Digital Humanities: Notes on Web Archiving

FRANCESCA ZANTEDESCHI
University of Amsterdam

Introduction: Making History in the Digital Revolution Era

In May 1968, in the French weekly magazine *Le Nouvel Observateur*, the French historian Emmanuel Le Roy-Ladurie published the article, 'La fin des érudits. L'historien de demain sera programmeur ou ne sera plus' ('The end of scholars. The historian of tomorrow will be a programmer or will not exist'). This telling title aimed at highlighting the new challenges that the ongoing technological revolution posed to historians. In the article, the prominent figure from the École des Annales pointed out the emergence of a new type of historian, a kind of 'engineer in history' – quite different from the erudite scholars of the past – capable of 'manipulating' the vast amounts of data that computers made it possible to store.[1]

During those years, only a small number of scholars used computers for historical research, mainly to apply quantitative methods for linguistic and literary analyses.[2] However, since the 1990s, the use of consultation and research techniques linked to computers and the web has pervaded historical research.[3] This has impacted upon the way historical sources are produced, published, stored and analysed.[4] As a consequence, historians must now not only use computational tools and computer techniques to process data,[5] but also learn to work with new sources such as websites and their archives. As Serge Noiret observes, while digital history, which uses the analytical potential of computers, 'concerns relatively few scholars', there are many more historians who rely on historical methods 'that have been revised and simplified by the

technological and communicative revolution of the digital age and the web'.[6]

The web has provided access to many documentary sources, owing to the massive digitalisation of documents and the creation of virtual archives and libraries. It has greatly improved communication and the sharing of research results. However, it has also contributed to extending 'in an unquantifiable way the primary sources, not only textual, that can be interrogated transversally'.[7] As a result, the web is not just a research and communication tool but has also become a primary source itself.

In this text, we will not discuss the profound changes that new technologies and the internet have brought about in historical research. Instead, we will focus on some methodological issues that are involved in web archiving. The preservation of websites is considered a cultural and historical necessity, and it requires a radical revision of traditional preservation practices, as pointed out by Julien Masanès.[8] In this regard, we will pay particular attention to the processes of selection and disposal, the use of redundancy to retrieve as much information as possible, and the problems that incorrect web archiving practices pose in terms of economic, social, and technological sustainability. Additionally, we will explore the benefits of correct metadata and the cataloguing of archived websites. This approach offers significant advantages in terms of the usability of archived web content, especially in fields of research like history, sociology, and anthropology. To illustrate our arguments, we will use the web archives of the European Union as a case study.

The World Wide Web: A New Historical Source?

The World Wide Web has certainly changed the way we obtain, access and exchange information. Websites have become indispensable sources for anyone interested in contemporary history.[9] On websites and social

media platforms, decisions are made, important news is disseminated and political campaigns are conducted. Institutional websites of public and private entities often serve as both the headquarters and archives of these organisations, as they store all official documents. In some cases, institutions are legally required to archive the content of their websites.[10]

However, the web is an intricate and constantly changing entity, which can cause the information to be lost or degraded just as quickly as it becomes readily accessible. Therefore, how can online information be saved and preserved accurately – that is, ensuring its authenticity,[11] integrity, and readability – over time? This is achieved through web archiving, which involves using specialised software, known as harvesters or crawlers, to scan and capture web pages, packaging the data in a format that is suitable for preservation.[12] Quality control must then be carried out in order to determine whether what was intended to be captured has been captured and whether the websites have been captured in their entirety. The addition of both bibliographic and semantic metadata completes the operation and facilitates information retrieval for users.

