archive-It has been providing a sophisticated and flexible solution to a broad range of organizations and institutions focused on creating and managing collections of Web content. Adapted from the source document.

Duncan, S. (2015). Preserving born-digital catalogues raisonnés: Web archiving at the New York Art Resources Consortium (NYARC). *Art Libraries Journal*, *40*(2), 50–55. Retrieved from https://search.proquest.com/docview/1693347798?accountid=27464

Dupuis, C. (2017). Web-Archivierung an der Saarländischen Universitäts- und Landesbibliothek (SULB). (German). *Web Archiving at the Saarland University and State Library (Saarländische Universitätsund Landesbibliothek, SULB). (English)*, *51*(6), 529. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Since 2008, websites are being archived at the Saarland University and State Library (SULB). The repository SaarDok is available for archiving electronic documents concerning regional studies. A legal basis for depositing non-physical works exists since December 2015. (English) [ABSTRACT FROM AUTHOR]

Geisler, F., Dannehl, W., Keitel, C., & Wolf, S. (2017). Zum Stand der Webarchivierung in Baden-Württemberg. (German). *Web Archiving - the Present Situation in Baden-Württemberg. (English)*, *51*(6), 481. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article describes the present situation of web archiving at the level of the Land of Baden-Württemberg. The essential legal basis for collecting and archiving websites and related single documents exists. At the same time, the selection of contents and technical realization of the offer is an extensive task that is shared among several state institutions. (English) [ABSTRACT FROM AUTHOR]

Gibby, R., & Brazier, C. (2012). Observations on the development of non-print legal deposit in the UK. *Library Review*, *61*(5), 362–377. https://doi.org/http://dx.doi.org/10.1108/00242531211280487

Purpose - The process of developing and implementing UK legislation for the legal deposit of electronic and other non-print publications has been lengthy and remains incomplete, although the Government has consulted on draft regulations for implementation in 2013. The purpose of this paper is to provide a short account of progress and review the experience, analysing several factors that have influenced the legislative process and helped shape the proposed regulations. It summarises the regulatory and non-regulatory steps taken by the UK legal deposit libraries to address the legitimate concerns of publishers and describes some of the practical implications of implementing legal deposit for non-print publications. Design/methodology/approach - The paper draws upon the personal experiences of the authors, who have been directly involved in the legislative process and negotiations with publishers and other stakeholders. Findings - The paper provides new information and a summary of key issues and outcomes, with explanations and some insights into the factors that have influenced them. Originality/value - This paper provides new information about the development of legal deposit in the UK and a review of the issues that have affected its progress.

Gmerek, K. (2012). Web Archives on Both Sides of the Atlantic Ocean -- Internet Archive, Wayback Machine and UK Web Archive TT - Archiwa internetowe po obu stronach Atlantyku Internet Archive, Wayback Machine oraz UK Web Archive. *Biuletyn EBIB*, (1). Retrieved from https://search.proquest.com/docview/1266143228?accountid=27464

The article is a comparison of two web archives -- from the US and UK -- which differ in terms of storage rules and ways of using resources. The Wayback Machine is a private initiative, based on a private foundation, using the cooperation of volunteers in many countries in the world. It gathers world websites without any selection or censorship. UK Web Archive is an initiative of British libraries, which wish to fulfill the idea of legal deposit of British websites (which is currently unworkable because of underdeveloped laws). The websites are carefully selected and their content is evaluated by the librarians according to the importance and usefulness of the site. Adapted from the source document.

Hagenah, U. (2017). Webarchivierung in der SUB Hamburg: kleine Schritte in der Region - Bausteine zu einem größeren Ganzen? (German). *Web Archiving at the Hamburg State and University Library (SUB): Small Steps in the Region - Components of a Larger Whole? (English)*, *51*(6), 500. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Since 2015, the Hamburg State and University Library has been collecting websites of Hamburg institutions or websites concerning regional topics as part of its responsibility as state and depository library. The article describes the legal and technical basis, selection principles, the processing workflow including cataloging as well as how to access the websites' archive copies. Technical and organizational limits and the status of web archiving are discussed in the state library context. (English) [ABSTRACT FROM AUTHOR]

Kimezawa, T., & Murayama, Y. (2017). Preservation of Database on Website -Accessibility Survey for "Dnavi (Database Navigation Service)" and Sustainability of Online

Databases. *Joho No Kagaku to Gijutsu*, *67*(9), 459. Retrieved from https://search.proquest.com/docview/2003801756?accountid=27464

For publicly accessible databases on the Internet, long-term accessibility is not necessarily secured in general, because of change in contents, relocation of URL, and/or even closure of the websites. In this paper, we report on the results of accessibility survey as of April 2017 for the databases registered in the National Diet Library (NDL) Database Navigation Service (Dnavi; operated in 2002-2014), showing that web access was rejected for 22% of total 17,470 databases after three years. We discuss sustainable access to databases published on web sites, from a view point of use of NDL's WARP (Web Archiving Project), and based on the OAIS reference model which is a standard of electronic information for long-term preservation.

Klebczyk, F. (2012). The Website "Archiwum Internetu" Against a Background of Problems with Archiving Web Resources TT - Serwis 'Archiwum Internetu' na tle ogolnych problemow archiwizacji zasobow sieciowych. *Biuletyn EBIB*, (1). Retrieved from https://search.proquest.com/docview/1266143222?accountid=27464

The article is an analysis of the possibilities and actions undertaken in the field of archiving Polish web resources and making them accessible. From the point of view of the portal, there are not only the financial and technical barriers, but also legal regulations are important. To a significant extent, the law puts constraints on this activity, especially when it comes to making the resources accessible. The article presents and overview of the international experience with web archiving and with the techniques used commonly in archiving and using open access for such resources. The Polish project "Archiwum Internetu" is also discussed. "Archiwum Internetu" is created by the National Digital Archives -- its legal foundation, present state and the development direction for this and similar projects are also included in the article. Adapted from the source document.

Kubilius, R. (2018). Bringing Your Physical Books to Digital Learners via the Open Library Project. *Against the Grain*, *30*(2), 63. Retrieved from https://search.proquest.com/docview/2077576451?accountid=27464

Kahle, the founder and digital librarian of Internet Archive, is a visionary, to be sure, and his plenary presentation in Charleston was sincere and enthusiastic. It was quite impressive to hear how many patrons visit Internet Archive each day (3-4 million), that there are 170 staff, and 500 libraries and university partners. It is not hard to believe that the average life of a web page is (only) 100 days before it is deleted or changed.

Kuzmin, E. I. (2013). Behind the Scenes of the Global Information Society: Libraries and Big-time Politics. *Bibliotekovedenie [Library and Information Science (Russia)]*, (2), 13–18. https://doi.org/10.25281/0869-608X-2013-0-2-13-18

The paper examines the challenges facing libraries in the new information environment. Accessibility and preservation of information, information ethics, promotion of media and information literacy and reading, the promotion of multilingualism and diversity in cyberspace are a reflection of the global problems, solving them libraries contribute to the creation of the information society.

Kvasnica, J., Rudišinová, B., & Kreibich, R. (2016). Vědecké využití dat z webových archivů. *Research Use of Web Archived Data.*, *27*(2), 24–34. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=120430576&lang=hu&site=ehost-live

Major part of our communication and media production has moved from traditional print media into digital universe. Digital content on the web is diverse and fluid; it emerges, changes and disappears every day. Such content is unique and valuable from academic perspective but as it disappears over time, we are losing the ability to study recent history. Web archives are now taking the responsibility to capture and preserve such content for future research. Web archives preserve vast amounts of data captured over the years and one of the main goals now is to improve research usability of their collections. This article describes the way the web archives store web content and related metadata and summarizes several recent studies that have dealt with research requirements for web archived data. Based on conclusion of these studies, it suggests further actions to establish cooperation with research community. (English) [ABSTRACT FROM AUTHOR]

Lasfargues, F., & Medjkoune, L. (2012). Archiving the Web, A Service Construct TT - Archiver le Web, un service en construction. *Documentaliste - Sciences de l'Information*, *49*(3), 8–9. Retrieved from https://search.proquest.com/docview/1283633770?accountid=27464

Archiving the Web is an old problem that is taking shape, accompanied by new businesses contours. This article gives a few reminders of technical, historical and legal issues of web archiving before discussing the tasks entrusted to a Web archivist. The article outlines the context of Web archiving. Several duties involved in the work of web archivists are outlined: enriching collections, managing a budget, controlling the quality of the collection, giving access to the archive and preserving web content. Adapted from the source document.

Maeda, N., & Oyama, S. (2017). The Technologies of Web Archiving. *Joho No Kagaku to Gijutsu*, *67*(2), 73. Retrieved from https://search.proquest.com/docview/2007445742?accountid=27464

In the last two decades, web archiving initiatives have spread around the world and have made substantial progresses in legislation, improvement of tools and standards, and fostering of human resources. Especially, international collaboration in the tool developments initiated by IIPC has achieved significant results that constitute the core of the web archiving technologies today. This paper shows how the tools were developed and to what extent they have been implemented into the archives, and briefly describes the mechanisms of the core technologies, Heritrix, WARC and Wayback. Furthermore, it gives an overview of full-text search tools such as NutchWAX and Solr, organization by generating metadata and Memento project which provides integrated access to open archives.

Marciniak, J. (2015). Growing an Archives Department: (And Other Concerns of a New Library Manager). *Computers in Libraries*, *35*(3), 16–19. Retrieved from https://search.proquest.com/docview/1680526979?accountid=27464

The difference between a librarian and an archivist was librarian will drill, glue, and tape a resource to get it back in the stacks. An archivist will seal, hide, and lock up a resource to preserve it. In other words, the difference between a librarian and an archivist is everything.

Librarians and archivists just have different professional philosophies. It comes down to access versus preservation. Although the archives department had existed in various ways for many decades, it was only given a permanent library home in 2009. The library's mission statement outlines the overall goal of the library: to provide quality resources, a high level of service, and innovative learning environments with leading-edge technology. Providing quality resources was where the author felt the archives department could fit into the overall mission of the library. The mission statement of your institution is an essential starting point for establishing common ground with a colleague when working on a project.

Nalewajska, L. (2012). Archiving Websites in the Nordic Countries TT - Archiwizowanie stron internetowych w krajach nordyckich. *Biuletyn EBIB*, (1). Retrieved from https://search.proquest.com/docview/1266143226?accountid=27464

The Nordic countries (Norway, Sweden, Finland, Denmark and Iceland) are the pioneers of web archiving. The process of collecting materials from the web requires arrangements concerning technical-technological, legal and organization issues, was started in these countries in the late 1990s or in the beginning of the 21st century. Archiving is being carried out mainly in national libraries, which also cooperate with International Internet Preservation Consortium and co-create Nordic Web Archive. The way of functioning and the difficulties which occur during archiving in Nordic countries show the complexity of the process and point out how important long-term planning is. Adapted from the source document.

Peach, M. (1998). Archiving the Internet - Web pages of political parties. *Assignation*, *15*(4), 54–58. Retrieved from https://search.proquest.com/docview/57443754?accountid=27464

The Internet has great potential as a source of grey literature. Describes the efforts of the Centro de Estudios Avanzados en Ciencias Sociales (CEACS) of the Instituto Juan March in Madrid, Spain, to take advantage of that potential as a source for researchers present and future. Discusses the following: public use of the Internet in Spain; profile of the CEACS project; nature of political party pages; current status of the project; problems and technical needs; and project expansion.

Phillips, M. E., & Phillips, K. K. (2018). End of Term 2016 Presidential Web Archive. *Against the Grain*, *29*(6), 27. Retrieved from https://search.proquest.com/docview/2077076158?accountid=27464

During every Presidential election in the US since 2008, a group of librarians, archivists, and technologists representing institutions across the nation can be found hard at work, preserving the federal web domain and documenting the changes that occur online during the transition. Anecdotally, evidence exists that the data available on the federal web changes after each election cycle, either as a new president takes office, or when an incumbent president changes messages during the transition into a new term of office. Until 2004, nothing had been done to document this change. Originally, the National Archives and Records Administration (NARA) conducted the first large-scale capture of the federal web at the end of George W. Bush's first term in office in 2004. This is noteworthy because, while institutions like the Library of Congress, the Government Publishing Office, and NARA itself have web archiving as part of their imperative, none of their mandates are so broad as to cover the capture and preservation of the entirety of the federal web.

Plaisance, C. (2016). Methods of Web Philology: Computer Metadata and Web Archiving in the Primary Source Documents of Contemporary Esotericism. *International Journal for the Study of New Religions*, *7*(1), 43–68. https://doi.org/10.1558/ijsnr.v7i1.26074

This article explores the issues surrounding the critical analysis of first generation electronic objects within the context of the study of contemporary esoteric discourse. This is achieved through a detailed case study of Benjamin Rowe's work, A Short Course in Scrying, which is solely exemplified by digital witnesses. This article demonstrates that the critical analysis of these witnesses is only possible by adapting the general methods of textual scholarship to the specific techniques of digital forensics-particularly the analysis of computer metadata and web archives. The resulting method, here termed web philology, is applicable to the critical analysis by the scholar of religion of any primary source documents originating on the web as electronic objects. [ABSTRACT FROM AUTHOR]

Schellnack-Kelly, I. (2014). Practical Digital Preservation: A How-to Guide for Organizations of Any Size2014 2 Practical Digital Preservation: A How-to Guide for Organizations of Any Size London Facet 2013 336 pp. ISBN978-1-85604-755-5 £49.95 soft cover. *The Electronic Library*, *32*(6), 924–925. https://doi.org/10.1108/EL-02-2014-0033

Slaska, K., & Wasilewska, A. (2012). Web Archiving -- the Situation in Polish Law from the Point of View of the Librarians TT - Archiwizacja Internetu -- sytuacja w polskim prawie z punktu widzenia bibliotekarzy. *Biuletyn EBIB*, (1). Retrieved from https://search.proquest.com/docview/1266143224?accountid=27464

The authors present the possibilities of archiving Polish web sites from the point of view of the librarians, focusing mainly on the legal aspects of this issue. The reflections are based on the Act on Legal Deposit Copies [Ustawa o obowiazkowych egzemplarzach bibliotecznych] currently in force in Poland and on selected legal interpretation. The authors present some materials on the situation in foreign countries and mention the international organization International Internet Preservation Consortium (IIPC). Adapted from the source document.

Velte, A. (2018). Ethical Challenges and Current Practices in Activist Social Media Archives. *The American Archivist*, *81*(1), 112–134. https://doi.org/10.17723/0360-9081-81.1.112

Social media (Web applications supporting communication between Internet users) empower current activist groups to create records of their activities. Recent digital collections, such as the digital archives of the Occupy Wall Street movement and the Documenting Ferguson Project, demonstrate archival interest in preserving and providing access to activist social media. Literature describing current practices exists for related topics such as Web and social media archives, privacy and access for digital materials, and activist archives. However, research on activist social media archives is scarce. These materials likely present subject- and format-specific challenges not yet identified in peer-reviewed research. Using a survey and semistructured interviews with archivists who collect activist social media, this study describes ethical challenges regarding acquisition and access. Specifically, the respondents were concerned about acquiring permission to collect and provide long-term access to activist groups' social media. When collecting social media as data sets, archivists currently intend to provide moderated access to the archives, whereas when dealing with social media accounts, archivists intend to seek permission to collect from the activist groups and provide access online. These current practices addressing ethical issues may serve as models for other institutions interested in collecting social media from activists. Understanding how to

approach activist social media ethically decreases the risk th a t these important records of modern activism will be left out of the historical narrative. [ABSTRACT FROM AUTHOR]

Bornand, N. J., Balakireva, L., de Sompel, H., & Van de Sompel, H. (2016). Routing Memento Requests Using Binary Classifiers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 63–72). New York, NY, USA: ACM. https://doi.org/10.1145/2910896.2910899

The Memento protocol provides a uniform approach to query individual web archives. Soon after its emergence, Memento Aggregator infrastructure was introduced that supports querying across multiple archives simultaneously. An Aggregator generates a response by issuing the respective Memento request against each of the distributed archives it covers. As the number of archives grows, it becomes increasingly challenging to deliver aggregate responses while keeping response times and computational costs under control. Ad-hoc heuristic approaches have been introduced to address this challenge and research has been conducted aimed at optimizing query routing based on archive profiles. In this paper, we explore the use of binary, archive-specific classifiers generated on the basis of the content cached by an Aggregator, to determine whether or not to query an archive for a given URI. Our results turn out to be readily applicable and can help to significantly decrease both the number of requests and the overall response times without compromising on recall. We find, among others, that classifiers can reduce the average number of requests by 77% compared to a brute force approach on all archives, and the overall response time by 42% while maintaining a recall of 0.847.

David Dubin, Joe Futrelle, Joel Plutchak, & Janet Eke. (2009). Preserving Meaning, Not Just Objects: Semantics and Digital Preservation. *Library Trends*, *57*(3), 595–610. https://doi.org/10.1353/lib.0.0054

The ECHO DEPository project is a digital preservation research and development project funded by the National Digital Information Infrastructure and Preservation Program (NDIIPP) and administered by the Library of Congress. A key goal of this project is to investigate both practical solutions for supporting digital preservation activities today, and the more fundamental research questions underlying the development of the next generation of digital preservation systems. To support on-the-ground preservation efforts in existing technical and organizational environments, we have developed tools to help curators collect and manage Web-based digital resources, such as the Web Archives Workbench (Kaczmarek et al., 2008), and to enhance existing repositories' support for interoperability and emerging preservation standards, such as the Hub and Spoke Tool Suite (Habing et al., 2008). In the longer term, however, we recognize that successful digital preservation activities will require a more precise and complete account of the meaning of relationships within and among digital objects. This article describes project efforts to identify the core underlying semantic issues affecting long-term digital preservation, and to model how semantic inference may help next-generation archives head off long-term preservation risks. [ABSTRACT FROM AUTHOR]

Häusner, E. M. (2017). Memory Entanglements and Collection Development in a Transnational Media Landscape. In *IFLA 2017* (p. 5). IFLA. Retrieved from http://library.ifla.org/1683/1/186-haeusner-en.pdf

Defining a national domain is the crux of the matter of every National Library's mission. The National Library of Sweden collects, preserves, registers, and guarantees access to all

materials published and distributed in Sweden, printed, audio-visual and since 2012, even electronic. Furthermore the National Library of Sweden collects Suecana, foreign publications which possess historical significance to Sweden and even Swedish literature in translation. Collection strategies have to be updated and developed to fit the times: Digitalization and media convergence presuppose a new concept and new definition of the national domain. How should the National Library work with selection and collection strategies in a way that to make sure that the Suecana-collection and the Swedish collection are truly representative and relevant? This paper describes difficulties inherent to defining a national domain in today's media landscape and presents s

Holzmann, H., Sperber, W., & Runnwerth, M. (2016). Archiving Software Surrogates on the Web for Future Reference. *Research & Advanced Technology for Digital Libraries: 20th International Conference on Theory & Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings*, 215. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Software has long been established as an essential aspect of the scientific process in mathematics and other disciplines. However, reliably referencing software in scientific publications is still challenging for various reasons. A crucial factor is that software dynamics with temporal versions or states are difficult to capture over time. We propose to archive and reference surrogates instead, which can be found on the Web and reflect the actual software to a remarkable extent. Our study shows that about a half of the webpages of software are already archived with almost all of them including some kind of documentation.

Schafer, V. V., Musiani, F., & Borelli, M. (2016). Negotiating the Web of the Past. *French Journal for Media Research*. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The material, practical, theoretical elements of Web archiving as an ensemble of practices and a terrain of inquiry are inextricably entwined. Thus, its processes and infrastructures – often discreet and invisible – are increasingly relevant. Approaches inspired by Science and Technology Studies (STS) can contribute to shed light on the shaping of Web archives.

Summers, E., & Punzalan, R. (2017). Bots, Seeds and People: Web Archives As Infrastructure. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 821–834). New York, NY, USA: ACM. https://doi.org/10.1145/2998181.2998345

The field of web archiving provides a unique mix of human and automated agents collaborating to achieve the preservation of the web. Centuries old theories of archival appraisal are being transplanted into the sociotechnical environment of the World Wide Web with varying degrees of success. The work of the archivist and bots in contact with the material of the web present a distinctive and understudied CSCW shaped problem. To investigate this space we conducted semi-structured interviews with archivists and technologists who were directly involved in the selection of content from the web for archives. These semi-structured interviews identified thematic areas that inform the appraisal process in web archives, some of which are encoded in heuristics and algorithms. Making the

infrastructure of web archives legible to the archivist, the automated agents and the future researcher is presented as a challenge to the CSCW and archival community.

Abbate, J. (2017). What and where is the Internet? (Re)defining Internet histories. *Internet Histories*, *1*(1–2), 8–14. https://doi.org/10.1080/24701475.2017.1305836

Both the Internet and the Web beat out numerous rivals to become today's dominant network and online system,11. "Online system" in this essay is used as a generic term for Web-like systems, i.e. systems for navigating information over networks. The origins of the term are in the 1960s oNLine System (NLS). This essay uses "online world" as a generic term for all of cyberspace. View all notes respectively. Many of those rival systems and networks had developed alternative solutions to issues that face us today, from micropayments to copyright. But few scholars, much less thought leaders, have a meaningful overview of the origins of our online world, or of the many systems which came before. This exclusivity is a problem, since as a society we are now making some of the permanent decisions that will determine how we deal with information for decades and even centuries to come. Those decisions are about regulatory structures, economic models, civil liberties, publishing and more. This essay argues for the need to comparatively study online information systems across all these axes, and to thus develop a "common language" of known precedents and concepts as a prerequisite for making informed discussions about the future of the online world. Doing so depends on two factors: (1) preservation of enough historical materials about earlier systems to be able to meaningfully examine them; (2) interdisciplinary, international attention to "meta" stories that emerge from considering the evolution of multiple networks and online systems.

Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A., & Ueda, S. (2014). Life Span of Web Pages: A Survey of 10 Million Pages Collected in 2001. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 463–464). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2740769.2740869

Identifying and tracking new information on the Web is important in sociology, marketing, and survey research, since new trends might be apparent in the new information. Such changes can be observed by crawling the Web periodically. In practice, however, it is impossible to crawl the entire expanding Web repeatedly. This means that the novelty of a page remains unknown, even if that page did not exist in previous snapshots. In this paper, we propose a novelty measure for estimating the certainty that a newly crawled page appeared between the previous and current crawls. Using this novelty measure, new pages can be extracted from a series of unstable snapshots for further analysis and mining to identify new trends on the Web. We evaluated the precision, recall, and miss rate of the novelty measure using our Japanese web archive, and applied it to a Web archive search engine.

Ainsworth, S. G., & Nelson, M. L. (2015). Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. *International Journal on Digital Libraries*, *16*(2), 129–144. https://doi.org/http://dx.doi.org/10.1007/s00799-014-0120-4

When viewing an archived page using the archive's user interface (UI), the user selects a datetime to view from a list. The archived web page, if available, is then displayed. From this display, the web archive UI attempts to simulate the web browsing experience by smoothly transitioning between archived pages. During this process, the target datetime changes with each link followed, potentially drifting away from the datetime originally selected. For

sparsely archived resources, this almost transparent drift can be many years in just a few clicks. We conducted 200,000 acyclic walks of archived pages, following up to 50 links per walk, comparing the results of two target datetime policies. The Sliding Target policy allows the target datetime to change as it does in archive UIs such as the Internet Archive's Wayback Machine. The Sticky Target policy, represented by the Memento API, keeps the target datetime the same throughout the walk. We found that the Sliding Target policy drift increases with the number of walk steps, number of domains visited, and choice (number of links available). However, the Sticky Target policy controls temporal drift, holding it to <30 days on average regardless of walk length or number of domains visited. The Sticky Target policy shows some increase as choice increases, but this may be caused by other factors. We conclude that based on walk length, the Sticky Target policy generally produces at least 30 days less drift than the Sliding Target policy.

Ainsworth, S. G., Nelson, M. L., & Van de Sompel, H. (2015). Only One Out of Five Archived Web Pages Existed as Presented. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15* (pp. 257–266). New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791044

When a user retrieves a page from a web archive, the page is marked with the acquisition datetime of the root resource, which effectively asserts "this is how the page looked at a that datetime." However, embedded resources, such as images, are often archived at different datetimes than the main page. The presentation appears temporally coherent, but is composed from resources acquired over a wide range of datetimes. We examine the completeness and temporal coherence of composite archived resources (composite mementos) under two selection heuristics. The completeness and temporal coherence achieved using a single archive was compared to the results achieved using multiple archives. We found that at most 38.7% of composite mementos are both temporally coherent and that at most only 17.9% (roughly 1 in 5) are temporally coherent and 100% complete. Using multiple archives increases mean completeness by 3.1-4.1% but also reduces temporal coherence.

Ainsworth, S. G., Alsum, A., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2011). How Much of the Web is Archived? In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 133–136). New York, NY, USA: ACM. https://doi.org/10.1145/1998076.1998100

The Memento Project's archive access additions to HTTP have enabled development of new web archive access user interfaces. After experiencing this web time travel, the in- evitable question that comes to mind is "How much of the Web is archived?" This question is studied by approximating the Web via sampling URIs from DMOZ, Delicious, Bitly, and search engine indexes and measuring number of archive copies available in various public web archives. The results indicate that 35%-90% of URIs have at least one archived copy, 17%-49% have two to five copies, 1%-8% have six to ten copies, and 8%-63% at least ten copies. The number of URI copies varies as a function of time, but only 14.6-31.3% of URIs are archived more than once per month.

Ainsworth, S. G., & Nelson, M. L. (2013). Evaluating Sliding and Sticky Target Policies by Measuring Temporal Drift in Acyclic Walks Through a Web Archive. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 39–48). New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467718

When a user views an archived page using the archive's user interface (UI), the user selects a datetime to view from a list. The archived web page, if available, is then displayed. From this display, the web archive UI attempts to simulate the web browsing experience by smoothly transitioning between archived pages. During this process, the target datetime changes with each link followed; drifting away from the datetime originally selected. When browsing sparsely-archived pages, this nearly-silent drift can be many years in just a few clicks. We conducted 200,000 acyclic walks of archived pages, following up to 50 links per walk, comparing the results of two target datetime policies. The Sliding Target policy allows the target datetime to change as it does in archive UIs such as the Internet Archive's Wayback Machine. The Sticky Target policy, represented by the Memento API, keeps the target datetime the same throughout the walk. We found that the Sliding Target policy drift increases with the number of walk steps, number of domains visited, and choice (number of links available). However, the Sticky Target policy controls temporal drift, holding it to less than 30 days on average regardless of walk length or number of domains visited. The Sticky Target policy shows some increase as choice increases, but this may be caused by other factors. We conclude that based on walk length, the Sticky Target policy generally produces at least 30 days less drift than the Sliding Target policy.

Alam, S., Kelly, M., & Nelson, M. L. (2016). InterPlanetary Wayback: The Permanent Web Archive. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 273–274). New York, NY, USA: ACM. https://doi.org/10.1145/2910896.2925467

To facilitate permanence and collaboration in web archives, we built InterPlanetary Wayback to disseminate the contents of WARC files into the IPFS network. IPFS is a peer-to-peer content-addressable file system that inherently allows deduplication and facilitates opt-in replication. We split the header and payload of WARC response records before disseminating into IPFS to leverage the deduplication, build a CDXJ index, and combine them at the time of replay. From a 1.0 GB sample Archive-It collection of WARCs containing 21,994 mementos, we found that on an average, 570 files can be indexed and disseminated into IPFS per minute. We also found that in our naive prototype implementation, replay took on an average 370 milliseconds per request.

Alam, S., Kelly, M., Weigle, M. C., & Nelson, M. L. (2017). Client-side Reconstruction of Composite Mementos Using Serviceworker. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 237–240). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=3200334.3200361

We use the ServiceWorker (SW) API to intercept HTTP requests for embedded resources and reconstruct Composite Mementos without the need for conventional URL rewriting typically per- formed by web archives. URL rewriting is a problem for archival replay systems, especially for URLs constructed by JavaScript, that frequently results in incorrect URI references. By intercept- ing requests on the client using SW, we are able to strategically reroute instead of rewrite. Our implementation moves rewrit- ing to clients, saving servers' computing resources and allowing servers to return responses more quickly. In our experiments, re- trieving the original instead of rewritten pages from the archive resulted in a one-third reduction in time overhead and a one-fifth reduction in data overhead. Our system, reconstructive.js , prevents the live web from leaking into Composite Mementos while being easy to distribute and maintain.

Alam, S., & Nelson, M. L. (2016). MemGator - A Portable Concurrent Memento Aggregator. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (pp. 243–244). New York, New York, USA: ACM Press. https://doi.org/10.1145/2910896.2925452

The Memento protocol makes it easy to build a uniform lookup service to aggregate the holdings of web archives. However, there is a lack of tools to utilize this capability in archiving applications and research projects. We created MemGator, an open source, easy to use, portable, concurrent, cross-platform, and self-documented Memento aggregator CLI and server tool written in Go. MemGator implements all the basic features of a Memento aggregator (e.g., TimeMap and TimeGate) and gives the ability to customize various options including which archives are aggregated. It is being used heavily by tools and services such as Mink, WAIL, OldWeb. today, and archiving research projects and has proved to be reliable even in conditions of extreme load.

Alam, S., Nelson, M. L., Van de Sompel, H., Balakireva, L. L., Shankar, H., & Rosenthal, D. S. H. (2016). Web archive profiling through CDX summarization. *International Journal on Digital Libraries*, *17*(3), 223–238. https://doi.org/http://dx.doi.org/10.1007/s00799-016-0184-4

Issue Title: Focused Issue on TPDL 2015 With the proliferation of public web archives, it is becoming more important to better profile their contents, both to understand their immense holdings as well as to support routing of requests in the Memento aggregator. To save time, the Memento aggregator should only poll the archives that are likely to have a copy of the requested URI. Using the crawler index files produced after crawling, we can generate profiles of the archives that summarize their holdings and can be used to inform routing of the Memento aggregator's URI requests. Previous work in profiling ranged from using full URIs (no false positives, but with large profiles) to using only top-level domains (TLDs) (smaller profiles, but with many false positives). This work explores strategies in between these two extremes. In our experiments, we correctly identified about 78 % of the URIs that were present or not present in the archive with less than 1 % relative cost as compared to the complete knowledge profile and 94 % URIs with less than 10 % relative cost without any false negatives. With respect to the TLD-only profile, the registered domain profile doubled the routing precision, while complete hostname and one path segment gave a tenfold increase in the routing precision.

Alam, S., Nelson, M. L., Van de Sompel, H., & Rosenthal, D. S. H. (2016). Web Archive Profiling Through Fulltext Search. In N. Fuhr, L. Kovács, T. Risse, & W. Nejdl (Eds.), *TPDL 2016: Research and Advanced Technology for Digital Libraries* (pp. 121–132). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-43997-6_10

An archive profile is a high-level summary of a web archive's holdings that can be used for routing Memento queries to the appropriate archives. It can be created by generating summaries from the CDX files (index of web archives) which we explored in an earlier work. However, requiring archives to update their profiles periodically is difficult. Alternative means to discover the holdings of an archive involve sampling based approaches such as fulltext keyword searching to learn the URIs present in the response or looking up for a sample set of URIs and see which of those are present in the archive. It is the fulltext search based discovery and profiling that is the scope of this paper. We developed the Random Searcher Model (RSM) to discover the holdings of an archive by a random search walk. We

measured the search cost of discovering certain percentages of the archive holdings for various profiling policies under different RSM configurations. We can make routing decisions of 80 % of the requests correctly while maintaining about 0.9 recall by discovering only 10 % of the archive holdings and generating a profile that costs less than 1 % of the complete knowledge profile.

Alasaadi, A., & Nelson, M. L. (2011). Persistent annotations deserve new URIs. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11* (p. 195). New York, New York, USA: ACM Press. https://doi.org/10.1145/1998076.1998113

Some digital libraries support annotations, but sharing these annotations with other systems or across the web is difficult because of the need of special applications to read and decode these annotations. Due to the frequent change of web resources, the annotation's meaning can change if the underlying resources change. This project concentrates on minting a new URI for every annotation and creating a persistent and independent archived version of all resources. Users should be able to select a segment of an image or a video to be part of the annotation. The media fragment URIs described in the Open Annotation Collaboration data model can be used, but in practice they have limits, and they face the lack of support by the browsers. So in this project the segments of images, and videos can be used in the annotations without using media fragment URIs.

Alkwai, L. M., Nelson, M. L., & Weigle, M. C. (2015). How Well Are Arabic Websites Archived? In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 223–232). New York, NY, USA: ACM. https://doi.org/10.1145/2756406.2756912

t is has long been anecdotally known that web archives and search engines favor Western and English-language sites. In this paper we quantitatively explore how well indexed and archived are Arabic language web sites. We began by sam- pling 15,092 unique URIs from three different website direc- tories: DMOZ (multi-lingual), Raddadi and Star28 (both primarily Arabic language). Using language identification tools we eliminated pages not in the Arabic language (e.g., English language versions of Al-Jazeera sites) and culled the collection to 7,976 definitely Arabic language web pages. We then used these 7,976 pages and crawled the live web and web archives to produce a collection of 300,646 Arabic lan- guage pages. We discovered: 1) 46% are not archived and 31% are not indexed by Google (www.google.com), 2) only 14.84% of the URIs had an Arabic country code top-level domain (e.g., .sa) and only 10.53% had a GeoIP in an Ara- bic country, 3) having either only an Arabic GeoIP or only an Arabic top-level domain appears to negatively impact archiving, 4) most of the archived pages are near the top level of the site and deeper links into the site are not well-archived, 5) the presence in a directory positively impacts indexing and presence in the DMOZ directory, specifically, positively impacts archiving.

Alkwai, L. M., Nelson, M. L., & Weigle, M. C. (2017). Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages. *ACM Trans. Inf. Syst.*, *36*(1), 1:1--1:34. https://doi.org/10.1145/3041656

It has long been suspected that web archives and search engines favor Western and English language webpages. In this article, we quantitatively explore how well indexed and archived Arabic language webpages are as compared to those from other languages. We began by

sampling 15,092 unique URIs from three different website directories: DMOZ (multilingual), Raddadi, and Star28 (the last two primarily Arabic language). Using language identification tools, we eliminated pages not in the Arabic language (e.g., English-language versions of Aljazeera pages) and culled the collection to 7,976 Arabic language webpages. We then used these 7,976 pages and crawled the live web and web archives to produce a collection of 300,646 Arabic language pages. We compared the analysis of Arabic language pages with that of English, Danish, and Korean language pages. First, for each language, we sampled unique URIs from DMOZ; then, using language identification tools, we kept only pages in the desired language. Finally, we crawled the archived and live web to collect a larger sample of pages in English, Danish, or Korean. In total for the four languages, we analyzed over 500,000 webpages. We discovered: (1) English has a higher archiving rate than Arabic, with 72.04% archived. However, Arabic has a higher archiving rate than Danish and Korean, with 53.36% of Arabic URIs archived, followed by Danish and Korean with 35.89% and 32.81% archived, respectively. (2) Most Arabic and English language pages are located in the United States; only 14.84% of the Arabic URIs had an Arabic country code top-level domain (e.g., sa) and only 10.53% had a GeoIP in an Arabic country. Most Danish-language pages were located in Denmark, and most Korean-language pages were located in South Korea. (3) The presence of a webpage in a directory positively impacts indexing and presence in the DMOZ directory, specifically, positively impacts archiving in all four languages. In this work, we show that web archives and search engines favor English pages. However, it is not universally true for all Western-language webpages because, in this work, we show that Arabic webpages have a higher archival rate than Danish language webpages.

Allegrezza, S. (2015). *Nuove prospettive per il web archiving: gli standard ISO 28500 (formato WARC) e ISO/TR 14873 sulla qualità del web archiving*. Italy, Europe. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Il Web archiving è un argomento di forte attualità in quanto, come è noto, se non si individuano in breve tempo soluzioni efficaci e sostenibili nel lungo periodo, si rischia di perdere per sempre quello che si è prodotto e pubblicato sul Web negli ultimi venti-trenta anni, dal momento che tale materiale è caratterizzato da un'estrema mutevolezza e dinamicità e spesso interi siti Web cambiano o scompaiono nel giro di poco tempo. Le soluzioni che sono state proposte fino ad oggi sono parziali e non sempre hanno raggiunto l'obiettivo. Tuttavia, recentemente ci sono state due novità che sembrerebbero poter assicurare prospettive migliori: si tratta da una parte della proposta di un formato elettronico specificatamente pensato per l'archiviazione del Web (il formato WARC), dall'altra della pubblicazione di una specifica norma ISO dedicata alla qualità nella conservazione del Web (ISO/TR 14873:2013). La rilevanza dell'argomento per il settore dei beni culturali è tale che è opportuno fare un po' di chiarezza su queste tematiche analizzando sia lo stato dell'arte che le prospettive future.

AlNoamany, Y. A., Weigle, M. C., & Nelson, M. L. (2013). Access Patterns for Robots and Humans in Web Archives. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 339–348). New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467722

Although user access patterns on the live web are well-understood, there has been no corresponding study of how users, both humans and robots, access web archives. Based on samples from the Internet Archive's public Wayback Machine, we propose a set of basic

usage patterns: Dip (a single access), Slide (the same page at different archive times), Dive (different pages at approximately the same archive time), and Skim (lists of what pages are archived, i.e., TimeMaps). Robots are limited almost exclusively to Dips and Skims, but human accesses are more varied between all four types. Robots outnumber humans 10:1 in terms of sessions, 5:4 in terms of raw HTTP accesses, and 4:1 in terms of megabytes transferred. Robots almost always access TimeMaps (95% of accesses), but humans predominately access the archived web pages themselves (82% of accesses). In terms of unique archived web pages, there is no overall preference for a particular time, but the recent past (within the last year) shows significant repeat accesses.

AlNoamany, Y., AlSum, A., Weigle, M. C., & Nelson, M. L. (2014). Who and what links to the Internet Archive. *International Journal on Digital Libraries*, *14*(3–4), 101–115. https://doi.org/10.1007/s00799-014-0111-5

Issue Title: 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013) The Internet Archive's (IA) Wayback Machine is the largest and oldest public Web archive and has become a significant repository of our recent history and cultural heritage. Despite its importance, there has been little research about how it is discovered and used. Based on Web access logs, we analyze what users are looking for, why they come to IA, where they come from, and how pages link to IA. We find that users request English pages the most, followed by the European languages. Most human users come to Web archives because they do not find the requested pages on the live Web. About 65 % of the requested archived pages no longer exist on the live Web. We find that more than 82 % of human sessions connect to the Wayback Machine via referrals from other Web sites, while only 15 % of robots have referrers. Most of the links (86 %) from Websites are to individual archived pages at specific points in time, and of those 83 % no longer exist on the live Web. Finally, we find that users who come from search engines browse more pages than users who come from external Web sites.[PUBLICATION ABSTRACT]

AlNoamany, Y., Weigle, M. C., & Nelson, M. L. (2017). Generating Stories From Archived Collections. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 309–318). New York, NY, USA: ACM. https://doi.org/10.1145/3091478.3091508

With the extensive growth of the Web, multiple Web archiving initiatives have been started to archive different aspects of the Web. Services such as Archive-It exist to allow institutions to develop, curate, and preserve collections of Web resources. Understanding the contents and boundaries of these archived collections is a challenge, resulting in the paradox of the larger the collection, the harder it is to understand. Meanwhile, as the sheer volume of data grows on the Web, "storytelling" is becoming a popular technique in social media for selecting Web resources to support a particular narrative or "story". We address the problem of understanding archived collections by proposing the Dark and Stormy Archive (DSA) framework, in which we integrate "storytelling" social media and Web archives. In the DSA framework, we identify, evaluate, and select candidate Web pages from archived collections that summarize the holdings of these collections, arrange them in chronological order, and then visualize these pages using tools that users already are familiar with, such as Storify. Inspired by the Turing Test, we evaluate the stories automatically generated by the DSA framework against a ground truth dataset of hand-crafted stories, generated by expert archivists from Archive-It collections. Using Amazon's Mechanical Turk, we found that the stories automatically generated by DSA are indistinguishable from those created by human

subject domain experts, while at the same time both kinds of stories (automatic and human) are easily distinguished from randomly generated stories.

AlNoamany, Y., Weigle, M. C., & Nelson, M. L. (2016). Detecting off-topic pages within TimeMaps in Web archives. *International Journal on Digital Libraries*, *17*(3), 203–221. https://doi.org/http://dx.doi.org/10.1007/s00799-016-0183-5

Web archives have become a significant repository of our recent history and cultural heritage. Archival integrity and accuracy is a precondition for future cultural research. Currently, there are no quantitative or content-based tools that allow archivists to judge the quality of the Web archive captures. In this paper, we address the problems of detecting when a particular page in a Web archive collection has gone off-topic relative to its first archived copy. We do not delete off-topic pages (they remain part of the collection), but they are flagged as off-topic so they can be excluded for consideration for downstream services, such as collection summarization and thumbnail generation. We propose different methods (cosine similarity, Jaccard similarity, intersection of the 20 most frequent terms, Web-based kernel function, and the change in size using the number of words and content length) to detect when a page has gone off-topic. Those predicted off-topic pages will be presented to the collection's curator for possible elimination from the collection or cessation of crawling. We created a gold standard data set from three Archive-It collections to evaluate the proposed methods at different thresholds. We found that combining cosine similarity at threshold 0.10 and change in size using word count at threshold −0.85 performs the best with accuracy = 0.987, $$F_{1}$$ score = 0.906, and AUC $$=$$ 0.968. We evaluated the performance of the proposed method on several Archive-It collections. The average precision of detecting off-topic pages in the collections is 0.89. [ABSTRACT FROM AUTHOR]

AlNoamany, Y., Weigle, M. C., & Nelson, M. L. (2015). Detecting Off-Topic Pages in Web Archives. In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), *Research and Advanced Technology for Digital Libraries. TPDL 2015. Lecture Notes in Computer Science, vol 9316.* (pp. 225–237). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24592-8_17

Web archives have become a significant repository of our recent history and cultural heritage. Archival integrity and accuracy is a precondition for future cultural research. Currently, there are no quantitative or content-based tools that allow archivists to judge the quality of the Web archive captures. In this paper, we address the problems of detecting off-topic pages in Web archive collections. We evaluate six different methods to detect when the page has gone off-topic through subsequent captures. Those predicted off-topic pages will be presented to the collection's curator for possible elimination from the collection or cessation of crawling. We created a gold standard data set from three Archive-It collections to evaluate the proposed methods at different thresholds. We found that combining cosine similarity at threshold 0.10 and change in size using word count at threshold $$-$$0.85 performs the best with accuracy = 0.987, $$F_{1}$$score = 0.906, and AUC = 0.968. We evaluated the performance of the proposed method on several Archive-It collections. The average precision of detecting the off-topic pages is 0.92.

Alonso, O., Kandylas, V., & Tremblay, S.-E. (2018). How it Happened. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18* (pp. 193–202). New York, New York, USA: ACM Press. https://doi.org/10.1145/3197026.3197034

Social networks like Twitter and Facebook are the largest sources of public opinion and real-time information on the Internet. If an event is of general interest, news articles follow and eventually a Wikipedia page. We propose the problem of automatic event story generation and archiving by combining social and news data to construct a new type of document in the form of a Wiki-like page structure. We introduce a technique that shows the evolution of a story as perceived by the crowd in social media, along with editorially authored articles annotated with examples of social media as supporting evidence. At the core of our research, is the temporally sensitive extraction of data that serve as context for retrieval purposes. Our approach includes a fine-grained vote counting strategy that is used for weighting purposes, pseudo-relevance feedback and query expansion with social data and web query logs along with a timeline algorithm as the base for a story. We demonstrate the effectiveness of our approach by processing a dataset comprising millions of English language tweets generated over a one year period and present a full implementation of our system.

Alshukri, A., & Coenen, F. (2017). Mining the information architecture of the WWW using automated website boundary detection. *Web Intelligence (2405-6456)*, *15*(4), 269–290. Retrieved from http://10.0.12.161/WEB-170365

The world wide web has two main forms of architecture, the first is that which is explicitly encoded into web pages, and the second is that which is implied by the web content, particularly pertaining to look and feel. The latter is exemplified by the concept of a website, a concept that is only loosely defined, although users intuitively understand it. The Website Boundary Detection (WBD) problem is concerned with the task of identifying the complete collection of web pages/resources that are contained within a single website. Whatever the case, the concept of a website is used with respect to a number of application domains including; website archiving, spam detection, and www analysis. In the context of such applications it is beneficial if a website can be automatically identified. This is usually done by identifying a website of interest in terms of its boundary, the so called WBD problem. In this paper seven WBD techniques are proposed and compared, four statistical techniques where the web data to be used is obtained apriori, and three dynamic techniques where the data to be used is obtained as the process progresses. All seven techniques are presented in detail and evaluated. [ABSTRACT FROM AUTHOR]

AlSum, A. (2015). Reconstruction of the US First Website. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15* (pp. 285–286). New York, New York, USA: ACM Press. https://doi.org/10.1145/2756406.2756954

The Web idea started on 1989 with a proposal from Sir Tim Berners-Lee. The first US website has been developed at SLAC on 1991. This early version of the Web and the subsequent updates until 1998 have been preserved by SLAC archive and history office for many years. In this paper, we discuss the strategy and techniques to reconstruct this early website and make it available through Stanford Web Archive Portal.

AlSum, A., & Nelson, M. L. (2013). ArcLink. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13* (pp. 377–378). New York, New York, USA: ACM Press. https://doi.org/10.1145/2467696.2467751

AlSum, A., Nelson, M. L., Sanderson, R., & Van de Sompel, H. (2013). Archival HTTP redirection retrieval policies. In *Proceedings of the 22nd International Conference on*

World Wide Web - WWW '13 Companion (pp. 1051–1058). New York, New York, USA: ACM Press. https://doi.org/10.1145/2487788.2488117

When retrieving archived copies of web resources (mementos) from web archives, the original resource's URI-R is typically used as the lookup key in the web archive. This is straightforward until the resource on the live web issues a redirect: R ->R`. Then it is not clear if R or R` should be used as the lookup key to the web archive. In this paper, we report on a quantitative study to evaluate a set of policies to help the client discover the correct memento when faced with redirection. We studied the stability of 10,000 resources and found that 48% of the sample URIs tested were not stable, with respect to their status and redirection location. 27% of the resources were not perfectly reliable in terms of the number of mementos of successful responses over the total number of mementos, and 2% had a reliability score of less than 0.5. We tested two retrieval policies. The first policy covered the resources which currently issue redirects and successfully resolved 17 out of 77 URIs that did not have mementos of the original URI, but did of the resource that was being redirected to. The second policy covered archived copies with HTTP redirection and helped the client in 58% of the cases tested to discover the nearest memento to the requested datetime.

AlSum, A., Weigle, M. C., Nelson, M. L., & Van de Sompel, H. (2014). Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries*, *14*(3–4), 149–166. https://doi.org/10.1007/s00799-014-0118-y

(ProQuest: ... denotes formulae and/or non-USASCII text omitted; see image) Issue Title: 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013) The Memento Aggregator currently polls every known public web archive when serving a request for an archived web page, even though some web archives focus on only specific domains and ignore the others. Similar to query routing in distributed search, we investigate the impact on aggregated Memento TimeMaps (lists of when and where a web page was archived) by only sending queries to archives likely to hold the archived page. We profile fifteen public web archives using data from a variety of sources (the web, archives' access logs, and fulltext queries to archives) and use these profiles as resource descriptor. These profiles are used in matching the URI-lookup requests to the most probable web archives. We define ..... as the percentage of a TimeMap that was returned using ..... web archives. We discover that only sending queries to the top three web archives (i.e., 80 % reduction in the number of queries) for any request reaches on average ..... If we exclude the Internet Archive from the list, we can reach ..... on average using only the remaining top three web archives.[PUBLICATION ABSTRACT]

Altman, M., Adams, M. O., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. H. (2009). Digital Preservation through Archival Collaboration: The Data Preservation Alliance for the Social Sciences. *American Archivist*, *72*(1), 170–184. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The Data Preservation Alliance for the Social Sciences (Data-PASS) is a partnership of five major U.S. institutions with a strong focus on archiving social science research. The Library of Congress supports the partnership through its National Digital Information Infrastructure and Preservation Program (NDIIPP). The goal of Data-PASS is to acquire and preserve data from opinion polls, voting records, large-scale surveys, and other social science studies at risk

of being lost to the research community. This paper discusses the agreements, processes, and infrastructure that provide a foundation for the collaboration. [ABSTRACT FROM AUTHOR]

Amanda, L. R. (2018). *Getting to Know Our Web Archive: A Pilot Project to Collaboratively Increase Access to Digital Cultural Heritage Materials in Wyoming*. United States, North America: Digital USD. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The University of Wyoming is the only four year higher education institution in the state, a unique position amongst colleges and universities in the United States. Given this unusual status it is especially important that the university libraries use their resources to identify and partner with communities around the state to build collections that preserve their cultural heritage. An Archive-It subscription was purchased in 2016, with an initial goal of capturing university related materials. In an effort to expand the scope and meaningfulness of the web archive, a project has been undertaken to use university and statewide relationships to build a Wyoming focused Native American digital cultural heritage collection comprised of web-based materials. This is an interdepartmental effort led by the Digital Collections Librarian and the Metadata Librarian that includes collaboration within the library, the university, and the state.

Anand, A., & Bailey, J. (2016). Exploring the Past of the Web: Alexandria & Archive-it Hackathon. In *Proceedings of the 8th ACM Conference on Web Science* (p. 14). New York, NY, USA: ACM. https://doi.org/10.1145/2908131.2908212

The Web has pervaded all walks of life and has become an important corpus for studying the humanities, social sciences, and for use by computer scientists and other disciplines. Web archives collect, preserve, and provide ongoing access to ephemeral Web pages and hence encode traces of human thought, activity, and history. This makes them a valuable resource for analysis and study. However, there have been only few concerted efforts to bring together tools, platforms, storage, processing frameworks, and existing collections for mining and analysing Web archives.

Anand, A., Bedathur, S., Berberich, K., & Schenkel, R. (2012). Index Maintenance for Time-travel Text Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 235–244). New York, NY, USA: ACM. https://doi.org/10.1145/2348283.2348318

Time-travel text search enriches standard text search by temporal predicates, so that users of web archives can easily retrieve document versions that are considered relevant to a given keyword query and existed during a given time interval. Different index structures have been proposed to efficiently support time-travel text search. None of them, however, can easily be updated as the Web evolves and new document versions are added to the web archive. In this work, we describe a novel index structure that efficiently supports time-travel text search and can be main- tained incrementally as new document versions are added to the web archive. Our solution uses a sharded index organiza- tion, bounds the number of spuriously read index entries per shard, and can be maintained using small in-memory buffers and append-only operations. We present experiments on two large-scale real-world datasets demonstrating that main- taining our novel index structure is an order of magnitude more efficient than

periodically rebuilding one of the existing index structures, while query-processing performance is not adversely affected.

Anand, A., Bedathur, S., Berberich, K., & Schenkel, R. (2010). Efficient Temporal Keyword Search over Versioned Text. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 699–708). New York, NY, USA: ACM. https://doi.org/10.1145/1871437.1871528

Modern text analytics applications operate on large volumes of temporal text data such as Web archives, newspaper archives, blogs, wikis, and micro-blogs. In these settings, searching and mining needs to use constraints on the time dimension in addition to keyword constraints. A natural approach to address such queries is using an inverted index whose entries are enriched with valid-time intervals. It has been shown that these indexes have to be partitioned along time in order to achieve efficiency. However, when the temporal predicate corresponds to a long time range, requiring the processing of multiple partitions, naive query processing incurs high cost of reading of redundant entries across partitions. We present a framework for efficient approximate processing of keyword queries over a temporally partitioned inverted index which minimizes this overhead, thus speeding up query processing. By using a small synopsis for each partition we identify partitions that maximize the number of final non-redundant results, and schedule them for processing early on. Our approach aims to balance the estimated gains in the final result recall against the cost of index reading required. We present practical algorithms for the resulting optimization problem of index partition selection. Our experiments with three diverse, large-scale text archives reveal that our proposed approach can provide close to 80% result recall even when only about half the index is allowed to be read.

Anand, A., Bedathur, S., Berberich, K., Schenkel, R., & Tryfonopoulos, C. (2009). EverLast: A Distributed Architecture for Preserving the Web. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 331–340). New York, NY, USA: ACM. https://doi.org/10.1145/1555400.1555455

The World Wide Web has become a key source of knowledge pertaining to almost every walk of life. Unfortunately, much of data on the Web is highly ephemeral in nature, with more than 50-80% of content estimated to be changing within a short time. Continuing the pioneering efforts of many national (digital) libraries, organizations such as the International Internet Preservation Consortium (IIPC), the Internet Archive (IA) and the European Archive (EA) have been tirelessly working towards preserving the ever changing Web. However, while these web archiving efforts have paid significant attention towards long term preservation of Web data, they have paid little attention to developing an global-scale infrastructure for collecting, archiving, and performing historical analyzes on the collected data. Based on insights from our recent work on building text analytics for Web Archives, we propose EverLast, a scalable distributed framework for next generation Web archival and temporal text analytics over the archive. Our system is built on a loosely-coupled distributed architecture that can be deployed over large-scale peer-to-peer networks. In this way, we allow the integration of many archival efforts taken mainly at a national level by national digital libraries. Key features of EverLast include support of time-based text search & analysis and the use of human-assisted archive gathering. In this paper, we outline the overall architecture of EverLast, and present some promising preliminary results.

Androvič, I. A., Bizík, B. A., Hausleitner, I. P., Katrincová, P. B., Lacková, M. I., & Matúšková, P. J. (2017). Digitálne pramene - národný projekt zberu a archivácie v roku 1. *Knihovna PLUS*, (1), 1–14. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

In 2015 the University Library in Bratislava put in the practice the national project Digital Resources -- Webharvesting and E-Born Content Archiving. The project was running in the framework of the Operational Program Informatisation of Society. Its ambition was to establish a technical, application and management infrastructure for systematical harvesting and long term preservation of web pages and e-Born resources. The implementation is based on open source software modules (Heritrix, OpenWayback, Invenio). The systems management is optimized for parallel webharvesting. This article presents the experiences and results of the operation of IS Digital Resources in 2016. It describes the workflow of webharvesting and acquisition of e-Born resources and discusses some methodological and practical problems in dealing with e-Born serials. The article brings the analytical and statistical overview of harvests realised in 2016 with a special highlight on the complex harvest of the national .sk domain. (English) [ABSTRACT FROM AUTHOR]

Androvič, I. A., & Prókai, M. (2007). Web-archívum made in Slovakia: Kísérleti projekt az elektronikus információforrások gyűjtésére és archiválására A web mint kulturális örökség. *Tudományos És Műszaki Tájékoztatás*, *54*(10). Retrieved from http://tmt-archive.omikk.bme.hu/show_news.html?id=4788&issue_id=487

A Cseh Nemzeti Könyvtár és a Pozsonyi Egyetemi Könyvtár (PEK) a CULTURE 2000 európai program keretében vállalta a web archiválási módszereinek, szempontjainak kidolgozását. Ezzel egyidejűleg Szlovákiában is megkezdték a web minőségi és mennyiségi felmérését, a szlovák nemzeti doménnel rendelkező weboldalak feltérképezését. 2006. májusi adatok szerint a szlovák nemzeti domén - .sk - keretében összesen 92 ezer doménnevet regisztrált mintegy 46 961 felhasználó.

Aniesa, A., & Bouchard, A. (2017). Constituer un réseau d'accès aux archives de l'internet : l'exemple français. In *IFLA Congress 2017, Wroclaw, Poland*. Retrieved from http://library.ifla.org/1655/

Depuis 2006, la BnF a pour mission de collecter l'internet français au titre du dépôt légal. Pour remplir cette mission au mieux, elle a progressivement mis en place un système d'archivage complet et ainsi collecté des milliards de pages web. Sur la base du décret d'application de la loi DADVSI, la BnF a cherché à rendre ses collections d'archives de l'internet, à l'origine uniquement consultables dans ses espaces Recherche, accessibles dans d'autres établissements en région. Cet article présente les différentes étapes de l'ouverture de ces accès : l'habilitation des bibliothèques de dépôt légal imprimeur ; les problématiques organisationnelles et techniques rencontrées et les solutions adoptées ; les enjeux au stade actuel du projet, alors que seize établissements sont déjà équipés d'un service d'accès aux archives de l'internet.

Anonymous. (2017). Discovery Happens Here: PW Talks with Wikipedia's Jake Orlowitz. *Publishers Weekly*, *264*(38), 28. Retrieved from https://search.proquest.com/docview/1940703367?accountid=27464

[...]we’re looking to provide a better experience for our users.[...]we’re working with partners like the Internet Archive to make sure more than a million URLs are properly archived and functioning; with OCLC to make it possible to cite books automatically, via an ISBN; and with OAdoi and OAbot to make free versions of paywalled sources cited on Wikipedia accessible and easy to find.[...]our hope is that readers who engage with Wikipedia will go on to explore the full-text resources cited there, whether in books, repositories, publisher websites, or, of course, in their public or university libraries.[...]those edits must pass through machine learning bots running on increasingly sophisticated neural networks looking for common vandalism patterns, through hundreds of language-matching RegEx filters catching bad words, through thousands of human “recent change” patrollers, and through tens of thousands of people’s personal article watch lists.There’s been tremendous evolution and flux around everything from peer review, to article levels and alternative metrics, open access and business models, creative commons licensing, social media, you name it.

Anonymous. (2017). The Business of Making E-books Free. *Publishers Weekly*, *264*(39), 4. Retrieved from https://search.proquest.com/docview/1943436218?accountid=27464

The work of scanning and preparing digital editions is being done by the Internet Archive, an online repository containing 11 million books, founded by Kahle in 1996 with the goal of making as much of the world’s written, visual, and audio content available for free as possible.Brand said she believes the language of the existing contracts used by the press allows it to digitize the books without seeking renewed permissions, “but out of courtesy to authors and their estates, we’re reaching out for every single book.”Brand said a small number of authors refused to give permission, but she added that asking them is part and parcel of the mission of the press to devise novel ways to protect the works it publishes and the authors who write them.

Antracoli, A., Duckworth, S., Silva, J., & Yarmey, K. (2014). Capture All the URLs: First Steps in Web Archiving. *Pennsylvania Libraries*, *2*(2), 155–170. https://doi.org/http://dx.doi.org/10.5195/palrap.2014.67

As higher education embraces new technologies, university activities--including teaching, learning, and research--increasingly take place on university websites, on university-related social media pages, and elsewhere on the open Web. Despite perceptions that “once it’s on the Web, it’s there forever,” this dynamic digital content is highly vulnerable to degradation and loss. In order to preserve and provide enduring access to this complex body of university records, archivists and librarians must rise to the challenge of Web archiving. As digital archivists at our respective institutions, the authors introduce the concept of Web archiving and articulate its importance in higher education. We provide our institutions’ rationale for selecting subscription service Archive-It as a preservation tool, outline the progress of our institutional Web archiving initiatives, and share lessons learned, from unexpected stumbling blocks to strategies for raising funds and support from campus stakeholders.

Arvidson, A., & Lettenström, F. (1998). The Kulturarw Project — The Swedish Royal Web Archive. *The Electronic Library*, *16*(2), 105–108. https://doi.org/10.1108/eb045623

KB (Kungliga biblioteket, The Royal Library), The National Library of Sweden, was founded in the 1500s. Since 1661, when the first Legal Deposit Law was introduced, it has functioned as the kingdom’s national memory. Today KB receives everything printed that is distributed to the public in the form of books, journals, posters, maps, advertisements, catalogues and so

on. Since 1994, following the latest version of the Legal Deposit Law, KB has also stored electronic publications ‘in fixed form’, i.e. published on CD‐ROM, tape or diskette. The total growth at KB is about 1.5 shelf‐kilometres per year. At that rate, KB’s underground storage will be completely full by the year 2050.

Asahara, M., Maekawa, K., Imada, M., Kato, S., & Konishi, H. (2014). Archiving and Analysing Techniques of the Ultra-Large-Scale Web-Based Corpus Project of NINJAL, Japan. *Alexandria: The Journal of National and International Library and Information Issues*, *25*(1–2), 129–148. https://doi.org/10.7227/ALX.0024

In 2011, the National Institute for Japanese Language and Linguistics (NINJAL) launched a corpus compilation project to construct a web corpus for linguistic research comprising ten billion words by 2016. The project is divided into four categories: Page Collection, Linguistic Annotation, Release and Preservation. For Page Collection, web crawlers are employed to collect web text by crawling 100 million pages every three months and retaining several versions of the text for three-month periods. For Linguistic Annotation, the linguistic studies web corpus contains annotated linguistic information. To improve the usability of these linguistic resources, normalization tasks such as tag removal, word segmentation, dependency parsing, and register estimation are performed. For Release, word lists and n-gram data are published based on the crawled and annotated text corpus. In addition, applications are being developed to enable searching for morphosyntax patterns in the ten-billion-word corpus. For Preservation, crawled web pages are preserved in chronological order as web archives primarily to support the survey of ongoing linguistic changes. In this paper, we present the basic design of the four categories. Additionally, we report the current status of the corpus using basic statistics of the crawled data and discuss the importance of deduplicating sentences. [ABSTRACT FROM AUTHOR]

Aturban, M., Kelly, M., Alam, S., Berlin, J. A., Nelson, M. L., & Weigle, M. C. (2018). ArchiveNow. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL ’18* (pp. 321–322). New York, New York, USA: ACM Press. https://doi.org/10.1145/3197026.3203880

ArchiveNow is a Python module for preserving web pages in on- demand web archives. This module allows a user to submit a URI of a web page for archiving at several configured web archives. Once the web page is captured, ArchiveNow provides the user with links to the archived copies of the web page. ArchiveNow is initially configured to use four archives but is easily configurable to add or remove other archives. In addition to pushing web pages to public archives, ArchiveNow , through the use of Wget and Squidwarc , allows users to generate local WARC files, enabling them to create their own personal and private archives.

Aubry, S. (2010). Introducing Web Archives as a New Library Service: the Experience of the National Library of France. *LIBER Quarterly*, *20*(2), 179. https://doi.org/10.18352/lq.7987

The collections held by the National Library of France (BnF) are part of the national heritage and include nearly 31 million documents of all types (books, journals, manuscripts, photographs, maps, etc.). New collection challenges have been posed by the emergence of the Internet. Within an international framework, the BnF is developing policy guidelines, workflows and tools to harvest relevant and representative segments of the French part of the Internet and organise their preservation and access. The Web archives of the French national

domain were developed as a new service, released as a new application and made available to the public in April 2008. Since then, strategies have been and continue to be developed to involve librarians and reach out end users. This article will discuss the BnF experiment and will focus specifically on four issues: * collection building: Web archives as a new and challenging collection, * resource discovery: access services and tools for end users, * usage: facts and figures, * involvement: strategies to build a librarian community and reach out end users.

BAILEY, S., Thomson, D., & Szalóki, G. (2006). Az első nyilvános webarchívum az Egyesült Királyságban. *Tudományos És Műszaki Tájékoztatás*, *53*(10). Retrieved from http://tmt-archive.omikk.bme.hu/show_news.html?id=4555&issue_id=476

Sokak számára a web az elsődleges információforrás, eddig mégis kevés figyelmet fordítottak a weboldalak hosszú távú megőrzésére, ami azzal a veszéllyel jár, hogy felbecsülhetetlen tudományos és kulturális értékek vesznek el a jövő generációi számára. A probléma megoldására hat vezető brit intézmény dolgozik közösen egy tesztelési környezet kidolgozásán, amely alapján kiválaszthatók az archiválni kívánt weboldalak. A hat intézmény: Brit Nemzeti Levéltár, Brit Nemzeti Könyvtár, Közös Információs Rendszerek Bizottsága (JISC), a skót és a walesi nemzeti könyvtárak és a Wellcome Könyvtár, megalakította az Egyesült Királyság Webarchiválási Konzorciumát (UK Web Archiving Consortium = UKWAC). Az archiválásra az Ausztrál Nemzeti Könyvtár által kifejlesztett PANDAS (PANDORA Digital Archival System = Pandora Digitális Archiváló Rendszer) szoftvert használják. A partnerek az adott intézmény szakterületéhez kapcsolódó oldalakat mentik el.

Baltussen, L. B., Blom, J., Medjkoune, L., Pop, R., Van Gorp, J., Huurdeman, H., & Haaijer, L. (2014). Hard Content, Fab Front-End: Archiving Websites of Dutch Public Broadcasters. *Alexandria: The Journal of National and International Library and Information Issues*, *25*(1–2), 69–91. https://doi.org/10.7227/ALX.0021

Although there are a great variety of web archiving projects around the world, there are not many that focus explicitly on websites of broadcasters. The reason is that funds are often lacking to do this, and that broadcaster websites are difficult to archive, due to their dynamic and audiovisual content. The Netherlands Institute for Sound and Vision, with its collection of over 800,000 hours of audiovisual content has been involved in a small-scale research project related to web archiving since 2008. When Sound and Vision was approached by Dutch public broadcaster NTR to archive four of its websites, it was decided to start a collaborative pilot project that focused both on learning more about archiving broadcaster websites and developing a clean and modern public access interface. The main lesson learned from this pilot is that to archive highly dynamic and AV-heavy broadcaster websites it is vital to use supplementary capture tools and manual archiving of this ‘difficult’ content. Furthermore, since the focus of web archiving projects is usually not on a good-looking front-end, the wheel had to be partly re-invented by involving various stakeholders and determining the most important requirements. The first version of the web archive was evaluated by various prospective target users. This evaluation revealed that the participants indeed appreciated the look and speed of the web archive, and that users needed to be made more aware of the web archive’s purpose and limitations. The work will be continued and scaled up, by archiving more broadcaster websites, continuing the research on how best to capture and make accessible dynamic and AV content, and by creating standard practices for making the web archive publicly available.

Banos, V., & Manolopoulos, Y. (2016). A quantitative approach to evaluate Website Archivability using the CLEAR+ method. *International Journal on Digital Libraries*, *17*(2), 119–141. https://doi.org/http://dx.doi.org/10.1007/s00799-015-0144-4

Website Archivability (WA) is a notion established to capture the core aspects of a website, crucial in diagnosing whether it has the potential to be archived with completeness and accuracy. In this work, aiming at measuring WA, we introduce and elaborate on all aspects of CLEAR+, an extended version of the Credible Live Evaluation Method for Archive Readiness (CLEAR) method. We use a systematic approach to evaluate WA from multiple different perspectives, which we call Website Archivability Facets. We then analyse archiveready.com, a web application we created as the reference implementation of CLEAR+, and discuss the implementation of the evaluation workflow. Finally, we conduct thorough evaluations of all aspects of WA to support the validity, the reliability and the benefits of our method using real-world web data.

Banos, V., & Manolopoulos, Y. (2015). Web Content Management Systems Archivability. In M. Tadeusz, P. Valduriez, & L. Bellatreche (Eds.), *Advances in Databases and Information Systems. ADBIS 2015* (pp. 198–212). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-23135-8_14

Web archiving is the process of collecting and preserving web content in an archive for current and future generations. One of the key issues in web archiving is that not all websites can be archived correctly due to various issues that arise from the use of different technologies, standards and implementation practices. Nevertheless, one of the common denominators of current websites is that they are implemented using a Web Content Management System (WCMS). We evaluate the Website Archivability (WA) of the most prevalent WCMSs. We investigate the extent to which each WCMS meets the conditions for a safe transfer of their content to a web archive for preservation purposes, and thus identify their strengths and weaknesses. More importantly, we deduce specific recommendations to improve the WA of each WCMS, aiming to advance the general practice of web data extraction and archiving.

Barik, T., Lubick, K., Smith, J., Slankas, J., & Murphy-Hill, E. (2015). Fuse: A Reproducible, Extendable, Internet-scale Corpus of Spreadsheets. In *Proceedings of the 12th Working Conference on Mining Software Repositories* (pp. 486–489). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2820518.2820594

Spreadsheets are perhaps the most ubiquitous form of end-user programming software. This paper describes a corpus, called Fuse, containing 2,127,284 URLs that return spreadsheets (and their HTTP server responses), and 249,376 unique spreadsheets, contained within a public web archive of over 26.83 billion pages. Obtained using nearly 60,000 hours of computation, the resulting corpus exhibits several useful properties over prior spreadsheet corpora, including reproducibility and extendability. Our corpus is unencumbered by any license agreements, available to all, and intended for wide usage by end-user software engineering researchers. In this paper, we detail the data and the spreadsheet extraction process, describe the data schema, and discuss the trade-offs of Fuse with other corpora.

Bartlett, V. (2014). New medium, old archives? Exploring archival potential in The Live Art Collection of the UK Web Archive. *International Journal of Performance Arts and Digital Media*, *10*(1), 91–103. https://doi.org/10.1080/14794713.2014.912504

This article speculates about the new kinds of historical information that performance scholars may be able to preserve as a result of recent innovations in web archiving. Using The Live Art Collection of the UK Web Archive as its case study, the article draws on influences from oral history, new media theory and the digital humanities. Beginning with an assertion that the Web has a tendency to aggregate existing media forms into one archival location, the article makes the case that online writing is key to web archiving's potential to document new kinds of knowledge about performance and live art. Subsequently it points to limitations in the current archival structures of the collection and concludes that further innovation is required in order to maximize the scholarly potential of the material contained within it. Interviews with the team who manage and curate the collection are used throughout to support assertions about the collections intended use and functions. [ABSTRACT FROM AUTHOR]

Beasley, S., & Kail, C. (2009). a2o: Access to Archives from the National Archives of Singapore. *Journal of Web Librarianship*, *3*(2), 149–155. https://doi.org/10.1080/19322900902896531

The article offers information about a2o that was created by the National Archives of Singapore in 2009. Accordingly, a2o is taken after the chemical symbol of water, which is considered as an essential element of life. It provides access to various databases, photographs, maps and plans, oral history audio files, and other audiovisual recordings in multiple ways. It also offers a variety of online exhibitions, including "Colours Behind Barbed Wires: A Prisoner of War's Story through Haxworth's Sketches" and "Colours in the Wind: Hill Street Police Station in Retrospect."

BELOVARI, S. (2017). Historians and Web Archives. *Archivaria*, (83), 59–79. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Since the 1990s, the Web has increasingly become the location where we carry out our activities and generate primary and secondary records. Increasingly, such records exist only on the Web, with no complementary or supplementary records available elsewhere. While web archives began to preserve this legacy in 1991, web history has not yet emerged as a fully developed field. One explanation may be historians' concerns that they will not be able to replicate their historical research process when using web archives, and may not find essential and authoritative records. The article's first section proposes a thought experiment in which a future historian in 2050 wants to research web history using web archives as they existed in 2015. She relies on the customary historical research process through which historians choose topics and search, browse, and contextualize sources in depth and iteratively. The experiment fails when our historian is unable to locate appropriate repositories and authoritative records without resorting to the live Web of 2015. The second section then analyzes 21 eminent web archives in 2015 and issues that may have an impact on historical research. Most web archives are apparently akin to libraries of information resources. Archivists and historians, however, need web repositories to contain and make accessible essential web records of enduring cultural, historical, and evidentiary value. The article suggests that historians may once again prove invaluable in figuring out basic archival issues related to web records and archives, just as they helped shape archival policies a couple of centuries ago. (English) [ABSTRACT FROM AUTHOR]

Ben Saad, M., & Gançarski, S. (2011). Archiving the Web Using Page Changes Patterns: A Case Study. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 113–122). New York, NY, USA: ACM. https://doi.org/10.1145/1998076.1998098

A pattern is a model or a template used to summarize and describe the behavior (or the trend) of a data having generally some recurrent events. Patterns have received a considerable attention in recent years and were widely studied in the data mining field. Various pattern mining approaches have been proposed and used for different applications such as network monitoring, moving object tracking, financial or medical data analysis, scientific data processing, etc. In these different contexts, discovered patterns were useful to detect anomalies, to predict data behavior (or trend), or more generally, to simplify data processing or to improve system performance. However, to the best of our knowledge, patterns have never been used in the context of web archiving. Web archiving is the process of continuously collecting and preserving portions of the World Wide Web for future generations. In this paper, we show how patterns of page changes can be useful tools to efficiently archive web sites. We first define our pattern model that describes the changes of pages. Then, we present the strategy used to (i) extract the temporal evolution of page changes, to (ii) discover patterns and to (iii) exploit them to improve web archives. We choose the archive of French public TV channels « France Télévisions » as a case study in order to validate our approach. Our experimental evaluation based on real web pages shows the utility of patterns to improve archive quality and to optimize indexing or storing.

Benczúr, A. A., Erdélyi, M., Masanés, J., & Siklósi, D. (2009). Web Spam Challenge Proposal for Filtering in Archives. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web* (pp. 61–62). New York, NY, USA: ACM. https://doi.org/10.1145/1531914.1531928

In this paper we propose new tasks for a possible future Web Spam Challenge motivated by the needs of the archival community. The Web archival community consists of several relatively small institutions that operate independently and possibly over different top level domains (TLDs). Each of them may have a large set of historic crawls. Efficient filtering would hence require (1) enhanced use of the time series of domain snapshots and (2) collaboration by transferring models across different TLDs. Corresponding Challenge tasks could hence include the distribution of crawl snapshot data for feature generation as well as classification of unlabeled new crawls of the same or even different TLDs.

Ben-David, A. (2016). What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, *18*(7), 1103. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

This article argues that the use of the Web as a primary source for studying the history of nations is conditioned by the structural ties between sovereignty and the Internet protocol, and by a temporal proximity between live and archived websites. The argument is illustrated by an empirical reconstruction of the history of the top-level domain of Yugoslavia (.yu), which was deleted from the Internet in 2010. The archival discovery method used four lists of historical .yu Uniform Resource Locators (URLs) that were captured from the live Web before the domain was deleted, and an automated hyperlink discovery script that retrieved their

snapshots from the Internet Archive and reconstructed their immediate hyperlinked environment in a network. Although a considerable portion of the historical .yu domain was found on the Internet Archive, the reconstructed space was predominantly Serbian. [ABSTRACT FROM AUTHOR]

Ben-David, A., Amram, A., & Bekkerman, R. (2018). The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries*, *19*(1), 95–106. https://doi.org/10.1007/s00799-016-0202-6

This study examines the use of visual data analytics as a method for historical investigation of national Webs, using Web archives. It empirically analyzes all graphically designed (non-photographic) images extracted from Websites hosted in the historical .yu domain and archived by the Internet Archive between 1997 and 2000, to assess the utility and value of visual data analytics as a measure of nationality of a Web domain. First, we report that only 23.5% of Websites hosted in the .yu domain over the studied years had their graphically designed images properly archived. Second, we detect significant differences between the color palettes of .yu sub-domains (commercial, organizational, academic, and governmental), as well as between Montenegrin and Serbian Websites. Third, we show that the similarity of the domains' colors to the colors of the Yugoslav national flag decreases over time. However, there are spikes in the use of Yugoslav national colors that correlate with major developments on the Kosovo frontier.

Ben-David, A., & Huurdeman, H. (2014). Web Archive Search as Research: Methodological and Theoretical Implications. *Alexandria: The Journal of National and International Library and Information Issues*, *25*(1–2), 93–111. https://doi.org/10.7227/ALX.0022

The field of web archiving is at a turning point. In the early years of web archiving, the single URL has been the dominant unit for preservation and access. Access tools such as the Internet Archive's Wayback Machine reflect this notion as they allowed consultation, or browsing, of one URL at a time. In recent years, however, the single URL approach to accessing web archives is being gradually replaced by search interfaces. This paper addresses the theoretical and methodological implications of the transition to search on web archive research. It introduces "search as research" methods, practices already applied in studies of the live web, which can be repurposed and implemented for critically studying archived web data. Such methods open up a variety of analytical practices that were so far precluded by the single URL entry point to the web archive, such as the re-assemblage of existing collections around a theme or an event, the study of archival artefacts and scaling the unit of analysis from the single URL to the full archive, by generating aggregate views and summaries. The paper introduces examples to "search as research" scenarios, which have been developed by the WebART project at the University of Amsterdam and the Centrum Wiskunde & Informática, in collaboration with the National Library of the Netherlands. The paper concludes with a discussion of current and potential limitations of "search as research" methods for studying web archives, and the ways with which they can be overcome in the near future.

Berberich, K., Bedathur, S., Neumann, T., & Weikum, G. (2007). FluxCapacitor: Efficient Time-travel Text Search. In *Proceedings of the 33rd International Conference on Very Large Data Bases* (pp. 1414–1417). VLDB Endowment. Retrieved from http://dl.acm.org/citation.cfm?id=1325851.1326029

An increasing number of temporally versioned text collections is available today with Web archives being a prime example. Search on such collections, however, is often not satisfactory and ignores their temporal dimension completely. Time-travel text search solves this problem by evaluating a keyword query on the state of the text collection as of a user-specified time point. This work demonstrates our approach to efficient time-travel text search and its implementation in the FLUXCAPACITOR prototype.

Berberich, K., Bedathur, S., & Weikum, G. (2006). Rank Synopses for Efficient Time Travel on the Web Graph. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (pp. 864–865). New York, NY, USA: ACM. https://doi.org/10.1145/1183614.1183769

Berlin, Andrew, J. (2018). To Relive the Web: A Framework for the Transformation and Archival Replay of Web Pages. United States, North America: ODU Digital Commons. https://doi.org/10.25777/n8m g-da06

When replaying an archived web page (known as a memento), the fundamental expectation is that the page should be viewable and function exactly as it did at archival time. However, this expectation requires web archives to modify the page and its embedded resources, so that they no longer reference (link to) the original server(s) they were archived from but back to the archive. Although these modifications necessarily change the state of the representation, it is understood that without them the replay of mementos from the archive would not be possible. Unfortunately, because the replay of mementos and the modifications made to them by web archives in order to facilitate replay varies between archives, the terminology for describing replay and the modification made to mementos for facilitating replay does not exist. In this thesis, we propose terminology for describing the existing styles of replay and the modifications made on the part of web archives to mementos in order to facilitate replay. This thesis also, in the process of defining terminology for the modifications made by client-side rewriting libraries to the JavaScript execution environment of the browser during replay, proposes a general framework for the auto-generation of client-side rewriting libraries. Finally, we evaluate the effectiveness of using a generated client-side rewriting library to augment the existing replay systems of web archives by crawling mementos replayed from the Internet Archive's Wayback Machine with and without the generated client-side rewriter. By using the generated client-side rewriter we were able to decrease the cumulative number of requests blocked by the content security policy of the Wayback Machine for 577 mementos by 87.5% and increased the cumulative number of requests made by 32.8%. Also by using the generated client-side rewriter, we were able to replay mementos that were previously not replayable from the Internet Archive.

Bingham, N. (2014). Quality Assurance Paradigms in Web Archiving Pre and Post Legal Deposit. *Alexandria: The Journal of National and International Library and Information Issues*, *25*(1–2), 51–68. https://doi.org/10.7227/ALX.0020

This article discusses quality assurance paradigms in the pre and post legal deposit environments, exploring how workflows and processes have adapted from a small-scale, selective model to domain-scale harvesting activity. It draws comparisons between the two approaches and discusses the trade-offs necessitated by the change in scale of web harvesting activity. The requirements of the non-print legal deposit legislation of 2013 and the change in scale in web archiving operations have necessitated new quality metrics for the web archive collection. Whereas it was possible to manually review every instance of a harvested website,

the new model requires that more automated methods are employed. The article looks at the tools employed in the selective web archiving model such as the Web Curator Tool and those designed for the legal deposit workflow such as the Annotation and Curation Tool. It examines the key technical issues in archiving websites and how content is prioritized for quality assurance. The article will be of interest to people employed in memory institutions including national libraries who are tasked with preserving online content as well as a wider general audience.

Black, M. L. (2016). The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research. *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities*, *10*(1), 95–109. Retrieved from http://10.0.13.38/ijhac.2016.0162

While intellectual property protections effectively frame digital humanities text mining as a field primarily for the study of the nineteenth century, the Internet offers an intriguing object of study for humanists working in later periods. As a complex data source, the World Wide Web presents its own methodological challenges for digital humanists, but lessons learned from projects studying large nineteenth century corpora offer helpful starting points. Complicating matters further, legal and ethical questions surrounding web scraping, or the practice of large scale data retrieval over the Internet, will require humanists to frame their research to distinguish it from commercial and malicious activities. This essay reviews relevant research in the digital humanities and new media studies in order to show how web scraping might contribute to humanities research questions. In addition to recommendations for addressing the complex concerns surrounding web scraping this essay also provides a basic overview of the process and some recommendations for resources. [ABSTRACT FROM AUTHOR]

Bolette, J., & Zierau, E. (2017). Data Management of Web Archive Research Data. In *"Researchers, pratictioners and their use of the archived web", London, School of Advanced Study, University of London*. Retrieved from https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-JurikZierau-Data_management_of_web_archive_research_data.pdf

This paper will provide recommendations to overcome various challenges for data management of web materials. The recommendations are based on results from two independent Danish research projects with different requirements to data management: The first project focuses on high precision on a par with traditional references for analogue material and with web materials found in different web archives. The second project focuses on large corpora (collections) of archived web references as basis for analysis.

Bonnel, S., & Oury, C. (2014). *Selecting websites in an encyclopaedic national library ; La sélection de sites web dans une bibliothèque nationale encyclopédique ; Selecting websites in an encyclopaedic national library : A shared collection policy for BnF internet legal deposit ; La s*. France, Europe: HAL CCSD. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

International audience ; En quelques années, le web est devenu l'un des principaux vecteurs d'expression et de consommation culturelles de la société française ; les publications en ligne ont rejoint notre patrimoine. Celui-ci est d'autant plus précieux qu'il est fragile. En France, il

a été décidé d'inscrire la mission de conservation de l'internet dans le sillage pluriséculaire du dépôt légal. Cependant, l'adaptation de ce dispositif juridique et scientifique à un espace de diffusion aussi vaste et étendu n'a rien d'évident. La BnF définit son périmètre de collecte par une série de restrictions successives : juridiques, techniques et économiques. Pour assurer la représentativité de son dépôt légal, la BnF a également adopté un modèle original d'archivage, qui associe des collectes « larges » du domaine national et des approches plus ciblées de sites identifiés par des bibliothécaires de la BnF ou par des partenaires. La BnF a ainsi été amenée à appliquer des logiques de sélection dans un cadre de dépôt légal. À cette fin, chaque département associé à la collecte du web a élaboré, au fil des expérimentations, sa propre stratégie documentaire. Les « correspondants » du dépôt légal du web ont adopté des logiques non pas contradictoires mais complémentaires : sélection / échantillonnage, continuité des collections / exploration de nouveaux territoires. C'est désormais à une synthèse de ces différentes politiques que la BnF doit s'atteler, dans le cadre de la refonte de sa charte documentaire et dans un contexte où les contraintes budgétaires appellent à la définition de priorités plus affirmées.

Boss, K., & Broussard, M. (2017). Challenges of archiving and preserving born-digital news applications. *IFLA Journal*, *43*(2), 150–157. https://doi.org/http://dx.doi.org/10.1177/0340035216686355

Born-digital news content is increasingly becoming the format of the first draft of history. Archiving and preserving this history is of paramount importance to the future of scholarly research, but many technical, legal, financial, and logistical challenges stand in the way of these efforts. This is especially true for news applications, or custom-built websites that comprise some of the most sophisticated journalism stories today, such as the "Dollars for Docs" project by ProPublica. Many news applications are standalone pieces of software that query a database, and this significant subset of apps cannot be archived in the same way as text-based news stories, or fully captured by web archiving tools such as Archive-It. As such, they are currently disappearing. This paper will outline the various challenges facing the archiving and preservation of born-digital news applications, as well as outline suggestions for how to approach this important work.

Böhm, T. (2014). Development of the National Library of the Czech Republic 2011–2016: Past, Present and Future. *Alexandria: The Journal of National and International Library and Information Issues*, *25*(3), 17–24. https://doi.org/10.7227/ALX.0028

The National Library of the Czech Republic, which was founded in 1773 by the Austrian Empress Maria Theresa, is one of the oldest National Libraries in Europe. It has been through various organizational changes incorporating other libraries and institutions. In addition to providing traditional library service, the library is active in such fields as digitization, paper documents restoration and preservation, refurbishment of its main seat in the baroque Klementinum building and international cooperation. The most important digitization project is the creation of the National Digital Library, which will also serve as the LTP (Long Term Preservation) repository for other digitization projects carried out by either the National Library or by other libraries and institutions in the Czech Republic. Other projects in this field are: the world's biggest digital manuscript library (Manuscriptorium), creation of the Web Archive, digitization of rare books in partnership with Google, formation of the repository for digitized Czech cultural heritage and, together with other main Czech libraries, work on the creation of the Czech Libraries Portal. The Library is further active in paper documents restoration and preservation where it is trying to tackle the problem of de-acidification as well

as the formation of the physical Czech Depository Library and the Interdisciplinary Methodological Centre for Book Restoration and Conservation. The Library continues to serve its users during the refurbishment of the Klementinum. It aims to create “a modern library in baroque walls” by the end of 2018. Furthermore, a new physical depository has been bulit on the outskirts of Prague. [ABSTRACT FROM AUTHOR]

Brack, M. (2012). The Future of the Past of the Web. *Ariadne*, (68). https://doi.org/http://www.ariadne.ac.uk/issue68/fpw11-rpt

We have all heard at least some of the extraordinary statistics that attempt to capture the sheer size and ephemeral nature of the Web. According to the Digital Preservation Coalition (DPC), more than 70 new domains are registered and more than 500,000 documents are added to the Web every minute. This scale, coupled with its ever-evolving use, present significant challenges to those concerned with preserving both the content and context of the Web. Co-organised by the DPC, the British Library and JISC, this workshop was the third in a series of discussions around the nature and potential of Web archiving. Following the key note address, two thematic sessions looked at “Using Web Archives” (as it is only recently that use cases have started to emerge) and “Emerging Trends” (acknowledging that Web archiving activities are on the increase, along with a corresponding rise in public awareness). Adapted from the source document.

Braman, S. (2017). Internet histories: the view from the design process. *Internet Histories*, *1*(1–2), 70–78. https://doi.org/10.1080/24701475.2017.1305716

The electrical engineers and computer scientists who have designed the Internet are among those who have written Internet history. They have done so within the technical document series created to provide a medium for and record of the design process, the Internet Requests for Comments (RFCs) as well as in other venues. Internet designers have explicitly written the network’s history in documents explicitly devoted to history as well as indirectly in documents focused on technical matters. The Internet RFCs also provide data for research on Internet history and on large-scale sociotechnical infrastructure written by outsiders to the design process. Incorporating the history of the Internet as understood by those responsible for its design, whether in their own words or by treating the design conversation as data, makes visible some elements of that history not otherwise available, corrects misperceptions of factors underlying some of its features, and provides fascinating details on the people and events involved that are of interest to those seeking to understand the Internet. Within the RFCs, history has served both technical and social functions.

Brightenburg, C. (2016). The digitization of early English books: A database comparison of Internet Archive and Early English Books Online. *Journal of Electronic Resources Librarianship*, *28*(1), 1–8. https://doi.org/http://dx.doi.org/10.1080/1941126X.2016.1130448

The use of digital books is diverse, ranging from casual reading to in-depth primary source research. Digitization of early English printed books in particular, has provided greater access to a previously limited resource for academic faculty and researchers. Internet Archive, a free, internet website and Early English Books Online, a subscription based database are two such resources. This study compares the scope, coverage and visual quality of the two book databases to determine the usability of each for faculty and researchers.

Brown, H. (2013). The Interconnected Web: A Paradigm for Managing Digital Preservation. *World Digital Libraries*, *6*(1), 1. https://doi.org/10.3233/WDL-120096

Digital preservation management has evolved from an initial emphasis on technological issues to a broader understanding of resourcing and organizational issues. Internationally, the trend has moved to a risk management framework that is common to both digital and physical worlds. There are a number of common "high level" principles and frameworks that intersect both digital and traditional (physical) preservation, and which in turn provide an opportunity to explore an integrated approach to preserving both digital and physical materials. This paper explores the opportunity for such an integrated approach through the paradigm of an interconnected web. [ABSTRACT FROM AUTHOR]

Brown, K. E. K. (2015). Personal Archiving: Preserving Our Digital Heritage. *Library Resources & Technical Services*, *59*(2), 94–95. Retrieved from https://search.proquest.com/docview/1684295944?accountid=27464

Extending from this, Danielle Conklin, author of chapter 2, "Personal Archiving for Individuals and Families" does a terrific job of outlining risks, such as obsolescence of the formats and software, the need to migrate information forward, the importance of keeping your collections organized, and distributing copies to assist preservation efforts. Richard Banks goes on in "Our Technology Heritage" in chapter 11 about devices that could, if ever fully developed, bring our digital lives into our physical lives (he values the idea of displaying our digital images, for example, in our homes), but ri ght now little boxes that sit around and salvage and store information don't seem exactly visionär)'.

Brunelle, J. F. (2016). *Scripts in a frame: A framework for archiving deferred representations*. *ProQuest Dissertations and Theses*. Old Dominion University, Ann Arbor. Retrieved from https://search.proquest.com/docview/1803306325?accountid=27464

Web archives provide a view of the Web as seen by Web crawlers. Because of rapid advancements and adoption of client-side technologies like JavaScript and Ajax, coupled with the inability of crawlers to execute these technologies effectively, Web resources become harder to archive as they become more interactive. At Web scale, we cannot capture client-side representations using the current state-of-the art toolsets because of the migration from Web pages to Web applications. Web applications increasingly rely on JavaScript and other client-side programming languages to load embedded resources and change client-side state. We demonstrate that Web crawlers and other automatic archival tools are unable to archive the resulting JavaScript-dependent representations (what we term deferred representations), resulting in missing or incorrect content in the archives and the general inability to replay the archived resource as it existed at the time of capture. Building on prior studies on Web archiving, client-side monitoring of events and embedded resources, and studies of the Web, we establish an understanding of the trends contributing to the increasing unarchivability of deferred representations. We show that JavaScript leads to lower-quality mementos (archived Web resources) due to the archival difficulties it introduces. We measure the historical impact of JavaScript on mementos, demonstrating that the increased adoption of JavaScript and Ajax correlates with the increase in missing embedded resources. To measure memento and archive quality, we propose and evaluate a metric to assess memento quality closer to Web users' perception. We propose a two-tiered crawling approach that enables crawlers to capture embedded resources dependent upon JavaScript. Measuring the performance benefits between

crawl approaches, we propose a classification method that mitigates the performance impacts of the two-tiered crawling approach, and we measure the frontier size improvements observed with the two-tiered approach. Using the two-tiered crawling approach, we measure the number of client-side states associated with each URI-R and propose a mechanism for storing the mementos of deferred representations. In short, this dissertation details a body of work that explores the following: why JavaScript and deferred representations are difficult to archive (establishing the term deferred representation to describe JavaScript dependent representations); the extent to which JavaSc…

Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2016). The impact of JavaScript on archivability. *International Journal on Digital Libraries*, *17*(2), 95–117. https://doi.org/10.1007/s00799-015-0140-8

Web Archiving Integration Layer (WAIL) is a desktop application writ- ten in Python that integrates Heritrix and OpenWayback. In this work we recreate and extend WAIL from the ground up to facilitate collection- based personal Web archiving. Our new iteration of the software, WAIL- Electron, leverages native Web technologies (e.g., JavaScript, Chromium) using Electron to open new potential for Web archiving by individ- uals in a stand-alone cross-platform native application. By replacing OpenWayback with PyWb, we provide a novel means for personal Web archivists to curate collections of their captures from their own per- sonal computer rather than relying on an external archival Web service. As extended features we also provide the ability for a user to monitor and automatically archive Twitter users' feeds, even those requiring authentication, as well as provide a reference implementation for in- tegrating a browser-based preservation tool into an OS native applica- tion.

Brunelle, J. F., Ferrante, K., Wilczek, E., Weigle, M. C., & Nelson, M. L. (2016). Leveraging Heritrix and the Wayback Machine on a Corporate Intranet: A Case Study on Improving Corporate Archives. *D-Lib Magazine*, *22*(1/2), 1. https://doi.org/10.1045/january2016-brunelle

In this work, we present a case study in which we investigate using open-source, web-scale web archiving tools (i.e., Heritrix and the Wayback Machine installed on the MITRE Intranet) to automatically archive a corporate Intranet. We use this case study to outline the challenges of Intranet web archiving, identify situations in which the open source tools are not well suited for the needs of the corporate archivists, and make recommendations for future corporate archivists wishing to use such tools. We performed a crawl of 143,268 URIs (125 GB and 25 hours) to demonstrate that the crawlers are easy to set up, efficiently crawl the Intranet, and improve archive management. However, challenges exist when the Intranet contains sensitive information, areas with potential archival value require user credentials, or archival targets make extensive use of internally developed and customized web services. We elaborate on and recommend approaches for overcoming these challenges.

Brunelle, J. F., Kelly, M., Salaheldeen, H., Weigle, M. C., & Nelson, M. L. (2015). Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries*, *16*(3–4), 283–301. https://doi.org/http://dx.doi.org/10.1007/s00799-015-0150-6

(ProQuest: ... denotes formulae and/or non-USASCII text omitted; see image) Issue Title: Focused Issue on Digital Libraries 2014 Web archives do not always capture every resource

on every page that they attempt to archive. This results in archived pages missing a portion of their embedded resources. These embedded resources have varying historic, utility, and importance values. The proportion of missing embedded resources does not provide an accurate measure of their impact on the Web page; some embedded resources are more important to the utility of a page than others. We propose a method to measure the relative value of embedded resources and assign a damage rating to archived pages as a way to evaluate archival success. In this paper, we show that Web users' perceptions of damage are not accurately estimated by the proportion of missing embedded resources. In fact, the proportion of missing embedded resources is a less accurate estimate of resource damage than a random selection. We propose a damage rating algorithm that provides closer alignment to Web user perception, providing an overall improved agreement with users on memento damage by 17 % and an improvement by 51 % if the mementos have a damage rating delta .....0.30. We use our algorithm to measure damage in the Internet Archive, showing that it is getting better at mitigating damage over time (going from a damage rating of 0.16 in 1998 to 0.13 in 2013). However, we show that a greater number of important embedded resources (2.05 per memento on average) are missing over time. Alternatively, the damage in WebCite is increasing over time (going from 0.375 in 2007 to 0.475 in 2014), while the missing embedded resources remain constant (13 % of the resources are missing on average). Finally, we investigate the impact of JavaScript on the damage of the archives, showing that a crawler that can archive JavaScript-dependent representations will reduce memento damage by 13.5 %.

Brunelle, J. F., & Nelson, M. L. (2013). An Evaluation of Caching Policies for Memento Timemaps. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 267–276). New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467717

As defined by the Memento Framework, TimeMaps are machine-readable lists of time-specific copies -- called "mementos" -- of an archived original resource. In theory, as an archive acquires additional mementos over time, a TimeMap should be monotonically increasing. However, there are reasons why the number of mementos in a TimeMap would decrease, for example: archival redaction of some or all of the mementos, archival restructuring, and transient errors of one or more archives. We study TimeMaps for 4,000 original resources over a three month period, note their change patterns, and develop a caching algorithm for TimeMaps suitable for a reverse proxy in front of a Memento aggregator. We show that TimeMap cardinality is constant or monotonically increasing for 80.2% of all TimeMap downloads in the observation period. The goal of the caching algorithm is to exploit the ideally monotonically increasing nature of TimeMaps and not cache responses with fewer mementos than the already cached TimeMap. This new caching algorithm uses conditional cache replacement and a Time To Live (TTL) value to ensure the user has access to the most complete TimeMap available. Based on our empirical data, a TTL of 15 days will minimize the number of mementos missed by users, and minimize the load on archives contributing to TimeMaps.

Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2017). Archival Crawlers and JavaScript: Discover More Stuff but Crawl More Slowly. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 1–10). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=3200334.3200336

The web is today's primary publication medium, making web archiving an important activity for historical and analytical purposes. Web pages are increasingly interactive, resulting in pages that are correspondingly difficult to archive. JavaScript enables interactions that can potentially change the client-side state of a representation. We refer to representations that load embedded resources via JavaScript as deferred representations. It is difficult to discover and crawl all of the resources in deferred representations and the result of archiving deferred representations is archived web pages that are either incomplete or erroneously load embedded resources from the live web. We propose a method of discovering and archiving deferred representations and their descendants (representation states) that are only reachable through client-side events. Our approach identified an average of 38.5 descendants per seed URI crawled, 70.9% of which are reached through an onclick event. This approach also added 15.6 times more embedded resources than Heritrix to the crawl frontier, but at a crawl rate that was 38.9 times slower than simply using Heritrix. If our method was applied to the July 2015 Common Crawl dataset, a web-scale archival crawler will discover an additional 7.17 PB (5.12 times more) of information per year. This illustrates the significant increase in resources necessary for more thorough archival crawls.

Brügger, N. (2016). Digital Humanities in the 21st Century:Digital Material as a Driving Force. *Digital Humanities Quarterly*, *10*(3). Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

In this article it is argued that one of the major transformative factors of the humanities at the beginning of the 21st century is the shift from analogue to digital source material, and that this shift will affect the humanities in a variety of ways. But various kinds of digital material are not digital in the same way, which a distinction between digitized, born-digital, and reborn-digital may help us acknowledge, thereby helping us to understand how each of these types of digital material affects different phases of scholarly work in its own way. This is illustrated by a detailed comparison of the nature of digitized collections and web archives. ; In this article it is argued that one of the major transformative factors of the humanities at the beginning of the 21st century is the shift from analogue to digital source material, and that this shift will affect the humanities in a variety of ways. But various kinds of digital material are not digital in the same way, which a distinction between digitized, born-digital, and reborn-digital may help us acknowledge, thereby helping us to understand how each of these types of digital material affects different phases of scholarly work in its own way. This is illustrated by a detailed comparison of the nature of digitized collections and web archives.

Brügger, N. (2013). Web historiography and Internet Studies: Challenges and perspectives. *New Media & Society*, *15*(5), 752–764. https://doi.org/10.1177/1461444812462852

I argue that web historiography should be placed higher on the Internet Studies' research agenda, since a better understanding of the web of the past is an important condition for gaining a more complete understanding of the web of today, regardless of our focus (e.g. political economy, language and culture, social interaction or everyday use). Building on reflections about 'historiography' and the 'web', I discuss several major challenges of web historiography vis-à-vis historiography in general, focusing on the characteristics of the archived website and the web sphere, and the consequences of these characteristics for web historians. I conclude by outlining future directions for web historiography. [ABSTRACT FROM AUTHOR]

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, *11*(1–2), 115–132. https://doi.org/10.1177/1461444808099574

Brügger, N. (2016). Introduction: The Web's first 25 years. *New Media & Society*, *18*(7), 1059–1065. https://doi.org/10.1177/1461444816643787

In August 2016, we can celebrate the 25th anniversary of the World Wide Web. Or can we? There is no doubt that the World Wide Web – or simply: the Web – has played an important role in the communicative infrastructure of most societies since the mid-1990s, but when did the Web actually start? And how has the Web developed from its beginning until today? The six articles in this Special Issue/section revolve around one of these questions in various ways.

Brügger, N., Goggin, G., Milligan, I., & Schafer, V. (2017). Introduction: Internet histories. *Internet Histories*, *1*(1–2), 1–7. https://doi.org/10.1080/24701475.2017.1317128

The ways in which historians define the Internet profoundly shape the histories we write. Many studies implicitly define the Internet in material terms, as a particular set of hardware and software, and consequently tend to frame the development of the Internet as the spread of these technologies from the United States. This essay explores implications of defining the Internet alternatively in terms of technology, use and local experience. While there is not a single "correct" definition, historians should be aware of the politics of the definitions they use.

Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History: Using Web Archives to Understand the Past and the Present* (1st ed.). United States, North America: UCL Press. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

London: UCL Press, c2017

Burda, D., & Teuteberg, F. (2015). Understanding Service Quality and System Quality Success Factors in Cloud Archiving from an End-User Perspective. *Information Systems Management*, *32*(4), 266–284. https://doi.org/http://dx.doi.org/10.1080/10580530.2015.1079998

This study seeks to explain the adoption of cloud storage services as a means of personal archiving thereby focusing on users' service and system quality perceptions and their drivers. The authors derive and empirically validate a model that incorporates users' perceptions of service/system quality as well as behavioral factors to explain usage. Finally, the authors highlight important determinants of system/service quality perceptions that cloud providers should pay attention to in their attempts to increase marketshare.