Archiving websites is a complex operation that involves numerous challenges. According to Matthew S. Weber, web archiving is 'a complicated space for research' and entails various issues, such as 'the need to continue developing a knowledge base, the importance of increased accessibility and scalability, the role of developing intersections with existing domains of research, and the need for approaches that aide in establishing the validity and reliability of research conducted via Web archives'.[13]

Archiving websites poses challenges not only for scholars who want to use the web as a research tool or source but also for archivists who aim to preserve web content because of the complexity of the issues involved. Firstly, it is impossible to save and preserve all the information available on the web; hence, there is a need to select and define the scope and field

of application. Secondly, technological challenges arise because websites are dynamic and frequently updated, and they may contain a plethora of pages and various media types including text, graphics, audio, video and links. Thirdly, archivists must take into account the legal copyright and intellectual property rights of website owners, as well as data protection and privacy issues, particularly for social media. As emphasised by Maureen Pennock, 'the primary non-technical problem that web archives must address' is the issue of legality. This issue is related to the 'legal right to make copies of content and provide access to it independently of the original site and without the explicit permission of the owner'. While some websites have resolved this issue by displaying licenses or providing information on copyright, the solution often depends on the country in question and the competencies of the collecting institution.[14]

In the field of web archiving, the website archiving project conducted by the Internet Archive (also known as Archive.org because of its domain name) has gained great visibility. It is a non-profit organisation that collects and preserves web content, not so much because of legislative requirements, but rather out of a 'social' interest whereby it records 'the evolution and content of the Internet in its entirety and makes it available to users'. It was founded in 1996 by Brewster Kahle and Bruce Gilliat in the United States with the goal of providing 'universal access to all knowledge'. The Internet Archive serves as a digital library that contains manuscripts, images, audio, video, and software programs in digital format. It also provides digitisation services to many institutions and acts as a platform for website archiving. To date, the Internet Archive has stored over 41 million books and texts, 14.7 million audio recordings, 8.4 million videos, 4.4 million images, 890,000 software programs, and 735 billion web pages.[15] In 2001, the Internet Archive was enriched with the Wayback Machine, a veritable digital time machine that captures websites at more or less regular intervals, enabling users to see not only how they change over time, but also the information they contain. A huge collection of snapshots of the web, Wayback Machine makes it possible to maintain the historicity of a website and retrieve old

versions of websites, owing also to collaboration with over a thousand libraries and other partners through the Archive-It program, a web archiving service for collecting and accessing cultural heritage on the web.[16]

Archive-It is used by various organisations, including the Publications Office of the European Union, to archive websites related to EU institutions, agencies, and bodies.[17] These archived websites are grouped into collections. Currently, there are five thematic collections, the main one being the European Union, which collects 250 websites hosted on the europa.eu domain.[18] Monica Steletti, along with Samir Musa, is one of the archivists who designed the pilot project for archiving EU websites at the Historical Archives of the European Union. She explained that initially, merit, urgency, legal requirements, and preservation needs were the criteria used for the selection of EU websites. In the short term, it was decided to archive the europa.eu domain along with other sites that were about to be decommissioned, as well as the websites of institutions, agencies, and entities associated with the activities of European institutions that did not fall under the europa.eu domain. In the medium term, institutional databases, intranet networks and social media platforms were also included in the archiving process.

Selection and disposal in the digital environment and web archiving

The internet and its archives have been at the heart of research (historical, sociological, anthropological, political, etc.) as primary sources since the 1990s. Web pages are important not only for the information they provide or the messages they deliver but also for the way they are designed. The structure and design of web pages reflect the thought that went into their creation and development. This is why both archivists and researchers agree on the importance of capturing and

preserving web content in order to maintain 'the integrity and continuity of historical, cultural, and academic documents'.[19]

However, it is impossible to collect and preserve everything. Therefore, how can one determine which websites to collect and preserve, whether they are institutional or focused on specific events? Also, how long should they be preserved and how can one determine what will be significant for future research? In other words, what criteria should be used for selection and when should they be applied? Should the selection be made during the document transmission phase, from the deposit archive to the historical archive (as with 'traditional' archives)? Or should a selection be made upstream, before capturing websites, to avoid being overwhelmed by an unmanageable amount of data? Alternatively, why not consider a process similar to pruning, as seen in the non-digital environment, which involves eliminating some documents from the file before transferring them to the deposit archive?[20]

According to Monica Steletti, the selection policy and method are crucial elements of a web archiving program and must be continuously updated. Steletti identifies four selection methods: non-selective, selective, thematic, and hybrid.[21] She also explains that selection criteria should be defined based on the lifecycle of a website. Some institutional websites, for example, retain content for a longer period, while others are more transient – especially websites dedicated to specific events (COVID-19, the Russo-Ukrainian War, and so on) or specific topics (fundamental rights, environment, health, science, etc.). It is also essential to consider the dynamic nature of website content, such as newspapers and social media, which change content multiple times a day. Additionally, monitoring websites that are at risk of decommissioning, such as pages with expired funding or completed projects, is important.