Burnhill, P. (2013). Tales from The Keepers Registry: Serial Issues About Archiving & the Web. *Serials Review*, *39*(1), 3–20. https://doi.org/10.1016/j.serrev.2013.02.003

Abstract: A key task for libraries is to ensure access for their patrons to the scholarly statements now found across the Internet. Three stories reveal progress towards success in that task. The context of these stories is the shift from print to digital format for all types of continuing resources, particularly journals, and the need to archive not just serials but also

ongoing ‘integrating resources’ such as databases and Web sites. The first story is about The Keepers Registry, an international initiative to monitor the extent of e-journal archiving. The second story is about the variety of ‘serial issues’ that have had to be addressed during the PEPRS (Piloting an E-journals Preservation Registry Service) project which was commissioned in the UK by JISC. These include identification, naming and identification of publishers, and the continuing need for a universal holdings statement. The role of the ISSN, and of the ISSN-L, has been a key. The third story looks beyond e-journals to new research objects and the dynamics of the Web, to the role of citation and fixity, and to broader matters of digital preservation. This story reflects upon seriality, as the Web becomes the principal arena and medium for scholarly discourse. Scientific discourse is now resident on the Web. Much that is issued on the Web is issued nowhere else: it is a digital native. Statistics that indicate the extent of archiving for e-journals to which major university libraries subscribe are also included in the article. [Copyright &y& Elsevier]

Bustillos, M., & Freshwater, S. (2018). Erasing history. (Cover story). *Columbia Journalism Review*, *57*(1), 112–118. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article discusses the digital journalism focusing on the failure of an online news outlet “The Honolulu Advertiser”. The author discusses the history of digital archiving systems, role played by U.S. government in protecting digital archival documents, and technological innovations that protects internet archives.

Cadavid, J. A. P. (2017). Evolution of legal deposit in New Zealand. *IFLA Journal*, *43*(4), 379–390. https://doi.org/http://dx.doi.org/10.1177/0340035217713763

The evolution of legal deposit shows changes and challenges in collecting, access to and use of documentary heritage. Legal deposit emerged in New Zealand at the beginning of the 20th century with the aim of preserving print publications mainly for the use of a privileged part of society. In the 21st century legal deposit has evolved to include the safeguarding of electronic resources and providing access to the documentary heritage for all New Zealanders. The National Library of New Zealand has acquired new functions for a proper stewardship of digital heritage. E-deposit and web harvesting are two new mechanisms for collecting New Zealand publications. The article proposes that legal deposit through human rights and multiculturalism should involve different communities of heritage in web curation.

Cadavid, J. A. P., & PABÓN CADAVID, J. A. (2014). Copyright Challenges of Legal Deposit and Web Archiving in the National Library of Singapore. *Alexandria*, *25*(1/2), 1–19. Retrieved from http://10.0.28.59/ALX.0017

This article discusses the development of web archiving in Singapore and its relationship to copyright law. The author describes legal deposit, its definition and historical development, the differences between voluntary and compulsory legal deposit, and the practices of such approaches within the National Library of Singapore. It highlights two main projects, the Singapore Memory Project and Web Archive Singapore (WAS). The paper analyses how the implementation of legal deposit for preserving web material creates a complex relationship between copyright and digital heritage, and describes difficulties that cover the information lifecycle of web archiving. Finally, the paper presents a set of conclusions and

recommendations regarding the need for modifying copyright legislation to foster research activities within Singapore's knowledge economy. [ABSTRACT FROM AUTHOR]

Caisley, J., Ball, J., & Phillips, M. (2016). Online British Official Publications from the University of Southampton. *Refer*, *32*(2), 27–32. Retrieved from https://search.proquest.com/docview/1803448852?accountid=27464

The Library at the University of Southampton has a particularly strong collection of printed British Official Publications, known as the Ford Collection. The collection is named after the late Professor Percy Ford and his wife Dr Grace Ford who brought the collection to the University of Southampton in the 1950s from the Carlton Club and conducted research based on the collection. Hoping to increase both the appreciation and the use of official publications, Ford, the Fords compiled breviates or select lists, in seven volumes covering the years 1833-1983. These were not catalogues of all British Official Publications. Instead the Fords identified and summarised documents which have been, or might have been, the subject of legislation or have dealt with public policy, Ford. Although funding sources were for specific tasks and periods, the Library continues to work unfunded with these valuable digital collections in 2016 to ensure that they are made fully accessible for readers worldwide.

Callón, M., Fdez-Glez, J., Ruano-Ordás, D., Laza, R., Pavón, R., Fdez-Riverola, F., & Méndez, J. (2017). WARCProcessor: An Integrative Tool for Building and Management of Web Spam Corpora. *Sensors*, *18*(2), 16. https://doi.org/10.3390/s18010016

In this work we present the design and implementation of WARC Processor, a novel multiplatform integrative tool aimed to build scientific datasets to facilitate experimentation in web spam research. The developed application allows the user to specify multiple criteria that change the way in which new corpora are generated whilst reducing the number of repetitive and error prone tasks related with existing corpus maintenance. For this goal, WARCProcessor supports up to six commonly used data sources for web spam research, being able to store output corpus in standard WARC format together with complementary metadata files. Additionally, the application facilitates the automatic and concurrent download of web sites from Internet, giving the possibility of configuring the deep of the links to be followed as well as the behaviour when redirected URLs appear. WARCProcessor supports both an interactive GUI interface and a command line utility for being executed in background. [ABSTRACT FROM AUTHOR]

Campos, R. (2007). Digital Libraries and Engines of Search: New Information Systems in the Context of the Digital Preservation. In *Proceedings of the 2007 Euro American Conference on Telematics and Information Systems* (p. 8:1--8:9). New York, NY, USA: ACM. https://doi.org/10.1145/1352694.1352703

The first's library projects occur some years ago with digitization, but just in 1996, the first's web archive initiatives start occurring. Such, was based in the Internet growth and in its increasing use, items that revealed to be an opportunity to transform and readapt the traditional library services. In this context, search engines play a fundamental role of support to the new paradigm of knowledge, by capturing, storing and providing access to the resources, allowing the existence of a digital library in each computer with internet access. In this article we analyze the ways of developing a digital library, taking higher attention to the web harvesting technique, and presenting digital libraries capabilities and limitations. Then

we fully summarize relevant projects and initiatives, to finally study the role of search engines in what concerns to, digital preservation, access and information diffusion.

Capell, L. (2015). Building the Foundation: Creating an Electronic-Records Program at the University of Miami. *Computers in Libraries*, *35*(9), 28–32. Retrieved from https://search.proquest.com/docview/1755071188?accountid=27464

Developing and implementing effective strategies to manage electronic records (e-records) is one of the biggest challenges facing the archives field today, as they acquire growing quantities of contemporary records generated by an increasingly digital society. However, jumping into e-records archiving can be a daunting task. As the author's continue to move through the pilot project and develop their policies and procedures for born-digital content, they're looking ahead at the next steps. First of all, they want to build more robust digital forensics workflows, including exploring methods for more extensive analysis of their digital content and developing workflows to handle a wider range of media and formats. Second, they want to use the results of their survey to start processing legacy media in their collections. Finally, they want to explore more options for providing access so that they can effectively make a wide range of born-digital content available for research.

Cardona Restrepo, J. C., & Stanojevic, R. (2012). A History of an Internet Exchange Point. *SIGCOMM Comput. Commun. Rev.*, *42*(2), 58–64. https://doi.org/10.1145/2185376.2185384

In spite of the tremendous amount of measurement efforts on understanding the Internet as a global system, little is known about the "local" Internet (among ISPs inside a region or a country) due to limitations of the existing measurement tools and scarce data. In this paper, empirical in nature, we characterize the evolution of one such ecosystem of local ISPs by studying the interactions between ISPs happening at the Slovak Internet eXchange (SIX). By crawling the web archive waybackmachine.org we collect 158 snapshots (spanning 14 years) of the SIX website, with the relevant data that allows us to study the dynamics of the Slovak ISPs in terms of: the local ISP peering, the traffic distribution, the port capacity/utilization and the local AS-level traffic matrix. Examining our data revealed a number of invariant and dynamic properties of the studied ecosystem that we report in detail.

Cartledge, C., & Nelson, M. (2015). When should I make preservation copies of myself? *International Journal on Digital Libraries*, *16*(3/4), 183–205. Retrieved from http://10.0.3.239/s00799-015-0155-1

We investigate how different replication policies ranging from least aggressive to most aggressive affect the level of preservation achieved by autonomic processes used by web objects (WOs). Based on simulations of small-world graphs of WOs created by the Unsupervised Small-World algorithm, we report quantitative and qualitative results for graphs ranging in order from 10 to 5000 WOs. Our results show that a moderately aggressive replication policy makes the best use of distributed host resources by not causing spikes in CPU resources nor spikes in network activity while meeting preservation goals. We examine different approaches that WOs can communicate with each other and determine the how long it would take for a message from one WO to reach a specific WO, or all WOs. [ABSTRACT FROM AUTHOR]

Celbová, L., & Prókai, M. (2009). A cseh web és a kötelespéldány-rendelet. *Könyvtári Figyelő : Külföldi Lapszemle*, (3), 518–520. Retrieved from http://epa.oszk.hu/00100/00143/00072/pdf/2009_3 szam_referatumok.pdf#page=14

Csehországban nincs jogszabály az elektronikus kötelespéldányok beszolgáltatásáról. 2000 óta foglalkoznak a nemzeti könyvtárban a webarchiválással, de a probléma nemzeti és nemzetközi szinten sem egyértelmű. Érinti a szerzői jogi törvényt, a nyomtatott kötelespéldányokról szóló szabályozást és a könyvtári törvény intézkedéseit.

Cerdeira-Pena, A., Farina, A., Fernandez, J. D., & Martinez-Prieto, M. A. (2016). Self-Indexing RDF Archives. *2016 Data Compression Conference (DCC)*, 526–535. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Although Big RDF management is an emerging topic in the so-called Web of Data, existing techniques disregard the dynamic nature of RDF data. These RDF archives evolve over time and need to be preserved and queried across it. This paper presents v-RDFCSA, an RDF archiving solution that extends RDFCSA (an RDF self-index) to provide versionbased queries on top of compressed RDF archives. Our experiments show that v-RDFCSA reduces space requirements up to 35 − 60 times over a state-of-the-art baseline, and gets more than one order of magnitude ahead over it for query resolution.

Ceroni, A., Georgescu, M., Gadiraju, U., Naini, K. D., & Fisichella, M. (2014). Information Evolution in Wikipedia. In *Proceedings of The International Symposium on Open Collaboration* (p. 24:1--24:10). New York, NY, USA: ACM. https://doi.org/10.1145/2641580.2641612

The Web of data is constantly evolving based on the dynamics of its content. Current Web search engine technologies consider static collections and do not factor in explicitly or implicitly available temporal information, that can be leveraged to gain insights into the dynamics of the data. In this paper, we hypothesize that by employing the temporal aspect as the primary means for capturing the evolution of entities, it is possible to provide entity-based accessibility to Web archives. We empirically show that the edit activity on Wikipedia can be exploited to provide evidence of the evolution of Wikipedia pages over time, both in terms of their content and in terms of their temporally defined relationships, classified in literature as events. Finally, we present results from our extensive analysis of a dataset consisting of 31,998 Wikipedia pages describing politicians, and observations from in-depth case studies. Our findings reflect the usefulness of leveraging temporal information in order to study the evolution of entities and breed promising grounds for further research.

Chambers, S., Mechant, P., Vandepontseele, S., Isbergue, N., & Depoortere, R. (2016). Towards a national web in a federated country : a Belgian case study. In *National Webs*. Retrieved from https://biblio.ugent.be/publication/8511255

Although the .be domain was introduced in June 1988, the Belgian web is currently not systematically archived. As of August 2016, 1.550.147 domains are registered by DNS Belgium. Without a Belgian web archive, the content of these websites will not be preserved for future generations and a significant portion of Belgian history will be lost forever. In this paper we present the initial findings of a research project exploring the policy, legal, technical

and scientific issues around archiving the Belgian web. The aim of this project is to a) identify current best practices in web-archiving b) pilot a Belgian web archive and c) identify research use cases for the scientific study of the Belgian web. This case study is seen as a first step towards implementing a long-term web archiving strategy for Belgium.

Chao, D., & Gill, S. (2015). The Design of a Cloud-Based Website Parallel Archiving System. *Issues in Information Systems*, *16*(1), 226. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Many business applications are designed and organized to support business activities for a period of time and to be renewed at the turn of the period. Design changes are typically implemented in a revision of the application that supports future periods to assure smooth operation. Very often the applications supporting the previous periods need to be operational continuously even after the application for the new period started. Parallel operation of current and previous periods' applications may be problematic for web-based applications due to the rapid change in Internet technologies. Cloud computing provides a solution to this problem with the capability of offering virtual servers with user-specified configurations. This paper proposes a parallel archiving scheme that uses virtual server to run each period's application in a cloud platform so that previous periods' applications will run in parallel with the current period system and forms an easy-to-access archive for historical data. [ABSTRACT FROM AUTHOR]

Chen, L., Bhowmick, S. S., & Nejdl, W. (2009). NEAR-Miner: Mining Evolution Associations of Web Site Directories for Efficient Maintenance of Web Archives. *Proc. VLDB Endow.*, *2*(1), 1150–1161. https://doi.org/10.14778/1687627.1687757

Web archives preserve the history of autonomous Web sites and are potential gold mines for all kinds of media and business analysts. The most common Web archiving technique uses crawlers to automate the process of collecting Web pages. However, (re)downloading entire collection of pages periodically from a large Web site is unfeasible. In this paper, we take a step towards addressing this problem. We devise a data mining-driven policy for selectively (re)downloading Web pages that are located in hierarchical directory structures which are believed to have changed significantly (e.g., a substantial percentage of pages are inserted to/removed from the directory). Consequently, there is no need to download and maintain pages that have not changed since the last crawl as they can be easily retrieved from the archive. In our approach, we propose an off-line data mining algorithm called near-Miner that analyzes the evolution history of Web directory structures of the original Web site stored in the archive and mines negatively correlated association rules (near) between ancestor-descendant Web directories. These rules indicate the evolution correlations between Web directories. Using the discovered rules, we propose an efficient Web archive maintenance algorithm called warm that optimally skips the subdirectories (during the next crawl) which are negatively correlated with it in undergoing significant changes. Our experimental results with real data show that our approach improves the efficiency of the archive maintenance process significantly while sacrificing slightly in keeping the "freshness" of the archives. Furthermore, our experiments demonstrate that it is not necessary to discover nears frequently as the mining rules can be utilized effectively for archive maintenance over multiple versions.

Chen, Y.-Y., Gan, Q., & Suel, T. (2004). Local Methods for Estimating Pagerank Values. In *Proceedings of the Thirteenth ACM International Conference on Information and*

Knowledge Management (pp. 381–389). New York, NY, USA: ACM. https://doi.org/10.1145/1031171.1031248

The Google search engine uses a method called PageRank, together with term-based and other ranking techniques, to order search results returned to the user. PageRank uses link analysis to assign a global importance score to each web page. The PageRank scores of all the pages are usually determined off-line in a large-scale computation on the en- tire hyperlink graph of the web, and several recent studies have focused on improving the efficiency of this computa- tion, which may require multiple hours on a workstation. However, in some scenarios, such as online analysis of link evolution and mining of large web archives such as the In- ternet Archive, it may be desirable to quickly approximate or update the PageRanks of individual nodes without per- forming a large-scale computation on the entire graph. We address this problem by studying several methods for effi- ciently estimating the PageRank score of a particular web page using only a small subgraph of the entire web. In our model, we assume that the graph is accessible remotely via a link database (such as the AltaVista Connectivity Server) or is stored in a relational database that performs lookups on disks to retrieve node and connectivity information. We show that a reasonable estimate of the PageRank value of a node is possible in most cases by retrieving only a moderate number of nodes in the local neighborhood of the node.

Condill, K. (2017). The Online Media Environment of the North Caucasus: Issues of Preservation and Accessibility in a Zone of Political and Ideological Conflict. *Preservation, Digital Technology & Culture*, *45*(4), 166–176. https://doi.org/http://dx.doi.org/10.1515/pdtc-2016-0022

As one of the world's most ethnolinguistically-diverse and conflict-prone regions, the North Caucasus presents particular challenges for librarians seeking to preserve its rich and varied online news media content. This content is generated in multiple languages in multiple political and ideological contexts, both within the North Caucasus region and abroad. While online news media content in general is ephemeral, poorly-preserved, and difficult to access via any single search interface or search strategy, content relating to the North Caucasus is at additional risk due to ongoing insurgency/counterinsurgency activity, as well as historical, political and linguistic factors. Various options for preserving and searching North Caucasus web content are explored.

Costa, M., Couto, F., & Silva, M. (2014). Learning temporal-dependent ranking models. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14* (pp. 757–766). New York, New York, USA: ACM Press. https://doi.org/10.1145/2600428.2609619

Web archives already hold together more than 534 billion files and this number continues to grow as new initiatives arise. Searching on all versions of these files acquired throughout time is challenging, since users expect as fast and precise answers from web archives as the ones provided by current web search engines. This work studies, for the first time, how to improve the search effectiveness of web archives, including the creation of novel temporal features that explore the correlation found between web document persistence and relevance. The persistence was analyzed over 14 years of web snapshots. Additionally, we propose a temporal-dependent ranking framework that exploits the variance of web characteristics over time influencing ranking models. Based on the assumption that closer periods are more likely to hold similar web characteristics, our framework learns multiple models simultaneously,

each tuned for a specific period. Experimental results show significant improvements over the search effectiveness of single-models that learn from all data independently of its time. Thus, our approach represents an important step forward on the state-of-the-art IR technology usually employed in web archives.

Cowls, J. (2017). Kultura brytyjskiej sieci web TT - British culture web. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951541478?accountid=27464

Autor przedstawia brytyjski projekt BUDDAH, który polegał na tym, że naukowcy korzystając ze zgromadzonych zasobów archiwalnych pobranych z sieci robili humanistyczne badania naukowe. Chodziło o stwierdzenie, czy jest sens w archiwizacji stron internetowych w celach badawczych. W artykule opisano wiele różnych badań, podejść metodologicznych, studiów przypadków oraz narzędzi technicznych, które stworzono, by zrealizować te badania.

Cox, D. (2018). Developing and raising awareness of the zine collections at the British Library. *Art Libraries Journal*, *43*(2), 77–81. https://doi.org/http://dx.doi.org/10.1017/alj.2018.5

This article presents a practice-based account of collection development related to zines in the British Library. Rather than making the case for the collecting of zines, it aims to describe the process of collection building in a specific time and place, so that researchers have a better understanding of why certain resources are offered to them and others are not, and to share experiences with other librarians with zine collections. Zines form an element of the cultural memory of activists and cultural creators, and for researchers studying them it would seem useful to make transparent the motivations, methods and limitations of collection building. Librarians in the USA have written about their collecting practices for some time, for instance at Barnard College1and New York Public Library2, there has been less written about the practices of UK libraries. The article aims to make a contribution as a case study alongside accounts of collection development in a range of other libraries with zine collections, and it is written primarily from my own perspective as a curator in Contemporary British Collections since 2015, focusing on current practice, with some reference to earlier collecting.

Coyle, K. (2009). Metadata mix and match. *Information Standards Quarterly*, *21*(1), 8–11. Retrieved from https://search.proquest.com/docview/1735033500?accountid=27464

The author was asked to consult with the Internet Archive's Open Library project primarily to lend her expertise in bibliographic data. To her dismay, the Open Library data did not look anything like library bibliographic data. She learned, however, that there were some good reasons for this. The first was that the Open Library was not limiting itself to library data. Another reason the Open Library does not limit itself to the more rigorous library data style was that the Open Library allows editing of its data by the general public: people with no particular bibliographic training. The most compelling reason to deviate from the standard view posited by library bibliographic data, however, has to do with the concept of linked data. It's an unfortunate fact that many systems combine data from different sources using only the "dumb down" method, reducing the metadata to the few matching elements and resulting in the least rich metadata record possible.

Dąbrowska, E. (2017). Problem archiwizacji internetu w kontekście egzemplarza obowiązkowego : sytuacja w Polsce i wybranych krajach europejskich TT - The problem

of archiving the Internet in the context of a mandatory copy: the situation in Poland and selected European countri. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951540566?accountid=27464

W artykule przedstawiono obowiązujące w Polsce i innych krajach przepisy prawne i przyjęte rozwiązania odnoszące się do możliwości archiwizowania zawartości internetu przez biblioteki narodowe oraz związane z tym wyzwania i problemy. Uprawnionym bibliotekom powinny być przekazywane publikacje o charakterze utworu zamieszczone w sieci. Nie jest to powszechnie przestrzegane, gdyż ta zasada nie została w polskim prawie wyraźnie wyartykułowana. Wobec rosnącego znaczenia komunikacji w środowisku cyfrowym, zwłaszcza komunikacji naukowej, w tworzonych przez biblioteki narodowe archiwach krajowego piśmiennictwa powstaje poważna luka. Problem ten powinien zostać rozwiązany w nowych przepisach o obowiązkowych egzemplarzach bibliotecznych.

Drótos, L., Kokas K. (2018). Webarchiválás és a történeti kutatások. *Digitális Bölcsészet*, *1*(1), 35-53. Retrieved from http://ojs.elte.hu/index.php/digitalisbolcseszet/article/view/129

A digitálisan születő tartalom sokkal részletesebb és teljesebb leképezése a jelennek, mint ami régebbi korokban a hagyományos információhordozó eszközökkel rögzíthető volt. A tanulmány első része arról ad áttekintést, hogy milyen próbálkozások és technológiák léteznek ennek a digitális jelennek a megőrzésére, illetve milyen korlátai vannak a már működő webarchívumoknak. A dolgozat második része azt vizsgálja, hogy a történeti szempontú kutatásoknak hogyan lehet hasznára mindez, s hogyan lesz elsősorban a közelmúlt történetének is elsőrangú forrása. A szerzők arra is rámutatnak, hogy a webaratások következtében előálló hatalmas adatsilók egészen új típusú forráskezelést és módszertant kívánnak majd meg, miközben azzal kecsegtetnek, hogy egészen új típusú eredményeket is fel lehet majd mutatni segítségükkel.

Dancs, S. (2011). Webarchiválási politikák. *Könyv, Könyvtár, Könyvtáros*, *20*(10), 14–20. Retrieved from http://epa.oszk.hu/01300/01367/00248/pdf/EPA01367_3K_2011_10_14-20.pdf

MatarkaID=1671000

Dang-Nguyen, D.-T., Riegler, M., Zhou, L., & Gurrin, C. (2018). Challenges and Opportunities within Personal Life Archives. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval - ICMR '18* (pp. 335–343). New York, New York, USA: ACM Press. https://doi.org/10.1145/3206025.3206040

Nowadays, almost everyone holds some form or other of a personal life archive. Automatically maintaining such an archive is an activity that is becoming increasingly common, however without automatic support the users will quickly be overwhelmed by the volume of data and will miss out on the potential benefits that lifelogs provide. In this paper we give an overview of the current status of lifelog research and propose a concept for exploring these archives. We motivate the need for new methodologies for indexing data, organizing content and supporting information access. Finally we will describe challenges to be addressed and give an overview of initial steps that have to be taken, to address the challenges of organising and searching personal life archives.

Davis, R. C. (2016). The Future of Web Citation Practices. *Behavioral & Social Sciences Librarian*, *35*(3), 128–134. https://doi.org/http://dx.doi.org/10.1080/01639269.2016.1241122

Citing webpages has been a common practice in scholarly publications for nearly two decades as the Web evolved into a major information source. But over the years, more and more bibliographies have suffered from "reference rot": Cited URLs are broken links or point to a page that no longer contains the content the author originally cited. In this column, I look at several studies showing how reference rot has affected different academic disciplines. I also examine citation styles' approach to citing Web sources. I then turn to emerging Web citation practices: Perma, a "freemium" Web archiving service specifically for citation; and the Internet Archive, the largest Web archive.

Davis, S. J. (2016). Disappearing News Archives. *Online Searcher*, *40*(6), 46. Retrieved from https://search.proquest.com/docview/1861822700?accountid=27464

Part of the preservation problem lies in the fact that newspapers are not official public records. According to the ProQuest title list, ProQuest News has the full text of the Milwaukee Journal Sentinel from April 1, 1995, to Dec. 31, 2009, a fraction of the full 123 years (1884-2007) formerly in Google News Archive.

De Baets, A. (2016). A historian's view on the right to be forgotten. *International Review of Law, Computers & Technology*, *30*(1–2), 57–66. https://doi.org/10.1080/13600869.2015.1125155

This essay explores the consequences for historians of the 'right to be forgotten'', a new concept proposed by the European Commission in 2012. I first explain that the right to be forgotten is a radical variant of the right to privacy and clarify the consequences of the concept for the historical study of public and private figures. I then treat the hard cases of spent and amnestied convictions and of internet archives. I further discuss the applicability of the right to be forgotten to dead persons as part of the problem of posthumous privacy, and finally point to the ambiguity of the impact of the passage of time. While I propose some compromise solutions, I also conclude that a generalized right to be forgotten would lead to the rewriting of history in ways that impoverish our insights not only into anecdotal lives but also into the larger trends of history. [ABSTRACT FROM AUTHOR]'

Debruyne, C., Beyan, O. D., Grant, R., Collins, S., Decker, S., & Harrower, N. (2016). A semantic architecture for preserving and interpreting the information contained in Irish historical vital records. *International Journal on Digital Libraries*, *17*(3), 159–174. https://doi.org/10.1007/s00799-016-0180-8

Decman, M. (2011). Problems of Long-Term Preservation of Web Pages TT - Problematika Dolgorocne Hrambe Spletnih Strani. *Knjiznica*, *55*(1), 193–208. Retrieved from https://search.proquest.com/docview/1266143497?accountid=27464

The World Wide Web is a distributed collection of web sites available on the Internet anywhere in the world. Its content is constantly changing: old data are being replaced which causes constant loss of a huge amount of information and consequently the loss of scientific, cultural and other heritage. Often, unnoticeably even legal certainty is questioned. In what way the data on the web can be stored and how to preserve them for the long term is a great

challenge. Even though some good practices have been developed, the question of final solution on the national level still remains. The paper presents the problems of long-term preservation of web pages from technical and organizational point of view. It includes phases such as capturing and preserving web pages, focusing on good solutions, world practices and strategies to find solutions in this area developed by different countries. The paper suggests some conceptual steps that have to be defined in Slovenia which would serve as a framework for all document creators in the web environment and therefore contributes to the consciousness in this field, mitigating problems of all dealing with these issues today and in the future. Adapted from the source document.

Demidova, E., Barbieri, N., Dietze, S., Funk, A., Holzmann, H., Maynard, D., … Spiliotopoulos, D. (2014). Analysing and Enriching Focused Semantic Web Archives for Parliament Applications. *Future Internet, Vol 6, Iss 3, Pp 433-456 (2014) VO - 6*, (3), 433. https://doi.org/10.3390/fi6030433

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.

Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2011). The SHARC framework for data quality in Web archiving. *The VLDB Journal*, *20*(2), 183–207. https://doi.org/10.1007/s00778-011-0219-9

Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2009). SHARC. *Proceedings of the VLDB Endowment*, *2*(1), 586–597. https://doi.org/10.14778/1687627.1687694

Web archives preserve the history of born-digital content and offer great potential for sociologists, business analysts, and legal experts on intellectual property and compliance issues. Data quality is crucial for these purposes. Ideally, crawlers should gather sharp captures of entire Web sites, but the politeness etiquette and completeness requirement mandate very slow, long-duration crawling while Web sites undergo changes. This paper presents the SHARC framework for assessing the data quality in Web archives and for tuning capturing strategies towards better quality with given resources. We define quality measures, characterize their properties, and derive a suite of quality-conscious scheduling strategies for archive crawling. It is assumed that change rates of Web pages can be statistically predicted based on page types, di- rectory depths, and URL names. We develop a stochastically optimal crawl algorithm for the offline case where all change rates are known. We generalize the approach into an online algorithm that detect information on a Web site while it is crawled. For dating a site capture and for assessing its quality, we propose several strategies that revisit pages after their initial downloads in a judiciously chosen order. All strategies are fully implemented in a testbed, and shown to be effective by experiments with both synthetically generated sites and a daily crawl series for a medium-sized site.

Derfert-Wolf, L. (2017). Archiwizacja internetu – wnioski i rekomendacje z kilku raportów TT - Internet archiving - conclusions and recommendations from several reports.

Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB, (172), 1. Retrieved from https://search.proquest.com/docview/1951539109?accountid=27464

W artykule omówiono trzy zagraniczne raporty dotyczące archiwizacji internetu. W materiale Web-Archiving z 2013 r. przedstawiono kluczowe problemy archiwizacji internetu, z punktu widzenia instytucji realizujących tego typu projekty, bez względu na to czy zlecają prace zewnętrznym firmom czy wykonują je we własnym zakresie. Raport Preserving Social Media, opracowany w 2016 r., dotyczy zabezpieczania zasobów mediów społecznościowych. Web Archiving Environmental Scan – stanowi analizę środowiskową, która przeprowadzono w 2015 r. na zlecenie Biblioteki Uniwersytetu Harvarda. Badaniem objęto 23 instytucje z całego świata, realizujące aktualnie tego typu projekty. W artykule przedstawiono również elementy dokumentu normalizacyjnego ISO/TR 14873:2013 Information and Documentation – Statistics and quality issues for web archiving. Na zakończenie nawiązano do prognoz dotyczących rozwoju archiwizacji internetu zaprezentowanych w raporcie Web Archives: The Future(s), opublikowanym w 2011 r.

Derfert-Wolf, L., & Wilkowski, M. (2017). Felieton "archiwalny" – ponownie po pięciu latach. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951541363?accountid=27464

Dietze, S., Maynard, D., Demidova, E., Risse, T., Peters, W., Doka, K., & Stavrakas, Y. (2012). *Entity Extraction and Consolidation for Social Web Content Preservation*. United States, North America. https://doi.org/10.1.1.423.3432

With the rapidly increasing pace at which Web content is evolving, particularly social media, preserving the Web and its evolution over time becomes an important challenge. Meaningful analysis of Web content lends itself to an entity-centric view to organise Web resources according to the information objects related to them. Therefore, the crucial challenge is to extract, detect and correlate entities from a vast number of heterogeneous Web resources where the nature and quality of the content may vary heavily. While a wealth of information extraction tools aid this process, we believe that, the consolidation of automatically extracted data has to be treated as an equally important step in order to ensure high quality and non-ambiguity of generated data. In this paper we present an approach which is based on an iterative cycle exploiting Web data for (1) targeted archiving/crawling of Web objects, (2) entity extraction, and detection, and (3) entity correlation. The long-term goal is to preserve Web content over time and allow its navigation and analysis based on well-formed structured RDF data about entities.

Dooley, J. (2017). Developing Web Archiving Metadata Best Practices to Meet User Needs. *Journal of Western Achives*, *8*(2). Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The OCLC Research Library Partnership Web Archiving Metadata Working Group was established to meet a widely recognized need for best practices for descriptive metadata for archived websites. The Working Group recognizes that development of successful best practices intended to ensure discoverability requires an understanding of user needs and behavior. We have therefore conducted an extensive literature review to build our knowledge and will issue a white paper summarizing what we have learned. We are also studying existing and emerging approaches to descriptive metadata in this realm and will publish a

second report recommending best practices. We will seek broad community input prior to publication.

Dougherty, M., & Meyer, E. T. (2014). Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. *Journal of the Association for Information Science and Technology*, *65*(11), 2195–2209. https://doi.org/http://dx.doi.org/10.1002/asi.23099

The web encourages the constant creation and distribution of large amounts of information; it is also a valuable resource for understanding human behavior and communication. To take full advantage of the web as a research resource that extends beyond the consideration of snapshots of the present, however, it is necessary to begin to take web archiving much more seriously as an important element of any research program involving web resources. The ephemeral character of the web requires that researchers take proactive steps in the present to enable future analysis. Efforts to archive the web or portions thereof have been developed around the world, but these efforts have not yet provided reliable and scalable solutions. This article summarizes the current state of web archiving in relation to researchers and research needs. Interviews with researchers, archivists, and technologists identify the differences in purpose, scope, and scale of current web archiving practice, and the professional tensions that arise given these differences. Findings outline the challenges that still face researchers who wish to engage seriously with web content as an object of research, and archivists who must strike a balance reflecting a range of user needs. [Copyright Wiley Periodicals Inc.]

Drótos, L. (2017). Webtörténetírás az Internet Archive-ból készített képernyővideókkal. *Tudományos És Műszaki Tájékoztatás*, *64*(7–8), 397–401. Retrieved from http://epa.oszk.hu/03000/03071/00109/pdf/EPA03071_tmt_2017_07_08_397-401.pdf

A globális Internet Archive és a nemzeti webarchívumok a digitális történelem kutatásának fő forrásai, mivel összegyűjtik és megőrzik az eleve digitális formában születő kultúrát, s így olyan tartalmakat lehet megtalálni bennük, amelyek sehol máshol nem kutathatók. Az 1990-es évek második felétől már elképzelhetetlen teljes körűen megírni valaminek a történetét kizárólag csak papírújságokra és -könyvekre alapozva, figyelmen kívül hagyva a téma internetes lenyomatait.

Drótos, L. (2017). Az internet archiválása mint könyvtári feladat. *Tudományos És Műszaki Tájékoztatás*, *64*(7–8), 361–371. Retrieved from http://epa.oszk.hu/03000/03071/00109/pdf/EPA03071_tmt_2017_07_08_361-371.pdf

A nyilvános internetről minden nap tömeges méretekben letörölt vagy máshová költöző dokumentumok és egyéb információforrások egyre nagyobb problémát jelentenek a tudományos publikációkban és a tananyagokban való hivatkozhatóság szempontjából, de az átlagos internetező is állandóan belefut az eltűnt weboldalakat jelző 404-es hibákba. A világháló alapvetően egy jelen idejű médium, de legalább egy részét érdemes lenne megőrizni és kutathatóvá tenni a jövő generációi számára. Ez a cikk arra a kérdésre keresi a választ, hogy ki, mit, hogyan, mivel és miért mentsen az internetről, és hol van itt a könyvtárak és a könyvtárosok feladata és felelőssége? Bemutat néhány hasznos eszközt és szolgáltatást, majd röviden ismerteti a nemzetközi helyzetet és az OSZK-ban 2017 tavaszán elindult kísérleti webarchiválási projektet.

Drótos, L., & Németh, M. (2018). Az OSZK-ban folyó kísérleti webarchiválási projekt első évének tapasztalatai. *Tudományos És Műszaki Tájékoztatás*, *65*(7–8), 389–400. Retrieved from http://tmt.omikk.bme.hu/tmt/article/view/7153/8156

Az Országos Széchényi Könyvtárban az OKR (Országos Könyvtári Rendszer) kifejlesztése keretében 2017−2018 között zajlik egy kísérleti projekt azzal céllal, hogy Magyarországon is megteremtsük a nyilvános webhelyek tömeges archiválásának és hosszú távú megőrzésének feltételeit, elsősorban az ehhez a munkához szükséges informatikai infrastruktúrát és szakértelmet. Ezen a téren több mint 20 éves lemaradást kell ledolgoznunk, mert például az amerikai nonprofit szervezet, az Internet Archive (IA) már 1996 óta foglalkozik ezzel, és azóta példáját számos országban követték, létrehoztak nemzeti, kormányzati vagy intézményi webarchívumokat, gyakran könyvtári, levéltári irányítással vagy közreműködéssel. Az OSZK-ban a 2000-es évek közepén merült fel egy magyar internet archívum (MIA) ötlete, de az ezt előkészítő munka feltételei csak 2017 tavaszán kezdtek megteremtődni. Az egri Networkshop első napján rendezett műhelymunka vitaindító előadásában a 2018 áprilisáig eltelt egy év fejleményeiről számoltunk be, s ezeket az eredményeket és tapasztalatokat foglaljuk össze ebben a cikkben.

Dulin, K., & Ziegler, A. (2017). Scaling Up Perma.cc: Ensuring the Integrity of the Digital Scholarly Record. *D - Lib Magazine*, *23*(5/6). Retrieved from https://search.proquest.com/docview/1925481174?accountid=27464

IMLS awarded the Harvard Library Innovation Lab a National Digital Platform grant to further develop the Lab's Perma.cc web archiving service. The funds will be used to provide technical enhancements to support an expanded user base, aid in outreach efforts to implement Perma.cc in the nation's academic libraries, and develop a commercial model for the service that will sustain the free service for the academic community. Perma.cc is a web archiving tool that puts the ability to archive a source in the hands of the author who is citing it. Once saved, Perma.cc assigns the source a new URL, which can be added to the original URL cited in the author's work, so that if the original link rots or is changed the Perma.cc URL will still lead to the original source. Perma.cc is being used widely in the legal community with great success; the IMLS grant will make the tool available to other areas of scholarship where link rot occurs and will provide a solution for those in the commercial arena who do not currently have one.

Dulong de Rosnay, M., & Guadamuz, A. (2017). Memory Hole or Right to Delist? Implications of the Right to be Forgotten for Web Archiving ; Trou mémoriel ou droit au déréférencement ? Les implications du droit à l'oubli pour l'archivage du Web. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

International audience ; This article studies the possible impact of the "right to be forgotten" (RTBF) on the preservation of native digital heritage. It analyses the extent to which archival practices may be affected by the new right, and whether the web may become impossible to preserve for future generations, risking to disappear from memories and history since no version would be available in public or private archives. Collective rights to remember and to memory, free access to information and freedom of expression, seem to clash with private individuals' right to privacy. After a presentation of core legal concepts of privacy, data protection and freedom of expression, we analyse the case of the European Union Court of

Justice vs. Google concerning the right to be forgotten, and look deeper into the controversies generated by the decision. We conclude that there is no room for concern for archives and for the right to remember given the restricted application of RTBF. ; Cet article étudie l'impact possible du « droit à l'oubli » (RTBF) sur la préservation du patrimoine numérique natif. Il analyse si les pratiques d'archivage sont susceptibles d'être affectées par le nouveau droit et s'il pourrait devenir impossible de préserver le Web pour les générations futures, avec le risque pour certains contenus de disparaître de la mémoire et de l'histoire si aucune version n'était disponible dans les archives publiques ou privées. Le droit collectif au souvenir et à la mémoire, l'accès libre à l'information et la liberté d'expression semblent entrer en conflit avec les droits individuels à la vie privée. Après une présentation des concepts juridiques fondamentaux de la vie privée, de la protection des données personnelles et de la liberté d'expression, nous analysons l'arrêt Google de la Cour de Justice de l'Union Européenne et le droit à l'oubli, et examinons les controverses qui ont été générées par la décision. On conclut que les archives et le droit au souvenir ne seront pas affectés par le droit à l'oubli, étant donné son application restreinte.

Eklund, P., Wray, T., & Ducrou, J. (2011). Linking Objects and their Stories: An API For Exploring Cultural Heritage Using Formal Concept Analysis. *Journal of Emerging Technologies in Web Intelligence*, *3*(3). https://doi.org/10.4304/jetwi.3.3.239-252

Erdélyi, M., Benczúr, A. A., Masanés, J., & Siklósi, D. (2009). Web Spam Filtering in Internet Archives. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web* (pp. 17–20). New York, NY, USA: ACM. https://doi.org/10.1145/1531914.1531918

While Web spam is targeted for the high commercial value of top-ranked search-engine results, Web archives observe quality deterioration and resource waste as a side effect. So far Web spam filtering technologies are rarely used by Web archivists but planned in the future as indicated in a survey with responses from more than 20 institutions worldwide. These archives typically operate on a modest level of budget that prohibits the operation of standalone Web spam filtering but collaborative efforts could lead to a high quality solution for them. In this paper we illustrate spam filtering needs, opportunities and blockers for Internet archives via analyzing several crawl snapshots and the difficulty of migrating filter models across different crawls via the example of the 13 .uk snapshots performed by UbiCrawler that include WEBSPAM-UK2006 and WEBSPAM-UK2007.

Espley, S., Carpentier, F., Pop, R., & Medjkoune, L. (2014). Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content. *Alexandria: The Journal of National and International Library and Information Issues*, *25*(1–2), 31–50. https://doi.org/10.7227/ALX.0019

It is The National Archives' responsibility to collect and secure the future of the public record in all its forms and to make it as accessible as possible. The UK Government Web Archive1 (UKGWA) effectively preserves the open digital record. This article will explore the challenges encountered, and the Application Programming Interface (API) based solutions developed, by The National Archives and the Internet Memory Foundation (IMF) in the completion of a pilot project to capture the record as it is published on the social media services Twitter and YouTube. An outline of the wider web archiving programme and its role within the management of the government web estate is provided. The legislative framework that guides web archiving at The National Archives is described as it has necessarily

influenced the policy decisions that shaped the solutions developed. A brief overview of some comparative approaches taken by other organizations and commercial services to capturing Twitter content is also presented as context to the policy and technical solutions arrived at by the authors. The National Archives has sought to develop the building blocks of a collection whose growth can be sustained over time. The publication of this part of the archive will be followed by further evaluation and improvements to the initial approach taken.

Fafalios, P., Holzmann, H., Kasturia, V., & Nejdl, W. (2018). Building and querying semantic layers for web archives (extended version). *International Journal on Digital Libraries*, *19*(1), 1–19. https://doi.org/10.1007/s00799-018-0251-0

 Web archiving is the process of collecting portions of the Web to ensure that the information is preserved for future exploitation. However, despite the increasing number of web archives worldwide, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning them into a usable and useful information source. In this paper, we focus on this problem and propose an RDF/S model and a distributed framework for building semantic profiles (layers) that describe semantic information about the contents of web archives. A semantic layer allows describing metadata information about the archived documents, annotating them with useful semantic information (like entities, concepts and events), and publishing all this data on the Web as Linked Data. Such structured repositories offer advanced query and integration capabilities and make web archives directly exploitable by other systems and tools. To demonstrate their query capabilities, we build and query semantic layers for three different types of web archives. An experimental evaluation showed that a semantic layer can answer information needs that existing keyword-based systems are not able to sufficiently satisfy.

Fafalios, P., Kasturia, V., & Nejdl, W. (2018). Ranking Archived Documents for Structured Queries on Semantic Layers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 155–164). New York, NY, USA: ACM. https://doi.org/10.1145/3197026.3197049

Archived collections of documents (like newspaper and web archives) serve as important information sources in a variety of disciplines, including Digital Humanities, Historical Science, and Journalism. However, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning them into usable sources of information. A semantic layer is an RDF graph that describes metadata and semantic information about a collection of archived documents, which in turn can be queried through a semantic query language (SPARQL). This allows running advanced queries by combining metadata of the documents (like publication date) and content-based semantic information (like entities mentioned in the documents). However, the results returned by such structured queries can be numerous and moreover they all equally match the query. In this paper, we deal with this problem and formalize the task of ranking archived documents for structured queries on semantic layers. Then, we propose two ranking models for the problem at hand which jointly consider: i) the relativeness of documents to entities, ii) the timeliness of documents, and iii) the temporal relations among the entities. The experimental results on a new evaluation dataset show the effectiveness of the proposed models and allow us to understand their limitations.

Faheem, M. (2012). Intelligent Crawling of Web Applications for Web Archiving. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 127–132). New York, NY, USA: ACM. https://doi.org/10.1145/2187980.2187996

The steady growth of the World Wide Web raises challenges regarding the preservation of meaningful Web data. Tools used currently by Web archivists blindly crawl and store Web pages found while crawling, disregarding the kind of Web site currently accessed (which leads to suboptimal crawling strategies) and whatever structured content is contained in Web pages (which results in page-level archives whose content is hard to exploit). We focus in this PhD work on the crawling and archiving of publicly accessible Web applications, especially those of the social Web. A Web application is any application that uses Web standards such as HTML and HTTP to publish information on the Web, accessible by Web browsers. Examples include Web forums, social networks, geolocation services, etc. We claim that the best strategy to crawl these applications is to make the Web crawler aware of the kind of application currently processed, allowing it to refine the list of URLs to process, and to annotate the archive with information about the structure of crawled content. We add adaptive characteristics to an archival Web crawler: being able to identify when a Web page belongs to a given Web application and applying the appropriate crawling and content extraction methodology.

Faheem, M., & Senellart, P. (2013). Demonstrating intelligent crawling and archiving of web applications. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 2481–2484). New York, NY, USA: ACM. https://doi.org/10.1145/2505515.2508197

We demonstrate here a new approach to Web archival crawling, based on an application-aware helper that drives crawls of Web applications according to their types (especially, according to their content management systems). By adapting the crawling strategy to the Web application type, one is able to crawl a given Web application (say, a given forum or blog) with fewer requests than traditional crawling techniques. Additionally, the application-aware helper is able to extract semantic content from the Web pages crawled, which results in a Web archive of richer value to an archive user. In our demonstration scenario, we invite a user to compare application-aware crawling to regular Web crawling on the Web site of their choice, both in terms of efficiency and of experience in browsing and searching the archive.

Fansler, C., Gilbertson, K., & Petersen, R. (2014). The Missing Link: Observations on the Evolution of a Web Archive. *Journal for the Society of North Carolina Archivists*, *11*(1), 46–59. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The web is vast and unorganized, making it difficult to collect and to curate for archival and research purposes. In this article, we discuss web archiving in the scope of a university archive, the challenges associated with such web archiving, and archival strategies for building and maintaining a web archive. This article chronicles our experience developing appropriate standards of practice for this medium, providing adequate metadata for the digital objects, constructing precise capturing protocols, and sharing access to these online collections. While some difficulties lie in transforming our archival modalities from print to digital, an equal share of obstacles relate to the speed, scale, and distribution technologies of the web itself. [ABSTRACT FROM AUTHOR]

Farag, M. M. G., Lee, S., & Fox, E. A. (2018). Focused crawler for events. *International Journal on Digital Libraries*, *19*(1), 3–19. https://doi.org/http://dx.doi.org/10.1007/s00799-016-0207-1

There is need for an Integrated Event Focused Crawling system to collect Web data about key events. When a disaster or other significant event occurs, many users try to locate the most up-to-date information about that event. Yet, there is little systematic collecting and archiving anywhere of event information. We propose intelligent event focused crawling for automatic event tracking and archiving, ultimately leading to effective access. We developed an event model that can capture key event information, and incorporated that model into a focused crawling algorithm. For the focused crawler to leverage the event model in predicting webpage relevance, we developed a function that measures the similarity between two event representations. We then conducted two series of experiments to evaluate our system about two recent events: California shooting and Brussels attack. The first experiment series evaluated the effectiveness of our proposed event model representation when assessing the relevance of webpages. Our event model-based representation outperformed the baseline method (topic-only); it showed better results in precision, recall, and F1-score with an improvement of 20% in F1-score. The second experiment series evaluated the effectiveness of the event model-based focused crawler for collecting relevant webpages from the WWW. Our event model-based focused crawler outperformed the state-of-the-art baseline focused crawler (best-first); it showed better results in harvest ratio with an average improvement of 40%. [ABSTRACT FROM AUTHOR]

Farag, M. M. G. (2016). *Intelligent Event Focused Crawling*. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

There is need for an integrated event focused crawling system to collect Web data about key events. When an event occurs, many users try to locate the most up-to-date information about that event. Yet, there is little systematic collecting and archiving anywhere of information about events. We propose intelligent event focused crawling for automatic event tracking and archiving, as well as effective access. We extend the traditional focused (topical) crawling techniques in two directions, modeling and representing: events and webpage source importance.We developed an event model that can capture key event information (topical, spatial, and temporal). We incorporated that model into the focused crawler algorithm. For the focused crawler to leverage the event model in predicting a webpage's relevance, we developed a function that measures the similarity between two event representations, based on textual content.Although the textual content provides a rich set of features, we proposed an additional source of evidence that allows the focused crawler to better estimate the importance of a webpage by considering its website. We estimated webpage source importance by the ratio of number of relevant webpages to non-relevant webpages found during crawling a website. We combined the textual content information and source importance into a single relevance score.For the focused crawler to work well, it needs a diverse set of high quality seed URLs (URLs of relevant webpages that link to other relevant webpages). Although manual curation of seed URLs guarantees quality, it requires exhaustive manual labor. We proposed an automated approach for curating seed URLs using social media content. We leveraged the richness of social media content about events to extract URLs that can be used as seed URLs for further focused crawling.We evaluated our system through four series of experiments, using recent events: Orlando shooting, Ecuador earthquake, Panama papers, California shooting, Brussels attack, Paris attack, and Oregon shooting. In the first experiment

series our proposed event model representation, used to predict webpage relevance, outperformed the topic-only approach, showing better results in precision, recall, and F1-score. In the second series, using harvest ratio to measure ability to collect relevant webpages, our event model-based focused crawler outperformed the state-of-the-art focused crawler (best-first search). The third series evaluated the effective…

Farag, M., Nakate, P., & Fox, E. A. (2016). Big Data Processing of School Shooting Archives. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (pp. 271–272). New York, New York, USA: ACM Press. https://doi.org/10.1145/2910896.2925466

Web archives about school shootings consist of webpages that may or may not be relevant to the events of interest. There are 3 main goals of this work; first is to clean the webpages, which involves getting rid of the stop words and non-relevant parts of a webpage. The second goal is to select just webpages relevant to the events of interest. The third goal is to upload the cleaned and relevant webpages to Apache Solr so that they are easily accessible. We show the details of all the steps required to achieve these goals. The results show that representative Web archives are noisy, with 2% - 40% relevant content. By cleaning the archives, we aid researchers to focus on relevant content for their analysis.

Fellows, G., Harvey, R., Lloyd, A., Pymm, B., & Wallis, J. (2008). Separating the Wheat from the Chaff: Identifying Key Elements in the NLA .Au Domain Harvest. *Australian Academic & Research Libraries*, *39*(3), 137–148. https://doi.org/10.1080/00048623.2008.10721346

In 2005 and 2006 the National Library of Australia (NLA) carried out two whole-domain web harvests which complement the selective web archiving approach taken by PANDORA. Web harvests of this size pose significant challenges to their use. Despite these challenges, such harvests present fascinating research opportunities. The NLA has provided Charles Sturt University's POA (Preservation for Ongoing Accessibility) research group with access to these web harvests and associated keyword indexes. This paper describes the 2006 harvest and uses the example of blogs to address how to identify material within the harvest and determine issues that need further investigation.

Fernando, Z. T., Marenzi, I., Nejdl, W., & Kalyani, R. (2016). ArchiveWeb: Collaboratively Extending and Exploring Web Archive Collections. In *Research & Advanced Technology for Digital Libraries: 20th International Conference on Theory & Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings* (pp. 107–118). https://doi.org/10.1007/978-3-319-43997-6_9

Ferreira, L. B., Martins, M. R., & Rockembach, M. (2018). Usos do Arquivamento da Web na Comunicação Científica. *Uses of Web Archiving in Scientific Communication.*, (36), 78–98. Retrieved from http://10.0.84.243/16463153/36a5

This research analyzes the web environment and the information produced in this medium, aiming to configure web archiving as an object of study, as a source of research data, along with scientific communication, as a practice of disseminating knowledge produced in universities. The methodology was delimited as exploratory research, based on an international bibliographic review on the subject, and analysis of the Initiatives of the International Consortium for the Preservation of the Internet (IIPC) related with Universities.

It uses qualitative analysis of the objectives and projects developed by these initiatives. It concludes that the Web archiving is a field still not explored enough, namely inside the universities, and it observes the lack of research in Latin America context, especially in Brazil. (English) [ABSTRACT FROM AUTHOR]

Finnemann, N. O. (2018). Web Arkiver ; Web archives. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

This article deals with general web archives and the principles for selection of materials to be preserved. It opens with a brief overview of reasons why general web archives are needed. Section two and three present major, long termed web archive initiatives and discuss the purposes and possible values of web archives and asks how to meet unknown future needs, demands and concerns. Section four analyses three main principles in contemporary web archiving strategies, topic centric, domain centric and time-centric archiving strategies and section five discuss how to combine these to provide a broad and rich archive. Section six is concerned with inherent limitations and why web archives are always flawed. The last sections deal with the question how web archives may fit into the rapidly expanding, but fragmented landscape of digital repositories taking care of various parts of the exponentially growing amounts of still more heterogeneous data materials. ; This article deals with general web archives and the principles for selection of materials to be preserved. It opens with a brief overview of reasons why general web archives are needed. Section two and three present major, long termed web archive initiatives and discuss the purposes and possible values of web archives and asks how to meet unknown future needs, demands and concerns. Section four analyses three main principles in contemporary web archiving strategies, topic centric, domain centric and time-centric archiving strategies and section five discuss how to combine these to provide a broad and rich archive. Section six is concerned with inherent limitations and why web archives are always flawed. The last sections deal with the question how web archives may fit into the rapidly expanding, but fragmented landscape of digital repositories taking care of various parts of the exponentially growing amounts of still more heterogeneous data materials.

Fox, E. A., & Farag, M. M. (2013). Report on the Workshop on Web Archiving and Digital Libraries (WADL 2013). *SIGIR Forum*, *47*(2), 128–133. https://doi.org/10.1145/2568388.2568408

This workshop explored the integration of Web archiving and digital libraries, so the complete life cycle involved is covered, from creation/authoring, uploading/publishing in the Web (including Web 2.0), (focused) crawling, curation, indexing, exploration (including searching and browsing), (text) analysis, archiving, and up through long-term preservation. It included particular coverage of current topics of interest: challenges facing archiving initiatives, archiving related to disasters, interaction with and use of archive data, applications on an international scale, working with big data, mobile Web archiving, temporal issues, Memento, and SiteStory.

Frederick, U. K. (2015). Glitch. *Journal of Contemporary Archaeology*, *2*(1), S28–S32. https://doi.org/10.1558/jca.v2i1.28244

The rapid and continual advancement of the internet as a platform for communication on archaeological topics has brought permanent changes to the methods through which we

present information from the sector to the public. This article discusses the potential for an exploration of the UK web archives for information about the history of archaeology online, and a case study undertaken as part of a Big Data project at the British Library by the author. The article concludes that we have a significant issue for media archaeologists in the future; the lack of material evidence for these iterations means we risk losing an understanding of our social, economic, cultural, and technological histories and our perception of these developments over time. It suggests that further exploration of these archives from an archaeological perspective could be beneficial both as an investigation of the iterations of digital archaeology (the creation of a history of public engagement with the subject), and as a study of the use of archaeological techniques for archival research. [ABSTRACT FROM AUTHOR]

Freeland, C., & Atiso, K. (2015). Determining Users' Motivations to Participate in Online Community Archives: A Preliminary Study of Documenting Ferguson. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (p. 106:1--106:4). Silver Springs, MD, USA: American Society for Information Science. Retrieved from http://dl.acm.org/citation.cfm?id=2857070.2857176

The shooting death of teenager Michael Brown in Ferguson, Missouri, spurred an immediate national and international response in the fall of 2014. Washington University Libraries in St. Louis, Missouri, established the Documenting Ferguson web archive to gather digital media documenting local protests and demonstrations as captured by community members in order to archive the materials for future research and scholarly use. This preliminary study identified the factors that motivated participants to contribute content to the Documenting Ferguson online community archive, uncovering themes of altruism, reciprocity, and personal development.

Galad, A. (2016). *ArchiveSpark - MS Independent Study Final Submission*. United States, North America: Virginia Tech. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

This project expands upon the work at the Internet Archive of researcher Vinay Goel and of Jefferson Bailey (co-PI on two NSF-funded collaborative projects with Virginia Tech: IDEAL, GETAR) on the ArchiveSpark project - a framework for efficient Web archive access, extraction, and derivation. The main goal of the project is to quantitatively and qualitatively evaluate ArchiveSpark against mainstream Web archive processing solutions and extend it as necessary with regard to the processing of testing collections. This also relates to an IMLS funded project. This report describes the efforts and contributions made as part of this project. The primary focus of these efforts lies in the comprehensive evaluation of ArchiveSpark against existing archive-processing solutions (pure Apache Spark with pre-installed Warcbase tools and HBase) in a variety of environments and setups in order to comparatively analyze performance improvements that ArchiveSpark brings to the table as well as understand the shortcomings and tradeoffs of its usage under varying scenarios. ; IMLS LG-71-16-0037-16: Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse ; NSF IIS-1619028, III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR) ; NSF IIS - 1319578: III: Small: Integrated Digital Event Archiving and Library (IDEAL) ; Included are the final report (PDF + Word), the final presentation (PPTX + PDF), the ArchiveSpark demo in the form of Jupyter Notebook, and the software developed during this project.

Garnar, M. (2018). Silencing Marginalized Voices: The Fragmentation of the Official Record - Library & Information Science Collection - ProQuest. *Reference & User Services Quarterly*, *57*(3), 193–195. Retrieved from https://search.proquest.com/libraryinformation/docview/2016963494/abstract/D0A3C983EC43401FPQ/2?accountid=27464

When researching historical topics, government statistics are often viewed as the most reliable source of information, lending credibility to the researchers' arguments by providing documentary evidence of how society is changing. In investigating issues related to equity, diversity, and inclusion, these statistics serve as benchmarks for the progress (or lack thereof) on how historic injustices are being addressed. Therefore, it is imperative that the information be reliable, verifiable, and available. In this case, the Internet Archive may have the missing pages on their website, but there's no guarantee that the desired information was captured, whether because pages were missed or snapshots missed important updates. There is also no guarantee that this nonprofit, nongovernmental website will continue to be available in the future. Without reliable access to government information, researchers will not be able to document what was available on governmental websites, and an important source of public policy data will be lost to future researchers.

Garzó, A., Daróczy, B., Kiss, T., Siklósi, D., & Benczúr, A. A. (2013). Cross-lingual Web Spam Classification. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 1149–1156). New York, NY, USA: ACM. https://doi.org/10.1145/2487788.2488139

While Web spam training data exists in English, we face an expensive human labeling procedure if we want to filter a Web domain in a different language. In this paper we overview how existing content and link based classification techniques work, how models can be "translated" from English into another language, and how language-dependent and independent methods combine. In particular we show that simple bag-of-words translation works very well and in this procedure we may also rely on mixed language Web hosts, i.e. those that contain an English translation of part of the local language text. Our experiments are conducted on the ClueWeb09 corpus as the training English collection and a large Portuguese crawl of the Portuguese Web Archive. To foster further research, we provide labels and precomputed values of term frequencies, content and link based features for both ClueWeb09 and the Portuguese data.

Gelfand, A. (2018). Web Archives for the Analog Archivist: Using Webpages Archived by the Internet Archive to Improve Processing and Description. *Journal of Western Archives*, *9*(1). Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Twenty years ago the Internet Archive was founded with the wide-ranging mission of providing universal access to all knowledge. In the two decades since, that organization has captured and made accessible over 150 billion websites. By incorporating the use of Internet Archive's Wayback Machine into their workflows, archivists working primarily with analog records may enhance their ability in such tasks as the construction of a processing plan, the creation of more accurate historical descriptions for finding aids, and potentially be able to provide better reference services to their patrons. This essay will look at some of the ways this may be accomplished.

Gillner, B. (2018). Offene Archive: Archive, Nutzer und Technologie im Miteinander. *OPEN ARCHIVES: ARCHIVES, USERS AND TECHNOLOGY INTERCONNECTED.*, *71*(1), 13–21. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The use of archives in the digital age is still a mostly analogue activity. This is not only due to the fact that the digitization of materials is costly and time-consuming, but also that there is a widely spread lack of interest in using the possibilities provided by the internet for the own agenda. For two decades the internet has primarily been a place for archives to present fixed (meta)data of archival materials. The concept of open archives strives to adapt the use of archives so far to the realities of the digital age. Its goal is to facilitate open data, focussing on users and using of digital tools. Only the interaction of those aspects can help show archives a way how to make the cultural heritage available to a large audience in a digital environment and how to make use of it in a variety of manners. [ABSTRACT FROM AUTHOR]

Glanville, L. (2010). Web archiving: ethical and legal issues affecting programmes in Australia and the Netherlands. *The Australian Library Journal*, *59*(3), 128–134. https://doi.org/10.1080/00049670.2010.10735999

Digital preservation is a major concern for libraries and organisations internationally. This paper will examine the barriers faced by web archiving programmes in national libraries, such as the Koninklijke Bibliotheek in the Netherlands and the National Library of Australia's PANDORA. The report will analyse how these programmes deal with the difficulties and limitations inherent in such programmes by examining how they approach issues of selection, access and copyright, while drawing comparisons between the programmes of the two institutions and the legal frameworks in which they function.

Golderman, G., & Connolly, B. (2018). Government Surveillance and Declassified Documents. *Library Journal*, *143*(1), 124–131. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Reviews are presented for several websites, including Digital National Security Archive at www.proquest.com/productsservices/databases/dnsa.html, ProQuest History Vault: Black Freedom Struggle in the 20th Century at www.proquest.com/productsservices/historyvault.html, and Secret Files from World.

Gomes, D., & Costa, M. (2014). The Importance of Web Archives for Humanities. *International Journal of Humanities and Arts Computing*, *8*(1), 106–123. https://doi.org/10.3366/ijhac.2014.0122

The web is the primary means of communication in developed societies. It contains descriptions of recent events generated through distinct perspectives. Thus, the web is a valuable resource for contemporary historical research. However, its information is extremely ephemeral. Several research studies have shown that only a small amount of information remains available on the web for longer than one year. Web archiving aims to acquire, preserve and provide access to historical information published online. In April 2013, there were at least sixty four web archiving initiatives worldwide. Altogether, these archived collections of web documents form a comprehensive picture of our cultural, commercial,

scientific and social history. Web archiving has also an important sociological impact because ordinary citizens are publishing personal information online without preservation concerns. In the future, web archives will probably be the only source of personal memories to many people. We provide some examples of tools that facilitate historical research over web archives highlighting their potential for Humanities. [ABSTRACT FROM AUTHOR]

Gomes, D., Costa, M., Cruz, D., Miranda, J., & Fontes, S. (2013). Creating a billion-scale searchable web archive. In *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion* (pp. 1059–1066). New York, New York, USA: ACM Press. https://doi.org/10.1145/2487788.2488118

Web information is ephemeral. Several organizations around the world are struggling to archive information from the web before it vanishes. However, users demand efficient and effective search mechanisms to access the already vast collections of historical information held by web archives. The Portuguese Web Archive is the largest full-text searchable web archive publicly available. It supports search over 1.2 billion files archived from the web since 1996. This study contributes with an overview of the lessons learned while developing the Portuguese Web Archive, focusing on web data acquisition, ranking search results and user interface design. The developed software is freely available as an open source project. We believe that sharing our experience obtained while developing and operating a running service will enable other organizations to start or improve their web archives.

Gomes, D., Santos, A. L., & Silva, M. J. (2006). Managing Duplicates in a Web Archive. In *Proceedings of the 2006 ACM Symposium on Applied Computing* (pp. 818–825). New York, NY, USA: ACM. https://doi.org/10.1145/1141277.1141465

Crawlers harvest the web by iteratively downloading documents referenced by URLs. It is frequent to find different URLs that refer to the same document, leading crawlers to download duplicates. Hence, web archives built through incremental crawls waste space storing these documents. In this paper, we study the existence of duplicates within a web archive and discuss strategies to eliminate them at storage level during the crawl. We present a storage system architecture that addresses the requirements of web archives and detail its implementation and evaluation. The system is now supporting an archive for the Portuguese web replacing previous NFS-based storage servers. Experimental results showed that the elimination of duplicates can improve storage throughput. The web storage system outperformed NFS based storage by 68% in read operations and by 50% in write operations.1

Gorsky, M. (2015). Into the Dark Domain: The UK Web Archive as a Source for the Contemporary History of Public Health. *Social History of Medicine*, *28*(3), 596. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

With the migration of the written record from paper to digital format, archivists and historians must urgently consider how web content should be conserved, retrieved and analysed. The British Library has recently acquired a large number of UK domain websites, captured 1996-2010, which is colloquially termed the Dark Domain Archive while technical issues surrounding user access are resolved. This article reports the results of an invited pilot project that explores methodological issues surrounding use of this archive. It asks how the relationship between UK public health and local government was represented on the web,

drawing on the “declinist” historiography to frame its questions. It points up some difficulties in developing an aggregate picture of web content due to duplication of sites. It also high lights their potential for thematic and discourse analysis, using both text and image, illustrated through an argument about the contradictory rationale for public health policy under New Labour. [ABSTRACT FROM AUTHOR]

Gossen, G., Demidova, E., & Risse, T. (2016). Analyzing web archives through topic and event focused sub-collections. In *Proceedings of the 8th ACM Conference on Web Science - WebSci ’16* (pp. 291–295). New York, New York, USA: ACM Press. https://doi.org/10.1145/2908131.2908175

Gossen, G., Demidova, E., & Risse, T. (2015). iCrawl: Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 75–84). New York, NY, USA: ACM. https://doi.org/10.1145/2756406.2756925

Researchers in the Digital Humanities and journalists need to monitor, collect and analyze fresh online content regarding current events such as the Ebola outbreak or the Ukraine crisis on demand. However, existing focused crawling approaches only consider topical aspects while ignoring temporal aspects and therefore cannot achieve thematically coherent and fresh Web collections. Especially Social Media provide a rich source of fresh content, which is not used by state-of-the-art focused crawlers. In this paper we address the issues of enabling the collection of fresh and relevant Web and Social Web content for a topic of interest through seamless integration of Web and Social Media in a novel integrated focused crawler. The crawler collects Web and Social Media content in a single system and exploits the stream of fresh Social Media content for guiding the crawler.

Grant, D., Debruyne, C., Grant, R., & Collins, S. (2015). Creating and Consuming Metadata from Transcribed Historical Vital Records for Ingestion in a Long-Term Digital Preservation Platform (pp. 445–450). https://doi.org/10.1007/978-3-319-26138-6_47

Gray, G., & Martin, S. (2013). Choosing a Sustainable Web Archiving Method: A Comparison of Capture Quality. *D-Lib Magazine*, *19*(5–6). https://doi.org/http://dx.doi.org/10.1045/may2013-gray

The UCLA Online Campaign Literature Archive has been collecting websites from Los Angeles and California elections since 1998. Over the years the number of websites created for these campaigns has soared while the staff manually capturing the websites has remained constant. By 2012 it became apparent that we would need to find a more sustainable model if we were to continue to archive campaign websites. Our ideal goal was to find an automated tool that could match the high quality captures produced by the Archive’s existing labor-intensive manual capture process. The tool we chose to investigate was the California Digital Library’s Web Archiving Service (WAS). To test the quality of WAS captures we created a duplicate capture of the June 2012 California election using both WAS and our manual capture and editing processes. We then compared the results from the two captures to measure the relative quality of the two captures. This paper presents the results of our findings and contributes a unique empirical analysis of the quality of websites archived using two divergent web archiving methods and sets of tools. Adapted from the source document.

Grimshaw, J. (2016). UK Official Publications: Managing the Transition to Electronic Deposit at the British Library. *Legal Information Management*, *16*(1), 3–9. https://doi.org/http://dx.doi.org/10.1017/S1472669616000037

This article by Jennie Grimshaw presents an overview of the transition of UK government publishing from print to electronic between the mid-1990s and 2016. It goes on to describe the tools being developed by the British Library in collaboration with the other five legal deposit libraries, to collect, preserve, organise and provide access to born digital government publications. This paradigm shift in official publishing gives the libraries a window of opportunity to improve their management of these materials and ensure that they can be found through their catalogues more easily than their print predecessors.

Grimshaw, J. (2015). The Digital Documents Harvesting and Processing Tool. *ALISS Quarterly*, *10*(2), 6–8. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article presents an overview of the Digital Documents Harvesting and Processing Tool (DDHAPT). Topics discussed include extension of legal deposit to cover electronic publications under the Legal Deposit Libraries (Non-Print Works) Regulations, DDHAPT web based application is an extension of the W3ACT tool used for web archiving by the British Library and the other legal deposit libraries and DDHAPT enabling a selector to set up a list of unique URLs to be crawled at set intervals.

Grotke, A. (2011). Web Archiving at the Library of Congress. *Computers in Libraries*, *31*(10), 15–19. Retrieved from https://search.proquest.com/docview/911079001?accountid=27464

The selection of sites is not something that the LC automates; recommending officers (ROs) do this work.n The goals of the consortium include collecting a rich body of internet content from around the world and fostering the development and use of common tools, techniques, and standards that enable the creation of international archives. IIPC members are currently engaged in a number of exciting projects: launching a worldwide education and training program that will feature technical and curatorial workshops and staff exchanges; planning an international collaborative collection project around the 2012 Summer Olympics; publishing information about the preservation of web archives in many institutional contexts; and establishing a technical program to fund exploratory projects and report about new techniques and tools to archives the fastchanging web.

Grotke, A. (2017). Getting Started in Web Archiving. In *IFLA Congress 2017, Wroclaw, Poland*. Retrieved from http://library.ifla.org/1637/

This purpose of this paper is to provide general information about how organizations can get started in web archiving, for both those who are developing new web archiving programs and for libraries that are just beginning to explore the possibilities. The paper includes an overview of considerations when establishing a web archiving program, including typical approaches that national libraries take when preserving the web. These include: collection development, legal issues, tools and approaches, staffing, and whether to do work in-house or outsource some or most of the work. The paper will introduce the International Internet Preservation Consortium and the benefits of collaboration when building web archives.

Gulyás, L., Jurányi, Z., Soós, S., & Kampis, G. (2014). Can web presence predict academic performance? In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* (pp. 1183–1188). New York, New York, USA: ACM Press. https://doi.org/10.1145/2567948.2579037

This paper reports the preliminary results of a project that aims at incorporating the analysis of the web presence (content) of research institutions into the scientometric analysis of these research institutions. The problem is to understand and predict the dynamics of academic activity and resource allocation using web presence. The present paper approaches this problem in two parts. First we develop a crawler and an archive of the web contents obtained from academic institutions, and present an early analysis of the records. Second, we use (currently off-line records to analyze the dynamics of resource allocation. Combination of the two parts is an ambition of ongoing work. The motivation in this study is twofold. First, we strongly believe that independent archiving, indexing and searching of (past) web content is an important task, even with regards to academic web presence. We are particularly interested in studying the dynamics of the "online scientific discourse", based on the assumption that the changing traces of web presence is an important factor that documents the intensity of activity. Second, we maintain that the trend-analysis of scientific activity represents a hitherto unused potential. We illustrate this by a pilot where, using 'offline' longitudinal datasets, we study whether past (i.e. cumulative) success can predict current (and future) activity in academia. Or, in short: do institutions invest and publish in areas where they have been successful? Answer to this question is, we believe, important to understanding and predicting research policies and their changes.

Hagedoorn, B., & Agterberg, B. (2016). The End of the Television Archive as We Know It? The National Archive as an Agent of Historical Knowledge in the Convergence Era. *Media and Communication, Vol 4, Iss 3, Pp 162-175 (2016) VO - 4*, (3), 162. https://doi.org/10.17645/mac.v4i3.595

Professionals in the television industry are working towards a certain future—rather than end—for the medium based on multi-platform storytelling, as well as multiple screens, distribution channels and streaming platforms. They do so rooted in institutional frameworks where traditional conceptualizations of television still persist. In this context, we reflect on the role of the national television archive as an agent of historical knowledge in the convergence era. Contextualisation and infrastructure function as important preconditions for users of archives to find their way through the enormous amounts of audio-visual material. Specifically, we consider the case of the Netherlands Institute for Sound and Vision, taking a critical stance towards the archive's practices of contextualisation and preservation of audio-visual footage in the convergence era. To do so, this article considers the impact of online circulation, contextualisation and preservation of audio-visual materials in relation to, first, how media policy complicates the re-use of material, and second, the archive's use by television professionals and media researchers. This article reflects on the possibilities for and benefits of systematic archiving, developments in web archiving, and accessibility of production and contextual documentation of public broadcasters in the Netherlands. We do so based on an analysis of internal documentation, best practices of archive-based history programmes and their related cross-media practices, as well as media policy documentation. We consider how audio-visual archives should deal with the shift towards multi-platform productions, and argue for both a more systematic archiving of production and contextual documentation in the Netherlands, and for media researchers who draw upon archival resources to show a greater awareness of an archive's history. In the digital age, even more

people are part of the archive's processes of selection and aggregation, affecting how the past is preserved through audio-visual images.

Halbert, M., Skinner, K., Wilson, M., & Zarndt, F. (2016). Here Today, Gone within a Month: The Fleeting Life of Digital News. In *IFLA WLIC 2016 – Columbus, OH – Connections. Collaboration. Community in Session S21 - Satellite Meeting: News Media. In: News, new roles & preservation advocacy: moving libraries into action, 10-12 August 2016, Lexington, KY, USA.* Lexington, KY, USA: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/id/eprint/2077

In 1989 on the shores of Montana's beautiful Flathead Lake, the owners of the weekly newspaper the Bigfork Eagle started TownNews.com to help community newspapers with developing technology. TownNews.com has since evolved into an integrated digital publishing and content management system used by more than 1600 newspaper, broadcast, magazine, and web-native publications in North America. TownNews.com is now headquartered on the banks of the mighty Mississippi river in Moline Illinois. Not long ago Marc Wilson, CEO of TownNews.com, noticed that of the 220,000+ e-edition pages posted on behalf of its customers at the beginning of the month, 210,000 were deleted by month's end. What? The front page story about a local business being sold to an international corporation that I read online September 1 will be gone by September 30? As well as the story about my daughter's 1st place finish in the district field and track meet? A 2014 national survey by the Reynolds Journalism Institute (RJI) of 70 digital-only and 406 hybrid (digital and print) newspapers conclusively showed that newspaper publishers also do not maintain archives of the content they produce. RJI found a dismal 12% of the "hybrid" newspapers reported even backing up their digital news content and fully 20% of the "digital-only" newspapers reported that they are backing up none of their content. Educopia Institute's 2012 and 2015 surveys with newspapers and libraries concur, and further demonstrate that the longstanding partner to the newspaper—the library—likewise is neither collecting nor preserving this digital content. This leaves us with a bitter irony, that today, one can find stories published prior to 1922 in the Library of Congress's Chronicling America and other digitized, out-of-copyright newspaper collections but cannot, and never will be able to, read a story published online less than a month ago. In this paper we look at how much news is published online that is never published in print or on more permanent media. We estimate how much online news is or will soon be forever lost because no one preserves it: not publishers, not libraries, not content management systems, and not the Internet Archive. We delve into some of the reasons why this content is not yet preserved, and we examine the persistent challenges of digital preservation and of digital curation of this content type. We then suggest a pathway forward, via some initial steps that journalists, producers, legislators, libr…

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK Webspace: Fifteen Years of British Universities on the Web. In *Proceedings of the 2014 ACM Conference on Web Science* (pp. 62–70). New York, NY, USA: ACM. https://doi.org/10.1145/2615569.2615691

This paper maps the national UK web presence on the basis of an analysis of the .uk domain from 1996 to 2010. It reviews previous attempts to use web archives to understand national web domains and describes the dataset. Next, it presents an analysis of the .uk domain, including the overall number of links in the archive and changes in the link density of different second-level domains over time. We then explore changes over time within a

particular second-level domain, the academic subdomain .ac.uk, and compare linking practices with variables, including institutional affiliation, league table ranking, and geographic location. We do not detect institutional affiliation affecting linking practices and find only partial evidence of league table ranking affecting network centrality, but find a clear inverse relationship between the density of links and the geographical distance between universities. This echoes prior findings regarding offline academic activity, which allows us to argue that real-world factors like geography continue to shape academic relationships even in the Internet age. We conclude with directions for future uses of web archive resources in this emerging area of research.

Harrison, T. L., & Nelson, M. L. (2006). Just-in-time Recovery of Missing Web Pages. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (pp. 145–156). New York, NY, USA: ACM. https://doi.org/10.1145/1149941.1149971

We present Opal, a light-weight framework for interactively locating missing web pages (http status code 404). Opal is an example of “in vivo” preservation: harnessing the collective behavior of web archives, commercial search engines, and research projects for the purpose of preservation. Opal servers learn from their experiences and are able to share their knowledge with other Opal servers by mutual harvesting using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Using cached copies that can be found on the web, Opal creates lexical signatures which are then used to search for similar versions of the web page. We present the architecture of the Opal framework, discuss a reference implementation of the framework, and present a quantitative analysis of the framework that indicates that Opal could be effectively deployed.

He, Ji., & Suel, T. (2012). Optimizing Positional Index Structures for Versioned Document Collections. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 245–254). New York, NY, USA: ACM. https://doi.org/10.1145/2348283.2348319

Versioned document collections are collections that contain multiple versions of each document. Important examples are Web archives, Wikipedia and other wikis, or source code and documents maintained in revision control systems. Versioned document collections can become very large, due to the need to retain past versions, but there is also a lot of redundancy between versions that can be exploited. Thus, versioned document collections are usually stored using special differential (delta) compression techniques, and a number of researchers have recently studied how to exploit this redundancy to obtain more succinct full-text index structures. In this paper, we study index organization and compression techniques for such versioned full-text index structures. In particular, we focus on the case of positional index structures, while most previous work has focused on the non-positional case. Building on earlier work in [zs:redun], we propose a framework for indexing and querying in versioned document collections that integrates non-positional and positional indexes to enable fast top-k query processing. Within this framework, we define and study the problem of minimizing positional index size through optimal substring partitioning. Experiments on Wikipedia and web archive data show that our techniques achieve significant reductions in index size over previous work while supporting very fast query processing.

He, J., Yan, H., & Suel, T. (2009). Compact Full-text Indexing of Versioned Document Collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge*

Management (pp. 415–424). New York, NY, USA: ACM. https://doi.org/10.1145/1645953.1646008

We study the problem of creating highly compressed full-text index structures for versioned document collections, that is, collections that contain multiple versions of each document. Important examples of such collections are Wikipedia or the web page archive maintained by the Internet Archive. A straightforward indexing approach would simply treat each document version as a separate document, such that index size scales linearly with the number of versions. However, several authors have recently studied approaches that exploit the significant similarities between different versions of the same document to obtain much smaller index sizes. In this paper, we propose new techniques for organizing and compressing inverted index structures for such collections. We also perform a detailed experimental comparison of new techniques and the existing techniques in the literature. Our results on an archive of the English version of Wikipedia, and on a subset of the Internet Archive collection, show significant benefits over previous approaches.

He, S., & Chan, E. (2012). Surfing Notes: An Integrated Web Annotation and Archiving Tool. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03* (pp. 301–305). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/WI-IAT.2012.174

Web archiving for preserving the valuable information online from disappearing due to the dynamic nature of the World Wide Web, and web annotation for promoting the development of the Web as a two-way information sharing platform are both active research fields. However, in spite of their common benefits to information management and intelligent learning, few attempts have been made to integrate web archiving and web annotation. This paper introduces Surfing Notes, a cloud-based system which allows the users to annotate and archive the web pages for personal use. The change detection algorithm as well as the change detection interval scheduler are discussed in detail and evaluated experimentally.

Heil, J. M., & Jin, S. (2017). Preserving Seeds of Knowledge: A Web Archiving Case Study. *Information Management Journal*, *51*(3), 20–22. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article presents a case study on the initiative of Queen's University in Ontario, Canada on a project which seeks to preserve website contents. Topics include the procedures involved in archiving web contents, the software tools used by archivists including its subscription to WayBackMachine, and the officers that make up the project.

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. *Alexandria*, *25*(1/2), 113–127. Retrieved from https://search.proquest.com/docview/1623365740?accountid=27464

Hockx-Yu, H. (2016). *Web Archiving at National Libraries Findings of Stakeholders' Consultation by the Internet Archive*.

Internet Archive conducted a stakeholders' consultation exercise between November 2015 and March 2016, with the aim to understand current practices, and then review Internet Archive's current services in this light and explore new aspects for national libraries. This document reports on the consultation and summarises the findings.

Hockx-Yu, H. (2011). The Past Issue of the Web. In *Proceedings of the 3rd International Web Science Conference* (p. 12:1--12:8). New York, NY, USA: ACM. https://doi.org/10.1145/2527031.2527050

This paper takes a critical look at the efforts since the mid-1990s in archiving and preserving websites by memory institutions around the world. It contains an overview of the approaches and practices to date, and a discussion of the various technical, curatorial and legal issues related to web archiving. It also looks at a number of current projects which take a different approach to dealing with the temporal aspects or persistence of the web. The paper argues for closer collaboration with the main stream web science research community and the use of technology developed for the live web, such as visualisation and data analytics, to advance the web archiving agenda.

HOCKX-Yu, H., & KAHLE, B. (2014). Forget Me Net, Not. *Newsweek Global*, *163*(2), 1–6. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article discusses web archiving, focussing on a private project, Internet Archive, founded by Brewster Kahle and the project of the British Library to capture and preserve every web-page in the British domain, .co.uk, led by Helen Hockx-Yu. Topics include estimates of the amount of digital data created each year, estimates of the amount of data lost or altered in a year and the evolution of the role of libraries as they branch out to web archiving.

Holub, K., & Rudomino, I. (2015). A decade of web archiving in the National and University Library in Zagreb. In *Preservation and Conservation with Information Technology. IFLA 2015 South Africa*. Cape Town: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/1092/1/090-holub-en.pdf

Due to the dynamic nature of the web, its explosive growth, short lifespan, instability and similar characteristics, the importance of its archiving has become priceless for future generations. The National and University Library in Zagreb (Nacionalna i sveučilišna knjižnica u Zagrebu, NSK), as a memory institution responsible for collecting, cataloguing, archiving and providing access to all types of resources, recognized the significance of collecting and storing online content as part of the NSK's core activities. This is supported by positive legal environment since 1997 when Croatia passed the Law on libraries which subjected online publications to legal deposit. In 2004 NSK established the Croatian Web Archive (Hrvatski arhiv weba, HAW) in collaboration with the University Computing Centre (Srce) and developed a system for capturing and archiving Croatian web resources. From 2004 to 2010 only selective archiving of web resources was conducted according to preestablished selection criteria. Taking into account NSK's responsibility to preserve resources on Croatian social, scientific and cultural history, the importance of taking a snapshot of all publicly available resources under the national top level domain (.hr) was been recognized in 2011. Since then national domain harvestings have been conducted annually. In addition, in 2011 NSK started to run thematic harvestings of national importance. The paper will present the NSK's ten years' experience in managing web resources with the emphasis on implementation of the system for selective and domain harvesting as well as the challenges for providing access to archived resources. Also, the harvested data from 2004 to 2014 will be analysed. The findings will illustrate the variability of URLs, frequency of harvesting and types of content. The data from the last four .hr harvestings will also be presented

Holzmann, H., & Anand, A. (2016). Tempas. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion* (pp. 207–210). New York, New York, USA: ACM Press. https://doi.org/10.1145/2872518.2890555

Holzmann, H., Goel, V., & Anand, A. (2016). ArchiveSpark. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (pp. 83–92). New York, New York, USA: ACM Press. https://doi.org/10.1145/2910896.2910902

Web archives are a valuable resource for researchers of various disciplines. However, to use them as a scholarly source, researchers require a tool that provides efficient access to Web archive data for extraction and derivation of smaller datasets. Besides efficient access we identify five other objectives based on practical researcher needs such as ease of use, extensibility and reusability. Towards these objectives we propose ArchiveSpark, a framework for efficient, distributed Web archive processing that builds a research corpus by working on existing and standardized data formats commonly held by Web archiving institutions. Performance optimizations in ArchiveSpark, facilitated by the use of a widely available metadata index, result in significant speed-ups of data processing. Our benchmarks show that ArchiveSpark is faster than alternative approaches without depending on any additional data stores while improving usability by seamlessly integrating queries and derivations with external tools.

Holzmann, H., Goel, V., & Gustainis, E. N. (2017). *Universal distant reading through metadata proxies with archivespark. 2017 IEEE International Conference on Big Data (Big Data), Big Data (Big Data), 2017 IEEE International Conference on*. IEEE. https://doi.org/10.1109/BigData.2017.8257958

Digitization and the large-scale preservation of digitized content have engendered new ways of accessing and analyzing collections concurrent with other data mining and extraction efforts. Distant reading refers to the analysis of entire collections instead of close reading individual items like a single physical book or electronic document. The steps performed in distant reading are often common across various types of data collections like books, journals, or web archives, sources that are very valuable and have often been neglected as Big Data. We have extended our tool ArchiveSpark, originally designed to efficiently process Web archives, in order to support arbitrary data collections being served from either local or remote data sources by using metadata proxies. The ability to share and reuse researcher workflows across disciplines with very different datasets makes ArchiveSpark a universal distant reading framework. In this paper, we describe ArchiveSpark's design extensions along an example of how it can be leveraged to analyze symptoms of Polio mentioned in journals from the Medical Heritage Library. Our experiments demonstrate how users can reuse large portions of their job pipeline to accomplish a specific task across diverse data types and sources. Migrating an ArchiveSpark job to process a different dataset introduces an additional average code complexity of only 4.8%. Its expressiveness, scalability, extensibility, reusability, and efficiency has the potential to advance novel and rich methods of scholarly inquiry.

Holzmann, H., Nejdl, W., & Anand, A. (2016). The Dawn of Today's Popular Domains. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (pp. 73–82). New York, New York, USA: ACM Press. https://doi.org/10.1145/2910896.2910901

The Web has been around and maturing for 25 years. The popular websites of today have undergone vast changes dur- ing this period, with a few being there almost since the beginning and many new ones becoming popular over the years. This makes it worthwhile to take a look at how these sites have evolved and what they might tell us about the future of the Web. We therefore embarked on a longitudinal study spanning almost the whole period of the Web, based on data collected by the Internet Archive starting in 1996, to retrospectively analyze how the popular Web as of now has evolved over the past 18 years. For our study we focused on the German Web, specifically on the top 100 most popular websites in 17 categories. This paper presents a selection of the most interesting findings in terms of volume, size as well as age of the Web. While re- lated work in the field of Web Dynamics has mainly focused on change rates and analyzed datasets spanning less than a year, we looked at the evolution of websites over 18 years. We found that around 70% of the pages we investigated are younger than a year, with an observed exponential growth in age as well as in size up to now. If this growth rate con- tinues, the number of pages from the popular domains will almost double in the next two years. In addition, we give insights into our data set, provided by the Internet Archive, which hosts the largest and most complete Web archive as of today.

Holzmann, H., Nejdl, W., & Anand, A. (2016). On the Applicability of Delicious for Temporal Search on Web Archives. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16* (pp. 929–932). New York, New York, USA: ACM Press. https://doi.org/10.1145/2911451.2914724

Holzmann, H., Nejdl, W., & Anand, A. (2017). Exploring Web Archives Through Temporal Anchor Texts. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 289–298). New York, NY, USA: ACM. https://doi.org/10.1145/3091478.3091500

Web archives have been instrumental in digital preservation of the Web and provide great opportunity for the study of the societal past and evolution. These Web archives are massive collections, typically in the order of terabytes and petabytes. Due to this, search and exploration of archives has been limited as full-text indexing is both resource and computationally expensive. We identify that for typical access methods to archives, which are navigational and temporal in nature, we do not always require indexing full-text. Instead, meaningful text surrogates like anchor texts already go a long way in providing meaningful solutions and can act as reasonable entry points to exploring Web archives. In this paper, we present a new approach to searching Web archives based on temporal link graphs and corresponding anchor texts. Departing from traditional informational intents, we show how temporal anchor texts can be effective in answering queries beyond purely navigational intents, like finding the most central webpages of an entity in a given time period. We propose indexing methods and a temporal retrieval model based on anchor texts. Further, we discuss several interesting search results as well as one experiment in which we demonstrate how such results can be integrated in a data processing workflow to scale up to thousands of pages. In this analysis we were able to replicate results reported by an offline study, showing that restaurant prices indeed increased in Germany when the Euro was introduced as Europe's currency.

Holzmann, H., & Risse, T. (2014). Named Entity Evolution Analysis on Wikipedia. In *Proceedings of the 2014 ACM Conference on Web Science* (pp. 241–242). New York, NY, USA: ACM. https://doi.org/10.1145/2615569.2615639

Accessing Web archives raises a number of issues caused by their temporal characteristics. Additional knowledge is needed to find and understand older texts. Especially entities mentioned in texts are subject to change. Most severe in terms of information retrieval are name changes. In order to find entities that have changed their name over time, search engines need to be aware of this evolution. We tackle this problem by analyzing Wikipedia in terms of entity evolutions mentioned in articles. We present statistical data on excerpts covering name changes, which will be used to discover similar text passages and extract evolution knowledge in future work.

Holzmann, H., & Runnwerth, M. (2018). Micro Archives as Rich Digital Object Representations. In *Proceedings of the 10th ACM Conference on Web Science - WebSci '18* (pp. 353–357). New York, New York, USA: ACM Press. https://doi.org/10.1145/3201064.3201110

Digital objects as well as real-world entities are commonly referred to in literature or on the Web by mentioning their name, linking to their website or citing unique identifiers, such as DOI and OR- CID, which are backed by a set of meta information. All of these methods have severe disadvantages and are not always suitable though: They are not very precise, not guaranteed to be persistent or mean a big additional effort for the author, who needs to collect the metadata to describe the reference accurately. Especially for complex, evolving entities and objects like software, pre-defined metadata schemas are often not expressive enough to capture its temporal state comprehensively. We found in previous work that a lot of meaningful information about software, such as a description, rich metadata, its documentation and source code, is usually avail- able online. However, all of this needs to be preserved coherently in order to constitute a rich digital representation of the entity. We show that this is currently not the case, as only 10% of the stud- ied blog posts and roughly 30% of the analyzed software websites are archived completely, i.e., all linked resources are captured as well. Therefore, we propose Micro Archives as rich digital object representations, which semantically and logically connect archived resources and ensure a coherent state. With Micrawler we present a modular solution to create, cite and analyze such Micro Archives. In this paper, we show the need for this approach as well as discuss opportunities and implications for various applications also beyond scholarly writing.

Holzmann, H., Tahmasebi, N., & Risse, T. (2015). Named entity evolution recognition on the Blogosphere. *International Journal on Digital Libraries*, *15*(2–4), 209–235. Retrieved from http://10.0.3.239/s00799-014-0135-x

Advancements in technology and culture lead to changes in our language. These changes create a gap between the language known by users and the language stored in digital archives. It affects user's possibility to firstly find content and secondly interpret that content. In a previous work, we introduced our approach for named entity evolution recognition (NEER) in newspaper collections. Lately, increasing efforts in Web preservation have led to increased availability of Web archives covering longer time spans. However, language on the Web is more dynamic than in traditional media and many of the basic assumptions from the newspaper domain do not hold for Web data. In this paper we discuss the limitations of existing methodology for NEER. We approach these by adapting an existing NEER method to work on noisy data like the Web and the Blogosphere in particular. We develop novel filters that reduce the noise and make use of Semantic Web resources to obtain more information about terms. Our evaluation shows the potentials of the proposed approach. [ABSTRACT FROM AUTHOR]

Huang, L., Zhu, J. J. H., & Li, X. (2008). Histrace: Building a Search Engine of Historical Events. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 1155–1156). New York, NY, USA: ACM. https://doi.org/10.1145/1367497.1367703

In this paper, we describe an experimental search engine on our Chinese web archive since 2001. The original data set contains nearly 3 billion Chinese web pages crawled from past 5 years. From the collection, 430 million "article-like" pages are selected and then partitioned into 68 million sets of similar pages. The titles and publication dates are determined for the pages. An index is built. When searching, the system returns related pages in a chronological order. This way, if a user is interested in news reports or commentaries for certain previously happened event, he/she will be able to find a quite rich set of highly related pages in a convenient way.

Huang, L., Zhu, J. J. H., & Li, X. (2008). Building a story tracer out of a web archive. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries - JCDL '08* (p. 455). New York, New York, USA: ACM Press. https://doi.org/10.1145/1378889.1379000

There are quite a few web archives around the world, such as Internet Archive and Web InfoMall (http://www.infomall.cn). Nevertheless, we have not seen substantial mechanism built on top of the archives to render the value of the data beyond what the Wayback machine offers. One of the reasons for this situation is the lack of a system vision and design which encompasses the oceanic data in a meaningful and cost-effective way. This paper describes an effort in this direction.

Hulser, R. P. (2015). The California Light and Sound Collection: Preserving Our Media Heritage. *Computers in Libraries*, *35*(3), 4–10. Retrieved from https://search.proquest.com/docview/1680527010?accountid=27464

While most of the focus on digital preservation and access has been on digitizing printed materials, there is an initiative underway in California to capture and make accessible audiovisual content in such a way that even libraries, museums, and archives with limited resources can participate. The California Light and Sound collection is the outgrowth of the California Preservation Program's California Audiovisual Preservation Project (CAVPP). CAVPP plays the lead role in helping participating partner organizations conserve and preserve their audiovisual collections according to best practices for the archiving and preservation of moving image and sound formats. It also established a low-cost and practical workflow for helping partner organizations efficiently digitize key media artifacts. CAVPP coordinates all digitization activities with the vendor doing the digitization work and helps the participating institution throughout the process. To optimize quality control, CAVPP prefers working with labs that can handle all audiovisual formats. This not only saves shipping costs but ensures that the appropriate standards and procedures are applied to all recordings.

Huurdeman, H. C. (2014). Adaptive search systems for web archive research. In *Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14* (pp. 354–356). New York, New York, USA: ACM Press. https://doi.org/10.1145/2637002.2637063

The wealth of digital information available in our time has become indispensable for a rich variety of tasks. We use data on the Web for work, leisure, and research, aided by various search systems, allowing us to find small needles in giant haystacks. Despite recent advances

in personalization and contextualization, however, various types of tasks, ranging from simple lookup tasks to complex, exploratory and ana- lytical ventures, are mainly supported in elementary, “one- size-fits-all” search interfaces. Web archives, keepers of our future cultural heritage, have gathered petabytes of valuable Web data, which characterize our times for future generations. Access to these archives, however, is surprisingly limited: online Web archives usu- ally provide a URL-based Wayback Machine interface, some- times extended with rudimentary search options. As a re- sult of limited access, Web archives have not been widely used for research so far. For emerging research using Web archives, there is a need to move beyond URL-based and simple search access, towards providing support for complex (re)search tasks. In my thesis, I am exploring ways to move beyond the “one-size-fits-all” approach for search systems, and I work on systems which can support the flow of complex search, also in the context of archived Web data. Rich models of search and research can be incorporated into adaptive search systems, supporting search strategies in various stages of complex search tasks. Concretely, I look at the use case of the Humanities researcher, for which the large, Terabyte- scale Web archives can be a valuable addition to existing sources utilized to perform research

Huurdeman, H. C., Ben-David, A., Kamps, J., Samar, T., & de Vries, A. P. (2014). Finding Pages on the Unarchived Web. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 331–340). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2740769.2740827

Web archives preserve the fast changing Web, yet are highly incomplete due to crawling restrictions, crawling depth and frequency, or restrictive selection policies---most of the Web is unarchived and therefore lost to posterity. In this paper, we propose an approach to recover significant parts of the unarchived Web, by reconstructing descriptions of these pages based on links and anchors in the set of crawled pages, and experiment with this approach on the Dutch Web archive. Our main findings are threefold. First, the crawled Web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of the Web archive. Second, the link and anchor descriptions have a highly skewed distribution: popular pages such as home pages have more terms, but the richness tapers off quickly. Third, the succinct representation is generally rich enough to uniquely identify pages on the unarchived Web: in a known-item search setting we can retrieve these pages within the first ranks on average.

Huurdeman, H. C., Ben-David, A., & Sammar, T. (2013). Sprint Methods for Web Archive Research. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 182–190). New York, NY, USA: ACM. https://doi.org/10.1145/2464464.2464513

Web archives provide access to snapshots of the Web of the past, and could be valuable for research purposes. However, access to these archives is often limited, both in terms of data availability, and interfaces to this data. This paper explores new methods to overcome these limitations. It presents “sprint-methods” for performing research using an archived collection of the Dutch news aggregator Website Nu.nl, and for developing and adapting a search system and interface to this data. The work aims to contribute to research in the humanities and social sciences, in particular New Media research employing digital methods to study the Web of the past. Secondly, this work aims to contribute to Computer Science, in the development of novel access tools for Web archives, that facilitate research.

Huurdeman, H. C., Kamps, J., Samar, T., de Vries, A. P., Ben-David, A., & Rogers, R. A. (2015). Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries*, *16*(3–4), 247–265. https://doi.org/10.1007/s00799-015-0153-3

Issue Title: Focused Issue on Digital Libraries 2014 Web archives attempt to preserve the fast changing web, yet they will always be incomplete. Due to restrictions in crawling depth, crawling frequency, and restrictive selection policies, large parts of the Web are unarchived and, therefore, lost to posterity. In this paper, we propose an approach to uncover unarchived web pages and websites and to reconstruct different types of descriptions for these pages and sites, based on links and anchor text in the set of crawled pages. We experiment with this approach on the Dutch Web Archive and evaluate the usefulness of page and host-level representations of unarchived content. Our main findings are the following: First, the crawled web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of a Web archive. Second, the link and anchor text have a highly skewed distribution: popular pages such as home pages have more links pointing to them and more terms in the anchor text, but the richness tapers off quickly. Aggregating web page evidence to the host-level leads to significantly richer representations, but the distribution remains skewed. Third, the succinct representation is generally rich enough to uniquely identify pages on the unarchived web: in a known-item search setting we can retrieve unarchived web pages within the first ranks on average, with host-level representations leading to further improvement of the retrieval effectiveness for websites.

Imafuji, N., & Kitsuregawa, M. (2002). Effects of Maximum Flow Algorithm on Identifying Web Community. In *Proceedings of the 4th International Workshop on Web Information and Data Management* (pp. 43–48). New York, NY, USA: ACM. https://doi.org/10.1145/584931.584941

In this paper, we describe the effects of using maximum flow algorithm on extracting web community from the web. A web community is a set of web pages having a common topic. Since the web can be recognized as a graph that consists of nodes and edges that represent web pages and hyperlinks respectively, so far various graph theoretical approaches have been proposed to extract web communities from the web graph. The method of finding a web community using maximum flow algorithm was proposed by NEC Research Institute in Princeton two years ago. However the properties of web communities derived by this method have been seldom known. To examine the effects of this method, we selected 30 topics randomly and experimented using Japanese web archives crawled in 2000. Through these experiments, it became clear that the method has both advantages and disadvantages. We will describe some strategies to use this method effectively. Moreover, by using same topics, we examined another method that is based on complete bipartite graphs. We compared the web communities obtained by those methods and analyzed those characteristics.

INOUE, S. (2018). Passing on the Lessons of the Great East Japan Earthquake to Future Generations—The National Diet Library Great East Japan Earthquake Archive. In *IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies in Session 233 - Government Information and Official Publications.* Kuala Lumpur. Retrieved from http://library.ifla.org/id/eprint/2217

In the aftermath of the Great East Japan Earthquake, which struck on March 11, 2011, the Japanese government recognized an urgent need to create a national archive of information

about this unprecedented natural disaster, so that the learned lessons from this experience would not be lost. Having an obligation as a national library to collect, preserve, and share materials that record all aspects of Japan's cultural heritage, the National Diet Library (NDL), in cooperation with other Japanese government agencies, has responded to this need by creating a portal site, called HINAGIKU, through which researchers can search and access a wide variety of earthquake archives. In this paper, I will report on our achievements as well as the challenges we face in configuring HINAGIKU to facilitate access to documentation published or archived primarily by the national and municipal government agencies. At present, HINAGIKU enables access to materials documenting both past experience and current disaster prevention planning via an integrated search functionality of multiple digital archives established by municipal governments, academic institutions, the Ministry of Internal Affairs and Communications, and other organizations as well as the NDL. Visitors to HINAGIKU are able to search records stored at the NDL and other institutions, and new knowledge generated from such research can also be integrated into HINAGIKU as new content. Over time, as interest in earthquake-related materials decreases, it becomes imperative that the NDL acquire and preserve these materials before such archives disappear. The NDL also has a role to play in handing down these most valuable records to future generations by managing issues related to copyright, personality rights, and secondary use, thereby making HINAGIKU even more useful.

Jackson, A., Lin, J., Milligan, I., & Ruest, N. (2016). Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 103–106). New York, NY, USA: ACM. https://doi.org/10.1145/2910896.2910912

Web archiving initiatives around the world capture ephemeral web content to preserve our collective digital memory. In this paper, we describe initial experiences in providing an exploratory search interface to web archives for humanities scholars and social scientists. We describe our initial implementation and discuss our findings in terms of desiderata for such a system. It is clear that the standard organization of a search engine results page (SERP), consisting of an ordered list of hits, is inadequate to support the needs of scholars. Shneiderman's mantra for visual information seeking ("overview first, zoom and filter, then details-on-demand") provides a nice organizing principle for interface design, to which we propose an addendum: "Make everything transparent". We elaborate on this by highlighting the importance of the temporal dimension of web pages as well as issues surrounding metadata and veracity.

Jacobsen, G. (2008). Web Archiving: Issues and Problems in Collection Building and Access. *Liber Quarterly: The Journal of European Research Libraries, Vol.18, No.3-4*, *18*(3–4). Retrieved from http://liber.library.uu.nl/

Denmark began web archiving in 2005 and the experiences are presented with a specific focus on collection-building and issues concerning access. In creating principles for what internet materials to collect for a national collection, one can in many ways build on existing practice and guidelines. The actual collection requires strategies for harvesting relevant segments of the internet in order to assure as complete a coverage as possible. Rethinking is also necessary when it comes to the issue of description, but cataloguing expertise can be utilised to find new ways for users to retrieve information. Technical problems in harvesting and archiving are identifiable and can be solved through international cooperation. Access to the archived materials, on the other hand, has become the major challenge to national libraries. Legal

obstacles prevent national libraries from offering generel access to their archived internet materials. In Europe the principal obstacles are the EU Directive on Data Protection (Directive 95/46/EC) and local data protection legislation based on this directive. LIBER is urged to take political action on this issue in order that the general public may have the same access to the collection of internet materials as it has to other national collections. Adapted from the source document.