For Maureen Pennock, 'the selection policies of web archives are generally consistent with broader organisational collection policies'. She distinguishes between two main types of collections: domain-based and

selective. Domain-based collections gather websites associated with a specific country (whether ending with the national domain suffix or focused on that country), while selective collections are thematic and tend to focus on a particular topic or event, like Brexit or the Olympics.[22] Lorenzana Bracciotti, while indicating the same selection criteria, nevertheless warns against their limitations: while domain collections, which are usually very large, run the risk of being incomplete for this very reason; selective (or thematic) collections, which are smaller, run the risk of 'being influenced by subjective collection criteria'.[23]

Be they domain-based or selective (or hybrid/mixed), websites are collected automatically by specifically programmed robots (by humans). This implies, among other things, the impossibility of knowing exactly what content will be available at the time of the robot's passage since 'the collection boundary is fixed a priori', as well as the value of the collected information. As various authors have noted, these factors significantly impact the work carried out in archives and libraries, as it requires the development of new skills and professional roles, from operators capable of handling these automated processes to experts who can oversee large-scale content indexing and address the issues related to preserving digital materials in the long term (technological and format obsolescence, etc.).[24]

Disposal in web archiving is equally if not more problematic than the selection operation.[25] This issue is thorny and has not yet been adequately addressed in the current literature. In 'traditional' archiving, disposal is usually performed on deposit archive documents and is (or should be) the result of a 'selective rationalisation' of documents, which leads to the physical elimination of transitional and instrumental documentation. Gilda Nicolai has identified four types of disposal of analogue documents: voluntary (the most common type and the only one where the archivist's 'selective' intention is evident), natural or involuntary, negligent and unintentional.[26] Despite the fact that digital documents are ill-suited to traditional modes of preservation, and therefore also to those of selection and disposal, for Nicolai 'the digital

dimension does not in principle change the nature of the activity of evaluation, selection and disposal, even if it imposes new modes of intervention and new tools'.[27] The highly volatile and fragile nature of the digital document, for instance, means that, for the purposes of a correct selection (and preservation), information 'relating to the context of a given set of documents or that can be inferred from the documents themselves' must be carefully collected and assessed.[28] Moreover, unlike in the analogue context, preserving a digital document requires a voluntary and deliberate act of preservation to avoid issues arising from technological and format obsolescence.[29]

The difficulties that characterise disposal in the digital environment are amplified when dealing with the archived web, due to the very nature of this type of source. It is plausible to think that by operating a voluntary (methodical and reasoned) selection upstream, disposal could be reduced to the physical elimination of 'transient and instrumental documents' (broken links, pages that are no longer accessible, etc.), which is similar to what occurs in non-digital contexts. While it would require significant human resources and time to implement selection criteria, it could ultimately make disposal in web archiving more manageable.

Web archiving: some open issues

Providing precise, clear, and straightforward metadata is of paramount importance while archiving web resources. This is crucial to ensure their proper preservation and usability. An accurate and comprehensive description is also required for the same purpose. This issue is relatively new in terms of treatment, but it is extremely important. The first guidelines for web archiving metadata, provided by the Web Archiving Metadata Working Group (WAM), date back only to 2018. These guidelines have clear objectives, including the development of neutral practices (in terms of community and standards) for descriptive

metadata for archived web content; this provides 'a bridge between bibliographic and archival approaches to description', and uses a 'scalable approach that requires neither in-depth description nor extensive changes to records over time'.[30]