Jassalini Jamain;, Yahya, A. L., Muhammad, N., & Musa Ayob Abdul Rahman. (2018). Web archiving issues and challenges in State Government of Sarawak (Malaysia): Do they really need their website to be archived? In IFLA (Ed.), *IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies in Session 160 - Preservation and Conservation with Information Technology*. Kuala Lumpur: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/id/eprint/2115

Sarawak State Web Archive (SSWA) is Sarawak State Library's (Pustaka) initiative. Website contents of the Sarawak State Civil Service (SSCS) entities obtained from World Wide Web (WWW), are archived for the purpose of preserving non-library resources, as part of the Legal Deposit requirements of Sarawak State Library Ordinance, 1999. Web preservation is considered as a common practice at international level, whereas in Malaysia this is still at a minimal level. Since 2009, Pustaka has been harvesting 132 websites of Sarawak State Government departments and agencies. However, Pustaka faced challenges in performing web archiving works. This paper focuses on the general issues and challenges in preserving corporate information heritage to make it available for future reference.

Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y., & Tanaka, K. (2006). Journey to the Past: Proposal of a Framework for Past Web Browser. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (pp. 135–144). New York, NY, USA: ACM. https://doi.org/10.1145/1149941.1149969

While the Internet community recognized early on the need to store and preserve past content of the Web for future use, the tools developed so far for retrieving information from Web archives are still difficult to use and far less efficient than those developed for the "live Web." We expect that future information retrieval systems will utilize both the "live" and "past Web" and have thus developed a general framework for a past Web browser. A browser built using this framework would be a client-side system that downloads, in real time, past page versions from Web archives for their customized presentation. It would use passive browsing, change detection and change animation to provide a smooth and satisfactory browsing experience. We propose a meta-archive approach for increasing the coverage of past Web pages and for providing a unified interface to the past Web. Finally, we introduce query-based and localized approaches for filtered browsing that enhance and speed up browsing and information retrieval from Web archives.

Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y., & Tanaka, K. (2006). A browser for browsing the past web. In *Proceedings of the 15th international conference on World Wide Web - WWW '06* (p. 877). New York, New York, USA: ACM Press. https://doi.org/10.1145/1135777.1135923

We describe a browser for the past web. It can retrieve data from multiple past web resources and f eatures a passive browsing style based on change detection and pr esentation. The

browser shows past pages one by one along a tim e line. The parts that were changed between consecutive page versions are animated to reflect their deletion or insertion, thereby drawing the user's attention to them. The browser enables automatic skipping of changeless periods and filtered br owsing based on user specified query

Jatowt, A., Kawai, Y., Ohshima, H., & Tanaka, K. (2008). What Can History Tell Us?: Towards Different Models of Interaction with Document Histories. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia* (pp. 5–14). New York, NY, USA: ACM. https://doi.org/10.1145/1379092.1379098

The current Web is a dynamic collection where little effort is made to version pages or to enable users to access historical data. As a consequence, they generally do not have sufficient temporal support when browsing the Web. However, we think that there are many benefits to be obtained from integrating documents with their histories. For example, a document's history can enable us to travel back through time to establish its trustworthiness. This paper discusses the possible types of interactions that users could have with document histories and it presents several examples of systems that we have implemented for utilizing this historical data. To support our view, we present the results of an online survey conducted with the objective of investigating user needs for temporal support on the Web. Although the results indicated quite low use of Web archives by users, they simultaneously emphasized their considerable interest in page histories.

Jatowt, A., Kawai, Y., & Tanaka, K. (2008). Visualizing Historical Content of Web Pages. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 1221–1222). New York, NY, USA: ACM. https://doi.org/10.1145/1367497.1367736

Jatowt, A., Kawai, Y., & Tanaka, K. (2007). Detecting Age of Page Content. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management* (pp. 137–144). New York, NY, USA: ACM. https://doi.org/10.1145/1316902.1316925

Web pages often contain objects cr eated at different times. The information about the age of such objects may provide useful context for understanding page content and may serve many potential uses. In this paper, we describe a novel concept for detecting approximate creation date s of content elements in Web pages. Our approach is based on dynamically reconstructing page histories using data extracted from external sources - Web archives and efficiently searching inside them to detect insertion dates of content elements. We di scuss various issues involving the proposed approach and demonstrate the example of an application that enhances browsing the Web by inserting annotations with temporal metadata into page content on user request.

Johnson, P. (2016). Where Did All the Information Go? Well at Least the Important Stuff. *Refer*, *32*(2), 8–12. Retrieved from https://search.proquest.com/docview/1803538845?accountid=27464

There is a lot of current concern about the sheer amount of web pages and digital documents being lost forever. In this definition lost implies destroyed. A report in 2011 by the Chesapeake Digital Preservation Group suggested that approximately 30% of a control group of 2,700 online law related materials disappeared in 3 years. Librarians have been highlighting this concern for many years and must take a lot of the credit for the introduction

of so many successful web archiving initiatives -- including the Non-Print Legal Deposit legislation enacted in the UK in 2013. However, for this article the author want to focus on a different kind of lost information, where the definition of lost refers to information that cannot be found. In December 2015 he gave a presentation at a Koha library systems event, which explored the changing environment of discovery services within the academic market. Clicking on the repeat the search link runs the search again with no apparent limits on the amount of results returned.

Joint, N. (2006). Legal deposit and collection development in a digital world. *Library Review*, *55*(8), 468–473. https://doi.org/10.1108/00242530610689310

Purpose – To compare and contrast national collection management principles for hard copy deposit collections and for digital deposit collections. Design/methodology/approach – A selective overview and summary of work to date on digital legal deposit and digital preservation. Findings – That the comprehensive nature of traditional print deposit collection often absolves national libraries from the more intractable problems of stock selection; whereas the difficulty of collecting the entire national digital web space means that intelligent selection is vital for the building of meaningful digital deposit collections. Research limitations/implications – These are indicative and partial insights based on small scale interrogation of trial digital deposit collections: the issue of collection development and selection biases in digital collection building needs greater in-depth research before hard and fast recommendations about collection management criteria can be arrived at. Practical implications – The principles outlined may offer practitioners in national libraries some useful insights into how to manage their digital deposit collections. Originality/value – This paper emphasises the social and political aspects of digital deposit issues, rather than the legal or technical aspects.

Jones, S. M., Nelson, M. L., & Van de Sompel, H. (2018). Avoiding spoilers: wiki time travel with Sheldon Cooper. *International Journal on Digital Libraries*, *19*(1), 77–93. https://doi.org/http://dx.doi.org/10.1007/s00799-016-0200-8

A variety of fan-based wikis about episodic fiction (e.g., television shows, novels, movies) exist on the World Wide Web. These wikis provide a wealth of information about complex stories, but if fans are behind in their viewing they run the risk of encountering “spoilers”—information that gives away key plot points before the intended time of the show’s writers. Because the wiki history is indexed by revisions, finding specific dates can be tedious, especially for pages with hundreds or thousands of edits. A wiki’s history interface does not permit browsing across historic pages without visiting current ones, thus revealing spoilers in the current page. Enterprising fans can resort to web archives and navigate there across wiki pages that were live prior to a specific episode date. In this paper, we explore the use of Memento with the Internet Archive as a means of avoiding spoilers in fan wikis. We conduct two experiments: one to determine the probability of encountering a spoiler when using Memento with the Internet Archive for a given wiki page, and a second to determine which date prior to an episode to choose when trying to avoid spoilers for that specific episode. Our results indicate that the Internet Archive is not safe for avoiding spoilers, and therefore we highlight the inherent capability of fan wikis to address the spoiler problem internally using existing, off-the-shelf technology. We use the spoiler use case to define and analyze different ways of discovering the best past version of a resource to avoid spoilers. We propose Memento as a structural solution to the problem, distinguishing it from prior content-based solutions to the spoiler problem. This research promotes the idea that content management

systems can benefit from exposing their version information in the standardized Memento way used by other archives. We support the idea that there are use cases for which specific prior versions of web resources are invaluable.

Jones, S. M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & Grover, C. (2016). Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. *PLoS ONE*, *11*(12), 1–32. Retrieved from http://10.0.5.91/journal.pone.0167475

Increasingly, scholarly articles contain URI references to "web at large" resources including project web sites, scholarly wikis, ontologies, online debates, presentations, blogs, and videos. Authors reference such resources to provide essential context for the research they report on. A reader who visits a web at large resource by following a URI reference in an article, some time after its publication, is led to believe that the resource's content is representative of what the author originally referenced. However, due to the dynamic nature of the web, that may very well not be the case. We reuse a dataset from a previous study in which several authors of this paper were involved, and investigate to what extent the textual content of web at large resources referenced in a vast collection of Science, Technology, and Medicine (STM) articles published between 1997 and 2012 has remained stable since the publication of the referencing article. We do so in a two-step approach that relies on various well-established similarity measures to compare textual content. In a first step, we use 19 web archives to find snapshots of referenced web at large resources that have textual content that is representative of the state of the resource around the time of publication of the referencing paper. We find that representative snapshots exist for about 30% of all URI references. In a second step, we compare the textual content of representative snapshots with that of their live web counterparts. We find that for over 75% of references the content has drifted away from what it was when referenced. These results raise significant concerns regarding the long term integrity of the web-based scholarly record and call for the deployment of techniques to combat these problems. [ABSTRACT FROM AUTHOR]

Jones, S., & Latzko-Toth, G. (2017). Out from the PLATO cave: uncovering the pre-Internet history of social computing. *Internet Histories*, *1*(1–2), 60–69. https://doi.org/10.1080/24701475.2017.1307544

Jordan, W., Kelly, M., Brunelle, J. F., Vobrak, L., Weigle, M. C., & Nelson, M. L. (2015). Mobile Mink. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15* (pp. 243–244). New York, New York, USA: ACM Press. https://doi.org/10.1145/2756406.2756956

We describe the mobile app emph{Mobile Mink} which extends Mink, a browser extension that integrates the live and archived web. Mobile Mink discovers mobile and desktop URIs and provides the user an aggregated TimeMap of both mobile and desktop mementos. Mobile Mink also allows users to submit mobile and desktop URIs for archiving at the Internet Archive and Archive.today. Mobile Mink helps to increase the archival coverage of the growing mobile web.

Kampis, G., & Gulyás, L. (2013). Big is small, and changes slowly in Hungary. In *Coginfo 2013 Conference*.

The Internet Archive is incomplete and national archives are necessary. We report a pilot study in Hungary, targeting the archiving of the public internet content of academic research

institutions, and present some early analysis results, indicating that the internet based “big data” is unexpectedly small for Hungary, and furthermore that this dataset changes at a low rate. We suggest that differences in the productivity of the institutions can be safely correlated with the differences in content refreshment in their internet presence.

Kanhabua, N. (2012). Time-aware Approaches to Information Retrieval. *SIGIR Forum*, *46*(1), 85. https://doi.org/10.1145/2215676.2215691

In this thesis, we address major challenges in searching temporal document collections. In such collections, documents are created and/or edited over time. Examples of temporal document collections are web archives, news archives, blogs, personal emails and enterprise documents. Unfortunately, traditional IR approaches based on term-matching only can give unsatisfactory results when searching temporal document collections. The reason for this is twofold: the contents of documents are strongly time-dependent, i.e., documents are about events happened at particular time periods, and a query representing an information need can be time-dependent as well, i.e., a temporal query. Our contributions in this thesis are different time-aware approaches within three topics in IR: content analysis, query analysis, and retrieval and ranking models. In particular, we aim at improving the retrieval effectiveness by 1) analyzing the contents of temporal document collections, 2) performing an analysis of temporal queries, and 3) explicitly modeling the time dimension into retrieval and ranking. Leveraging the time dimension in ranking can improve the retrieval effectiveness if information about the creation or publication time of documents is available. In this thesis, we analyze the contents of documents in order to determine the time of non-timestamped documents using temporal language models. We subsequently employ the temporal language models for determining the time of implicit temporal queries, and the determined time is used for re-ranking search results in order to improve the retrieval effectiveness. We study the effect of terminology changes over time and propose an approach to handling terminology changes using time-based synonyms. In addition, we propose different methods for predicting the effectiveness of temporal queries, so that a particular query enhancement technique can be performed to improve the overall performance. When the time dimension is incorporated into ranking, documents will be ranked according to both textual and temporal similarity. In this case, time uncertainty should also be taken into account. Thus, we propose a ranking model that considers the time uncertainty, and improve ranking by combining multiple features using learning-to-rank techniques. Through extensive evaluation, we show that our proposed time-aware approaches outperform traditional retrieval methods and improve the retrieval effectiveness in searching temporal document c…

Kanhabua, N., Kemkes, P., Nejdl, W., Nguyen, T. N., Reis, F., & Tran, N. K. (2016). How to Search the Internet Archive Without Indexing It. In *Research & Advanced Technology for Digital Libraries: 20th International Conference on Theory & Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings* (pp. 147–160). https://doi.org/10.1007/978-3-319-43997-6_12

Significant parts of cultural heritage are produced on the web during the last decades. While easy accessibility to the current web is a good baseline, optimal access to the past web faces several challenges. This includes dealing with large-scale web archive collections and lacking of usage logs that contain implicit human feedback most relevant for today’s web search. In this paper, we propose an entity-oriented search system to support retrieval and analytics on the Internet Archive. We use Bing to retrieve a ranked list of results from the current web. In addition, we link retrieved results to the WayBack Machine; thus allowing keyword search on

the Internet Archive without processing and indexing its raw archived content. Our search system complements existing web archive search tools through a user-friendly interface, which comes close to the functionalities of modern web search engines (e.g., keyword search, query auto-completion and related query suggestion), and provides a great benefit of taking user feedback on the current web into account also for web archive search. Through extensive experiments, we conduct quantitative and qualitative analyses in order to provide insights that enable further research on and practical applications of web archives

Kavcic-Colic, A., & Klasinc, J. (2011). Web Archiving in the National and University Library. *Knjiznica*, *55*(1), 209–232. Retrieved from https://search.proquest.com/docview/1266143501?accountid=27464

The National and University Library (NUK) of Slovenia has been investigating web archiving methods and techniques since 2001. Under the new Legal Deposit Law adopted in 2006, NUK is the responsible institution for harvesting and archiving the Slovenian web. In 2008 NUK started archiving the Slovenian web by making use of the web harvesting and access tools developed by the IIPC International Internet Preservation Consortium (IIPC). The paper presents the complexity of web harvesting and gives an overview of the international practice and NUK's cooperation in the IIPC consortium. Special attention is given to the analysis of public sector web content, harvested since 2008. Main goals of future development of the web archive are an increase of harvested Slovenian web sites, the development of a user interface for public access and development of improved methods for harvesting technically problematic content. Adapted from the source document.

Kelly, M., Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2013). A Method for Identifying Personalized Representations in Web Archives. *D-Lib Magazine*, *19*(11–12). https://doi.org/http://dx.doi.org/10.1045/november2013-kelly

Web resources are becoming increasingly personalized - two different users clicking on the same link at the same time can see content customized for each individual user. These changes result in multiple representations of a resource that cannot be canonicalized in Web archives. We identify characteristics of this problem by presenting a potential solution to generalize personalized representations in archives. We also present our proof-of-concept prototype that analyzes WARC (Web ARChive) format files, inserts metadata establishing relationships, and provides archive users the ability to navigate on the additional dimension of environment variables in a modified Wayback Machine. Adapted from the source document.

Kelly, M., Nelson, M. L., & Weigle, M. C. (2018). A Framework for Aggregating Private and Public Web Archives. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18* (pp. 273–282). New York, New York, USA: ACM Press. https://doi.org/10.1145/3197026.3197045

Personal and private Web archives are proliferating due to the increase in the tools to create them and the realization that Internet Archive and other public Web archives are unable to capture personalized (e.g., Facebook) and private (e.g., banking) Web pages. We introduce a framework to mitigate issues of aggregation in private, personal, and public Web archives without compromising potential sensitive information contained in private captures. We amend Memento syntax and semantics to allow TimeMap enrichment to account for additional attributes to be expressed inclusive of the requirements for dereferencing private Web archive captures. We provide a method to involve the user further in the negotiation of

archival captures in dimensions beyond time. We introduce a model for archival querying precedence and short-circuiting, as needed when aggregating private and personal Web archive captures with those from public Web archives through Memento. Negotiation of this sort is novel to Web archiving and allows for the more seamless aggregation of various types of Web archives to convey a more accurate picture of the past Web.

Kelly, M., Nelson, M. L., & Weigle, M. C. (2014). The Archival Acid Test: Evaluating Archive Performance on Advanced HTML and JavaScript. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 25–28). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2740769.2740774

When preserving web pages, archival crawlers sometimes produce a result that varies from what an end-user expects. To quantitatively evaluate the degree to which an archival crawler is capable of comprehensively reproducing a web page from the live web into the archives, the crawlers' capabilities must be evaluated. In this paper, we propose a set of metrics to evaluate the capability of archival crawlers and other preservation tools using the Acid Test concept. For a variety of web preservation tools, we examine previous captures within web archives and note the features that produce incomplete or unexpected results. From there, we design the test to produce a quantitative measure of how well each tool performs its task.

Kelly, M., & Weigle, M. C. (2012). WARCreate. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL '12* (p. 437). New York, New York, USA: ACM Press. https://doi.org/10.1145/2232817.2232930

The Internet Archive's Wayback Machine is the most common way that typical users interact with web archives. The Internet Archive uses the Heritrix web crawler to transform pages on the publicly available web into Web ARChive (WARC) files, which can then be accessed using the Wayback Machine. Because Heritrix can only access the publicly available web, many personal pages (e.g. password-protected pages, social media pages) cannot be easily archived into the standard WARC format. We have created a Google Chrome extension, WARCreate, that allows a user to create a WARC file from any webpage. Using this tool, content that might have been otherwise lost in time can be archived in a standard format by any user. This tool provides a way for casual users to easily create archives of personal online content. This is one of the first steps in resolving issues of "long term storage, maintenance, and access of personal digital assets that have emotional, intellectual, and historical value to individuals".

Kelly, M., Alkwai, L. M., Alam, S., Van de Sompel, H., Nelson, M. L., & Weigle, M. C. (2017). Impact of URI Canonicalization on Memento Count. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1–2). United States, North America: IEEE. https://doi.org/10.1109/JCDL.2017.7991601

Quantifying the captures of a URI over time is useful for researchers to identify the extent to which a Web page has been archived. Memento TimeMaps provide a format to list mementos (URI-Ms) for captures along with brief metadata, like Memento-Datetime, for each URI-M. However, when some URI-Ms are dereferenced, they simply provide a redirect to a different URI-M (instead of a unique representation at the datetime), often also present in the TimeMap. This infers that confidently obtaining an accurate count quantifying the number of non-forwarding captures for a URI-R is not possible using a TimeMap alone and that the magnitude of a TimeMap is not equivalent to the number of representations it identifies. In

this work we discuss this particular phenomena in depth. We also perform a breakdown of the dynamics of counting mementos for a particular URI-R (google.com) and quantify the prevalence of the various canonicalization patterns that exacerbate attempts at counting using only a TimeMap. For google.com we found that 84.9% of the URI-Ms result in an HTTP redirect when dereferenced. We expand on and apply this metric to TimeMaps for seven other URI-Rs of large Web sites and thirteen academic institutions. Using a ratio metric DI for the number of URI-Ms without redirects to those requiring a redirect when dereferenced, five of the eight large web sites' and two of the thirteen academic institutions' TimeMaps had a ratio of ratio less than one, indicating that more than half of the URI-Ms in these TimeMaps result in redirects when dereferenced.

Kennedy, S. D. (2015). Now You See It, Now You Don't. Unless ... *Information Today*, *32*(10), 8. Retrieved from https://search.proquest.com/docview/1761628166?accountid=27464

According to Jill Lepore, the average life of a webpage is 100 days. As she notes, the embarrassing stuff seems to stick around a lot longer, but it's an indisputable fact that web-based content often goes missing: corporate reports, scholarly articles, government documents, working papers, maps, and creative works of all sorts. The Internet Archive and its Wayback Machine are pretty much universally loved by information professionals. You already know this, but aside from the Wayback Machine's valuable research function, the Internet Archive itself is a major time suck. Entertainment value aside, in late October, the Internet Archive announced on its blog that "with generous support from the Laura and John Arnold Foundation," it was planning to build "the Next Generation Wayback Machine".

Kim, K., Jung, Y., & Myaeng, S.-H. (2016). A Topic Transition Map for Query Expansion: A Semantic Analysis of Click-Through Data and Test Collections. In B. H. Kang & Q. Bai (Eds.), *AI 2016: Advances in Artificial Intelligence* (pp. 648–664). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50127-7_57

Term mismatching between queries and documents has long been recognized as a key problem in information retrieval (IR). Based on our analysis of a large-scale web query log and relevant documents in standard test collections, we attempt to detect topic transitions between the topical categories of a query and those of relevant documents (or clicked pages) and create a Topic Transition Map (TTM) that captures how query topic categories are linked to those of relevant or clicked documents. TTM, a kind of click-graph at the semantic level, is then used for query expansion by suggesting the terms associated with the document categories strongly related to the query category. Unlike most other query expansion methods that attempt to either interpret the semantics of queries based on a thesaurus-like resource or use the content of a small number of relevant documents, our method proposes to retrieve documents in the semantic affinity of multiple categories of the documents relevant for the queries of a similar kind. Our experiments show that the proposed method is superior in effectiveness to other representative query expansion methods such as standard relevance feedback, pseudo relevance feedback, and thesaurus-based expansion of queries.

Kleiber, E. (2014). Gathering the 'Net: Efforts and Challenges in Archiving Pacific Websites. *The Contemporary Pacific*, *26*(1), 158–166. https://doi.org/10.1353/cp.2014.0017

In addition to more traditional material -- books, journals and other serial publications, brochures, music, films, manuscripts, photographs, postcards and archives -- the University of

Hawai'i-Manoa (UHM) Library's Hawaiian and Pacific Collections are now actively collecting websites. With so many new websites being created in and about the Pacific Islands region, and so much more information being made available online -- and at times exclusively so -- it has become increasingly clear to the librarians of these collections that to adequately document this period in history it is necessary to collect and preserve websites. The UHM Library has been attempting to archive websites in one form or another since 2001. This essay will discuss the importance of collecting Pacific websites, describe how the Hawaiian and Pacific Collections are finding solutions for the inherent challenges of preserving websites, and explore some potential future directions that would strengthen the project and meet the information and research needs of the Pacific Islands region. Adapted from the source document.

Klein, M., Balakireva, L., & de Sompel, H. (2018). Focused Crawl of Web Archives to Build Event Collections. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 333–342). New York, NY, USA: ACM. https://doi.org/10.1145/3201064.3201085

Event collections are frequently built by crawling the live web on the basis of seed URIs nominated by human experts. Focused web crawling is a technique where the crawler is guided by reference content pertaining to the event. Given the dynamic nature of the web and the pace with which topics evolve, the timing of the crawl is a concern for both approaches. We investigate the feasibility of performing focused crawls on the archived web. By utilizing the Memento infrastructure, we obtain resources from 22 web archives that contribute to building event collections. We create collections on four events and compare the relevance of their resources to collections built from crawling the live web as well as from a manually curated collection. Our results show that focused crawling on the archived web can be done and indeed results in highly relevant collections, especially for events that happened further in the past

Klein, M., Shankar, H., & de Sompel, H. (2018). Robust Links in Scholarly Communication. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 357–358). New York, NY, USA: ACM. https://doi.org/10.1145/3197026.3203885

Web resources change over time and many ultimately disappear. While this has become an inconvenient reality in day-to-day use of the web, it is problematic when these resources are referenced in scholarship where it is expected that referenced materials can reliably be revisited. We introduce Robust Links, an approach aimed at maintaining the integrity of the scholarly record in a dynamic web environment. The approach consists of archiving web resources when referencing them and decorating links to convey information that supports accessing referenced resources both on the live web and in web archives.

Konopa, B. (2017). Witryna internetowa – dokumentacja czy publikacja? TT - Website - documentation or publication? *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951539091?accountid=27464

W artykule podjęto rozważania teoretyczne nad charakterem witryn internetowych i próbą przypisania ich do definicji publikacji lub dokumentacji. Autor wskazuje na punkty, które mogą zostać wykorzystane w dyskusji nad tym zagadnieniem, między innymi obowiązujące normatywy, dorobek naukowy, praktykę innych państw oraz jednolite rzeczowe wykazy akt.

Kowal, K. C., & Meehan, S. (2018). Partnerships on Campus: Roles and Impacts in Developing a New Online Research Resource at Boston College. *The Catholic Library World*, *88*(3), 177–184. Retrieved from https://search.proquest.com/docview/2024455820?accountid=27464

A partnership at Boston College (BC) between the Libraries and the Institute for Advanced Jesuit Studies resulted in a blossoming of services and resources, made possible via a combination of discipline-focused scholarship and library digital expertise. With a shared, mission, the last three years have produced a number of programs and projects that relied upon a relationship of reciprocation, support, and ultimately the strategic directions guiding this Jesuit university.

Kragelj, M., & Kovačič, M. (2017). Uporabna vrednost podatkov spletnih zajemov: arhiviranje spletnih mest in analiza spletnih vsebin TT - The practical value of web capture data: archiving Web sites and Web content analysis. *Knjiznica*, *61*(1/2), 235–250. Retrieved from https://search.proquest.com/docview/1966852571?accountid=27464

Zakon o obveznem izvodu publikacij (2006) Narodni in univerzitetni knjižnici (NUK) nalaga skrb za zajem, ohranjanje in nudenje dostopa uporabnikom do zajetih spletnih publikacij, spletnih mest in vsebin. Leta 2015 je NUK opravil prvi zajem slovenske domene .si, naslove spletnih domen je priskrbel Arnes (Akademska in raziskovalna mreža Slovenije). V príspevku se osredotočamo na pomen zajema spletnih vsebin zaradi vsakodnevnega propadanja spletnih domen. Poleg zajema in dejavnosti za zagotavljanje ohranjanja zajetih vsebin je v prispevku tematizirano tudi pridobivanje informacij iz nestrukturiranih vsebin (spletnih dokumentov). Omenjeni so primeri in delovanje aplikacij za zajemanje specifičnih informacij iz različnih spletnih dokumentov, npr. zajem cene določenega artikla v določeni trgovini z namenom obveščanja končnega uporabnika o najugodnejši ponudbi na trgu. Večji del prispevka je namenjen analizi zajetih spletnih vsebin in možnosti luščenja ter uteževanja besedišča, pridobljenega iz spletnih dokumentov. Z algoritmi in statistikami za označevanje in razvrščanje terminov v množici spletnih vsebin se spletni arhiv iz pasivne podatkovne zbirke spremeni v okolje, ki omogoča dodano vrednost povezovanja podatkov, iskanja sorodnosti znotraj podatkov spletnega arhiva in s podatki zunaj njega.Alternatív absztrakt:The Legal Deposit Act imposes to the National and University Library the concern and rights for capturing, preserving and providing access to online publications, web sites, and other content to library users. In 2015, the Library started the first capture of Slovenian .si internet domain. The domain addresses were provided by ARNES (the Academic and Research Network of Slovenia). The article focuses on the importance of covering the web content due to the deterioration of daily web domains. In addition to covering and activities to ensure the conservation of web contents, the paper also covers the subject of how to obtain information from unstructured content (documents on the web). The article shows some examples and applications to capture specific information from a variety of online documents (scraping), like the price of a selected item in a particular web store in order to inform the end user about the best offer on the market. The major part of the article is devoted to the analysis of captured web content and the possibility of scaling and ranking the vocabulary derived from web documents. The algorithms and statistics for marking and docum…

KRAGELJ, M., & KOVAČIČ, M. (. (2015). First crawling of the Slovenian National web domain *.si: pitfalls, obstacles and challenges. In *Preservation and Conservation with Information Technology. IFLA 2015 South Africa*. Cape Town: IFLA -- International

Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/1191/1/090-kragelj-en.pdf

The National and University Library (NUK) has been archiving the web for almost fifteen years. During the last six years, we have been trying to act on different levels of harvesting. For most of the time, we have dealt with harvesting of selected web sites that might be significant for future generations. The harvesting process runs smoothly, with the exception of some technical difficulties resulting from the use of scripted languages (for instance Ajax, Flash, Java script, asynchronous transmissions, real time streaming protocols, etc.). The number of archived web pages keeps growing very fast. We are also very successful in harvesting social media web sites with tools developed in NUK. Being aware that the amount of the web pages cannot be compared with the harvested one - it is much more extensive – we decided to start the Slovenian domain (*.si) harvesting. The first domain harvesting was successful; however, we realized that much deeper and broader levels should be harvested by using heuristic methods. Our experiences showed that most informative web contents are hidden beneath the *.si domain's data provided by ARNES (Academic Research Network of Slovenia), therefore, the contents are not accessible. The paper presents the results of the first harvesting iteration of the Slovenian web. Further, on a sample of the first hundred domains, the results of the first and second harvesting iteration will be compared and analysed. At the end, the relevance of data acquired in the harvested web pages as a digital library complementary data source will be presented.

Kugler, A., Beinert, T., & Schoger, A. (2017). Archiwizacja internetu jako usługa naukowa TT - Internet archiving as a scientific service. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951541162?accountid=27464

Gromadzenie i archiwizowanie stron internetowych istotnych dla nauki to jak dotąd bardzo zaniedbana sfera aktywności bibliotek niemieckich. Aby zapobiec groźnym stratom oraz zapewnić pracownikom naukowym stały dostęp do stron internetowych ponad dwa lata temu Bavarian State Library (BSB) stworzyła system archiwizacji stron internetowych. Głównym celem projektu zaakceptowanym przez German Research Foundation (DFG) był rozwój i realizacja kooperacyjnego modelu usługowego. Usługa ta ma wspierać inne instytucje dziedzictwa kulturowego w ich aktywności archiwizacyjnej i ułatwiać budowanie rozproszonego niemieckiego systemu archiwizacji naukowych stron internetowych. Dzięki temu projektowi biblioteka bawarska chce poprawić zarówno ilość, jak i jakość zarchiwizowanych treści oraz promować ich wykorzystanie w obszarze nauki.

Kumar, A., & Xie, Z. (2018). Acquiring Web Content From In-Memory Cache. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18* (pp. 359–360). New York, New York, USA: ACM Press. https://doi.org/10.1145/3197026.3203868

Web content acquisition forms the foundation of value extraction of web data. Two main categories of acquisition methods are crawler based methods and transactional web archiving or server-side acquisition methods. In this poster, we propose a new method to acquire web content from web caches. Our method provides improvement in terms of reduced penalty on HTTP transaction, flexibility to accommodate peak web server loads and minimal involvement of System Administrator to set up the system.

Kumar, D. V., & Kumar, B. T. S. (2017). Recovery of vanished URLs: Comparing the efficiency of Internet Archive and Google. *Malaysian Journal of Library & Information Science*, *22*(2), 31. Retrieved from https://search.proquest.com/docview/1925123736?accountid=27464

This article examines the vanishing nature of URLs and recovery of vanished URLs through Internet Archive and Google search engine. For that purpose study investigates the URLs cited in the articles of two LIS journals published during 2009-2013. A total of 226 articles published in two open access LIS journals were selected. Of 5197 citations cited in 226 articles, 21.05 percent were URLs (1094). Study found that 38.12 percent (417 out of 5197) URLs were found missing and remaining 61.88 percent of URLs were active at the time of URL check with W3C link checker. The HTTP 404 error message – "page not found" was the overwhelming message encountered and represented 54.2 percent of all HTTP error message. Internet Archive and Google search engine were used to recover vanished URLs. However, the Internet Archive recovered 66.19 percent of the total vanished URLs, whereas, Google manages to recover only 30.70 percent of the total vanished URLs. The recovery of vanishing URLs through Internet Archive and Google increased the active URL's rate from 61.88 per cent to 87.11 per cent and 73.58 per cent respectively. Study found that Internet Archive is a most efficient tool to recover vanished URLs compared to Google search engine.

Kumar, V. D., & Sampath Kumar, B. T. (2017). Finding the unfound: Recovery of missing URLs through Internet Archive. *Annals of Library and Information Studies*, *64*(3), 165. Retrieved from https://search.proquest.com/docview/2073135310?accountid=27464

The study investigated the accessibility and permanency of citations containing URLs in the articles published in DESIDOC Journal of Library and Information Technology journal during 2006-2015. A total of 2133 URL citations were identified out of which 823 were found to be incorrect or missing. HTTP-404 was the most common error message associated with the missing URLs. The study also tried to recover the incorrect or URL citations using Internet Archive and recovered a total of 484 (58.81%) missing URL citations.

Kvasnica, J., & Kreibich, R. (2013). Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. (Czech). *File Format Recognition of Data Harvested by Web Archiving Project of National Library of the Czech Republic. (English)*, (2), 1. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

National Library of the Czech Republic just begun to ingest harvested data from web archiving project into Long-term Preservation System. This article is output of Institutional Science and Research project aiming to implement retrospective file format recognition framework for harvested data and map tools related to file format recognition. Precise knowledge of archived data is cornerstone for building Long-term Preservation Strategy. Such analysis may also improve conditions of end-user access. (English) [ABSTRACT FROM AUTHOR]

Lakshmi Tulasi, R., Rao, M. S., Ankita, K., & Hgoudar, R. (2017). Ontology-Based Automatic Annotation: An Approach for Efficient Retrieval of Semantic Results of Web Documents. In S. C. Satapathy, V. K. Prasad, B. P. Rani, S. K. Udgata, & K. S. Raju (Eds.), *Proceedings of the First International Conference on Computational Intelligence*

and Informatics (pp. 331–339). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-2471-9_32

The Web contains large amount of data of unstructured nature which gives the relevant as well as irrelevant results. To remove the irrelevancy in results, a methodology is defined which would retrieve the semantic information. Semantic search directly deals with the knowledge base which is domain specific. Everyone constructs ontology knowledge base in their own way, which results in heterogeneity in ontology. The problem of heterogeneity can be resolved by applying the algorithm of ontology mapping. All the documents are collected by Web crawler from the Web and a document base is created. The documents are then given as an input for performing semantic annotation on the updated ontology. The results against the users query are retrieved from semantic information retrieval system after applying searching algorithm on it. The experiments conducted with this methodology show that the results thus obtained provide more accurate and precise information.

Lampert, C. K., & Vaughan, J. (2018). Preparing to Preserve: Three Essential Steps to Building Experience with Long-Term Digital Preservation. In *IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies in Session 160 - Preservation and Conservation with Information Technology*. Kuala Lumpur: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/id/eprint/2114

Many organizations face complex questions of how to implement affordable and sustainable digital preservation practices. One strategic priority at the University Libraries at the University of Nevada-Las Vegas, United States, is increased focus toward preservation of unique digital assets, whether digitized from physical originals or born digital. A team comprised of experts from multiple functional library departments (including the special collections/archives area and the technology area) was established to help address this priority, and efforts are beginning to translate into operational practice. This work outlines a three-step approach: Partnership, Policy, Pilot taken by one academic research library to strategically build experience utilizing a collaborative team approach. Our experience included the formation of a team, education of all members, and a foundational attitude that decisions would be undertaken as partners rather than competing departments or units. The team's work included the development of an initial digital preservation policy, helping to distill the organizational priority and values associated with digital preservation. Several pilot projects were initiated and completed, which provided realistic, first-person experience with digital preservation activities, surfaced questions, and set the stage for developing and refining sustainable workflows. This work will highlight key activities in our journey to date, with the hope that experience gained through this effort could be applicable, in whole or part, to other organizations regardless of their size or capacity.

Lamphere, C. (2017). For Old Times' Sake: Technostalgia's Greatest Hits. *Online Searcher*, *41*(5), 27–29. Retrieved from https://search.proquest.com/docview/1942462381?accountid=27464

Nostalgia is a powerful feeling/emotion. In my case, chasing childhood nostalgia caused me to lug around an almost obsolete format for years before reluctantly parting with it- but only for practical reasons. Naturally, nostalgia's strong emotional pull makes it a driving force in consumption and marketing today. Nostalgia marketing is everywhere, from foods and advertising to technology. When it comes to technology, the coined word "technostalgia"

describes a “fond reminiscence of, or longing for, outdated technology” (en. wiktionary.org/wiki/technostalgia).

Lasfargues, F., Martin, C., & Medjkoune, L. (2012). Archiving before Loosing Valuable Data? Development of Web Archiving in Europe. *Bibliothek Forschung Und Praxis*, *36*(1), 117–124. https://doi.org/10.1515/bfp-2012-0014

Web content is, by nature, ephemeral: sites are updated regularly and disappear, which involves the loss of unique value information. The importance of this media grows continuously in our society and institutions are developing websites with a variety of content creating a large media-centric Web sphere. Like any media, it is essential to preserve it as a key part of our heritage.

László, D., & Crook, E. (2010). Crook, Edgar: Webarchiválás a webkettes világban. *Tudományos És Műszaki Tájékoztatás*, *57*(2), 78–81. Retrieved from http://epa.oszk.hu/03000/03071/00029/pdf/EPA03071_tmt_2010_02_068-081.pdf#page=11

A National Library of Australia vezető szerepet játszik az ausztrál web begyűjtésében és megőrzésében 1996, a PANDORA archívum (pandora. nla.gov.au) létrehozása óta. Emellett léteznek más, szűkebb körű projektek is, mint például a tasmániai Our Digital Island (odi.statelibrary. tas.gov.au), vagy a kontinens Northern Territory nevű részén működő Territory Stories (territorystories.nt.gov.au). A nemzeti könyvtár jelenleg már háromféle módon archivál: a PANDORA gyűjteménybe szelektíven válogat online forrásokat, továbbá az Internet „Archive” segítségével a teljes .au domént learatja, valamint elkezdte használni az „Archive-It” szolgáltatást is. Elmondható tehát, hogy az ausztrál online tartalom jelentős részét sikerül így megmenteni a jövő számára. De a technológiai változások miatt a könyvtárnak folyamatosan alkalmazkodnia kell: fejleszteni az archiváló eszközeit, bővíteni a gyűjtött tartalmak körét és újabb partnerekkel szövetkezni, hogy eredményesen tudja folytatni ezt a fontos munkát.

Law, M. T., Thome, N., Gançarski, S., & Cord, M. (2012). Structural and Visual Comparisons for Web Page Archiving. In *Proceedings of the 2012 ACM Symposium on Document Engineering* (pp. 117–120). New York, NY, USA: ACM. https://doi.org/10.1145/2361354.2361380

In this paper, we propose a Web page archiving system that combines state-of-the-art comparison methods based on the source codes of Web pages, with computer vision techniques. To detect whether successive versions of a Web page are similar or not, our system is based on: (1) a combination of structural and visual comparison methods embedded in a statistical discriminative model, (2) a visual similarity measure designed for Web pages that improves change detection, (3) a supervised feature selection method adapted to Web archiving. We train a Support Vector Machine model with vectors of similarity scores between successive versions of pages. The trained model then determines whether two versions, defined by their vector of similarity scores, are similar or not. Experiments on real archives validate our approach.

Le Béchec, M., & Hare, I. (2015). *Open data as political web archives : citizen involvement or reputation’s elected in a « digital public sphere » ?* France, Europe: HAL CCSD. Retrieved from

http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

International audience ; The access to digital data is an economic, social and political issue. Accessibility does not only focus on the online publication of these data in a database, but as well in the discourses made by stakeholders on the web, such as the French association Regards citoyens. Since 2009, this group aggregates data of the activity of French Deputies in the French National Assembly, on the website nosdeputes.fr. In this case, political people allow the circulation of data that are arranged by actors without professional requirements unlike journalists. We are here interested in the enrichment of public data by citizens who participate in the public sphere in a form that differs from the mass media. We do not want to comment this public sphere but to describe it from the devices, the mediations that connect institutions and citizens. Therefore, we discuss the opportunity that a website like nosdeputes.fr can become the holder of a “digital public sphere” and interrogate the form of citizen oversight it induces. The frame of data on nosdeputes.fr questions the relationship between citizens, media and elected officials. On the one hand, these devices change the relationship between citizens and political action. On the other hand, we can assume that these devices bring politicians to adapt some of their practices in the French National Assembly according to the electoral agenda. We do not focus on the influence of some actors but on the oversight of citizens induced by this device. For example, nosdeputes.fr has listed activities of the 577 French Deputies since 2009. This survey provides detailed analysis of political activity in National Assembly but it is also interested in the look of the “citizen”, by the comments he leaves on MPs’ action.

Leach-Murray, S. (2018). Archive-It. *Technical Services Quarterly*, *35*(2), 214. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article reviews the website “Archive-It” located at https://archive-it.org, which is a subscription web archiving service that collects and assesses cultural heritage on the Internet.

Ledoux, T. (2012). Long-term preservation at the National Library of France (BnF): Scalable Preservation and Archiving Repository (SPAR). *International Preservation News*, (57), 18–20. Retrieved from https://search.proquest.com/docview/1124539611?accountid=27464

The National Library of France (BnF) has the mission to collect, preserve and give access to all the published material in France. To this aim, the legal deposit has been extended to the different forms of publishing from the printed material in 1537, to electronic documents in 1992, as well as the Internet in 2006. To preserve all this digital cultural heritage, the BnF has designed a Scalable Preservation and Archiving Repository (SPAR). This central repository has to handle the diversity (media, formats, departments) by taking inspiration from good practices and standards. The key requirements of the system where: 1. OAIS compliance, 2. modularity and scalability, 3. abstraction, 4. use of well known formats and standards, 5. use of open-source technical building blocks.

Lee, J. (2017). Where should the culture of our lives and memory be preserved? - Rethinking the role of the library. In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity.*

Society. in Session 189 - Asia and Oceania. Wrocław: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/1691/

The abilities to store and transfer memory, to learn from others' experiences, or to share one's knowledge with the world are the drivers of social development. This driving force derives from the library's unique function and role to collect and service cultural assets. The National Library of Korea has recently expanded its scope of collection from printed media to online materials and broadcasting contents, and it opened its Memory Museum. The National Library of Korea has successfully demonstrated the example of a sustainable library in the new paradigm by strengthening its ability to preserve cultural memories. Meanwhile, public libraries in Korea have taken an initiative to preserve and transfer memory of a local community, which encourages locals' participation and revitalizes community spirit that has disappeared as a result of rapid economic growth. In this paper, cases of integrating a museum's archiving function into a library that led to social integration and community revitalization will be introduced; in addition, the paper will argue where and how the culture of our lives and memory should be preserved and utilized.

LEPORE, J. (2015). The COBWEB. *New Yorker*, *90*(45), 34–41. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article discusses efforts to archive historic Internet content, highlighting the Internet Archive nonprofit library in California. Topics addressed include the views and career history of Internet Archive founder Brewster Kahle, the legal aspects of the archive in terms of legal-deposit laws and copyright, and the Internet Archive's operations out of an old church building.

Lerner, A., Kohno, T., & Roesner, F. (2017). Rewriting History. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17* (pp. 1741–1755). New York, New York, USA: ACM Press. https://doi.org/10.1145/3133956.3134042

The Internet Archive's Wayback Machine is the largest modern web archive, preserving web content since 1996. We discover and analyze several vulnerabilities in how the Wayback Machine archives data, and then leverage these vulnerabilities to create what are to our knowledge the first attacks against a user's view of the archived web. Our vulnerabilities are enabled by the unique interaction between the Wayback Machine's archives, other websites, and a user's browser, and attackers do not need to compromise the archives in order to compromise users' views of a stored page. We demonstrate the effectiveness of our attacks through proof-of-concept implementations. Then, we conduct a measurement study to quantify the prevalence of vulnerabilities in the archive. Finally, we explore defenses which might be deployed by archives, website publishers, and the users of archives, and present the prototype of a defense for clients of the Wayback Machine, ArchiveWatcher.

Levy, D. C. (2017). *Using a Web-Archiving Service - How to ensure your cited web-references remain available and valid*. Australia, Australia/Oceania. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

In today's electronic information age, academic authors increasingly cite online resources such as blog posts, news articles, online policies and reports in their scholarly publications. Citing such webpages, or their URLs, poses long-term accessibility concern due to the ephemeral nature of the Internet: webpages can (and do!) change or disappear1 over time. When looking up cited web references, readers of scholarly publications might thus find content that is different from what author/s originally referenced; this is referred to as 'content drift'. Other times, readers are faced with a '404 Page Not Found' message, a phenomenon known as 'link rot'2. A recent Canadian study3 for example found a 23% link rot when examining 11,437 links in 664 doctoral dissertations from 2011-2015. Older publications are likely to face even higher rates of invalid links. Luckily, there are a few things you can do to make your cited web references more stable. The most common method is to use a web archiving service. Using a web archiving service means your web references and links are more likely to connect the reader to the content accessed at the time of writing/citing. In other words, references are less likely to "rot" or "drift" over time. As citing authors, we have limited influence on preserving web content that we don't own. We are generally at the mercy of the information custodians who tend to adjust, move or delete their web content to keep their site(s) current and interesting. All we can do to keep web content that we don't own but want to cite intact so that our readers can still access it in years to come is to create a "representative memento" of the online material as it was at the time of citing. This can be achieved by submitting the URL of the webpage we want to cite to a web archiving service which will generate a static ('cached') copy of it and allocate it a new, unique and permanent link, also called 'persistent identifier'. We can then use this new link to the archived webpage rather than the ephemeral link to the original webpage for our citation purposes. There are a range of web archives available. This guide contains a list of trusted web archiving services.

Lin, J. (2015). Scaling Down Distributed Infrastructure on Wimpy Machines for Personal Web Archiving. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1351–1355). New York, NY, USA: ACM. https://doi.org/10.1145/2740908.2741695

Warcbase is an open-source platform for storing, managing, and analyzing web archives using modern "big data" infrastructure on commodity clusters---specifically, HBase for storage and Hadoop for data analytics. This paper describes an effort to scale "down" Warcbase onto a Raspberry Pi, an inexpensive single-board computer about the size of a deck of playing cards. Apart from an interesting technology demonstration, such a design presents new opportunities for personal web archiving, in enabling a low-cost, low-power, portable device that is able to continuously capture a user's web browsing history---not only the URLs of the pages that a user has visited, but the contents of those pages---and allowing the user to revisit any previously-encountered page, as it appeared at that time. Experiments show that data ingestion throughput and temporal browsing latency are adequate with existing hardware, which means that such capabilities are already feasible today.

Lin, J., Gholami, M., & Rao, J. (2014). Infrastructure for Supporting Exploration and Discovery in Web Archives. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 851–856). New York, NY, USA: ACM. https://doi.org/10.1145/2567948.2579045

Web archiving initiatives around the world capture ephemeral web content to preserve our collective digital memory. However, unlocking the potential of web archives requires tools that support exploration and discovery of captured content. These tools need to be scalable

and responsive, and to this end we believe that modern "big data" infrastructure can provide a solid foundation. We present Warcbase, an open-source platform for managing web archives built on the distributed datastore HBase. Our system provides a flexible data model for storing and managing raw content as well as metadata and extracted knowledge. Tight integration with Hadoop provides powerful tools for analytics and data processing. Relying on HBase for storage infrastructure simplifies the development of scalable and responsive applications. We describe a service that provides temporal browsing and an interactive visualization based on topic models that allows users to explore archived content.

Lin, J., Milligan, I., Wiebe, J., & Zhou, A. (2017). Warcbase. *Journal on Computing and Cultural Heritage*, *10*(4), 1–30. https://doi.org/10.1145/3097570

Web archiving initiatives around the world capture ephemeral Web content to preserve our collective digital memory. However, unlocking the potential of Web archives for humanities scholars and social scientists requires a scalable analytics infrastructure to support exploration of captured content. We present Warcbase, an open-source Web archiving platform that aims to fill this need. Our platform takes advantage of modern open-source "big data" infrastructure, namely Hadoop, HBase, and Spark, that has been widely deployed in industry. Warcbase provides two main capabilities: support for temporal browsing and a domain-specific language that allows scholars to interrogate Web archives in several different ways. This work represents a collaboration between computer scientists and historians, where we have engaged in iterative codesign to build tools for scholars with no formal computer science training. To provide guidance, we propose a process model for scholarly interactions with Web archives that begins with a question and proceeds iteratively through four main steps: filter, analyze, aggregate, and visualize. We call this the FAAV cycle for short and illustrate with three prototypical case studies. This article presents the current state of the project and discusses future directions.

Lin, J., Tu, Z., Rose, M., & White, P. (2016). Prizm: A Wireless Access Point for Proxy-Based Web Lifelogging. In *Proceedings of the First Workshop on Lifelogging Tools and Applications* (pp. 19–25). New York, NY, USA: ACM. https://doi.org/10.1145/2983576.2983581

We present Prizm, a prototype lifelogging device that comprehensively records a user's web activity. Prizm is a wireless access point deployed on a Raspberry Pi that is designed to be a substitute for the user's normal wireless access point. Prizm proxies all HTTP(S) requests from devices connected to it and records all activity it observes. Although this particular design is not entirely novel, there are a few features that are unique to our approach, most notably the physical deployment as a wireless access point. Such a package allows capture of activity from multiple devices, integration with web archiving for preservation, and support for offline operation. This paper describes the design of Prizm, the current status of our project, and future plans.

Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., … Wrubel, L. (2018). API-based social media collecting as a form of web archiving. *International Journal on Digital Libraries*, *19*(1), 21–38. https://doi.org/http://dx.doi.org/10.1007/s00799-016-0201-7

Social media is increasingly a topic of study across a range of disciplines. Despite this popularity, current practices and open source tools for social media collecting do not

adequately support today's scholars or support building robust collections for future researchers. We are continuing to develop and improve Social Feed Manager (SFM), an open source application assisting scholars collecting data from Twitter's API for their research. Based on our experience with SFM to date and the viewpoints of archivists and researchers, we are reconsidering assumptions about API-based social media collecting and identifying requirements to guide the application's further development. We suggest that aligning social media collecting with web archiving practices and tools addresses many of the most pressing needs of current and future scholars conducting quality social media research. In this paper, we consider the basis for these new requirements, describe in depth an alignment between social media collecting and web archiving, outline a technical approach for effecting this alignment, and show how the technical approach has been implemented in SFM.

Lnenicka, M., Hovad, J., & Komarkova, J. (2015). A Proposal of a Big Web Data Application and Archive for the Distributed Data Processing with Apache Hadoop. In M. Núñez, N. T. Nguyen, D. Camacho, & B. Trawiński (Eds.), *Computational Collective Intelligence. Lecture Notes in Computer Science, vol 9330* (pp. 285–294). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24306-1_28

In recent years, research on big data, data storage and other topics that represent innovations in the analytics field has become very popular. This paper describes a proposal of a big web data application and archive for the distributed data processing with Apache Hadoop, including the framework with selected methods, which can be used with this platform. It proposes a workflow to create a web content mining application and a big data archive, which uses modern technologies like Python, PHP, JavaScript, MySQL and cloud services. It also shows the overview about the architecture, methods and data structures used in the context of web mining, distributed processing and big data analytics.

LOC Library Services Collection Development Office. (2017). Collecting Digital Content at the Library of Congress. *Digital Publishing Report, 5*(11), 2. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

In January 2017, the Library of Congress adopted a set of strategic steps related to its future acquisition of digital content. The purpose of this document is to provide background information and a high-level description of the strategy. The Library has been steadily increasing its digital collecting capacity and capability over the past two decades. This has come as the product of numerous independent efforts pointed to the same goal – acquire as much selected digital content as technically possible and make that content as broadly accessible to users as possible. In the past few years, much progress has been made, and an impressive amount of content has been acquired through several acquisitions methods. Further expansion of the Library's digital collecting program is seen as an essential part of the institution's strategic goal to: Acquire, preserve, and provide access to a universal collection of knowledge and the record of America's creativity. The scope of the newly-adopted strategy is limited to actions directly involved with acquisitions and collecting. It does not cover other related actions that are essential to a successful digital collections program. These primarily include the following. • Further development of the Library's technical infrastructure • Development of various access policies and procedures appropriate to different categories of digital content • Preservation of acquired digital content • Training and development of staff • Eventual realignment of resources to match an environment where a greater portion of the Library's collection building program focuses on digital materials The strategy also does not

cover digitization, which is the process by which the Library's physical collections materials (printed text, images, sound on tangible formats, etc.) are converted into digital formats that can be stored and accessed via a computer.

Locatelli, E. (2017). The role of Internet Wayback Machine in a multi-method research project. In *"Researchers, pratictioners and their use of the archived web", London, School of Advanced Study, University of London*. London. Retrieved from https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-BruggerLocatelliWeberNanni-Web25.pdf

If, on the one side, the web offers us a platform where content is searchable and replicable, on the other one, it cannot be forgotten that web content is perishable, unstable and subject to continuous change. This is a challenge for scholarly research about the historical development of web. The research here presented analyzed the historical development of weblogs in Italy investigating their technological, cultural, economic, and institutional dimensions. The approach chosen mixed participant observation, in-depth interviews, and semiotic analysis of blogs and blog posts. Since an important part of the research was about the development of platforms, graphics, layouts, and technology, beside interviews older versions of blogs were retrieved using Internet Wayback Machine. Even if partial versions of the blogs were archived, this part of the research was important to complete data obtained with interviews and blogs' analysis, since individual memory is not always accurate or some blogs were in the meanwhile closed and original posts were not accessible anymore.

Lollini, M. (2018). Hypertext and "Twitterature". *Profession*, 1. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article offer information on the Oregon Petrarch Open Book (OPOB), a database-driven hypertext version of the poetry collection "Rerum vulgarium fragmentata" (Rvf) by Francesco Petrarca. Topics discussed include the use of features of Web archive and hypertext for the creation of the database; the use of technology in teaching Petrarchism; and the archive of separate editions of Rvf in the database.

Lomborg, S. (2012). Researching Communicative Practice: Web Archiving in Qualitative Social Media Research. *Journal of Technology in Human Services*, *30*(3–4), 219–231. https://doi.org/http://dx.doi.org/10.1080/15228835.2012.744719

This article discusses the method of web archiving in qualitative social media research. While presenting a number of methodological challenges, social media archives (i.e., complete recordings of posts and comments on given social media) are also highly useful data corpuses for studying the social media users' communicative practices. Through a theoretical examination of web archiving as a new method enabled by the web itself, and an example-based discussion of the methodological, technical, and ethical challenges of harvesting social media archives, the article discusses the merits and limitations of using social media archives in empirical social media research. Adapted from the source document.

Lopes, R., Gomes, D., & Carriço, L. (2010). Web Not for All: A Large Scale Study of Web Accessibility. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (p. 10:1--10:4). New York, NY, USA: ACM. https://doi.org/10.1145/1805986.1806001

The Web accessibility discipline strives for the study and improvement of front-end Web design towards people with disabilities. Best practices such as WCAG dictate how Web pages should be created accordingly. On top of WCAG, several evaluation procedures enable the measurement of the quality level of a Web page. We leverage these procedures in an automated evaluation of a nearly 30 million Web page collection provided by the Portuguese Web Archive. Our study shows that there is high variability regarding the accessibility level of Web pages, and that few pages reach high accessibility levels. The obtained results show that there is a correlation between accessibility and complexity (i.e., number of HTML elements) of a Web page. We have also verified the effect of the interpretation of evaluation warnings towards the perception of accessibility.

Macnaught, B. (2018). Social Media Collecting at the National Library of New Zealand. In *IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies Session 93 - National Libraries and Social Media - Meeting the Challenges of Acquiring, Preserving and Proving Long-Term Access - National Libraries*. Kuala Lumpur: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/id/eprint/2274

Collecting content from the internet is an increasingly significant part of collection building at the National Library of New Zealand. Social media collecting is a new aspect of our digital collecting. We currently collect social media both under our legal deposit legislation and through donation as part of personal papers or archives. Social media offers unique content and voices, not always available in other formats. While this gives us new opportunities to diversify our collections, it isn't without challenges. Content is shifting away from traditional websites to social media. This is understandable – it's easier to post content, quick to circulate and cheaper. However, it is also comes with new collecting challenges.

Maemura, E., Becker, C., & Milligan, I. (2016). Understanding computational web archives research methods using research objects. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 3250–3259). IEEE. https://doi.org/10.1109/BigData.2016.7840982

Use of computational methods for exploration and analysis of web archives sources is emerging in new disciplines such as digital humanities. This raises urgent questions about how such research projects process web archival material using computational methods to construct their findings. This paper aims to enable web archives scholars to document their practices systematically to improve the transparency of their methods. We adopt the Research Object framework to characterize three case studies that use computational methods to analyze web archives within digital history research. We then discuss how the framework can support the characterization of research methods and serve as a basis for discussions of methods and issues such as reuse and provenance. The results suggest that the framework provides an effective conceptual perspective to describe and analyze the computational methods used in web archive research on a high level and make transparent the choices made in the process. The documentation of the research process contributes to a better understanding of the findings and their provenance, and the possible reuse of data, methods, and workflows.

Maharshi, S. (2017). *Performance Measurement and Analysis of Transactional Web Archiving*. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Web archiving is necessary to retain the history of the World Wide Web and to study its evolution. It is important for the cultural heritage community. Some organizations are legally obligated to capture and archive Web content. The advent of transactional Web archiving makes the archiving process more efficient, thereby aiding organizations to archive their Web content. This study measures and analyzes the performance of transactional Web archiving systems. To conduct a detailed analysis, we construct a meaningful design space defined by the system specifications that determine the performance of these systems. SiteStory, a state-of-the-art transactional Web archiving system, and local archiving, an alternative archiving technique, are used in this research. We experimentally evaluate the performance of these systems using the Greek version of Wikipedia deployed on dedicated hardware on a private network. Our benchmarking results show that the local archiving technique uses a Web server's resources more efficiently than SiteStory for one data point in our design space. Better performance than SiteStory in such scenarios makes our archiving solution favorable to use for transactional archiving. We also show that SiteStory does not impose any significant performance overhead on the Web server for the rest of the data points in our design space.

Mantratzis, C., & Orgun, M. (2004). Towards a Peer2Peer World-wide-web for the Broadband-enabled User Community. In *Proceedings of the 2004 ACM Workshop on Next-generation Residential Broadband Challenges* (pp. 42–49). New York, NY, USA: ACM. https://doi.org/10.1145/1026763.1026772

This paper aims to study the concept of a distributed World Wide Web archive that complements the existing WWW and "lives" across a vast Peer-to-Peer network of broadband-connected user nodes. It proposes the sharing of a web browser's cached data with other peers in an effort to provide an alternative resource to "discontinued" web documents with [normally] short life spans such as video and audio content as well as frequently restructured text pages. We have based this study on the success of existing file-sharing Peer-to-Peer networks and aim to extend their use further to facilitate content-oriented usage more appropriately while at the same time, addressing some of the major problems that arise from this.

Marill, J. L., Boyko, A., Ashenfelder, M., & Graham, L. (2004). Tools and techniques for harvesting the world wide web. In *Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries - JCDL '04* (p. 403). New York, New York, USA: ACM Press. https://doi.org/10.1145/996350.996469

Recently the Library of Congress began developing a strategy for the preservation of digital content. Efforts have focused on the need to select, harvest, describe, access and preserve Web resources. This poster focuses on the Library's initial investigation and evaluation of Web harvesting software tools.

Martínez-García, A., & Corti, L. (2012). Supporting student research with semantic technologies and digital archives. *Technology, Pedagogy and Education*, *21*(2), 273–288. https://doi.org/10.1080/1475939X.2012.704320

Martinez-Romo, J., & Araujo, L. (2009). Retrieving Broken Web Links Using an Approach Based on Contextual Information. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (pp. 351–352). New York, NY, USA: ACM. https://doi.org/10.1145/1557914.1557984

In this short note we present a recommendation system for automatic retrieval of broken Web links using an approach based on contextual information. We extract information from the context of a link such as the anchor text, the content of the page containing the link, and a combination of the cache page in some search engine and web archive, if it exists. Then the selected information is processed and submitted to a search engine. We propose an algorithm based on information retrieval techniques to select the most relevant information and to rank the candidate pages provided for the search engine, in order to help the user to find the best replacement. To test the different methods, we have also defined a methodology which does not require the user judgements, what increases the objectivity of the results.

Masanès, J. (2005). Web Archiving Methods and Approaches: A Comparative Study. *Library Trends*, *54*(1), 72–90. Retrieved from https://search.proquest.com/docview/220467286?accountid=27464

The Web is a virtually infinite information space, and archiving its entirety, all its aspects, is a utopia. The volume of information presents a challenge, but it is neither the only nor the most limiting factor given the continuous drop in storage device costs. Significant challenges lie in the management and technical issues of the location and collection of Web sites. As a consequence of this, archiving the Web is a task that no single institution can carry out alone. This article will present various approaches undertaken today by different institutions; it will discuss their focuses, strengths, and limits, as well as a model for appraisal and identifying potential complementary aspects amongst them. A comparison for discovery accuracy is presented between the snapshot approach done by the Internet Archive (IA) and the event-based collection done by the Bibliothèque Nationale de France (BNF) in 2002 for the presidential and parliamentary elections. The balanced conclusion of this comparison allows for identification of future direction for improvement of the former approach. [PUBLICATION ABSTRACT]

Massis, B. (2016). Libraries and digital memory. *New Library World*, *117*(9/10), 673–676. Retrieved from https://search.proquest.com/docview/1830312026?accountid=27464

Purpose The purpose of this column is to consider the role of libraries in an effort to preserve and protect a collective digital memory. Design/methodology/approach This paper addresses literature review and commentary on this topic that has been addressed by professionals, researchers and practitioners. Findings Libraries and library consortia will help go forward into the future and expand as trusted repositories where digital memory can be preserved and shared. Originality/value The value in exploring this topic is to examine the library environment for collection, storage and dissemination of digital information.

Mayagoitia, A., & González Aguilar, J. M. (2017). “Internet Archive”: la conservación de lo efímero TT - “Internet Archive”: the conservation of the ephemeral. *Documentación de las Ciencias de la Información*, *40*, 157–167. https://doi.org/http://dx.doi.org/10.5209/DCIN.57196

The ephemeral tends to be discarded, finding little room in traditional museums or archives. The emergence of digital archives and the acceptance of a sector in academia have helped to slowly modify the perception of ephemeral content. This article aims to analyze the evolution of the Internet Archive, a digital repository specialized in the compilation and conservation of ephemeral media. To conclude, a reflection is made about the future of digital preservation and the possibility of creating similar digital archives in Spanish-speaking countries.

Mayr, M., & Predikaka, A. (2016). Nationale Grenzen im World Wide Web – Erfahrungen bei der Webarchivierung in der Österreichischen Nationalbibliothek. *Bibliothek Forschung Und Praxis*, *40*(1), 90–95. https://doi.org/10.1515/bfp-2016-0007

Since 2009, the Austrian National Library performed four broad crawls, based on the Austrian Media Act, which focused primarily on the top level domain .at. The analysis of the crawls indicates that the aspect of national borders for the cultural heritage within the World Wide Web plays an important role for collection methods.

Mcclure, M. (2006). Archive-It 2: Internet Archive Strives to Ensure Preservation and Accessibility. *EContent*, *29*(8), 14–15. Retrieved from https://search.proquest.com/docview/213815870?accountid=27464

Preserving seemingly ephemeral Web content is a daunting task. The problem is even more difficult because the content of Web pages changes and the pages themselves come and go with great frequency, which means simply collecting URLs is not enough to keep tabs on valuable content. To help make digital content preservation possible, Internet Archive, a San Francisco-based nonprofit has led a charge to effectively capture and store Web content. The project recently released Archive-It 2 in its continued effort to archive the Web. Version 2 of Archive-It offers several new features not available in Version 1. Subscribers can now conduct test crawls, which enable them to see the type of Web material that would populate a specific collection before it is archived permanently. There is also a metadata search capability, which allows metadata to be included in the text searches of materials in a collection.

McCown, F., Diawara, N., & Nelson, M. L. (2007). Factors Affecting Website Reconstruction from the Web Infrastructure. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 39–48). New York, NY, USA: ACM. https://doi.org/10.1145/1255175.1255182

When a website is suddenly lost without a backup, it maybe reconstituted by probing web archives and search engine caches for missing content. In this paper we describe an experiment where we crawled and reconstructed 300 randomly selected websites on a weekly basis for 14 weeks. The reconstructions were performed using our web-repository crawler named Warrick which recovers missing resources from the Web Infrastructure (WI), the collective preservation effort of web archives and search engine caches. We examine several characteristics of the websites over time including birth rate, decay and age of resources. We evaluate the reconstructions when compared to the crawled sites and develop a statistical model for predicting reconstruction success from the WI. On average, we were able to recover 61% of each website's resources. We found that Google's PageRank, number of hops and resource age were the three most significant factors in determining if a resource would be recovered from the WI.

McCown, F., & Nelson, M. L. (2008). Usage Analysis of a Public Website Reconstruction Tool. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 371–374). New York, NY, USA: ACM. https://doi.org/10.1145/1378889.1378955

The Web is increasingly the medium by which information is published today, but due to its ephemeral nature, web pages and sometimes entire websites are often "lost" due to server crashes, viruses, hackers, run-ins with the law, bankruptcy and loss of interest. When a

website is lost and backups are unavailable, an individual or third party can use Warrick to recover the website from several search engine caches and web archives (the Web Infrastructure). In this short paper, we present Warrick usage data obtained from Brass, a queueing system for Warrick hosted at Old Dominion University and made available to the public for free. Over the last six months, 520 individuals have reconstructed more than 700 websites with 800K resources from the Web Infrastructure. Sixty-two percent of the static web pages were recovered, and 41% of all website resources were recovered. The Internet Archive was the largest contributor of recovered resources (78%).

McCown, F., & Nelson, M. L. (2008). Recovering a Website's Server Components from the Web Infrastructure. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 124–133). New York, NY, USA: ACM. https://doi.org/10.1145/1378889.1378911

Our previous research has shown that the collective behavior of search engine caches (e.g., Google, Yahoo, Live Search) and web archives (e.g., Internet Archive) results in the uncoordinated but large-scale refreshing and migrating of web resources. Interacting with these caches and archives, which we call the Web Infrastructure (WI), allows entire websites to be reconstructed in an approach we call lazy preservation. Unfortunately, the WI only captures the client-side view of a web resource. While this may be useful for recovering much of the content of a website, it is not helpful for restoring the scripts, web server configuration, databases, and other server-side components responsible for the construction of the website's resources. This paper proposes a novel technique for storing and recovering the server-side components of a website from the WI. Using erasure codes to embed the server-side components as HTML comments throughout the website, we can effectively reconstruct all the server components of a website when only a portion of the client-side resources have been extracted from the WI. We present the results of a preliminary study that baselines the lazy preservation of ten EPrints repositories and then examines the preservation of an EPrints repository that uses the erasure code technique to store the server-side EPrints software throughout the website. We found nearly 100% of the EPrints components were recoverable from the WI just two weeks after the repository came online, and it remained recoverable four months after it was "lost".

McCown, F., & Nelson, M. L. (2009). What Happens when Facebook is Gone? In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 251–254). New York, NY, USA: ACM. https://doi.org/10.1145/1555400.1555440

Web users are spending more of their time and creative energies within online social networking systems. While many of these networks allow users to export their personal data or expose themselves to third-party web archiving, some do not. Facebook, one of the most popular social networking websites, is one example of a "walled garden" where users' activities are trapped. We examine a variety of techniques for extracting users' activities from Facebook (and by extension, other social networking systems) for the personal archive and for the third-party archiver. Our framework could be applied to any walled garden where personal user data is being locked.

McDermott, I. E. (2016). Archives of the Americas, (Mostly) Free Online. *Online Searcher*, *40*(3), 27–29. Retrieved from https://search.proquest.com/docview/1818627659?accountid=27464

Established in 2008 to archive the transcribed texts of seminal documents in law, history, and diplomacy, the collection makes freely available important documents from ancient times, e.g., Agrarian Law, 111 BCE, right up to 2003, with "A Performance-Based Roadmap to a Permanent Two-State Solution to the Israeli-Palestinian Conflict." [...]visit the Digital Public Library of America (dp.la). According to Maura Marx, director of the DPLA Secretariat, "The DPLA's goal is to bring the entire nation's rich cultural collections off the shelves and into the innovative environment of the Internet for people to discover, download, remix, reuse and build on in ways we haven't yet begun to imagine" (cyber. law.harvard.edu/node/95550).

McKenna, L., Debruyne, C., & O'Sullivan, D. (2018). Understanding the Position of Information Professionals with regards to Linked Data. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18* (pp. 7–16). New York, New York, USA: ACM Press. https://doi.org/10.1145/3197026.3197041

The aim of this study was to explore the benefits and challenges to using Linked Data (LD) in Libraries, Archives and Museums (LAMs) as perceived by Information Professionals (IPs). The study also aimed to gain an insight into potential solutions for overcoming these challenges. Data was collected via a questionnaire which was completed by 185 Information Professionals (IPs) from a range of LAM institutions. Results indicated that IPs find the process of integrating and interlinking LD datasets particularly challenging, and that current LD tooling does not meet their needs. The study showed that LD tools designed with the workflows and expertise of IPs in mind could help overcome these challenges.

Méchoulan, É. (2011). Archiving in the Age of Digital Conversion: Notes for a Politics of "Remains." *Substance: A Review of Theory & Literary Criticism*, *40*(2), 92–104. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article focuses on archiving in the digital age. The author notes that caught in between materiality of the means of preservation and communication of documents and the relationships of power and of the institutions of the past is archiving. The archive is a form of social transmission, a process that transforms a text, image or sound into a document, an authorization to endure beyond ephemerality. This article was translated by Roxanne Lapidus.

Meimaris, M., Papastefanatos, G., Viglas, S., Stavrakas, Y., Pateritsas, C., & Anagnostopoulos, I. (2016). A query language for multi-version data web archives. *Expert Systems*, *33*(4), 383–404. https://doi.org/10.1111/exsy.12157

The Data Web refers to the vast and rapidly increasing quantity of scientific, corporate, government and crowd-sourced data published in the form of Linked Open Data, which encourages the uniform representation of heterogeneous data items on the web and the creation of links between them. The growing availability of open linked datasets has brought forth significant new challenges regarding their proper preservation and the management of evolving information within them. In this paper, we focus on the evolution and preservation challenges related to publishing and preserving evolving linked data across time. We discuss the main problems regarding their proper modelling and querying and provide a conceptual model and a query language for modelling and retrieving evolving data along with changes affecting them. We present in details the syntax of the query language and demonstrate its

functionality over a real-world use case of evolving linked dataset from the biological domain. [ABSTRACT FROM AUTHOR]

Milligan, I. (2016). Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities*, *10*(1), 78–94. Retrieved from http://10.0.13.38/ijhac.2016.0161

Contemporary and future historians need to grapple with and confront the challenges posed by web archives. These large collections of material, accessed either through the Internet Archive's Wayback Machine or through other computational methods, represent both a challenge and an opportunity to historians. Through these collections, we have the potential to access the voices of millions of non-elite individuals (recognizing of course the cleavages in both Web access as well as method of access). To put this in perspective, the Old Bailey Online currently describes its monumental holdings of 197,745 trials between 1674 and 1913 as the "largest body of texts detailing the lives of non-elite people ever published." GeoCities.com, a platform for everyday web publishing in the mid-to-late 1990s and early 2000s, amounted to over thirty-eight million individual webpages. Historians will have access, in some form, to millions of pages: written by everyday people of various classes, genders, ethnicities, and ages. While the Web was not a perfect democracy by any means - it was and is unevenly accessed across each of those categories - this still represents a massive collection of non-elite speech. Yet a figure like thirty-eight million webpages is both a blessing and a curse. We cannot read every website, and must instead rely upon discovery tools to find the information that we need. Yet these tools largely do not exist for web archives, or are in a very early state of development: what will they look like? What information do historians want to access? We cannot simply map over web tools optimized for discovering current information through online searches or metadata analysis. We need to find information that mattered at the time, to diverse and very large communities. Furthermore, web pages cannot be viewed in isolation, outside of the networks that they inhabited. In theory, amongst corpuses of millions of pages, researchers can find whatever they want to confirm. The trick is situating it into a larger social and cultural context: is it representative? Unique? In this paper, "Lost in the Infinite Archive," I explore what the future of digital methods for historians will be when they need to explore web archives. Historical research of periods beginning in the mid-1990s will need to use web archives, and right now we are not ready. This article draws on first-hand research with the Internet Archive and Archive-It web archiving teams. It draws upon three exh…

Milligan, I., Ruest, N., & Lin, J. (2016). Content Selection and Curation for Web Archiving. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (pp. 107–110). New York, New York, USA: ACM Press. https://doi.org/10.1145/2910896.2910913

Moulahi, B., Tamine, L., & Yahia, S. Ben. (2016). When time meets information retrieval: Past proposals, current plans and future trends. *Journal of Information Science*, *42*(6), 725. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

With the advent of Web search and the large amount of data published on the Web sphere, a tremendous amount of documents become strongly time-dependent. In this respect, the time dimension has been extensively exploited as a highly important relevance criterion to improve

the retrieval effectiveness of document ranking models. Thus, a compelling research interest is going on the temporal information retrieval realm, which gives rise to several temporal search applications. In this article, we intend to provide a scrutinizing overview of time-aware information retrieval models. We specifically put the focus on the use of timeliness and its impact on the global value of relevance as well as on the retrieval effectiveness. First, we attempt to motivate the importance of temporal signals, whenever combined with other relevance features, in accounting for document relevance. Then, we review the relevant studies standing at the crossroads of both information retrieval and time according to three common information retrieval aspects: the query level, the document content level and the document ranking model level. We organize the related temporal-based approaches around specific information retrieval tasks and regarding the task at hand, we emphasize the importance of results presentation and particularly timelines to the end user. We also report a set of relevant research trends and avenues that can be explored in the future. [ABSTRACT FROM AUTHOR]

Murray, G. P. (2016). Featured Web Resource: Theological Commons. *Theological Librarianship*, *9*(2), 1. Retrieved from https://search.proquest.com/docview/1842842888?accountid=27464

In late 2010, Dr Iain Torrance, at that time the President of Princeton Theological Seminary, asked a small subset of library staff to consider how to improve discoverability and access to the thousands of volumes on theology and religion that Princeton Seminary and other institutions had digitized through the Internet Archive, to facilitate research by students, scholars, and pastors both locally and globally. However, because the goal was to provide access to relevant resources, not to showcase Princeton's digital content, the digital library team subsequently took a detailed list of Library of Congress subject headings provided by Don Vorp, at that time Collection Development Librarian at Princeton Seminary, and performed searches in the Internet Archive system for digitized books with those subjects, irrespective of library of origin. Those items were then harvested in the same manner. This procedure soon amassed tens of thousands of digital texts, and in March 2012, the Theological Commons was publicly released as a free, web-accessible digital library.

Musiani, F., & Schafer, V. (2017). Digital Heritage and Heritagization ; Patrimoine et patrimonialisation numériques. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Introduction to a special issueThe six articles and the introduction composing this issue fully situate themselves within the interdisciplinary dimension of digital heritage analyses, including perspectives from history, information and communication sciences, sociology of innovation, digital humanities or juridical sciences.

Musso, M., & Merletti, F. (2016). This is the future: A reconstruction of the UK business web space (1996–2001). *New Media & Society*, *18*(7), 1120–1142. https://doi.org/10.1177/1461444816643791

The Internet and the World Wide Web in particular have dramatically changed the way in which many companies operate. On the Web, even the smallest and most localised business has a potential global reach, and the development of online payment has redefined the selling market in most sectors. Boundaries and borders are being radically rediscussed. This article

reconstructs the early approach of UK businesses to the World Wide Web between 1996 and 2001, a period in which the Web started to spread but it was not as engrained in everyday life as it would be in the following decade. While the fast and dispersed nature of the Web makes it almost impossible to accurately reconstruct the Web sphere in its historical dimension, this article proposes a methodology based on the usage of historical Web directories to access and map past Web spheres.

Nanni, F., Ponzetto, S. P., & Dietz, L. (2018). Toward comprehensive event collections. *International Journal on Digital Libraries*. https://doi.org/10.1007/s00799-018-0246-x

Web archives, such as the Internet Archive, preserve an unprecedented abundance of materials regarding major events and transformations in our society. In this paper, we present an approach for building event-centric sub-collections from such large archives, which includes not only the core documents related to the event itself but, even more importantly, documents describing related aspects (e.g., premises and consequences). This is achieved by identifying relevant concepts and entities from a knowledge base, and then detecting their mentions in documents, which are interpreted as indicators for relevance. We extensively evaluate our system on two diachronic corpora, the New York Times Corpus and the US Congressional Record; additionally, we test its performance on the TREC KBA Stream Corpus and on the TREC-CAR dataset, two publicly available large-scale web collections.

Nanni, F., Ponzetto, S. P., & Dietz, L. (2017). Building Entity-centric Event Collections. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 199–208). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=3200334.3200356

Web archives preserve an unprecedented abundance of materials regarding major events and transformations in our society. In this paper, we present an approach for building event-centric sub-collections from such large archives, which includes not only the core documents related to the event itself but, even more importantly, documents describing related aspects (e.g., premises and consequences). This is achieved by 1) identifying relevant concepts and entities from a knowledge base, and 2) detecting their mentions in documents, which are interpreted as indicators for relevance. We extensively evaluate our system on two diachronic corpora, the New York Times Corpus and the US Congressional Record, and we test its performance on the TREC KBA Stream corpus, a large and publicly available web archive.

Németh, M. (2017). Nemzetközi körkép a webarchiválás gyakorlatáról. *Könyvtári Figyelő*, *63*(4), 575–582. Retrieved from http://epa.oszk.hu/00100/00143/00349/pdf/EPA00143_konyvtari_figyelo_2017_04_575-582.pdf

A webarchiválás olyan dinamikusan fejlődő terület, mely számos vonatkozásban már a korábbiakban is felbukkant a Könyvtári Figyelő hasábjain, különösen a nemzetközi szakirodalom szemlézése kapcsán. (Például 2014-ben Hegyközi Ilona tekintette át a webarchiválással kapcsolatos nemzetközi trendeket.) Úgy éreztük, eljött az ideje egy újabb összegzésnek. Ennek különös hangsúlyt ad, hogy számos korábbi kezdeményezést követően, idén tavasztól megteremtődtek az alapjai az OSZK fejlesztési projektjén belül egy olyan kísérleti projekt elindításának, melyben felmérjük a webarchiváláshoz szükséges hardver és szoftver igényeket, valamint szakmai ismereteket. A fő cél, hogy jól megalapozott módon integrálni tudjuk e területet hosszú távon is az OSZK szolgáltatási tevékenységei közé. Az

OSZK Elektronikus Könyvtári Szolgáltatások Osztályán létrehoztunk egy Magyar Internet Archívum honlapot (http://mekosztaly.oszk.hu/mia), melyen tanulmányozhatók a webarchiválás különféle módszerei, alapfogalmai, meg a nemzetközi szakirodalom. Továbbá a projekttel kapcsolatos aktuális információkkal is szolgálunk és fel lehet iratkozni a webarchiválás szakmai kérdéseit tárgyaló levelezőlistára is. Ennek a cikknek nem az a célja tehát, hogy a webarchiválási tevékenységek szakmai alapjait járja körül (amelyre a honlapot böngészve nyílik lehetőség), hanem, hogy áttekintést adjunk a webarchiválási szolgáltatásokat megalapozó nemzetközi jó gyakorlatokból.

Németh, M. (2018). A webarchiválásról történeti megközelítésben. *Könyv, Könyvtár, Könyvtáros*, *27*(2), 48–52. Retrieved from http://ki2.oszk.hu/3k/2018/06/a-webarchivalasrol-torteneti-megkozelitesben/

A tanulmánykötet esettanulmányok formájában az elsők között tesz kísérletet arra, hogy felvillantsa a webarchiválás történeti, illetve széles társadalomtudományi kontextusának számos fontos elemét. A szerkesztők előszava is kitér rá, hogy eddig inkább az volt a jellemző, hogy magáról a webarchiválási folyamatról, annak technikai részleteiről, a világháló archiválásához kapcsolódó kurátori tevékenységekről szóltak az összefoglalók. A szerkesztők az előszóban ez alkalommal is a legfrissebb szakirodalom segítségével vázolják fel a webarchiválás általánosabb kontextusát, eddigi történetének kronológiáját és főbb szereplőit. Jellemzik a főbb intézményeket, melyek e tevékenységeket végzik. Az Internet Archive úttörő szerepe mellett rámutatnak arra, hogy míg egyes országokban egyetlen vezető intézmény köré csoportosul e tevékenység (például Dániában), addig máshol intézményi koordináció tapasztalható világosan elkülönülő szerepkörökkel (pl. Franciaország, Nagy-Britannia). Rövid tájékoztatást kapunk arról, hogy milyen szoftverháttérrel történik az anyagok begyűjtése, s milyen módszerekkel lehet az eltárolt webes információk visszakeresését biztosítani (pl. az Internet Archive által fejlesztett Wayback Machine szoftverrel URL címekre kereshetünk, ezt egészíti ki a teljesszövegű index szolgáltatás, már amelyik gyűjteményben éppen elérhető).

Németh, M. (2017). 404 Not Found - Ki őrzi meg az internetet; Webarchiválás workshop az Országos Széchényi Könyvtárban. *Tudományos És Műszaki Tájékoztatás*, *64*(11), 577–582. Retrieved from http://epa.oszk.hu/03000/03071/00112/pdf/EPA03071_tmt_2017_11_577-582.pdf

2017. október 13-án első alkalommal került sor kifejezetten a számítógépes világháló archiválásával foglalkozó rendezvényre az Országos Széchényi Könyvtárban (OSZK). Az intézmény és a Kormányzati Informatikai Ügynökség (KIFÜ) keretei között zajló Országos Könyvtári Rendszer (OKR) projekt egyik munkacsoportjaként idén tavasztól kezdhettünk el egy kísérleti projekt keretében foglalkozni a webarchiválással. (Bővebb információt a http://mekosztaly.oszk.hu/mia oldalon lehet találni erről.) Célunk az, hogy a projektidőszak végére egy olyan koncepcióval álljunk elő, mely lehetővé teszi, számos európai nemzeti könyvtár mintájára, az üzemszerű munkafolyamatként zajló webarchiválási tevékenység ellátását, illetve szervezését az OSZK részéről. Egy olyan rendszert kívánunk létrehozni, amely a kulturális örökség hosszú távú megőrzésének feladata mellett képes kiszolgálni az oktatás, a tudományos kutatás, az állami szervek, az üzleti szféra és az egyes internethasználók igényeit is. Az archívum megvalósulásával a most csak jelen időben létező magyar internetnek „múltja” is lenne, és olyan lehetőségek nyílnak meg a mai és a jövőbeli felhasználói számára, amelyek jelenleg nem, vagy csak nehézkesen valósíthatók meg (pl. megszűnt weboldalak megtalálása, webhelyek időbeli változásának elemzése és vizualizálása, stabil hivatkozhatóság, idődimenziót is tartalmazó szöveg- és adatbányászati alkalmazások

futtatása, internettörténeti kutatások, hiteles másolatok szolgáltatása). A projekt első fél évét mintegy lezárva került sor rendezvényünkre. A program összeállításakor különös gondot fordítottunk a meglévő külföldi szakmai tapasztalatok, illetve az itthoni előzmények bemutatására. A workshop hangsúlyos céljaként szerepelt továbbá a teljes közgyűjteményi szféra (

Nemeth, M., & Drotos, L. (2017). Hungarian web archiving pilot project in the National Széchényi Library. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 000209–000212). IEEE. https://doi.org/10.1109/CogInfoCom.2017.8268244

This demo paper introduces the web archiving pilot project in the Hungarian National Széchényi Library. Basic conception and goals are being described.

Nguyen, T. N., Kanhabua, N., Nejdl, W., & Niederée, C. (2015). Mining Relevant Time for Query Subtopics in Web Archives. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1357–1362). New York, NY, USA: ACM. https://doi.org/10.1145/2740908.2741702

With the reflection of nearly all types of social cultural, societal and everyday processes of our lives in the web, web archives from organizations such as the Internet Archive have the potential of becoming huge gold-mines for temporal content analytics of many kinds (e.g., on politics, social issues, economics or media). First hand evidences for such processes are of great benefit for expert users such as journalists, economists, historians, etc. However, searching in this unique longitudinal collection of huge redundancy (pages of near-identical content are crawled all over again) is completely different from searching over the web. In this work, we present our first study of mining the temporal dynamics of subtopics by leveraging the value of anchor text along the time dimension of the enormous web archives. This task is especially useful for one important ranking problem in the web archive context, the time-aware search result diversification. Due to the time uncertainty (the lagging nature and unpredicted behavior of the crawlers), identifying the trending periods for such temporal subtopics relying solely on the timestamp annotations of the web archive (i.e., crawling times) is extremely difficult. We introduce a brute-force approach to detect a time-reliable sub-collection and propose a method to leverage them for relevant time mining of subtopics. This is empirically found effective in solving the problem.

Nguyen, T. N., Kanhabua, N., Niederée, C., & Zhu, X. (2015). A Time-aware Random Walk Model for Finding Important Documents in Web Archives. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15* (pp. 915–918). New York, New York, USA: ACM Press. https://doi.org/10.1145/2766462.2767832

Due to their first-hand, diverse and evolution-aware reflection of nearly all areas of life, web archives are emerging as gold-mines for content analytics of many sorts. However, supporting search, which goes beyond navigational search via URLs, is a very challenging task in these unique structures with huge, redundant and noisy temporal content. In this paper, we address the search needs of expert users such as journalists, economists or historians for discovering a topic in time: Given a query, the top-k returned results should give the best representative documents that cover most interesting time-periods for the topic. For this purpose, we propose a novel random walk-based model that integrates relevance, temporal authority, diversity and

time in a unified framework. Our preliminary experimental results on the large-scale real-world web archival collection shows that our method significantly improves the state-of-the-art algorithms (i.e., PageRank) in ranking temporal web pages.

Nielsen, J. (2016). *Using web archives in research – an introduction* (1.). Aarhus: NetLab. Retrieved from http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf

This book has been written in connection with the development of NetLab's workshops on web archiving for researchers. These workshops provide the participants with an introduction to working with archived web materials in research, including a description of what web archiving is, the challenges of using archived web materials as an object of research, knowledge of existing web archives, and tools for micro archiving, so that researchers can themselves archive web materials. The purpose of this book is to gather and make available knowledge about the use of web archives for research. It is written in a Danish context and adapted to the needs of Danish researchers but can also be useful for other researchers. The

Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, *18*(3–4). https://doi.org/10.1045/march2012-niu1

This overview is a study of the methods used at a variety of universities, and international government libraries and archives, to select, acquire, describe and access web resources for their archives. Creating a web archive presents many challenges, and library and information schools should ensure that instruction in web archiving methods and skills is made part of their curricula, to help future practitioners meet those challenges. In preparation for developing a web archiving course, the author conducted a comprehensive literature review. The findings are reported in this paper, along with the author's views on some of the methods in use, such as how traditional archive management concepts and theories can be applied to the organization and description of archived web resources. Adapted from the source document.

Niu, J. (2012). Functionalities of Web Archives. *D-Lib Magazine*, *18*(3–4). https://doi.org/10.1045/march2012-niu2

The functionalities that are important to the users of web archives range from basic searching and browsing to advanced personalized and customized services, data mining, and website reconstruction. The author examined ten of the most established English language web archives to determine which functionalities each of the archives supported, and how they compared. A functionality checklist was designed, based on use cases created by the International Internet Preservation Consortium (IIPC), and the findings of two related user studies. The functionality review was conducted, along with a comprehensive literature review of web archiving methods, in preparation for the development of a web archiving course for Library and Information School students. This paper describes the functionalities used in the checklist, the extent to which those functionalities are implemented by the various archives, and discusses the author's findings. Adapted from the source document.

Notess, G. R. (2018). Search Engine Update. *Online Searcher*, *42*(1), 8–9. Retrieved from https://search.proquest.com/docview/1989831908?accountid=27464

Searchers can change the region in the settings, by adding ?gl=TLD (replace TLD with the country top level domain) to a Google search results URL, or use a VPN to instead mimic being in the other country.The prefix commands of info: and id: followed by a URL no longer display links to the cache copy, related pages, incoming links (not surprising since the link search capability in Google was abandoned previously), site search, and term matches.The partnership aims to increase the number of trained fact checkers, expand fact-checking capabilities to more countries, provide access to various fact-checking tools for free, and develop new fact-checking software tools to improve efficiency.

Nwala, A. C., Weigle, M. C., & Nelson, M. L. (2018). Bootstrapping Web Archive Collections from Social Media. In *Proceedings of the 29th on Hypertext and Social Media - HT '18* (pp. 64–72). New York, New York, USA: ACM Press. https://doi.org/10.1145/3209542.3209560

Human-generated collections of archived web pages are expensive to create, but provide a critical source of information for researchers studying historical events. Hand-selected collections of web pages about events shared by users on social media offer the opportunity for bootstrapping archived collections. We investigated if collections generated automatically and semi-automatically from social media sources such as Storify, Reddit, Twitter, and Wikipedia are similar to Archive-It human-generated collections. This is a challenging task because it requires comparing collections that may cater to different needs. It is also challenging to compare collections since there are many possible measures to use as a baseline for collection comparison: how does one narrow down this list to metrics that reflect if two collections are similar or dissimilar? We identified social media sources that may provide similar collections to Archive-It human-generated collections in two main steps. First, we explored the state of the art in collection comparison and defined a suite of seven measures (Collection Characterizing Suite - CCS) to describe the individual collections. Second, we calculated the distances between the CCS vectors of Archive-It collections and the CCS vectors of collections generated automatically and semi-automatically from social media sources, to identify social media collections most similar to Archive-It collections. The CCS distance comparison was done for three topics: "Ebola Virus," "Hurricane Harvey," and "2016 Pulse Nightclub Shooting." Our results showed that social media sources such as Reddit, Storify, Twitter, and Wikipedia produce collections that are similar to Archive-It collections. Consequently, curators may consider extracting URIs from these sources in order to begin or augment collections about various news topics.

Nwala, A. C., Weigle, M. C., & Nelson, M. L. (2018). Scraping SERPs for Archival Seeds: It Matters When You Start. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 263–272). New York, NY, USA: ACM. https://doi.org/10.1145/3197026.3197056

Event-based collections are often started with a web search, but the search results you find on Day 1 may not be the same as those you find on Day 7. In this paper, we consider collections that originate from extracting URIs (Uniform Resource Identifiers) from Search Engine Result Pages (SERPs). Specifically, we seek to provide insight about the retrievability of URIs of news stories found on Google, and to answer two main questions: first, can one "refind" the same URI of a news story (for the same query) from Google after a given time? Second, what is the probability of finding a story on Google over a given period of time? To answer these questions, we issued seven queries to Google every day for over seven months (2017-05-25 to 2018-01-12) and collected links from the first five SERPs to generate seven

collections for each query. The queries represent public interest stories: “healthcare bill,” “manchester bombing,” “london terrorism,” “trump russia,” “travel ban,” “hurricane harvey,” and “hurricane irma.” We tracked each URI in all collections over time to estimate the discoverability of URIs from the first five SERPs. Our results showed that the daily average rate at which stories were replaced on the default Google SERP ranged from 0.21 - 0.54, and a weekly rate of 0.39 - 0.79, suggesting the fast replacement of older stories by newer stories. The probability of finding the same URI of a news story after one day from the initial appearance on the SERP ranged from 0.34 - 0.44. After a week, the probability of finding the same news stories diminishes rapidly to 0.01 - 0.11. In addition to the reporting of these probabilities, we also provide two predictive models for estimating the probability of finding the URI of an arbitrary news story on SERPs as a function of time. The web archiving community considers link rot and content drift important reasons for collection building. Similarly, our findings suggest that due to the difficulty in retrieving the URIs of news stories from Google, collection building that originates from search engines should begin as soon as possible in order to capture the first stages of events, and should persist in order to capture the evolution of the events, because it becomes more difficult to find the same news stories with the same queries on Google, as time progresses.

Nwala, A. C., Weigle, M. C., Nelson, M. L., Ziegler, A. B., & Aizman, A. (2017). Local Memory Project: Providing Tools to Build Collections of Stories for Local Events from Local Sources. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 219–228). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=3200334.3200358

The national (non-local) news media has different priorities than the local news media. If one seeks to build a collection of stories about local events, the national news media may be insufficient, with the exception of local news which “bubbles” up to the national news media. If we rely exclusively on national media, or build collections exclusively on their reports, we could be late to the important milestones which precipitate major local events, thus, run the risk of losing important stories due to link rot and content drift. Consequently, it is important to consult local sources affected by local events. Our goal is to provide a suite of tools (beginning with two) under the umbrella of the Local Memory Project (LMP) to help users and small communities discover, collect, build, archive, and share collections of stories for important local events by leveraging local news sources. The first service (Geo) returns a list of local news sources (newspaper, TV and radio stations) in order of proximity to a user-supplied zip code. The second service (Local Stories Collection Generator) discovers, collects and archives a collection of news stories about a story or event represented by a user-supplied query and zip code pair. We evaluated 20 pairs of collections, Local (generated by our system) and non-Local, by measuring archival coverage, tweet index rate, temporal range, precision, and sub-collection overlap. Our experimental results showed Local and non-Local collections with archive rates of 0.63 and 0.83, respectively, and tweet index rates of 0.59 and 0.80, respectively. Local collections produced older stories than non-Local collections, at a higher precision (relevance) of 0.84 compared to a non-Local precision of 0.72. These results indicate that Local collections are less exposed, thus less popular than their nonLocal counterpart.

Nyvang, C., Kromann Hvid, T., & Zierau, E. (2017). Capturing the Web at Large A Critique of Current Web Referencing Practices. In *“Researchers, pratictioners and their use of the archived web”, London, School of Advanced Study, University of London*. Retrieved

from https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-NyvangKromannZierau-Capturing_the_web_at_large.pdf

The Internet and the cultural phenomena that exist online are increasingly attracting academic awareness, and e-materials both supplement and replace physical materials. These new opportunities come with a range of challenges. Websites are connected in new and unfamiliar ways, the amount of data easily surpasses what we have experienced previously, and we do not yet have an infrastructure that can lend prober support to the increased scholarly use of web resources [1-2]. This paper is an attempt to grapple with one of the core challenges, namely our ability to provide precise and persistent references to web material.1 The paper charts prevailing ideals and practices regarding web references within the Humanities. We highlight the challenges based on an analysis of web references in two case studies – a selection of Danish master's theses from 2015 and academic books on contemporary Danish literature. We propose a new best practice that is consistent with good scientific practice in terms of both precision and persistency, which cannot be obtained following the existing standards.

OCLC. (2018). Descriptive metadata for web archiving: Review of harvesting tools. United States, North America: OCLC Research. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

OCLC Research Library Partnership Web Archiving Working Group, Tools Subgroup's objective analysis of 11 tools designed to extract descriptive metadata from harvested web content. Selected tools included those tools that harvest or replay web content, are actively under development and/or are actively supported, and appeared to include descriptive metadata capture features. Tools reviewed include: Archive-It, Heritrix, HTTrack, Memento, Netarchive Suite, SiteStory, Social Feed Manager, Wayback Machine, Web Archive.

OCLC. (2018). Descriptive metadata for web archiving: Literature review of user needs. United States, North America: OCLC Research. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Under the auspices of the OCLC Research Library Partnership Web Archiving Metadata Working Group, this document is a literature review to inform the development of descriptive metadata best practices for archived web content that would meet end-user needs, enhance discovery, and improve metadata consistency. Selected readings include -- at minimum -- a substantive section related to metadata, but most covered a wider swath of issues. This helped the Working Group to learn much else about who the users of web archives are, the strategies they use and the challenges they face.

Ogden, J., Halford, S., & Carr, L. (2017). Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 299–308). New York, NY, USA: ACM. https://doi.org/10.1145/3091478.3091506

This paper makes the case for studying the work of web archivists, in an effort to explore the ways in which practitioners shape the preservation and maintenance of the archived Web in its various forms. An ethnographic approach is taken through the use of observation,

interviews and documentary sources over the course of several weeks in collaboration with web archivists, engineers and managers at the Internet Archive - a private, non-profit digital library that has been archiving the Web since 1996. The concept of web archival labour is proposed to encompass and highlight the ways in which web archivists (as both networked human and non-human agents) shape and maintain the preserved Web through work that is often embedded in and obscured by the complex technical arrangements of collection and access. As a result, this engagement positions web archives as places of knowledge and cultural production in their own right, revealing new insights into the performative nature of web archiving that have implications for how these data are used and understood.1

Oh, H.-J., Dong-Hyun, W., Kim, C., Park, S.-H., & Kim, Y. (2018). Design and implementation of crawling algorithm to collect deep web information for web archiving. *Data Technologies and Applications*, *52*(2), 266–277. https://doi.org/http://dx.doi.org/10.1108/DTA-07-2017-0053

PurposeThe purpose of this paper is to describe the development of an algorithm for realizing web crawlers that automatically collect dynamically generated webpages from the deep web.Design/methodology/approachThis study proposes and develops an algorithm to collect web information as if the web crawler gathers static webpages by managing script commands as links. The proposed web crawler actually experiments with the algorithm by collecting deep webpages.FindingsAmong the findings of this study is that if the actual crawling process provides search results as script pages, the outcome only collects the first page. However, the proposed algorithm can collect deep webpages in this case.Research limitations/implicationsTo use a script as a link, a human must first analyze the web document. This study uses the web browser object provided by Microsoft Visual Studio as a script launcher, so it cannot collect deep webpages if the web browser object cannot launch the script, or if the web document contains script errors.Practical implicationsThe research results show deep webs are estimated to have 450 to 550 times more information than surface webpages, and it is difficult to collect web documents. However, this algorithm helps to enable deep web collection through script runs.Originality/valueThis study presents a new method to be utilized with script links instead of adopting previous keywords. The proposed algorithm is available as an ordinary URL. From the conducted experiment, analysis of scripts on individual websites is needed to employ them as links.

O'Leary, M. (2003). Internet Archive joins history's great libraries. *Information Today*, *20*(10), 41. Retrieved from https://search.proquest.com/docview/214817883?accountid=27464

Brewster Kahle is a man of many roles: a famous Internet pioneer, a successful dot-com entrepreneur, a digital visionary, and a darned good librarian. Right now, he's best-known as the founder of Alexa and the WAIS system. However, with Kahle's creation of the Internet Archive (IA), the future may well ascribe greater importance to his work as a librarian. IA is the largest archival project in history. Kahle compares it - without presumption or exaggeration - to the ancient Library of Alexandria. It intends to do for the Internet what that great library did for antiquity: to capture and preserve the world's knowledge for everyone's benefit. IA has been hard at work for several years creating the largest database in the world. At first, it concentrated on preservation. Now, with that task well in hand, it's working on access tools for this unique information resource.

Olston, C., & Najork, M. (2010). Web Crawling. *Foundations and Trends® in Information Retrieval*, *4*(3), 175–246. https://doi.org/10.1561/1500000017

Opalinski, A., Nawarecki, E., & Kluska-Nawarecka, S. (2015). Agent-based Approach to WEB Exploration Process. *Procedia Computer Science*, *51*(International Conference On Computational Science, ICCS 2015), 1052–1061. https://doi.org/10.1016/j.procs.2015.05.263

The paper contains the concept of agent-based search system and monitoring of Web pages. It is oriented at the exploration of limited problem area, covering a given sector of industry or economy. The proposal of agent-based (modular) structure of the system is due to the desire to ease the introduction of modifications or enrichment of its functionality. Commonly used search engines do not offer such a feature. The second part of the article presents a pilot version of the WEB mining system, represent- ing a simplified implementation of the previously presented concept. Testing of the implemented application was executed by referring to the problem area of foundry industry.

Oury, C., Blumenthal, K.-R., & Peyrard, S. (2016). Digital Preservation Metadata Practice for Web Archives. In *Digital Preservation Metadata for Practitioners* (pp. 59–82). Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Oury, C., & Poll, R. (2013). Counting the uncountable: statistics for web archives. *Performance Measurement and Metrics*, *14*(2), 132–141. https://doi.org/http://dx.doi.org/10.1108/PMM-05-2013-0014

Purpose - The purpose of this paper is to describe the aims and contents of the ISO Report ISO/TR 14873. Design/methodology/approach - For more than a decade, libraries have started to “collect the web”. National libraries in particular select, collect and store publications and websites from their national domain, seeing this as a task similar to traditional legal deposit. The collection policies and collecting methods vary, so that it is difficult to compare the quantity and quality of the respective web archives. Findings - In order to harmonize the evaluation of web archives, ISO TC 46 SC 8 has produced a Technical Report that standardizes the terminology and statistics and offers tested indicators for assessing the quality of web archiving. Originality/value - This paper describes the shortly to be published ISO/TR 14873, a potentially vital guide to harmonize web archive collection internationally.

Oury, C., Steinke, T., & Jones, G. (2012). Ensuring Long-Term Access to the Memory of the Web Preservation Working Group of the International Internet Preservation Consortium. *International Preservation News*, (58), 34–37. Retrieved from https://search.proquest.com/docview/1272325401?accountid=27464

Archiving the Web is the process through which documents and objects on the World Wide Web are captured and stored. There are and have been a number of ways through which this has been accomplished, but the end result is archived Web content (Web site, page, or part of a Website) that is preserved for future researchers, historians and the general public. Preservation involves maintaining the ability to present meaningful access to information over time. In the context of Web archives, the intention of preservation is to retain access to archived Web resources, so they can continue to be used and understood despite changes in

access technologies or without unacceptable loss of integrity or meaning. The International Internet Preservation Consortium, chartered in 2003, is made up of institutions with basically similar goals of preserving Web content for heritage purposes and which generally share the same harvesting and access tools.