It becomes evident that metadata and description are fundamental elements for the proper storage and usability of archived websites when consulting the archives of the websites of the European Union – where these elements are severely lacking. For example, one of the first things that stands out is the lack of any description, even a concise one, of the collections. Additionally, descriptions for the captured sites are not always available either. The frequency of website captures is unclear, and there are fluctuations from capturing web pages for two consecutive days to having gaps of several weeks between captures. The metadata provided by the Publications Office (OP) is quite basic, falls short of the ideal proposed by WAM[31] and is often limited to Title, URL, number of captures, time span, Group, Subject, and the Institution to which the site belongs (such as the European Commission, European Data Protection Supervisor, or a generic 'Agencies and other bodies'). Particularly troubling is the frequent absence of the Description metadata, which instead could aid users in navigating between and within collections. According to WAM, Description is a crucial element as it allows for a clear explanation of the content and context of the site or collection. Description can include information about the source, historical or biographical information about the organisation or person responsible for creating the web content, objectives pursued, selection criteria, and reasons behind archiving a particular website. By investing more in descriptive activity and taking time to reflect on the use of metadata, many of the issues regarding user consultation and archive usability could be reduced.

It is possible that the reason for the methodological deficiencies in describing and providing metadata for EU websites is that the archiving operations are managed by the Publications Office of the European Union, which is the official provider of editorial services for all EU

institutions, organs, and agencies, rather than an archival or library institution. It is plausible to assume that the adoption of a more archival approach – and also greater interaction with the world of research – would resolve many of the remaining issues. As pointed out by Lorenzana Bracciotti, the archival discipline indeed pays particular attention to contextualising the resource (through descriptions of the producing entity and archival history), documenting relationships, and recording acquisition and preservation processes.[32]

The aforementioned issues are not exclusive to archiving European Union websites but are widespread in web archiving in general. Unfortunately, the situation has not improved much since Molly Bragg, Kristine Hanna and other authors pointed out ten years ago that institutions face difficulties in developing best practices and methodologies for web archiving programs. This is due in part to some organisation stakeholders not fully recognising the importance of web archiving for their digital preservation activities, resulting in limited or no funding.[33]

One also might think that in the absence of effective selection or disposal criteria, and in the face of a massive risk of loss or degradation of websites (and the information they contain), redundancy could prove to be a valuable tool for safeguarding and retrieving the maximum amount of information. In web archiving, redundancy manifests as capturing content multiple times and is inevitable as it results from the replication of documents. While this ensures the availability of information, it also leads to storage overabundance, making it difficult to analyse content and reducing the quality of search results. Additionally, this raises sustainability issues at environmental, social and economic levels.

In a workshop organised by Archive-it, archivist Jillian Lohndorf emphasised that considering sustainability when it comes to web archiving should not be an 'extra', but rather a '"way of thinking" that we apply towards planning and implementation of all our web archiving program activities'. Sustainability is about planning and documenting

roles, responsibilities, processes, and management and preservation policies. It involves 'responsible planning around how a resource (or project, program, or other entity) should be managed over time. To that end, every sustainability plan needs to include information not only about how to manage transitions and ensure persistence but also about how and when the resource might be discontinued and sunset'.[34]

In a study conducted a few years ago, it was found that the issue of sustainability for archives, repositories, and digital libraries is rarely discussed in the Library and Information Science (LIS). The study's authors define sustainability as 'the continued operation of a collection, service, or organization related to digital libraries, archives, and repositories, over time and in relation to ongoing challenges. We include, but do not limit ourselves to, bit- or item-level preservation or within the context of an organisation or project'.[35] They emphasise that the complexity and multidimensional nature of the sustainability concept makes it difficult to develop methodologies to analyse it, as well as the lack of conceptual models. In another study, also conducted on LIS literature, Eschenfelder and Shankar also highlight the importance of paying attention to the sustainability of the institutions that curate, preserve, organise, and provide access to archived data in order to ensure their persistence. The research revealed a varied and complex landscape, where the concept of sustainability applied by these institutions ranges from their internal capabilities to environmental monitoring, external environmental turbulence, governance and relationships, and changes in scientific communities and their data.[36]

In 2010, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access produced a final report which identified six conditions required for the economic sustainability of digital preservation. These conditions include 'recognition of the benefits of preservation by decision makers, selection of materials with long-term value, incentives for decision makers to act in the public interest, appropriate organisation and governance of preservation activities, ongoing and efficient allocation of resources for preservation, and timely actions to

ensure access'.[37] Even though the BRTF mainly focused on the issue of economic sustainability, it is clear that preserving valuable digital materials today is crucial for ensuring access to them in the future. Ongoing and efficient allocation of resources for preservation is key to achieving this.[38]