Oyelude, A. A. (2016). What's trending in libraries from the internet cybersphere - bookless libraries - 02 - 2016. *Library Hi Tech News*, *33*(6), 19–20. Retrieved from https://search.proquest.com/docview/1823127353?accountid=27464

Purpose Sean Follmer with his colleagues, Daniel Leithinger and Hiroshi Ishii have created inFORM, where the computer interface can actually come off the screen and one can physically manipulate it. Design/methodology/approach One can visualize 3D information physically and touch it and feel it to understand it in new ways. Findings The interface also allows one to interact through gestures and direct deformations to sculpt digital clay, and interface elements can arise out of the surface and change on demand. Their idea is that for each individual application, the physical form can be matched to the application. Urban planners and architects can use it to explore their designs in detail; using inFORM, one can reach out from the screen and manipulate things at a distance and also manipulate and collaborate on 3D sets, gesture around them and manipulate also. Originality/value It allows collaboration of people in ways hitherto not done. Posted on February 10, 2016, the Ted talk has received over one million views as at June 9, 2016. It is trending! The researchers are thinking of "new ways that we can bring people together, and bring our information into the world, and think about smart environments that can adapt to us physically".

Padgett, L. (2016). No Copies, No Comments. *Information Today*, *33*(10), 19. Retrieved from https://search.proquest.com/docview/1861789618?accountid=27464

MIA-Missing in Archives "Disappearing News Archives," an Online Searcher feature by Sarah Jane Davis, contains, in part, text Davis cites from a March 16, 2016, ResearchBuzz blog post by Tara Calishain, as well as additional comments Calishain emailed to Online Searcher editor-in-chief Marydee Ojala. There is an irreplaceable connection that comes from holding and reading an ink-lined paper, with a few crossed-out words and some smudges, or fingering a faded snapshot, yellowing and curling up at the edges, that was lovingly pressed into a page by hand, not automatically done with perfect precision via Shutterfly.

Paul-Choudhury, S. (2011). Your digital legacy. *New Scientist*, *210*(2809), 40–43. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The article discusses how individual's digital legacies, or the collection of posts from social networking websites, are being stored long term. The article notes that while Internet companies like Google store people's information on servers for research and advertising purposes, some historians feel this kind of digital preservation is not permanent enough, and caution individuals not to trust corporations to save this data. The article notes archive methods are being researched.

Peet, L. (2017). Technology: LC's New Born-Digital Archives. *Library Journal*, *142*(15), 14. Retrieved from https://search.proquest.com/docview/1937841993?accountid=27464

Technology: LC's New Born-Digital Archives The American Folklife Center at the Library of Congress (LC) announced June 15 the creation of two new born-digital collections: the Web Cultures Web Archive (WCWA), which will feature memes, GIFs, and image macros that surface in online pop culture, and the Webcomics Web Archive (WWA), which will collect comics created for an online audience. WCWA's goal, to document the creation and sharing of web culture, means that it features such online phenomena as Lolspeak and Leet, emoji, reaction GIFs, memes, and digital urban legends--along with sites such as Urban Dictionary, Giphy, Metafilter, Cute Overload, and the LOLCat Bible Translation Project. At some point, said LC Digital Library project manager Abbie Grotke, she would like to see more content included, such as the potential for full-text search or derivative data sets--ways to help users dig deeper into the archive.

Pendse, L. R. (2016). Collecting and preserving the Ukraine conflict (2014-2015): a web archive at University of California, Berkeley. *Collection Building*, *35*(3), 64–72. Retrieved from https://search.proquest.com/docview/1829452180?accountid=27464

Purpose The purpose of this paper is to highlight the web-archiving as a tool for possible collection development in a research level academic library. The paper highlights the web-archiving project that dealt with the contemporary Ukraine conflict. Currently, as the conflict in Ukraine drags on, the need for collecting and preserving the information from various web-based resources with different ideological orientations acquires a special importance. The demise of the Soviet Union in 1991 and the emergence of independent republics were heralded by some as a peaceful transition to the "free-market" style economies. This transition was nevertheless nuanced and not seamless. Besides the incomplete market liberalization, rent-seeking behaviors of different sort, it was also accompanied by the almost ubiquitous use of and access to the internet and the internet communication technologies. Now 24 years later, the ongoing conflict in Ukraine also appears to be unfolding on the World Wide Web. With the Russian annexation of Crimea and its unification to the Russian Federation, the governmental and non-governmental websites of the Ukrainian Crimea suddenly came to represent a sort of "an endangered archive". Design/methodology/approach The main purpose of this project was to make the information that is contained in Ukrainian and Russia websites available to the wider body of scholars and students over the longer period of time in a web archive. The author does not take any ideological stance on the legal status of Crimea or on the ongoing conflict in Ukraine. There are currently several projects that are devoted to the preservation of these websites. This article also focuses on providing a survey of the landscape of these projects and highlights the ongoing web-archiving project that is entitled, "the Ukraine Crisis: 2014-2015" at the UC Berkeley Library. Findings The UC Berkeley's Ukraine Conflict Archive was made available to public in March of 2015 after enough materials were archived. The initial purpose of the archive was to selectively harvest, and archive those websites that are bound to either disappear or change significantly during the evolution of Crimea's accession to Russia. However, in the aftermath of the Crimean conflict, the ensuing of military conflict in Ukraine had forced to reevaluate the web-archiving strategy. The project was never envisioned to be a competing project to the Ukraine Conflict project. Instead, it was supposed to capture…

Pendse, L. R. (2014). Archiving the Russian and East European Lesbian, Gay, Bisexual, and Transgender Web, 2013: A Pilot Project. *Slavic & East European Information Resources*, *15*(3), 182–196. https://doi.org/10.1080/15228886.2014.930973

This article focuses on the conceptualization and implementation of a web archiving pilot project of selected Russian and East European lesbian, gay, bisexual, and transgender (LGBT) websites by the University of California, Berkeley. It introduces the use of the Web Archiving Services (WAS) platform developed by the California Digital Library. While identifying the criteria used to harvest these websites, the paper also describes various complexities associated with the viability of projects related to such complex social and political issues as the Russian and Eastern European LGBT rights movements. The article does not take an ideological stance with respect to legal issues, but rather strives to preserve information for academic research. [ABSTRACT FROM AUTHOR]

Pennock, M., & Kelly, B. (2006). Archiving Web Site Resources: A Records Management View. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 987–988). New York, NY, USA: ACM. https://doi.org/10.1145/1135777.1135978

In this paper, we propose the use of records management principles to identify and manage Web site resources with enduring value as records. Current Web archiving activities, collaborative or organisational, whilst extremely valuable in their own right, often do not and cannot incorporate requirements for proper records management. Material collected under such initiatives therefore may not be reliable or authentic from a legal or archival perspective, with insufficient metadata collected about the object during its active life, and valuable materials destroyed whilst ephemeral items are maintained. Education, training, and collaboration between stakeholders are integral to avoiding these risks and successfully preserving valuable Web-based materials.

Pichlak, M. L. (2017). Fotografia cyfrowa i technologia 360o – zastosowanie w projektach realizowanych przez Politechnikę Wrocławską TT - Digital photography and 360o technology - applied in projects realized by Wroclaw University of Technology. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951540134?accountid=27464

W artykule została krótko przytoczona historia aparatu cyfrowego oraz podstawowe różnice między fotografią cyfrową a tradycyjną. Opisano działanie studia fotograficznego 360o oraz jego wykorzystanie w projektach Politechniki Wrocławskiej.

Plachouras, V., Carpentier, F., Faheem, M., Masanès, J., Risse, T., Senellart, P., … Stavrakas, Y. (2014). ARCOMEM Crawling Architecture. *Future Internet*, *6*(3), 518–541. https://doi.org/10.3390/fi6030518

The World Wide Web is the largest information repository available today. However, this information is very volatile and Web archiving is essential to preserve it for the future. Existing approaches to Web archiving are based on simple definitions of the scope of Web pages to crawl and are limited to basic interactions with Web servers. The aim of the ARCOMEM project is to overcome these limitations and to provide flexible, adaptive and intelligent content acquisition, relying on social media to create topical Web archives. In this article, we focus on ARCOMEM's crawling architecture. We introduce the overall architecture and we describe its modules, such as the online analysis module, which computes a priority for the Web pages to be crawled, and the Application-Aware Helper which takes into account the type of Web sites and applications to extract structure from crawled content. We also describe a large-scale distributed crawler that has been developed, as well as the modifications we have implemented to adapt Heritrix, an open source crawler, to the needs of

the project. Our experimental results from real crawls show that ARCOMEM's crawling architecture is effective in acquiring focused information about a topic and leveraging the information from social media.

Portilla, O., Aguilar, J., & León, C. (2015). Semantic recommender system for the recovery of the preserved web heritage. *2015 Latin American Computing Conference (CLEI), Computing Conference (CLEI), 2015 Latin American*. IEEE. https://doi.org/10.1109/CLEI.2015.7359467

This paper presents a prototype of a semantic personalized recommender system for a repository of preserved web files. To do this, we design and implement a semantic repository of preserved web files, containing metadata associated with each preserved site. The knowledge stored in the metadata of the semantic repository is used for the recommender system, in order to give prioritized recommendations of the different preserved web files (or web heritage) that meet certain search criteria. The proposed recommender also considers semantic associations, in order to recommend not only the websites matched to the search criteria, but also semantically related.

Post, C. (2017). Building a Living, Breathing Archive:A Review of Appraisal Theories and Approaches for Web Archives. *Preservation, Digital Technology & Culture*, *46*(2), 69–77. https://doi.org/http://dx.doi.org/10.1515/pdtc-2016-0031

The paper provides a review of published literature on the collection and development of Web archives, focusing specifically on the theories, techniques, tools, and approaches used to appraise Web-based materials for inclusion in collections. Facing an enormous amount of Web-based materials, archival institutions and other cultural heritage institutions need to devise methods to actively select Webpages for preservation, creating Web archives that constitute a cultural record of the Web for the benefit of users. This review outlines the challenges of collecting and appraising Web-based materials, places the theories and activities of collecting Web-based materials within the broader discourse of archival appraisal, and points out directions for future research and critical discourse for Web archives.

Poursardar, F. (2018). How Perceptions of Web Resource Boundaries Differ for Institutional and Personal Archives. *2018 IEEE International Conference on Information Reuse and Integration (IRI), Information Reuse and Integration (IRI), 2018 IEEE International Conference on, IRI*. IEEE. https://doi.org/10.1109/IRI.2018.00026

What is and is not part of a web resource does not have a simple answer. Exploration of web resource boundaries have shown that people's assessments of resource bounds rely on understanding relationships between content fragments on the same web page and between content fragments on different web pages. This study explores whether such perceptions change based on whether the archive is for personal use or is institutional in nature. This survey explores user expectations when accessing archived web resources. Participants in the study were asked to assume they are making use of an archive provided by an institution tasked with preserving online resources, such as a digital archive that is part of the Library of Congress. Groups of pair web pages presented to the participants. Each group has a primary web page that is the resource being saved by the institutional archive. Each group has several subsequent parts or pages, which we will ask about. Consistent with our previous study on personal archiving, the primary-page content in the study comes from multi-page stories, multi-image collections, product pages with reviews and ratings on separate pages, and short

single page writings. Participants were asked to assume the institutional archive wants to preserve the primary page and then answer what else they would expect to be saved along with the primary page. The results show that there are similar expectations for preserving continuations of the main content in personal and institutional archiving scenarios, institutional archives are more likely to be expected to preserve the context of the main content, such as additional linked content, advertisements, and author information.

Poursardar, F., & Shipman, F. (2016). On Identifying the Bounds of an Internet Resource. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR '16* (pp. 305–308). New York, New York, USA: ACM Press. https://doi.org/10.1145/2854946.2854982

Systems for retrieving or archiving Internet resources often assume a URI acts as a delimiter for the resource. But there are many situations where Internet resources do not have a one-to-one mapping with URIs. For URIs that point to the first page of a document that has been broken up over multiple pages, users are likely to consider the whole article as the resource, even though it is spread across multiple URIs. Comments, tags, ratings, and advertising might or might not be perceived as part of the resource whether they are retrieved as part of the primary URI or accessed via a link. Similarly, whether content accessible via links, tabs, or other navigation av ailable at the primary URI is perceived as part of the resource may depend on the design of the website. We are examining what people believe are the bounds of Internet resources with the hope of informing systems that better match user perceptions. To unders tand this challenge we explore a situation where the user is assumed to have identified a resource by a URI, particularly for archiving. To begin to answer these questions, we asked 110 participan ts how desirable it would be for web contents related to an id entified archived resource to also be archived. Results indicate that the features important to this decision likely vary considerably from resource to resource.

Poursardar, F., & Shipman, F. (2017). What is Part of That Resource?: User Expectations for Personal Archiving. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 229–238). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=3200334.3200359

Users wish to preserve Internet resources for later use. But wh at is part of and what is not part of an Internet resource remains an open question. In this paper we examine how specific relationships between web pages affect user perceptions of thei r being part of the same resource. T his study presented participa nts with pairs of pages and asked a bout their expectation for havin g access to the second page after t hey save the first. The primar y- page content in the study comes from multi-page stories, multi- image collections, product page s with reviews and ratings on separate pages, and short single page writings. Participants we re asked to agree or disagree with three statements regarding thei r expectation for later access. Nearly 80% of participants agreed in the case of articles spread across multiple pages, images in th e same collection, and additional details or assessments of produ ct information. About 50% agreed for related content on pages linked to by the original page or related items while only abou t 30% thought advertisements or wish lists linked to were part of the resource. Differences in responses to the same page pairs f or the three statements regarding later access indicate some users distinguish between what would be valuable to them and their expectations of syst ems saving or archiving web content

Price, H. V. and G. (2017). Ereviews. *Library Journal*, *142*(19), 100. Retrieved from https://search.proquest.com/docview/1964143235?accountid=27464

According to IA founder Brewster Kahle, the BPL collection includes "hillbilly music, early brass bands, and accordion recordings from the turn of the last century, offering an authentic audio portrait of how America sounded a century ago. The Presidential Records Act has in the past been understood to mean that executive branch administrative communication must be archived, but the U.S. Justice Department is moving to dismiss the lawsuit, saying that the president has authority over what is saved in accordance with the act. [...]FCW , a publication for federal technology executives, quotes Jason R. Baron, formerly chief litigator for the National Archives and Records Administration: “If White House counsel reads [the statute] narrowly...resulting in White House staff not being required to copy or transfer presidential records to an official electronic account before individual communications self-destruct, is that decision reviewable?” For further information on this case, see ow.ly/RAnE30fT8Yc.

Radzicka, J. (2017). Mobilny pracownik – sprawozdanie z międzynarodowych warsztatów. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951541346?accountid=27464

Rani, S., Goodkin, J., Cobb, J., Habing, T., Urban, R., Eke, J., & Pearce-Moses, R. (2006). Technical Architecture Overview: Tools for Acquisition, Packaging and Ingest of Web Objects into Multiple Repositories. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (p. 360). New York, NY, USA: ACM. https://doi.org/10.1145/1141753.1141855

This poster d escribes a mod el fo r acquiring, packag ing and ingesting web objec ts for arc hiving in m ultiple repos itories. T his ongoing work is part of t he ECHO DEPosit ory Project [1], a 3-year ND IIPP-p artner digital pr eserva tion project a t the Unive rsity of Illinois at Urbana -Champa ign with partners O CLC, a consortium of content provider partners, and the Library of Congress

Ras, M., & Sierman, B. (2015). Building a Future for Our Digital Memory: A Collaborative Infrastructure for Permanent Access to Digital Heritage in The Netherlands. *New Review of Information Networking*, *20*(1–2), 219–228. https://doi.org/10.1080/13614576.2015.1114828

This article describes the developments in The Netherlands to establish a national Network for Digital Heritage. This network is based on three pillars: to make the digital heritage visible, usable, and sustainably preserved. Three working programs will have their own but integrated set of dedicated actions in order to create a national infrastructure in The Netherlands, based on an optimal use of existing facilities. In this article the focus is on the activities related to the sustainable preservation of the Dutch national digital heritage.

Reilly, W., Wolfe, R., & Smith, M. (2006). MIT’s CWSpace project: packaging metadata for archiving educational content in DSpace. *International Journal on Digital Libraries*, *6*(2), 139–147. https://doi.org/10.1007/s00799-005-0131-2

Reyes Ayala, B. (2018). A Grounded Theory of Information Quality for Web Archives. United States, North America. Retrieved from

http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Presentation for the dissertation defense of Brenda Reyes Ayala. This presentation builds a theory of information quality for web archives that is grounded in human-centered data.

Risse, T., Demidova, E., Dietze, S., Peters, W., Papailiou, N., Doka, K., … Spiliotopoulos, D. (2014). The ARCOMEM Architecture for Social- and Semantic-Driven Web Archiving. *Future Internet, Vol 6, Iss 4, Pp 688-716 (2014) VO - 6*, *6*(4), 688. https://doi.org/10.3390/fi6040688

The constantly growing amount ofWeb content and the success of the SocialWeb lead to increasing needs for Web archiving. These needs go beyond the pure preservationo of Web pages. Web archives are turning into “community memories” that aim at building a better understanding of the public view on, e.g., celebrities, court decisions and other events. Due to the size of the Web, the traditional “collect-all” strategy is in many cases not the best method to build Web archives. In this paper, we present the ARCOMEM (From Future Internet 2014, 6 689 Collect-All Archives to Community Memories) architecture and implementation that uses semantic information, such as entities, topics and events, complemented with information from the Social Web to guide a novel Web crawler. The resulting archives are automatically enriched with semantic meta-information to ease the access and allow retrieval based on conditions that involve high-level concepts.

Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., & Senellart, P. (2012). Exploiting the Social and Semantic Web for Guided Web Archiving (pp. 426–432). Germany, Europe: Heidelberg : Springer Verlag. https://doi.org/10.1007/978-3-642-33290-6_47

The constantly growing amount of Web content and the success of the Social Web lead to increasing needs for Web archiving. These needs go beyond the pure preservation of Web pages. Web archives are turning into “community memories” that aim at building a better understanding of the public view on, e.g., celebrities, court decisions, and other events. In this paper we present the ARCOMEM architecture that uses semantic information such as entities, topics, and events complemented with information from the social Web to guide a novel Web crawler. The resulting archives are automatically enriched with semantic meta-information to ease the access and allow retrieval based on conditions that involve high-level concepts. The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-642-33290-6_47. ; German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety/0325296 ; Solland Solar Cells BV ; SolarWorld Innovations GmbH ; SCHOTT Solar AG ; RENA GmbH ; SINGULUS TECHNOLOGIES AG

Risse, T., & Peters, W. (2012). ARCOMEM: From Collect-all ARchives to COmmunity MEMories. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 275–278). New York, NY, USA: ACM. https://doi.org/10.1145/2187980.2188027

The ARCOMEM project is about memory institutions like archives, museums and libraries in the age of the Social Web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM’s aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make Web archiving a more selective and meaning-based process. ARCOMEM (FP7-IST-270239) is an Integrating Project in the FP7 program of the European Commission,

which involves twelve partners from academia, industry and public sector. The project will run from January 1, 2011 to December 31, 2013.

Roberto, R. (2008). Technology Intersecting Culture: The British Slave Trade Legacies Project. *Journal of the Society of Archivists*, *29*(2), 207–232. https://doi.org/10.1080/00379810902916274

‘British Slave Trade Legacies’ is a web archiving project that collected websites and online material related to and generated from the 2007 bicentenary of Parliament abolishing the British slave trade. The Internet Archive donated their Archive-It service to harvest websites for this collection, and now provides public access to digital objects within it. This paper describes two issues that the project raised: firstly, the validity of the 2007 anniversary as marked by cultural stakeholders; secondly, the challenges of documenting it, thereby adding to historical legacy material of this topic. The archivist’s role in the 21st century will also be discussed in the context of new digital age challenges.

Rockembach, M. (2018). Políticas e tecnologias de preservação digital no arquivamento da web ; Policies and technologies to digital preservation in web archiving ; Política y tecnologías de preservación digital en el archivo de la web. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

O objetivo do artigo foi analisar a preservação digital a partir da abordagem de arquivamento da web, desde as tecnologias envolvidas no processo de arquivamento, bem como políticas de seleção, preservação e disponibilização destes conteúdos, além do estudo de instituições internacionais que atuam na preservação da web. A metodologia utiliza pesquisa bibliográfica e documental sobre iniciativas internacionais de arquivamento da web e objetiva fomentar a discussão no Brasil, assim como servir de subsídio para estudos aplicados. Analisa as publicações científicas na base de periódicos Scopus dos últimos cinco anos (2012-2016) que versam sobre o arquivamento da web, políticas de seleção dos conteúdos web e tecnologias aplicadas à coleta, armazenamento e acesso aos websites arquivados. Traz também um panorama das tecnologias utilizadas pela comunidade de iniciativas de arquivamento da web, a partir da identificação dos dados disponibilizados no site do Consórcio Internacional de Preservação da Internet. Conclui que países que ainda não possuem iniciativas próprias, como o Brasil, com o estabelecimento de políticas de seleção com enfoques específicos (institucionais, temáticas, por domínio, etc.), assim como uma gestão do ciclo de vida do arquivamento da web e a adoção de tecnologias no formato código aberto (open source) podem não só preservar sua memória digital, mas também contribuir com a comunidade internacional de arquivamento da web. ; The objective of this paper was to analyze digital preservation from the web archiving approach, addressing the technologies involved in the archiving process, as well as policies for the selection, preservation and availability of these contents, as well as the study of international institutions that work on preservation of the web. The methodology uses bibliographic and documentary research on international archival web initiatives and aims to foment the discussion in Brazil, as well as to serve as a subsidy for applied studies. It analyzes the scientific publications based on Scopus journals of the last five years (2012-2016) that deal with web archiving, web content selection policies and technologies applied to the harvest, storage and access to archived websites. It also provides an overview of the technologies used by the community of web archiving initiatives, based on the identification of the data available on the website of the International Internet Preservation Consortium. It concludes that countries that …

Rogers, R. (2017). Doing Web history with the Internet Archive: screencast documentaries. *Internet Histories*, *1*(1–2), 160–172. https://doi.org/10.1080/24701475.2017.1307542

This short article explores the challenges involved in demonstrating the value of web archives, and the histories that they embody, beyond media and Internet studies. Given the difficulties of working with such complex archival material, how can researchers in the humanities and social sciences more generally be persuaded to integrate Internet histories into their research? How can institutions and organisations be sufficiently convinced of the worth of their own online histories to take steps to preserve them? And how can value be demonstrated to the wider general public? It touches on public attitudes to personal and institutional Internet histories, barriers to access to web archives – technical, legal and methodological - and the cultural factors within academia that have hindered the penetration of new ways of working with new kinds of primary source. Rather than providing answers, this article is intended to provoke discussion and dialogue between the communities for whom Internet histories can and should be of significance.

Rollason-Cass, S., & Reed, S. (2015). Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest. *New Review of Information Networking*, *20*(1–2), 241–247. https://doi.org/http://dx.doi.org/10.1080/13614576.2015.1114839

The ease of creating and sharing content on the web has had a profound impact on the scope, pace, and mobility of social movements, as well as on how the documents and evidence of these movements are collected and preserved. This article will focus on the process of creating a web based archive around the #blacklivesmatter movement while exploring the concept of the "living archive" through collaborative collection building around social movements. By examining this and other event-based web collections, best practices and strategies to improve the process of selection and capture of web content in Living Archives are presented.

Romaniuk, L. M. (2014). Metadata for a Web Archive: PREMIS and XMP as Tools for the Task. *Library Philosophy & Practice*, 1–20. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=97212804&lang=hu&site=ehost-live

In a time where websites are ever changing, what metadata standards and tools are best for ensuring that web archive objects (such as snapshots of websites) are readable for users of the future? Can the evolution of web interfaces be documented? Initiatives that explore these questions already exist such as the Internet Archive's Wayback Machine (which stores source code from websites along with images); however, other archive building solutions are also available but have yet to be explored. The field of digital asset management (DAM), for example, has long examined how assets (digital files) are stored, organized, retrieved, and preserved. Best practices related to the use of metadata standards and tools found in digital asset management are useful and relevant to web archive building. In order to better understand the practicality of implementing DAM best practices in building a web archive, a small project was performed which involved cross-walking two metadata standards, Adobe's eXtensible Metadata Platform (XMP) and PREservation Metadata: Implementation Strategies (PREMIS), and recording metadata related to snapshots of a website, the Perseus Digital Library, over a span of over a decade. The findings of this project showed that it is

impossible, at least in part, to encode PREMIS within XMP. [ABSTRACT FROM AUTHOR]

Ruusalep, R. (2017). Preserving digital legal deposit - new challenges and opportunities. In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. in Session 210 - Preservation and Conservation (PAC) Strategic Programme.* Wrocław: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/1677/1/210-ruusalepp-en.pdf

Born digital content has always been considered to be a bigger challenge for preservation than digitised content. Higher volume and technical complexity, dynamism as well as a complex surrounding rights space are frequently cited as aspects that make born digital content 'special' to memory institutions. This paper builds on the Estonian case of introducing digital legal deposit which has led to an exercise of reconceptualising the digital preservation function of the national library. The rapid increase in volume, file size and new file formats have led to making the library's preservation service levels explicit, an update to the preservation policy and automation of archiving workflows. The new demands on preservation are pushing the current digital repository system of the national library to its limits and the library needs to embark on migrating to a new preservation solution. This response to a sudden change in digital preservation workload is typical in the heritage sector – upgrading the ingest component is the first instinctive reaction of most memory institutions. This paper proposes that increasing the throughput of ingest component needs to be combined with a modular concept of a preservation system that sets interoperability as its core principle. When digital preservation is conceptualised as an exercise of resilience rather than sustainability, the interoperability requirement for systems architecture and service design follows logically.

Saad, M. Ben, & Gançarski, S. (2012). Archiving the web using page changes patterns: a case study. *International Journal on Digital Libraries*, *13*(1), 33–49. https://doi.org/10.1007/s00799-012-0094-z

Issue Title: Focused Issue on Joint Conference on Digital Libraries (JCDL) 2011 A pattern is a model or a template used to summarize and describe the behavior (or the trend) of data having generally some recurrent events. Patterns have received a considerable attention in recent years and were widely studied in the data mining field. Various pattern mining approaches have been proposed and used for different applications such as network monitoring, moving object tracking, financial or medical data analysis, scientific data processing, etc. In these different contexts, discovered patterns were useful to detect anomalies, to predict data behavior (or trend) or, more generally, to simplify data processing or to improve system performance. However, to the best of our knowledge, patterns have never been used in the context of Web archiving. Web archiving is the process of continuously collecting and preserving portions of the World Wide Web for future generations. In this paper, we show how patterns of page changes can be useful tools to efficiently archive Websites. We first define our pattern model that describes the importance of page changes. Then, we present the strategy used to (i) extract the temporal evolution of page changes, (ii) discover patterns, to (iii) exploit them to improve Web archives. The archive of French public TV channels France Télévisions is chosen as a case study to validate our approach. Our experimental evaluation based on real Web pages shows the utility of patterns to improve archive quality and to optimize indexing or storing.[PUBLICATION ABSTRACT]

Saad, M. Ben, & Gançarski, S. (2010). Using Visual Pages Analysis for Optimizing Web Archiving. In *Proceedings of the 2010 EDBT/ICDT Workshops* (p. 43:1--43:7). New York, NY, USA: ACM. https://doi.org/10.1145/1754239.1754287

SalahEldeen, H. M., & Nelson, M. L. (2013). Carbon Dating the Web: Estimating the Age of Web Resources. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 1075–1082). New York, NY, USA: ACM. https://doi.org/10.1145/2487788.2488121

In the course of web research it is often necessary to estimate the creation datetime for web resources (in the general case, this value can only be estimated). While it is feasible to manually establish likely datetime values for small numbers of resources, this becomes infeasible if the collection is large. We present “carbon date”, a simple web application that estimates the creation date for a URI by polling a number of sources of evidence and returning a machine-readable structure with their respective values. To establish a likely datetime, we poll bitly for the first time someone shortened the URI, topsy for the first time someone tweeted the URI, a Memento aggregator for the first time it appeared in a public web archive, Google’s time of last crawl, and the Last-Modified HTTP response header of the resource itself. We also examine the backlinks of the URI as reported by Google and apply the same techniques for the resources that link to the URI. We evaluated our tool on a gold standard data set of 1200 URIs in which the creation date was manually verified. We were able to estimate a creation date for 75.90% of the resources, with 32.78% having the correct value. Given the different nature of the URIs, the union of the various methods produces the best results. While the Google last crawl date and topsy account for nearly 66% of the closest answers, eliminating the web archives or Last-Modified from the results produces the largest overall negative impact on the results. The carbon date application is available for download or use via a web API.

Samar, T., Huurdeman, H. C., Ben-David, A., Kamps, J., & de Vries, A. (2014). Uncovering the Unarchived Web. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1199–1202). New York, NY, USA: ACM. https://doi.org/10.1145/2600428.2609544

Many national and international heritage institutes realize the importance of archiving the web for future culture heritage. Web archiving is currently performed either by harvesting a national domain, or by crawling a pre-defined list of websites selected by the archiving institution. In either method, crawling results in more information being harvested than just the websites intended for preservation; which could be used to reconstruct impressions of pages that existed on the live web of the crawl date, but would have been lost forever. We present a method to create representations of what we will refer to as a web collection’s (aura): the web documents that were not included in the archived collection, but are known to have existed --- due to their mentions on pages that were included in the archived web collection. To create representations of these unarchived pages, we exploit the information about the unarchived URLs that can be derived from the crawls by combining crawl date distribution, anchor text and link structure. We illustrate empirically that the size of the aura can be substantial: in 2012, the Dutch Web archive contained 12.3M unique pages, while we uncover references to 11.9M additional (unarchived) pages.

Samar, T., Traub, M. C., van Ossenbruggen, J., & de Vries, A. P. (2016). Comparing Topic Coverage in Breadth-First and Depth-First Crawls Using Anchor Texts. *Research &*

Advanced Technology for Digital Libraries: 20th International Conference on Theory & Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings, 133. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Samouelian, M. (2009). Embracing Web 2.0: Archives and the Newest Generation of Web Applications. *American Archivist*, *72*(1), 42–71. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Archivists are converting physical collections to digital formats and displaying surrogates of these primary sources on their websites. Simultaneously, the Web is moving toward a shared environment that embraces collective intelligence and participation, which is often called Web 2.0. This paper investigates the extent to which Web 2.0 features have been integrated into archival digitization projects. Although the use of Web 2.0 features has not yet been widely discussed in the professional archival literature, this exploratory study of college and university repository websites in the United States suggests that archival professionals are embracing Web 2.0 to promote their digital content and redefine relationships with their patrons. [ABSTRACT FROM AUTHOR]

Sanders, S., Sanka, G., Aikat, J., & Kaur, J. (2015). The Influence of Client Platform on Web Page Content: Measurements, Analysis, and Implications. In J. Wang, W. Cellary, D. Wang, H. Wang, S.-C. Chen, T. Li, & Y. Zhang (Eds.), *WISE 2015: Web Information Systems Engineering – WISE 2015* (pp. 1–16). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26187-4_1

Modern web users have access to a wide and diverse range of client platforms to browse the web. While it is anecdotally believed that the same URL may result in a different web page across different client platforms, the extent to which this occurs is not known. In this work, we systematically study the impact of different client platforms (browsers, operating systems, devices, and vantage points) on the content of base HTML pages. We collect and analyze the base HTML page downloaded for 3876 web pages composed of the top 250 web sites using 32 different client platforms for a period of 30 days — our dataset includes over 3.5 million web page downloads. We find that client platforms have a statistically significant influence on web page downloads in both expected and unexpected ways. We discuss the impact that these results will have in several application domains including web archiving, user experience, social interactions and information sharing, and web content sentiment analysis.

Sanderson, R. (2012). Global Web Archive Integration with Memento. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 379–380). New York, NY, USA: ACM. https://doi.org/10.1145/2232817.2232900

In this poster, we describe the approach taken to designing and implementing a tera-scale multi-repository index of archived web resources using massively parallel processing.

Sanderson, R., de Sompel, H., Burnhill, P., & Grover, C. (2013). Hiberlink: Towards Time Travel for the Scholarly Web. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts* (p. 21). New York, NY, USA: ACM. https://doi.org/10.1145/2499583.2500370

The preservation of traditional, digital scholarly output, such as PDF or HTML journal articles, is relatively well understood, and adequately organized through systems such as Portico and LoCKSS. However, the scholarly record is expanding with a wide variety of materials for which no established archival approaches exist. This includes, for example, workflows and software, project descriptions, demonstrations, datasets, and videos published on the web. Some of these resources are referenced in traditional papers and the lack of archival infrastructure yields a scholarly record with many loose ends. The Hiberlink project aims to quantify the extent to which such referenced resources are preserved in web archives, and propose solutions to ensure the longevity of the context of the research, along side the formal publication. The Hiberlink project regards the problem of preserving web resources referenced in scholarly papers as a special case of the more general problem of preserving scholarly compound objects, aka Research Objects, which consist of resources with a variety of relationships and dependencies.

Sanoja, A., & Gançarski, S. (2017). Migrating Web Archives from HTML4 to HTML5: A Block-Based Approach and Its Evaluation. In M. Kirikova, K. Nørvåg, & G. A. Papadopoulos (Eds.), *ADBIS 2017: Advances in Databases and Information Systems* (pp. 375–393). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-66917-5_25

Web archives (and the Web itself) are likely to suffer from format obsolescence. In a few years or decades, future Web browsers will no more be able to properly render Web pages written in HTML4 format. Thus we propose a migration tool from HTML4 to HTML5. This is challenging, because it requires to generate HTML5 semantic elements that do not exist in HTML4 pages. To solve this issue, we propose to use a Web page segmenter. Indeed, blocks generated by a segmenter are good candidates for being semantic elements as both reflect the content structure of the page. We use an evaluation framework for Web page segmentation, that helps defining and computing relevant metrics to measure the quality of the migration process. We ran experiments on a sample of 40 pages. The migrated pages we produce are compared to a ground truth. The automatic labeling of blocks is quite similar to the ground truth, though its quality depends on the type of page we migrate. When comparing the rendering of the original page and the rendering of its migrated version, we note some differences, mainly due to the fact that rendering engines do not (yet) properly render the content of semantic elements.

Santipantakis, G. M., Kotis, K. I., Vouros, G. A., & Doulkeridis, C. (2018). RDF-Gen. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics - WIMS '18* (pp. 1–10). New York, New York, USA: ACM Press. https://doi.org/10.1145/3227609.3227658

Recent state-of-the-art approaches and technologies for generating RDF graphs from non-RDF data, use languages designed for specifying transformations or mappings to data of various kinds of format. This paper presents a new approach for the generation of ontology-annotated RDF graphs, linking data from multiple heterogeneous streaming and archival data sources, with high throughput and low latency. To support this, and in contrast to existing approaches, we propose embedding in the RDF generation process a close-to-sources data processing and linkage stage, supporting the fast template-driven generation of triples in a subsequent stage. This approach, called RDF-Gen, has been implemented as a SPARQL-based RDF generation approach. RDF-Gen is evaluated against the latest related work of RML and SPARQL-Generate, using real world datasets.

Schneider, R., & McCown, F. (2013). First Steps in Archiving the Mobile Web: Automated Discovery of Mobile Websites. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 53–56). New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467735

Smartphones and tablets are increasingly used to access the Web, and many websites now provide alternative sites tailored specifically for these mobile devices. Web archivists are in need of tools to aid in archiving this equally ephemeral Mobile Web. We present Findmobile, a tool for automating the discovery of mobile websites. We tested our tool in an experiment examining 10K popular websites and found that the most frequently used technique used by popular websites to direct mobile users to mobile sites was by automated client and server-side redirection. We found that nearly half of mobile web pages differ dramatically from their stationary web counterparts and that the most popular websites are those most likely to have mobile-specific pages.

Schostag, S., & Fonss-Jorgensen, E. (2012). Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective. *Microform & Digitization Review*, *41*(3–4), 110–120. Retrieved from https://search.proquest.com/docview/1520327503?accountid=27464

Since 2005 archiving the dynamic Internet has been required by law in Denmark. This article tells the story of the last seven years of experience with archiving the Internet in Denmark: What is covered by the law? How do we organize the work? How do we collect the web in practice? Who has access to the web archive? And finally, what are the challenges and future perspectives? The article focuses on the curatorial aspects and does not go into technical details. Adapted from the source document.

Schuler, A. (2017). *Collection & community building through web archiving: engaging with faculty and students in a collaborative web archiving project*. United States, North America: Digital USD. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Tisch Library at Tufts University has recently begun a pilot web archiving project, aiming to deepen Tufts' collections in areas of strategic importance and support more "traditional" library collection development activities, while collecting material that is not known to be comprehensively collected by other institutions. Additionally, the project offers an opportunity for collaborative collection building with faculty and students that serves as a unique way to deepen our community‘s engagement with the library. The initial pilot collection focuses on environmental justice, selected due to its relevance to the Tufts community and curriculum and to build on existing Tisch Library collection strengths. Two undergraduate courses related to environmental justice were identified and invited to partner in the pilot project. This partnership would leverage student research to expand the initial collection while introducing students to concepts of web archiving and information literacy around websites and providing them with the opportunity to contribute to shaping the scholarly record. Both courses added a brief assignment to their syllabus: while doing research on their chosen topics, students would identify 3-7 web sites they felt would benefit from preservation and submit the sites to the library, to be evaluated and added to the web archive as appropriate. This presentation discusses the process of beginning a subject-based web archiving project, focusing on the collaborative project with two undergraduate classes. It addresses decisions made when starting and scoping the project; collection development

issues; the logistics, benefits, and outcomes of the student and faculty collaboration; and future directions.

Seadle, M. (2011). Archiving in the networked world: preserving plagiarized works. *Library Hi Tech, 29*(4), 655–662. https://doi.org/10.1108/07378831111189750

Purpose – Plagiarism has become a salient issue for universities and thus for university libraries in recent years. This paper aims to discuss three interrelated aspects of preserving plagiarized works: collection development issues, copyright problems, and technological requirements. Too often these three are handled separately even though in fact each has an influence on the other. Design/methodology/approach – The paper looks first at the ingest process (called the Submission Information Package or SIP), then at storage management in the archive (the AIP or Archival Information Package), and finally at the retrieval process (the DIP or Distribution Information Package). Findings – The chief argument of this paper is that works of plagiarism and the evidence exposing them are complex objects, technically, legally and culturally. Merely treating them like any other work needing preservation runs the risk of encountering problems on one of those three fronts. Practical implications – This is a problem, since currently many public preservation strategies focus on ingesting large amounts of self-contained content that resembles print on paper, rather than on online works that need special handling. Archival systems also often deliberately ignore the cultural issues that affect future usability. Originality/value – The paper discusses special handling and special considerations for archiving works of plagiarism. [ABSTRACT FROM AUTHOR]

Seadle, M. (2009). Archiving in the networked world: betting on the future. *Library Hi Tech, 27*(2), 319–325. https://doi.org/10.1108/07378830910968326

Purpose – The goal of this column is not to argue the pros and cons of digital archiving, or to propose solutions to its problems, but to describe it as a research subject and a social phenomenon. Design/methodology/approach – This column relies on cultural anthropology, in particular the approach that Clifford Geertz championed, and for cultural anthropology, language and its social context matter. Findings – Archiving systems abound with competing claims about effectiveness. Transparency and evidence of public testing is rare, with a few exceptions. The lack of public testing does not mean that systems do less than they claim, but it does mean that libraries, archives and museums need to press for proof if they want to have confidence in the product. Originality/value – When betting on the future, these cannot be certainty, but bets placed should be based on knowledge.

Seadle, M. (2001). Copyright in the networked world: digital legal deposit. *Library Hi Tech, 19*(3), 299–303. https://doi.org/10.1108/EUM0000000005893

Legal deposit is the requirement that particular types of material be deposited with a national library or designated research libraries. US law does not at present include any requirement for the deposit of works that exist solely in the form of Web pages. For digital materials, it makes no sense to write rules for legal deposit based on the medium. Nations and national libraries that ignore legal deposit for digital works will find themselves missing a significant and unrecoverable portion of their cultural heritage

Senellart, P., & Oita, M. (2011). Deriving Dynamics of Web Pages: A Survey. HAL CCSD. Retrieved from

http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

The World Wide Web is dynamic by nature: content is continuously added, deleted, or changed, which makes it challenging for Web crawlers to keep up-to-date with the current version of a Web page, all the more so since not all apparent changes are significant ones. We review major approaches to change detection in Web pages and extraction of temporal properties (especially, timestamps) of Web pages. We focus our attention on techniques and systems that have been proposed in the last ten years and we analyze them to get some insight into the practical solutions and best practices available. We aim at providing an analytical view of the range of methods that can be used, distinguishing them on several dimensions, especially, their static or dynamic nature, the modeling of Web pages, or, for dynamic methods relying on comparison of successive versions of a page, the similarity metrics used. We advocate for more comprehensive studies of the effectiveness of Web page change detection methods, and finally highlight open issues.

Setty, V., Bedathur, S., Berberich, K., & Weikum, G. (2010). InZeit: Efficiently Identifying Insightful Time Points. *Proc. VLDB Endow.*, *3*(1–2), 1605–1608. https://doi.org/10.14778/1920841.1921050

Web archives are useful resources to find out about the temporal evolution of persons, organizations, products, or other topics. However, even when advanced text search functionality is available, gaining insights into the temporal evolution of a topic can be a tedious task and often requires sifting through many documents. The demonstrated system named InZeit (pronounced “insight”) assists users by determining insightful time points for a given query. These are the time points at which the top-k time-travel query result changes substantially and for which the user should therefore inspect query results. InZeit determines the m most insightful time points efficiently using an extended segment tree for in-memory bookkeeping.

Shan, D., Zhao, W. X., Chen, R., Shu, B., Wang, Z., Yao, J., … Li, X. (2012). EventSearch: A System for Event Discovery and Retrieval on Multi-type Historical Data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1564–1567). New York, NY, USA: ACM. https://doi.org/10.1145/2339530.2339781

We present EventSearch, a system for event extraction and retrieval on four types of news-related historical data, i.e., Web news articles, newspapers, TV news program, and micro-blog short messages. The system incorporates over 11 million web pages extracted from “Web InfoMall”, the Chinese Web Archive since 2001. The newspaper and TV news video clips also span from 2001 to 2011. The system, upon a user query, returns a list of event snippets from multiple data sources. A novel burst model is used to discover events from time-stamped texts. In addition to offline event extraction, our system also provides online event extraction to further meet the user needs. EventSearch provides meaningful analytics that synthesize an accurate description of events. Users interact with the system by ranking the identified events using different criteria (scale, recency and relevance) and submitting their own information needs in different input fields.

Shein, E. (2015). Preserving the internet. *Communications of the ACM*, *59*(1), 26–28. https://doi.org/10.1145/2843553

The article looks at efforts to preserve the contents of the Internet for future generations. Particular focus is given to the Global Database of Events, Language, and Tone (GDELT) project, led by computer scientist Kaylev Leetaru, and to the not-for-profit digital library known as the Internet Archive. Topics include the alteration of online documents such as government press releases and the digitization of books and other museum and library collections.

Shiozaki, R., & Eisenschitz, T. (2009). Role and justification of web archiving by national libraries. *Journal of Librarianship and Information Science, 41*(2), 90–107. https://doi.org/10.1177/0961000609102831

This paper reports on a questionnaire survey of 16 national libraries designed to clarify how national libraries attempt to justify their web archiving activities. Results indicate they envisage that a) the benefits brought about by their initiatives are greater than the overall costs, b) the costs imposed on libraries are greater than the costs imposed on stakeholders, and c) all of them are making efforts to respond to legal risks in various ways (e.g. legislation, contracting and opt-out policies) although there are trade-off relations in terms of costs for negotiation, scope of access and size and scope of the web archive. The paper discusses whether a basic logic for justification of their web archiving is valid from the perspective of balancing cost—benefit. Further, it highlights the potential, underlying premises of the logic that motivates the intervention of national libraries as public sector organizations.

Shveiky, R., & Bar-Ilan, J. (2013). National Libraries' Traditional Collection Policy Facing Web Archiving. *Alexandria, 24*(3), 37–72. Retrieved from https://search.proquest.com/docview/1548796786?accountid=27464

One of the main missions of a national library is to preserve the national creative works in printed and non-printed formats. In the 1990s, national libraries began to harvest and archive the national body of creative work that was published on the internet. The aim of the study was to examine to what extent national libraries implement their general collection policy when they establish a national web archive. The study, which was based on a qualitative approach, had three phases: examining the characteristics of a traditional collection policy of a national library; identifying the characteristics of a collection policy of a national library's web archive; and comparing the traditional collection characteristics with the national library's web archive characteristics. The results showed that although the libraries that were studied were from different regions of the world and various cultures, the characteristics of their traditional collections are similar. In contrast, the difference between their web archives is more significant. National libraries do not apply the traditional policy to the internet, and struggle to shape new rules for coping with web contents.

Sierman, B., & Teszelszky, K. (2017). How can we improve our web collection? An evaluation of webarchiving at the KB National Library of the Netherlands (2007-2017). https://doi.org/10.1177/0955749017725930

The Koninklijke Bibliotheek, the Dutch National Library (KB-NL), started in 2007 the project "web archiving" based on a selection of Dutch websites. The initial selection of 1,000 websites has currently grown into over 12,000 selected web sites, crawled on different intervals. Although due to legal restrictions the current use is limited to the KB-NL reading room, it is important that the KB-NL includes the requirements of the (future) users in her approach of creating a web collection. With respect to the long term preservation of the

collection, we also need to incorporate the requirements for long term archiving in our approach, as described in the Open Archival Information Model (OAIS)1. This article describes the results of a research project on web archiving and the web collection of archived sites in the KB-NL, investigating the following questions. What is web archiving in the Netherlands? What are the selection criteria of KB-NL and how are these related to what can be found on the Dutch web by the contemporary user? What is the influence of the choice of tools we use to harvest on the final archived website? Do we know enough of the value of the web collection and the potential usage of it by researchers and how can we improve this value? This article will describe the outcomes of the research, the conclusions and advice that can be drawn from it and will hopefully inspire broader discussions about the essence of creating web collections for long term preservation as part of cultural heritage.

Signori, B. (2017). Preserving cultural heritage: Better together! In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. in Session S08 - Satellite Meeting: Preservation and Conservation Section joint with the Association International Francophone des Bibliothécaires et Documentalistes (AIFBD) in collaborati*. Wrocław: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/id/eprint/1801

The Swiss National Library has a mandate to collect, catalogue, store and disseminate the cultural heritage created in Switzerland and abroad by and about the Swiss, both in print and digital. This sounds like a clear enough mission, but dig deeper and this mandate raises all sorts of tough questions. What exactly is cultural heritage? Obviously, it goes far beyond e-books and e-journals of well-established Swiss publishers. It is Swiss websites, newsletters of Swiss societies, and so on. However, what about all the digital data that is created by Swiss people every waking moment? The selfies, blogs, tweets, social media, personal digital archives. Surely not everything can be considered cultural heritage. But who decides what is and what isn't? And then how do we cope with the enormous quantity of information being produced? How can we decide what to keep for future generations when we cannot even cope with the output of the current generation? Not to mention the costs. With budgets being cut all the time, what does that mean for our cultural heritage? Amidst all these tough questions, one thing is clear: no single institution can possibly cope with collecting all that information nor be tasked with the decision on what to preserve and what not. This paper will use the example of Web Archive Switzerland to show how trust and interoperability have led to constructive collaboration. Web Archive Switzerland was born in 2008 following 5 years of discussion with the cantonal libraries. Since then websites with a bearing on Switzerland have been selected, documented, preserved and disseminated collaboratively among 30 Swiss institutions. The key lesson learned over the past 14 years is that to answer the tough questions and challenges we had to look beyond our own walls and borders. We learned to let go of the idea that we can do it alone, that we can control the world of content through clever curation. We learned how to create partnerships and strong networks of institutions, how to engage new sorts of curators, how to trust each other and share synergies and costs, all with the common goal of saving as much digital heritage as possible. In summary, this paper is a call to arms to join forces, to forge partnerships, to bundle competences, and to build collaborative networks! It will show that curating collaboration between institutions is as important as curating cultural heritage and it will suggest ways forward to create more collaborative collections of…

Simes, L., & Pymm, B. (2009). Legal Issues Related to Whole-of-Domain Web Harvesting in Australia. *Journal of Web Librarianship*, *3*(2), 129–142. https://doi.org/10.1080/19322900902787227

Selective archiving of Web sites in Australia has been under way since 1996. This approach has seen carefully selected sites preserved after site owners granted permission. The labor-intensive nature of this process means only a small number of sites can ever be acquired in this manner. An alternate approach is an automated "whole-of-domain" capture of sites, which has been undertaken in a number of countries, including Australia. This article considers the existing legal position in taking this approach and looks at how legal deposit and copyright legislation constrains the process. It also considers recent amendments to the Copyright Act to provide more flexibility along the lines of the U.S. fair-use approach and the possible impact these new provisions may have for those involved with large-scale Web archiving in Australia

Slater, K. (2017). Who Gets to Die of Dysentery?: Ideology, Geography, and The Oregon Trail. *Children's Literature Association Quarterly*, *42*(4), 374–395. https://doi.org/http://dx.doi.org/10.1353/chq.2017.0040

This article examines the co-constitutive relationship between ideology and geography in three editions of the educational computer game The Oregon Trail, arguing that the game reinforces a colonialist worldview through representations of place, space, and time. Despite seeming to accommodate players of any race or gender, The Oregon Trail imagines its protagonist-the "you" traveling the Trail-as white and male, a construct that reinforces the supremacist narrative of nineteenth-century settlement. Through a rapid in-game compression of time and space that urges progress, the game encourages child players to perform the spatialized worldview that codifies manifest destiny.