Conducting a thorough examination of digital preservation methods can help improve digital preservation's environmental sustainability. This can be achieved by reducing the adverse environmental effects such as using electronic devices and resources, the support infrastructure required for cloud and network storage, the raw materials and energy needed for these infrastructures, and other factors.[39]

Conclusion

As Stefano Allegrezza pointed out, the selecting and disposing of content in digital environments, particularly web archiving, are complex operations requiring scientific and methodological reflection. This reflection should also encompass aspects relating to web archiving, such as copyright and personal data protection. Furthermore, it is essential to constantly reassess the goals of web archiving, particularly in the context of long-term projects. Websites are not the only things that evolve; the questions that guide their capture evolve and so do our objectives.

Creating well-curated and representative collections in web archiving should be a universally pursued goal. However, there is no single criterion that can be followed to achieve that goal. In this regard, Gilda Nicolai emphasises the failure of efforts to provide universal educational guidelines or criteria. However, she also rightly reminds us that 'all acts of evaluation are conditioned by the context and constrained by social elements, by both international and national laws and regulations and, for digital archives, by technology' and that, for an evaluation to be successful, it is necessary to carry it out 'in full knowledge of the

contextual conditions'.[40] I would also add that a successful evaluation is only possible through close and ongoing collaboration between archivists, computer scientists, legal experts, and scholars from different disciplines and research areas. Without this collaboration and long-term methodological and scientific reflection, it is impossible to carry out even a modest web archiving project.

One final consideration before concluding: the loss of websites and web pages is inevitable, just as it was and continues to be with analogue and digital documents and information.[41] Julien Masanès has also highlighted this aspect, pointing out that at the end of the nineteenth century, the rise of serial publications such as newspapers and periodicals led to doubts among the librarian community about their intellectual scope, and reactions (for example about the difficulty in cataloguing these publications), similar to our approach to dealing with the vast amount of information available on the web today. Of course, there are clear differences between the publishing boom of the past and the vast amount of information and documents produced and available today on the web. The 'bulimia' that often characterises web archiving initiatives does not lead to well-curated and representative collections. For example, the web archiving operation carried out by the PO for the EU's websites is futile because they 'capture' and collect as many websites as possible without any methodological reflection, criteria defining its objectives, and metadata that enable the usability of the archived websites. Furthermore, concerning the issue of information loss in the digital age, Masanès has emphasised how the 'expansion of the online publication's sphere' has also led to the 'mechanical drop in average number of readers of each unit of published content. Some pages are even not read by any human nor indexed by any robot at all'.[42] This is a 'physiological' loss of documents and information, proportional to the quantity (and quality?) of documents and information produced today.

In conclusion, there is a recurring theme in the specialist literature regarding digital archives that cannot be ignored and can easily be applied to the archiving of websites. To proceed with web archiving, it is

essential to reflect on the purpose of archiving, define conceptual models, methodological and scientific frameworks, and prepare forward-looking yet flexible actions to adapt to the needs of the moment. The specialists in this field appeal to all those involved in the creation, management and preservation of digital archives and archived websites. It is a warning that needs to be heard now more than ever before.

Endnotes

[1] Emmanuel Le Roy Ladurie, 'La fin des érudits. L'historien de demain sera programmeur ou ne sera plus', *Le Nouvel Observateur* (8 May 1968).

[2] Serge Noiret, 'Storia contemporanea digitale', in *Il web e gli studi storici: guida critica all'uso della rete*, ed. by R. Minuti (Roma: Carocci, 2015), 267–300.

[3] Rolando Minuti, 'Introduzione', in *Il web e gli studi storici: guida critica all'uso della rete*, ed. by Rolando Minuti (Roma: Carocci, 2015), 11–19.

[4] Minuti, 'Introduzione'; Serge Noiret, 'Homo digitalis', in *La storia in digitale. Teorie e metodologie*, ed. by Deborah Paci (Milano: Unicopli, 2019), 9–18.