Souza, T., Demidova, E., Risse, T., Holzmann, H., Gossen, G., & Szymanski, J. (2015). Semantic URL Analytics to Support Efficient Annotation of Large Scale Web Archives. In J. Cardoso, F. Guerra, G.-J. Houben, A. M. Pinto, & Y. Velegrakis (Eds.), *IKC 2015: Semantic Keyword-based Search on Structured Data Sources* (pp. 153–166). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27932-9_14

Long-term Web archives comprise Web documents gathered over longer time periods and can easily reach hundreds of terabytes in size. Semantic annotations such as named entities can facilitate intelligent access to the Web archive data. However, the annotation of the entire archive content on this scale is often infeasible. The most efficient way to access the documents within Web archives is provided through their URLs, which are typically stored in dedicated index files. The URLs of the archived Web documents can contain semantic information and can offer an efficient way to obtain initial semantic annotations for the archived documents. In this paper, we analyse the applicability of semantic analysis techniques such as named entity extraction to the URLs in a Web archive. We evaluate the precision of the named entity extraction from the URLs in the Popular German Web dataset and analyse the proportion of the archived URLs from 1,444 popular domains in the time interval from 2000 to 2012 to which these techniques are applicable. Our results demonstrate that named entity recognition can be successfully applied to a large number of URLs in our Web archive and provide a good starting point to efficiently annotate large scale collections of Web documents.

Spaniol, M., Denev, D., Mazeika, A., Weikum, G., & Senellart, P. (2009). Data Quality in Web Archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web* (pp. 19–26). New York, NY, USA: ACM. https://doi.org/10.1145/1526993.1526999

Web archives preserve the history of Web sites and have high long-term value for media and business analysts. Such archives are maintained by periodically re-crawling entire Web sites of interest. From an archivist's point of view, the ideal case to ensure highest possible data quality of the archive would be to "freeze" the complete contents of an entire Web site during the time span of crawling and capturing the site. Of course, this is practically infeasible. To comply with the politeness specification of a Web site, the crawler needs to pause between subsequent http requests in order to avoid unduly high load on the site's http server. As a consequence, capturing a large Web site may span hours or even days, which increases the risk that contents collected so far are incoherent with the parts that are still to be crawled. This paper introduces a model for identifying coherent sections of an archive and, thus, measuring the data quality in Web archiving. Additionally, we present a crawling strategy that aims to ensure archive coherence by minimizing the diffusion of Web site captures. Preliminary experiments demonstrate the usefulness of the model and the effectiveness of the strategy.

Spaniol, M., & Weikum, G. (2012). Tracking Entities in Web Archives: The LAWA Project. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 287–290). New York, NY, USA: ACM. https://doi.org/10.1145/2187980.2188030

Web-preservation organization like the Internet Archive not only capture the history of born-digital content but also reflect the zeitgeist of different time periods over more than a decade. This longitudinal data is a potential gold mine for researchers like sociologists, politologists, media and market analysts, or experts on intellectual property. The LAWA project (Longitudinal Analytics of Web Archive data) is developing an Internet-based experimental testbed for large-scale data analytics on Web archive collections. Its emphasis is on scalable methods for this specific kind of big-data analytics, and software tools for aggregating, querying, mining, and analyzing Web contents over long epochs. In this paper, we highlight our research on {em entity-level analytics} in Web archive data, which lifts Web analytics from plain text to the entity-level by detecting named entities, resolving ambiguous names, extracting temporal facts and visualizing entities over extended time periods. Our results provide key assets for tracking named entities in the evolving Web, news, and social media.

Sparks, S., Look, H., Bide, M., & Muir, A. (2010). A registry of archived electronic journals. *Journal of Librarianship and Information Science*, *42*(2), 111–121. https://doi.org/10.1177/0961000610361552

Stirling, P., Chevallier, P., & Illien, G. (2012). Web Archives for Researchers: Representations, Expectations and Potential Uses. *D-Lib Magazine*, *18*(3–4). https://doi.org/10.1045/march2012-stirling

The Internet has been covered by legal deposit legislation in France since 2006, making web archiving one of the missions of the Bibliotheque nationale de France (BnF). Access to the web archives has been provided in the library on an experimental basis since 2008. In the context of increasing interest in many countries in web archiving and how it may best serve the needs of researchers, especially in the expanding field of Internet studies for social sciences, a qualitative study was performed, based on interviews with potential users of the

web archives held at the BnF, and particularly researchers working in various areas related to the Internet. The study aimed to explore their needs in terms of both content and services, and also to analyse different ways of representing the archives, in order to identify ways of increasing their use. While the interest of maintaining the "memory" of the web is obvious to the researchers, they are faced with the difficulty of defining, in what is a seemingly limitless space, meaningful collections of documents. Cultural heritage institutions such as national libraries are perceived as trusted third parties capable of creating rationally-constructed and well-documented collections, but such archives raise certain ethical and methodological questions. Adapted from the source document.

Stirling, P., Illien, G., Sanz and, P., & Sepetjan, S. (2012). The state of e-legal deposit in France: Looking back at five years of putting new legislation into practice and envisioning the future. *IFLA Journal*, *38*(1), 5–24. Retrieved from http://10.0.4.153/0340035211435323

The article describes the legal situation in France regarding the legal deposit of digital material, and shows how it has been implemented in practice at the Bibliothèque nationale de France (BnF). The focus is on web archiving, where the BnF has experience going back almost 10 years, but other aspects of digital legal deposit are discussed, with possible future developments and challenges. Throughout comparisons are made with the situations in other countries. [ABSTRACT FROM PUBLISHER]

Strodl, S., Becker, C., Neumayer, R., & Rauber, A. (2007). How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 29–38). New York, NY, USA: ACM. https://doi.org/10.1145/1255175.1255181

An increasing number of institutions throughout the world face legal obligations or business needs to collect and preserve digital objects over several decades. A range of tools exists today to support the variety of preservation strategies such as migration or emulation. Yet, different preservation requirements across institutions and settings make the decision on which solution to implement very diffcult. This paper presents the PLANETS Preservation Planning approach. It provides an approved way to make informed and accountable decisions on which solution to implement in order to optimally preserve digital objects for a given purpose. It is based on Utility Analysis to evaluate the performance of various solutions against well-defined requirements and goals. The viability of this approach is shown in a range of case studies for different settings. We present its application to two scenarios of web archives, two collections of electronic publications, and a collection of multimedia art. This work focuses on the different requirements and goals in the various preservation settings.

Suebchua, T., Manaskasemsak, B., Rungsawang, A., & Yamana, H. (2018). Efficient Topical Focused Crawling Through Neighborhood Feature. *New Generation Computing*, *36*(2), 95–118. https://doi.org/10.1007/s00354-017-0029-8

A focused web crawler is an essential tool for gathering domain-specific data used by national web corpora, vertical search engines, and so on, since it is more efficient than general Breadth-First or Depth-First crawlers. The problem in focused crawling research is the prioritization of unvisited web pages in the crawling frontier followed by crawling these web pages in the order of their priority. The most common feature, adopted in many focused crawling researches, to prioritize an unvisited web page is the relevancy of the set of its

source web pages, i.e., its in-linked web pages. However, this feature is limited, because we cannot estimate the relevancy of the unvisited web page correctly if we have few source web pages. To solve this problem and enhance the efficiency of focused web crawlers, we propose a new feature, called the '"neighborhood feature"'. This enables the adoption of additional already-downloaded web pages to estimate the priority of a target web page. The additionally adopted web pages consist both of web pages located at the same directory as that of the target web page and web pages whose directory paths are similar to that of the target web page. Our experimental results show that our enhanced focused crawlers outperform the crawlers not utilizing the neighborhood feature as well as the state-ofthe-art focused crawlers, including HMM crawler.

Suomela, T. (2015). Growing a web archiving program: A case study for evolving an organization-management plan. In *Preservation and Conservation with Information Technology. IFLA 2015 South Africa*. Cape Town: IFLA. Retrieved from http://library.ifla.org/1088/1/090-suomela-en.pdf

Web archiving presents a number of technical and organizational challenges for libraries. The University of Alberta Libraries has been using Archive-IT to manage a web archiving program for since 2009. This presentation will describe the history of web archiving at the University of Alberta and show the evolution of those services over time. Web archiving is not just technically challenging, it can also be organizationally challenging. Alberta has elected to use a distributed model for collection management by spreading the work for collection development and maintenance across subject librarians and library support staff. Some of the challenges of such a management plan include collection scoping, skill transfer, quality assurance, and metadata creation. The libraries also collaborate with regional and national consortia while working to expand services to researchers and casual users of the library. Attendees will takeaway lessons about collection management, collaboration, and research services for web archives.

Tajedini, O., Sadatmoosavi, A., Ghazizade, A., & Tajedini, A. (2018). Investigation of the Currency, Disappearance and Half-Life of Urls of Web Resources Cited In Iranian Researchers: A Comparative Study. *International Journal of Information Science & Management*, *16*(1), 27–47. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

This research was intended to comparatively investigate the currency, disappearance and half-life of URLs of web resources cited in Iranian researchers' articles indexed in ISI in information science, psychology and management from 2009 to 2011. The research method was citation analysis. The statistical population of this research was all articles by Iranian researchers in psychology, information science and management from 2009 to 2011 which were indexed in SSCI. In order to extract bibliographic information of articles, ISI database was searched and the titles of the articles were extracted. After investigating the currency and disappearance of cited URLs and calculating the half-life of web resources, collected data were analyzed in accordance with research questions by means of Excel Software. The results of this research revealed that in articles written by Iranian researchers indexed in ISI in information science, psychology and management there were 6152, 3639 and 8926 citations, respectively, of which 13.7, 44.8 and 14.23 percent were online citations, respectively. The most frequently used domain in all three fields was .org. The most stable and persistent domain in psychology was .com, in information science was .org and in management was for

those domains other than the mentioned domains. The most frequent file format was pdf in all three fields. In information science, pdf. Files were the most stable while in management, rtf files and in psychology, ppt files were the most stable ones, respectively. In the initial search for online citations in psychology, information science and management, respectively, 58, 82 and 88 percent of citations were accessible which were even increased after second check with due measurements to 95, 98 and 97 percent, respectively. The research results also demonstrated that most accessible internet addresses in investigated articles of all three fields were found in the cited internet address. The status of inaccessible internet addresses in all investigated articles regarding error messages also indicated that in psychology and management 404 error message (Not found) was the most frequent error with 34 and 22 percent, respectively and in information science, 403 error message (forbidden) was the most frequent error message with 21 percent. The average half-life of online citations calculated in all investigated articles was 2.6 years which was calculated as 3 years and 4 months in information science, 2 years …

Takata, Y., Akiyama, M., Yagi, T., Hato, K., & Goto, S. (2018). POSTER: Predicting Website Abuse Using Update Histories. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (pp. 9–10). New York, New York, USA: ACM Press. https://doi.org/10.1145/3184558.3186903

Threats of abusing websites that webmasters have stopped updating have increased. In this poster, we propose a method of predicting potentially abusable websites by retrospectively analyzing updates of software that composes websites. The method captures webmaster behaviors from archived snapshots of a website and analyzes the changes of web servers and web applications used in the past as update histories. A classifier that predicts website abuses is finally built by using update histories from snapshots of known malicious websites before the detections. Evaluation results showed that the classifier could predict various website abuses, such as drive-by downloads, phishes, and defacements, with accuracy: a 76% true positive rate and a 26% false positive rate.

Tan, Q., Zhuang, Z., Mitra, P., & Giles, C. L. (2007). Designing Efficient Sampling Techniques to Detect Webpage Updates. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 1147–1148). New York, NY, USA: ACM. https://doi.org/10.1145/1242572.1242738

Due to resource constraints, Web archiving systems and search engines usually have difficulties keeping the entire local repository synchronized with the Web. We advance the state-of-art of the sampling-based synchronization techniques by answering a challenging question: Given a sampled webpage and its change status, which other webpages are also likely to change? We present a study of various downloading granularities and policies, and propose an adaptive model based on the update history and the popularity of the webpages. We run extensive experiments on a large dataset of approximately 300,000 webpages to demonstrate that it is most likely to find more updated webpages in the current or upper directories of the changed samples. Moreover, the adaptive strategies outperform the non-adaptive one in terms of detecting important changes.

Theobald, M., Siddharth, J., & Paepcke, A. (2008). SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information*

Retrieval (pp. 563–570). New York, NY, USA: ACM. https://doi.org/10.1145/1390334.1390431

Motivated by our work with political scientists who need to manually analyze large Web archives of news sites, we present SpotSigs, a new algorithm for extracting and matching signatures for near duplicate detection in large Web crawls. Our spot signatures are designed to favor natural-language portions of Web pages over advertisements and navigational bars. The contributions of SpotSigs are twofold: 1) by combining stopword antecedents with short chains of adjacent content terms, we create robust document signatures with a natural ability to filter out noisy components of Web pages that would otherwise distract pure n-gram-based approaches such as Shingling; 2) we provide an exact and efficient, self-tuning matching algorithm that exploits a novel combination of collection partitioning and inverted index pruning for high-dimensional similarity search. Experiments confirm an increase in combined precision and recall of more than 24 percent over state-of-the-art approaches such as Shingling or I-Match and up to a factor of 3 faster execution times than Locality Sensitive Hashing (LSH), over a demonstrative "Gold Set" of manually assessed near-duplicate news articles as well as the TREC WT10g Web collection.

Thomas Habing, Janet Eke, Matthew A. Cordial, William Ingram, & Robert Manaster. (2009). Developments in Digital Preservation at the University of Illinois: The Hub and Spoke Architecture for Supporting Repository Interoperability and Emerging Preservation Standards. *Library Trends*, *57*(3), 556–579. https://doi.org/10.1353/lib.0.0052

Funded by the National Digital Information Infrastructure and Preservation Program (NDIIPP), the ECHO DEPository Project supports the digital preservation efforts of the Library of Congress by contributing research and software to help society GET, SAVE, and KEEP its digital cultural heritage. Project activities include building Web archiving tools, evaluating existing repository software, developing architectures to enhance existing repositories' interoperability and preservation features, and modeling next-generation repositories for supporting long-term preservation. This article describes the development of the Hub and Spoke (HandS) Tool Suite, built to help curators of digital objects manage content in multiple repository systems while preserving valuable preservation metadata. Implementing METS and PREMIS, HandS provides a standards-based method for packaging content that allows digital objects to be moved between repositories more easily while supporting the collection of technical and provenance information crucial for long-term preservation. Related project work investigating the more fundamental semantic issues underlying the preservation of the meaning of digital objects over time is profiled separately in this issue (Dubin et al., 2009). [ABSTRACT FROM AUTHOR]

Thomson, S. D., & Kilbride, W. (2015). Preserving Social Media: The Problem of Access. *New Review of Information Networking*, *20*(1/2), 261–275. Retrieved from http://10.0.4.56/13614576.2015.1114842

This article is part of a 12-month study commissioned by the UK Data Service as part of the "Big Data Network" program funded by the Economic and Social Research Council (ESRC). The larger study focuses on the potential uses and accompanying challenges of data generated by social networking applications. This article, "Preserving Social Media: The Problem of Access," comprises an excerpt of that longer study, allowing the authors a space to explore in closer detail the issue of making social media archives accessible to researchers and students

now and in the future.© Sara Day Thomson and William Kilbride [ABSTRACT FROM AUTHOR]

Thornton, J. B. (2012). Archiving of Comprehensive Annual Financial Reports (CAFRs) on State Government Web Sites. *Behavioral & Social Sciences Librarian*, *31*(2), 87–95. https://doi.org/http://dx.doi.org/10.1080/01639269.2012.686244

Rising cost and declining revenues have hampered the financial affairs of state governments, forcing many to curtail services, reduce employee benefits, and trim the workforce, calling into question the fiscal sustainability of many state governments. As a result, stakeholders are demanding greater accountability and increased transparency into state government finances. An important link or communication tool between state governments and stakeholders is the comprehensive annual financial report. The comprehensive annual financial report (CAFR), produced by state governments, provides some insight into how taxpayer dollars are spent and the benefits derived therefrom. This article analyzes the extent to which the states electronically archive the CAFR on their websites and the accessibility of the reports to users searching state government websites. Adapted from the source document.

Tokarska, A., & Wilkowski, M. (2017). Seminarium “Archiwizacja Internetu” w LaCH UW TT - Seminar “Archiving the Internet” at LaCH UW. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951539759?accountid=27464

Sprawozdanie z seminarium poświęconego zagadnieniom archiwizacji Webu, zorganizowanego w siedzibie Laboratorium Cyfrowego Humanistyki UW 2 marca 2017 roku.

Toyoda, M. (2014). Multiple Media Analysis and Visualization for Understanding Social Activities. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 825–826). New York, NY, USA: ACM. https://doi.org/10.1145/2567948.2579040

The Web has involved diverse media services, such as blogs, photo/video/link sharing, social networks, and microblogs. These Web media react to and affect realworld events, while the mass media still has big influence on social activities. The Web and mass media now affect each other. Our use of media has evolved dynamically in the last decade, and this affects our societal behavior. For instance, the first photo of a plane crash landing during the “Miracle on the Hudson” on January 15, 2009 appeared and spread on Twitter and was then used in TV news. During the “Chelyabinsk Meteor” incident on February 15, 2013, many people reported videos of the incident on YouTube then mass media reused them on TV programs. Large scale collection, analysis, and visualization of those multiple media are strongly required for sociology, linguistics, risk management, and marketing researches. We are building a huge scale Japanese web archive, and various analytics engines with a large-scale display wall. Our archive consists of 30 billion web pages crawled for 14 years, 1 billion blog posts for 7 years, and 15 billion tweets for 3 years. In this talk, I present several analysis and visualization systems based on network analysis, natural language processing, image processing, and 3 dimensional visualization.

Toyoda, M., & Kitsuregawa, M. (2003). Extracting Evolution of Web Communities from a Series of Web Archives. In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia* (pp. 28–37). New York, NY, USA: ACM. https://doi.org/10.1145/900051.900059

Recent advances in storage technology make it possible to store a series of large Web archives. It is now an exciting challenge for us to observe evolution of the Web. In this paper, we propose a method for observing evolution of web communities. A web community is a set of web pages created by individuals or associations with a common interest on a topic. So far, various link analysis techniques have been developed to extract web communities. We analyze evolution of web communities by comparing four Japanese web archives crawled from 1999 to 2002. Statistics of these archives and community evolution are examined, and the global behavior of evolution is described. Several metrics are introduced to measure the degree of web community evolution, such as growth rate, novelty, and stability. We developed a system for extracting detailed evolution of communities using these metrics. It allows us to understand when and how communities emerged and evolved. Some evolution examples are shown using our system.

Toyoda, M., & Kitsuregawa, M. (2005). A System for Visualizing and Analyzing the Evolution of the Web with a Time Series of Graphs. In *Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia* (pp. 151–160). New York, NY, USA: ACM. https://doi.org/10.1145/1083356.1083387

We propose WebRelievo, a system for visualizing and analyzing the evolution of the web structure based on a large Web archive with a series of snapshots. It visualizes the evolution with a time series of graphs, in which nodes are web pages, and edges are relationships between pages. Graphs can be clustered to show the overview of changes in graphs. WebRelievo aligns these graphs according to their time, and automatically determines their layout keeping positions of nodes synchronized over time, so that the user can keep track pages and clusters. This visualization enables us to understand when pages appeared, how their relationships have evolved, and how clusters are merged and split over time. Current implementation of WebRelievo is based on six Japanese web archives crawled from 1999 to 2003. The user can interactively browse those graphs by changing the focused page and by changing layouts of graphs. Using WebRelievo we can answer historical questions, and to investigate changes in trends on the Web. We show the feasibility of WebRelievo by applying it to tracking trends in P2P systems and search engines for mobile phones, and to investigating link spamming.

Toyoda, M., & Kitsuregawa, M. (2006). What's Really New on the Web?: Identifying New Pages from a Series of Unstable Web Snapshots. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 233–241). New York, NY, USA: ACM. https://doi.org/10.1145/1135777.1135815

Tracy Seneca. (2009). The Web-at-Risk at Three: Overview of an NDIIPP Web Archiving Initiative. *Library Trends*, *57*(3), 427–441. https://doi.org/10.1353/lib.0.0045

The Web-at-Risk project is a multi-year National Digital Information Infrastructure and Preservation Program (NDIIPP) funded effort to enable librarians and archivists to capture, curate, and preserve political and government information on the Web, and to make the resulting Web archives available to researchers. The Web-at-Risk project is a collaborative effort between the California Digital Library, New York University Libraries, the Stanford School of Computer Sci- ence, and the University of North Texas Libraries. Web-at-Risk is a multifaceted project that involves software development, integration of open-source solutions, and extensive needs assessment and collec- tion planning work with the project's curatorial partners. A major outcome of this project is the Web Archiving Service (WAS), a Web

archiving curatorial tool developed at the California Digital Library. This paper will examine both the Web-at-Risk project overall, how Web archiving fits into existing collection development practices, and the Web Archiving Ser vice workflow, features, and technical approach. Issues addressed will include how the reliance on exist- ing technologies both benefited and hindered the project, and how curator feedback shaped WAS design. Finally, the challenges faced and future directions for the project will be examined

Triebsees, T., & Borghoff, U. M. (2007). Towards Automatic Document Migration: Semantic Preservation of Embedded Queries. In *Proceedings of the 2007 ACM Symposium on Document Engineering* (pp. 209–218). New York, NY, USA: ACM. https://doi.org/10.1145/1284420.1284472

Archivists and librarians face an ever increasing amount of digital material. Their task is to preserve its authentic content. In the long run, this requires periodic migrations (from one format to another or from one hardware/software platform to another). Document migrations are challenging tasks where tool-support and a high degree of automation are important. A central aspect is that documents are often mutually related and, hence, a document's semantics has to be considered in its whole context. References between documents are usually formulated in graph- or tree-based query languages like URL or XPath. A typical scenario is web-archiving where websites are stored inside a server infrastructure that can be queried from HTML-files using URLs. Migrating websites will often require link adaptation in order to preserve link consistency. Although automated and "trustworthy" preservation of link consistency is easy to postulate, it is hard to carry out, in particular, if "trustworthy" means "provably working correct". In this paper, we propose a general approach to semantically evaluating and constructing graph queries, which at the same time conform to a regular grammar, appear as part of a document's content, and access a graph structure that is specified using First- Order Predicate Logic (FOPL). In order to do so, we adapt model checking techniques by constructing suitable query automata. We integrate these techniques into our preservation framework [12] and show the feasibility of this approach using an example. We migrate a website to a specific archiving format and demonstrate the automated preservation of link-consistency. The approach shown in this paper mainly contributes to a higher degree of automation in document migration while still maintaining a high degree of "trustworthiness", namely "provable correctness".

Tryon, J. R. (2016). The Rosarium Project. *Digital Library Perspectives*, *32*(3), 209–222. https://doi.org/http://dx.doi.org/10.1108/DLP-01-2016-0001

PurposeThis paper aims to describe the Rosarium Project, a digital humanities project being undertaken at the Phillips Memorial Library + Commons of Providence College in Providence, Rhode Island. The project focuses on a collection of English language non-fiction writings about the genus Rosa. The collection will comprise books, pamphlets, catalogs and articles from popular magazines, scholarly journals and newspapers written on the rose published before 1923. The source material is being encoded using the Text Encoding Initiative (TEI) Consortium's P5 guidelines and the extensible markup language (XML) editor software <oXygen/>.Design/methodology/approachThis paper outlines the Rosarium Project and describes its workflow. This paper demonstrates how to create TEI-encoded files for digital curation using the XML editing software <oXygen/> and the TEI Archiving Publishing and Access Service (TAPAS) Project. The paper provides information on the purpose, scope, audience and phases of the project. It also identifies the resources – hardware, software and membership – needed for undertaking such a project.FindingsThis paper shows how

straightforward it is to encode transcriptions of primary sources using the TEI and XML editing software and to make the resulting digital resources available on the Web.Originality/valueThis paper presents a case study of how a research project transitioned from traditional printed bibliography to a web-accessible resource by capitalizing on the tools in the TEI toolkit using specialized XML editing software. The details of the project can be a guide for librarians and researchers contemplating digitally curating primary resources and making them available on the Web.

Tsou, J., & Vallier, J. (2016). Ether Today, Gone Tomorrow: 21st Century Sound Recording Collection in Crisis. *Music Library Association. Notes*, *72*(3), 461–483. Retrieved from https://search.proquest.com/docview/1761140761?accountid=27464

Today's music industry increasingly favors online-only, direct-to-consumer distribution. No longer can librarians expect to collect recordings on tangible media where first-sale doctrine applies. Instead, at an ever-increasing rate, librarians are discovering that music recordings are available only via such online distribution sites as iTunes or Amazon.com. These distributors require individual purchasers to agree to restrictive end-user license agreements (EULAs) that explicitly forbid institutional ownership and such core library functions as lending. What does this mean for the future of music libraries? The coauthors present an overview of an Institute of Museum and Library Services (IMLS) funded project tasked with investigating the issue, and recommend a series of next steps designed to build our professional capacity toward addressing the challenge.

Tuck, J. (2008). Web Archiving in the UK: Cooperation, Legislation and Regulation. *Liber Quarterly: The Journal of European Research Libraries, Vol 18, Iss 3-4, Pp 357-365 (2008) VO - 18*, (3–4), 357. https://doi.org/10.18352/lq.7935

The author presents an overview of web archiving in an international context, focussing on web archiving initiatives in the United Kingdom from 2001 onwards.

Turner, F. (2017). Can we write a cultural history of the Internet? If so, how? *Internet Histories*, *1*(1–2), 39–46. https://doi.org/10.1080/24701475.2017.1307540

Turner, S. (2012). Case Studies in Web Sustainability. *Ariadne*, (70). Retrieved from https://search.proquest.com/docview/1680141236?accountid=27464

At the moment organisations often make significant investments in producing Web-based material, often funded through public money, for example from JISC. We are seeing cuts in funding or changes in governmental policy, which is resulting in the closure of some of these organisations. What happens to those Web resources when the organisations are no longer in existence? Public money has often been used to develop these resources - from that perspective it would be a shame to lose them. Moreover, the resources might be needed or someone may actually want to take over the maintenance of the site at a later date. JISC previously funded three projects to look at this area through a programme called Sustaining at risk online resources [1]. One of these projects, which ran at The University of Northampton, looked into rescuing one of the recently closed East Midlands Universities Associations online resources. This resource, called East Midlands Knowledge Network (EMKN), lists many of the knowledge transfer activities of 10 of the East Midlands universities. The project looked at options on how to migrate the site to a free hosting option to make it make it more sustainable even when it is no longer available on the original host's servers. This article

looks at this work as a case study on Web sustainability and also included a case study of another project where Web sustainability was central. Adapted from the source document.

Tweedy, H., McCown, F., & Nelson, M. L. (2013). A Memento Web Browser for iOS. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 371–372). New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467764

The Memento framework allows web browsers to request and view archived web pages in a transparent fashion. However, Memento is still in the early stages of adoption, and browser-plugins are often required to enable Memento support. We report on a new iOS app called the Memento Browser, a web browser that supports Memento and gives iPhone and iPad users transparent access to the world's largest web archives.

U, L. H., Mamoulis, N., Berberich, K., & Bedathur, S. (2010). Durable Top-k Search in Document Archives. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (pp. 555–566). New York, NY, USA: ACM. https://doi.org/10.1145/1807167.1807228

We propose and study a new ranking problem in versioned databases. Consider a database of versioned objects which have different valid instances along a history (e.g., documents in a web archive). Durable top-k search finds the set of objects that are consistently in the top-k results of a query (e.g., a keyword query) throughout a given time interval (e.g., from June 2008 to May 2009). Existing work on temporal top-k queries mainly focuses on finding the most representative top-k elements within a time interval. Such methods are not readily applicable to durable top-k queries. To address this need, we propose two techniques that compute the durable top-k result. The first is adapted from the classic top-k rank aggregation algorithm NRA. The second technique is based on a shared execution paradigm and is more efficient than the first approach. In addition, we propose a special indexing technique for archived data. The index, coupled with a space partitioning technique, improves performance even further. We use data from Wikipedia and the Internet Archive to demonstrate the efficiency and effectiveness of our solutions.

Van de Sompel, H., & Davis, S. (2015). From a System of Journals to a Web of Objects. *The Serials Librarian*, *68*(1–4), 51–63. https://doi.org/10.1080/0361526X.2015.1026748

The article focuses on the web-based research process presented by Herbert Van de Sompel, Prototyping Team Leader at the Research Library of the Los Alamos National Laboratory in New Mexico, in which he explored the transition from a paper-based system to a web-based scholarly communication system. Topics discussed include de Sompel's current and ongoing projects, the core functions of the scholarly communication system, and the possibility of a long-term access to the scholarly record.

Veikkolainen, P., & Lager, L. (2016). Long-Term Preservation of the Finnish Web Archive. Finland, Europe. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Presentation at the IIPC General Assembly, Reykjavik, 12 April, 2016

VIDEIRA, T. G., & ROSA, J. M. (2017). A New Online Archive of Encoded Fado Transcriptions. *Empirical Musicology Review*, *12*(3/4), 229–243. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

A new online archive of encoded fado transcriptions is presented. This dataset is relevant as the first step towards a cultural heritage archive and as source material for the study of songs associated with fado practice using empirical, analytical and systematic methodologies (namely MIR techniques). It is also relevant as a source for artistic purposes, namely the creation of new songs. We detail the constitution of this symbolic music corpus and present how we conceived of and implemented a methodology for testing its internal consistency using a supervised classification system. [ABSTRACT FROM AUTHOR]

Vo, K. D., Tran, T., Nguyen, T. N., Zhu, X., & Nejdl, W. (2016). Can we find documents in web archives without knowing their contents? In *Proceedings of the 8th ACM Conference on Web Science - WebSci '16* (pp. 173–182). New York, New York, USA: ACM Press. https://doi.org/10.1145/2908131.2908165

Recent advances of preservation technologies have led to an increasing number of Web archive systems and collections. These collections are valuable to explore the past of the Web, but their value can only be uncovered with effective access and exploration mechanisms. Ideal search and ranking methods must be robust to the high redundancy and the temporal noise of contents, as well as scalable to the huge amount of data archived. Despite several attempts in Web archive search, facilitating access to Web archive still remains a challenging problem. In this work, we conduct a first analysis on different ranking strategies that exploit evidences from metadata instead of the full content of documents. We perform a first study to compare the usefulness of non-content evidences to Web archive search, where the evidences are mined from the metadata of file headers, links and URL strings only. Based on these findings, we propose a simple yet surprisingly effective learning model that combines multiple evidences to distinguish "good" from "bad" search results. We conduct empirical experiments quantitatively as well as qualitatively to confirm the validity of our proposed method, as a first step towards better ranking in Web archives taking metadata into account.

Wang, L., Chen, P., & Huang, L. (2009). An Efficient Clustering Algorithm for Large-scale Topical Web Pages. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1851–1854). New York, NY, USA: ACM. https://doi.org/10.1145/1645953.1646247

The clustering of topic-related web pages has been recognized as a foundational work in exploiting large sets of web pages such as the cases in search engines and web archive systems, which collect and preserve billions of web pages. However, this task faces great challenges both in efficiency and accuracy. In this paper we present a novel clustering algorithm for large scale topical web pages which achieves high efficiency together with considerately high accuracy. In our algorithm, a two-phase divide and conquer framework is developed to solve the efficiency problem, in which both link analysis and content analysis are utilized in mining the topical similarity between pages to achieve a high accuracy. A comprehensive experiment was conducted to evaluate our method in terms of its effectiveness, efficiency, and quality of result.

Weber, M. (2017). A common language. *Internet Histories*, *1*(1–2), 26–38. https://doi.org/10.1080/24701475.2017.1317118

What would a cultural history of the Internet look like? The question almost makes no sense: the Internet spans the globe and traverses any number of completely distinct human groups. It simply cannot have a single culture. And yet, like the railroad, the telegraph and the highway system before it, the Internet has been an extraordinary agent for cultural change. How should we study that process? To begin to answer that question, this essay returns to four canonical studies of earlier technologies and cultures: Carolyn Marvin's When Old Technologies Were New; Leo Marx's The Machine in the Garden; Ruth Schwarz Cowan's More Work for Mother and Lynn Spigel's Make Room for TV. In each case, the essay mines the earlier works for research tactics and uses them as jumping-off points to explore the ways in which the Internet requires new and different approaches. It concludes by speculating on the ways that the American-centric nature of much earlier work will need to be replaced with a newly global focus and research tactics to match.

Weber, M. S. (2018). Methods and Approaches to Using Web Archives in Computational Communication Research. *Communication Methods and Measures*, *12*(2–3), 200–215. https://doi.org/10.1080/19312458.2018.1447657

This article examines the role of web archives as a critical source of data for conducting computational communication research. Web archives are large-scale databases containing comprehensive records of websites showing how those websites have evolved over time. Recent communication scholarship using web archives is reviewed, demonstrating the breadth of research conducted in this space. Subsequently, a methodological framework is proposed for using web archives in computational communication research. As a source of data, web archives present a number of methodological challenges, particularly with regards to the accuracy and completeness of web archives. These problems are addressed in order to better inform future work in this area. The closing sections outline a forward-looking trajectory for computational communication research using web archives.

Webster, P. (2017). Digital Contemporary History Sources, Tools, Methods, Issues. *Temp - Tidsskrift for Historie*, *7*(14), 30–38. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Digital contemporary history: sources, tools, methods, issuesThis essay suggests that there has been a relative lack of digitally enabled historical research on the recent past, when compared to earlier periods of history. It explores why this might be the case, focussing in particular on both the obstacles and some missing drivers to mass digitisation of primary sources for the 20th century. It suggests that the situation is likely to change, and relatively soon, as a result of the increasing availability of sources that were born digital, and of Web archives in particular. The article ends with some reflections on several shifts in method and approach which that changed situation is likely to entail.

West, J. (2016). Avoiding Courseware With Slack. *Computers in Libraries*, *36*(8), 14–15. Retrieved from https://search.proquest.com/docview/1830247744?accountid=27464

Slack is a cloud-based software tool for team collaboration. The author used it as the primary tool to teach an asynchronous graduate level course called Tools for Community Advocacy at

the University of Hawaii's library and information science (UHLIS) program, and it went well. UHLIS uses courseware that is some of the best out there -- Laulima, based on Sakai -- but similar to all courseware, it has a steep learning curve and some limitations. As an adjunct who was teaching a single 6-week class, she didn't have the time available to learn to use the tool well. She decided to stick with what she knew -- which was Web sites, Google Docs, Skype, and Slack -- using Slack as the activity hub. Slack's pricing model is also attractive, which is why she mention it as a real option for libraries.

West, J. (2017). Managing Your Digital Afterlife. *Computers in Libraries*, *37*(5), 23–25. Retrieved from https://search.proquest.com/docview/1918332139?accountid=27464

More and more, people's lives are lived online. When the author's father died 6 years ago, they were pleased to find a Google Docs file with the usernames and passwords to every account he owned. He was an engineer, so this was not terribly surprising. Most of these were things such as bank accounts and cable subscriptions, but a few were email accounts and (small) social media profiles. This made a complicated time much simpler. What if they hadn't been able to access his information? Jan Zastrow has written a great article in this issue on digital estate planning, which touches on these same ideas. In this article are some specific tech tools you can use to help archive and prepare your legacy on social media sites and in content repositories.

Wilkowski, M. (2017). Wayback Machine - podstawy wykorzystania TT - Wayback Machine - the basis of use. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy : EBIB*, (172), 1. Retrieved from https://search.proquest.com/docview/1951539799?accountid=27464

Autor analizuje zaprojektowane już w 1996 r., a udostępnione publicznie pięć lat później Wayback Machine, które jest internetowym archiwum zasobów World Wide Web. Celem artykułu jest przedstawienie podstawowych metod wykorzystywania Wayback Machine do pracy z archiwalnymi wersjami stron www zabezpieczanych w tej usłudze. Wersja beta archiwum została udostępniona w październiku 2016 r. Fundacja Internet Archive obchodziła wówczas 20-lecie działań na polu archiwizacji webu.

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140

New and emerging forms of data, including posts harvested from social media sites such as Twitter, have become part of the sociologist's data diet. In particular, some researchers see an advantage in the perceived 'public' nature of Twitter posts, representing them in publications without seeking informed consent. While such practice may not be at odds with Twitter's terms of service, we argue there is a need to interpret these through the lens of social science research methods that imply a more reflexive ethical approach than provided in 'legal' accounts of the permissible use of these data in research publications. To challenge some existing practice in Twitter-based research, this article brings to the fore: (1) views of Twitter users through analysis of online survey data; (2) the effect of context collapse and online disinhibition on the behaviours of users; and (3) the publication of identifiable sensitive classifications derived from algorithms.

Winters, J. (2017). Breaking in to the mainstream: demonstrating the value of internet (and web) histories. *Internet Histories*, *1*(1–2), 173–179. https://doi.org/10.1080/24701475.2017.1305713

Woodward, N. J., Xu, W., & Norsworthy, K. (2012). On Automatically Tagging Web Documents from Examples. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1111–1112). New York, NY, USA: ACM. https://doi.org/10.1145/2348283.2348494

An emerging need in information retrieval is to identify a set of documents conforming to an abstract description. This task presents two major challenges to existing methods of document retrieval and classification. First, similarity based on overall content is less effective because there may be great variance in both content and subject of documents produced for similar functions, e.g. a presidential speech or a government ministry white paper. Second, the function of the document can be defined based on user interests or the specific data set through a set of existing examples, which cannot be described with standard categories. Additionally, the increasing volume and complexity of document collections demands new scalable computational solutions. We conducted a case study using web-archived data from the Latin American Government Documents Archive (LAGDA) to illustrate these problems and challenges. We propose a new hybrid approach based on Naïve Bayes inference that uses mixed n-gram models obtained from a training set to classify documents in the corpus. The approach has been developed to exploit parallel processing for large scale data set. The preliminary work shows promising results with improved accuracy for this type of retrieval problem.

Xiaoming, L., & Lianen, H. (2007). From Web Archive to WebDigest: Concept and Examples. In *Proceedings of the Nineteenth Conference on Australasian Database - Volume 75* (p. 11). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1378307.1378313

Much like a black hole, the Web, since its birth, has been absorbing all sorts of data (information) around the globe, ever generated along the path of human civilization. On the other hand, the digitized and networked (webbed) nature of web data, which generally means “easy to access”, gives rise to much imagination on re-discovering, re-engineering, and re-using of the oceanic information. Nevertheless, lunch is not free. The same time when we see the grand opportunities, tremendous challenges are ahead. In this talk, I’ll first introduce Web InfoMall (http://www.infomall.cn), the Chinese web archive we have been constructing since 2001. Along with the activities, we observe some useful capabilities have been developed, such as large scale web crawling and very large scale data organization. In addition, we discuss a step beyond the WebArchive, called WebDigest, which is an effort aimed at making use of the data in the web archive. With a web archive and associated capability, “web mining” here has a more or less different meaning, which spans from the structure analysis of the web to named entity and relation extractions, from spatial (if we consider URL as a space) information discovery to temporal information exhibition. The main challenge for us is around the theme of achieving reasonably good performance with affordable cost. As we are from a university lab, the underlying question is: what can be done (and how) in a university lab environment with modest resource. After all, most of the researches started from university lab. We need to understand the feasibilities and compromises while seeing the promises.

Xie, Z. (2015). A UWS Case for 200-Style Memento Negotiations ; Bulletin of IEEE Technical Committee on Digital Libraries. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Uninterruptible web service (UWS) is a web archiving application that handles server errors using the most recently archived representation of the requested web resource. The application is developed as an Apache module. It leverages the transactional web archiving tool SiteStory, which archives all previously accessed representations of web resources originating from a website. This application helps to improve the websites quality of service by temporarily masking server errors from the end user and gaining precious time for the system administrator to debug and recover from server failures. By providing value-added support to website operations, we aim to reduce the resistance to transactional web archiving, which in turn may lead to a better coverage of web history.

Xie, Z., Nayyar, K., Fox, E. A., , & 3. (2016). *Nearline Web Archiving*. United States, North America. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

In this paper, we propose a modified approach to realtime transactional web archiving. It leverages the web caching infrastructure that is already prevalent on web servers. Instead of archiving web content at HTTP transaction time, in our approach the archiving happens when the cached copy expires and is about to be expunged. Before the deletion, all expired cache copies are combined and then sent to the web archive in small batches. Since the cache is purged at much lower frequency than HTTP transactions, the archival workload is also much lower than that for transactional archiving. To further decrease the processing load at the origin server, archival copy deduplication is carried out at the archive instead of at the origin server. It is crucial to note that the cache purging process is separate from those that serve the HTTP requests. It can be, and usually is set to lower priority. The archiving therefore occurs only when the server is not busy fulfilling its more mission critical tasks; this is much less disruptive to the origin server. This approach, however, does not guarantee that the freshest copy is archived, although the cache purging policy may be adjusted to attempt to bound the freshness of the archive.

Xie, Z., Chandrasekar, P., & Fox, E. A. (2015). Using Transactional Web Archives To Handle Server Errors. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15* (pp. 241–242). New York, New York, USA: ACM Press. https://doi.org/10.1145/2756406.2756955

We describe a web archiving application that handles server errors using the most recently archived representation of the requested web resource. The application is developed as an Apache module. It leverages the transactional web archiving tool SiteStory, which archives all previously accessed representations of web resources originating from a website. This application helps to improve the website's quality of service by temporarily masking server errors from the end user and gaining precious time for the system administrator to debug and recover from server failures. By providing pertinent support to website operations, we aim to reduce the resistance to transactional web archiving, which in turn may lead to a better coverage of web history.

Xie, Z., de Sompel, H., Liu, J., van Reenen, J., & Jordan, R. (2013). Archiving the relaxed consistency web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 2119–2128). New York, NY, USA: ACM. https://doi.org/10.1145/2505515.2505551

The historical, cultural, and intellectual importance of archiving the web has been widely recognized. Today, all countries with high Internet penetration rate have established high-profile archiving initiatives to crawl and archive the fast-disappearing web content for long-term use. As web technologies evolve, established web archiving techniques face challenges. This paper focuses on the potential impact of the relaxed consistency web design on crawler driven web archiving. Relaxed consistent websites may disseminate, albeit ephemerally, inaccurate and even contradictory information. If captured and preserved in the web archives as historical records, such information will degrade the overall archival quality. To assess the extent of such quality degradation, we build a simplified feed-following application and simulate its operation with synthetic workloads. The results indicate that a non-trivial portion of a relaxed consistency web archive may contain observable inconsistency, and the inconsistency window may extend significantly longer than that observed at the data store. We discuss the nature of such quality degradation and propose a few possible remedies.

Yang, S., Chitturi, K., Wilson, G., Magdy, M., & Fox, E. A. (2012). A Study of Automation from Seed URL Generation to Focused Web Archive Development: The CTRnet Context. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 341–342). New York, NY, USA: ACM. https://doi.org/10.1145/2232817.2232881

In the event of emergencies and disasters, massive amounts of web resources are generated and shared. Due to the rapidly changing nature of those resources, it is important to start archiving them as soon as a disaster occurs. This led us to develop a prototype system for constructing archives with minimum human intervention using the seed URLs extracted from tweet collections. We present the details of our prototype system. We applied it to five tweet collections that had been developed in advance, for evaluation. We also identify five categories of non- relevant files and conclude with a discussion of findings from the evaluation.

Ye, Y., Ye, D., Zeljak, C., Kerchner, D., He, Y., & Littman, J. (2017). Web-Archiving Chinese Social Media: Final Project Report August 2017. *Journal of East Asian Libraries*, (165), 93–112. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

Yelmi, P., Ku\cscu, H., & Yantaç, A. E. (2016). Towards a Sustainable Crowdsourced Sound Heritage Archive by Public Participation: The Soundsslike Project. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (p. 71:1--71:9). New York, NY, USA: ACM. https://doi.org/10.1145/2971485.2971492

This paper explains how user-centered design approach shapes a cultural heritage project in the sustainability con - text. The project aims to protect urban sounds as intangible cultural heritage elements and turn the action of protecting sounds into a collaborative work. Sounds are of great sig - nificance in daily urban life and in culture as they carry emotions and awaken cultural memories. Thus, they de - serve to be protected and transferred to next

generations. In this paper, we first evaluate soundscapes as an intangible cultural heritage element, second we explore the presenta - tion techniques in soundscape studies in the literature, then we explain how the methods implemented step by step, and finally we introduce the two outcomes: the library archive (The Soundscape of Istanbul project) and the crowdsourced web archive (The Soundsslike project). The Soundscape of Istanbul project aims to collect and archive cultural and urban sounds of the city while The Soundsslike project is basically a crowdsourced online sound archive which in - vites people to record symbolic urban sounds and upload them to the online sound archive. This online platform was built and displayed in an exhibition by means of an interac - tive tabletop interface to learn more from users and contrib - utors, and to enrich the archive content by raising public awareness of urban sounds

Yunpeng, Q. (2012). *Web Archiving Effort in National Library of China: Paper - iPRES 2012 - Digital Curation Institute, iSchool, Toronto*. Austria, Europe: Digital Curation Institute, iSchool University of Toronto. Retrieved from http://search.ebscohost.com/login.aspx?authtype=ip,cookie,cpid&custid=s6213251&groupid=main&profile=eds

In this paper, we introduce the effort in National Library of China in recent years, including resources accumulation, software development and works in Promotion Project in China. We have developed a platform for Chinese web archiving. And we are building some sites to propagate our works to the nation. At last we figure out some questions about the web archiving in China.

Zarndt, F., Carner, D., & McCain, E. (2017). Born Digital Legal Deposit Policies and Practices. In *IFLA WLIC 2017 – Wrocław, Poland – Libraries. Solidarity. Society. in Session S18 - Satellite Meeting: News Media Section.* Wrocław: IFLA -- International Federation of Library Associations and Institutions. Retrieved from http://library.ifla.org/1905/

In 2014, the authors surveyed the born digital content legal deposit policies and practices in 17 different countries and presented the results of the survey at the 2015 International News Media Conference hosted by the National Library of Sweden in Stockholm, Sweden, April 15-16, 2015. Three years later, the authors expanded their team and updated the survey in order to assess progress in creating or improving national policies and in implementing practices for preserving born digital content. The 2017 survey reach has been broadened to include countries that did not participate in the 2014 survey. To optimise survey design, and allow for comparability of results with previous surveys, the authors briefly review 17 efforts over the last 12 years to understand the state of digital legal deposit and broader digital preservation policies (a deeper analysis will be provided in a future paper), and then set out the logic behind the current survey.

Zhang, Y., Jatowt, A., & Tanaka, K. (2017). Temporal Analog Retrieval Using Transformation over Dual Hierarchical Structures. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 717–726). New York, NY, USA: ACM. https://doi.org/10.1145/3132847.3132917

In recent years, we have witnessed a rapid increase of text con- tent stored in digital archives such as newspaper archives or web archives. Many old documents have been converted to digital form and made accessible online. Due to the passage of time, it is however difficult to

effectively perform search within such collections. Users, especially younger ones, may have problems in finding appropriate keywords to perform effective search due to the terminology gap arising between their knowledge and the unfamiliar domain of archival collections. In this paper, we provide a general framework to bridge different domains across-time and, by this, to facilitate search and comparison as if carried in user's familiar domain (i.e., the present). In particular, we propose to find analogical terms across temporal text collections by applying a series of transformation procedures. We develop a cluster-biased transformation technique which makes use of hierarchical cluster structures built on the temporally distributed document collections. Our methods do not need any specially prepared training data and can be applied to diverse collections and time periods. We test the performance of the proposed approaches on the collections separated by both short (e.g., 20 years) and long time gaps (70 years), and we report improvements in range of 18%-27% over short and 56%-92% over long periods when compared to state-of-the-art baselines.

Zhao, Y., & Hauff, C. (2015). Sub-document Timestamping of Web Documents. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15* (pp. 1023–1026). New York, New York, USA: ACM Press. https://doi.org/10.1145/2766462.2767803

Knowledge about a (Web) document's creation time has been shown to be an important factor in various temporal information retrieval settings. Commonly, it is assumed that such documents were created at a single point in time. While this assumption may hold for news articles and similar document types, it is a clear oversimplification for general Web documents. In this paper, we investigate to what extent (i) this simplifying assumption is violated for a corpus of Web documents, and, (ii) it is possible to accurately estimate the creation time of individual Web documents' components (so-called sub-documents).

Zhou, K., Grover, C., Klein, M., & Tobin, R. (2015). No More 404s. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15* (pp. 233–236). New York, New York, USA: ACM Press. https://doi.org/10.1145/2756406.2756940

The citation of resources is a fundamental part of scholarly discourse. Due to the popularity of the web, there is an increasing trend for scholarly articles to reference web resources (e.g. software, data). However, due to the dynamic nature of the web, the referenced links may become inaccessible ('rotten') sometime after publication, returning a "404 Not Found" HTTP error. In this paper we first present some preliminary findings of a study of the persistence and availability of web resources referenced from papers in a large-scale scholarly repository. We reaffirm previous research that link rot is a serious problem in the scholarly world and that current web archives do not always preserve all rotten links. Therefore, a more pro-active archival solution needs to be developed to further preserve web content referenced in scholarly articles. To this end, we propose to apply machine learning techniques to train a link rot predictor for use by an archival framework to prioritise pro-active archiving of links that are more likely to be rotten. We demonstrate that we can obtain a fairly high link rot prediction AUC (0.72) with only a small set of features. By simulation, we also show that our prediction framework is more effective than current web archives for preserving links that are likely to be rotten. This work has a potential impact for the scholarly world where publishers can utilise this framework to prioritise the archiving of links for digital preservation, especially when there is a large quantity of links to be archived.

Zhou, L., Dang-Nguyen, D.-T., & Gurrin, C. (2017). A Baseline Search Engine for Personal Life Archives. In *Proceedings of the 2nd Workshop on Lifelogging Tools and Applications - LTA '17* (pp. 21–24). New York, New York, USA: ACM Press. https://doi.org/10.1145/3133202.3133206

In lifelogging, as the volume of personal life archive data is ever increasing, we have to consider how to take advantage of a tool to extract or exploit valuable information from these personal life archives. In this work we motivate the need for, and present, a baseline search engine for personal life archives, which aims to make the personal life archive searchable, organizable and easy to be updated. We also present some preliminary results, which illustrate the feasibility of the baseline search engine as a tool for getting insights from personal life archives.

Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Legal Information Management*, *14*(2), 88–99. https://doi.org/http://dx.doi.org/10.1017/S1472669614000255

Abstract It has become increasingly common for a reader to follow a URL cited in a court opinion or a law review article, only to be met with an error message because the resource has been moved from its original online address. This form of reference rot, commonly referred to as "linkrot", has arisen from the disconnect between the transience of online materials and the permanence of legal citation, and will only become more prevalent as scholarly materials move online. The present paper*, written by Jonathan Zittrain, Kendra Albert and Lawrence Lessig, explores the pervasiveness of linkrot in academic and legal citations, finding that more than 70% of the URLs within the Harvard Law Review and other journals, and 50% of the URLs within United States Supreme Court opinions, do not link to the originally cited information. In light of these results, a solution is proposed for authors and editors of new scholarship that involves libraries undertaking the distributed, long-term preservation of link contents. [PUBLICATION ABSTRACT]