[5] Deborah Paci, 'Introduzione', in *La storia in digitale. Teorie e metodologie*, ed. by Deborah Paci (Milano: Unicopli, 2019),19–24.

[6] Serge Noiret, 'Storia contemporanea digitale', in *Il web e gli studi storici: guida critica all'uso della rete*, ed. by Rolando Minuti (Roma: Carocci, 2015), 267–300.

[7] Noiret, 'Storia contemporanea', 269.

[8] Julien Masanès, 'Web Archiving: Issues and Methods', in *Web Archiving*, ed. by Julien Masanès (Berlin/ Heidelberg: Springer, 2006), 1.

[9] See Niels Brügger, 'Understanding the Archived Web as a Historical Source', in *The SAGE Handbook of Web History*, ed. by Niels Brügger & Ian Milligan (London: SAGE publications, 2019), 16–29; Niels Brügger, 'When the Present Wb is Later the Past: Web Historiography, Digital History, and Internet Studies', *Historical Social Research/Historische Sozialforschung*, 37/4 (2012), 102–117.

[10] Presentation by Stefano Allegrezza at the 'Primavera archivistica 2022. La selezione in ambiente digitale', 16 June 2022, https://www.youtube.com/watch?v=TNmo2OIfm4U

[11] As Maureen Pennock, from the Digital Preservation Coalition, observes, 'The form of an "authentic" experience however, is anything but clear, and some academic issues remain. What for example, is an "authentic" archived website? Is an archived website still authentic if some of the links are broken, or content missing? How can criteria for authenticity even be determined when sites do not exist as static objects but are generated dynamically and rendered differently for different users? These issues are still being explored'; Maureen Pennock, *Web-Archiving* (Great Britain: Digital Preservation Coalition, 2013), 5; http://dx.doi.org/10.7207/twr13-01.

[12] *Web crawling* is the most widespread, but also the most complex form of web archiving. There are other, simpler ways to archive the web, for instance by creating an image, either in the form of a screenshot or using specific software; or by downloading individual files from the web; Brügger, 'Understanding the Archived Web', 19.

[13] Matthew S. Weber, 'Web Archives: A Critical Method for the Future of Digital Research', *WARCnet Papers* (2020), 5.

[14] 'Copyright can pose further problems when additional or altered copies of the work are generated as part of a long-term preservation strategy'; Pennock, *Web-Archiving*, 9.

[15] https://archive.org/about/. Over the years, web archiving has evolved from a 'traditional' (documentary) approach 'to a "temporal archive" logic that seeks to fully capture the instability of the web, developing dynamic archiving methods, just as the Web itself is dynamic'; Francesca Musiani, Camille Paloque-Bergès, Valérie Schafer and Benjamin G. Thierry, *Qu'est-ce qu'une archive du web?*, Collection 'Encyclopédie numérique' (2019), 32: http://books.openedition.org/oep/8713.

[16] There are numerous tools and services for web archiving, both commercial and open source; in addition to those for retrieving websites, there are also solutions for consulting archived websites; see C. Landino, 'Strumenti per il Web Archiving: alcune soluzioni', *Il mondo degli archivi*, 6 July 2018:

http://www.ilmondodegliarchivi.org/rubriche/archivi-digitali/650-strumenti-per-il-web-archiving-alcune-soluzioni.

[17] The initiative to preserve the websites of European institutions is actually due to the Historical Archives of the European Union (HAEU), which began capturing them in 2013. In 2018, the Publication Office of the EU (OP) took over this task.

[18] According to the website's presentation, the other four collections include Brexit archive, Horizon 2020, Presidencies of the Council of the EU, and Publications. The websites are acquired in all available linguistic versions. However, as of today (December 2023), there is no longer any trace of these five collections (visible at the beginning of 2023). After choosing whether to search for 'European institutions, agencies, and bodies' or by 'Topic', the only possible search is by keywords; https://op.europa.eu/en/web/euwebarchive/about-eu-web-archive.

[19] Jackie Dooley & Kate Bowers, *Descriptive Metadata for Web Archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*, Dublin, OH, OCLC Research (2018): https://doi.org/10.25333/C3005C.

[20] However, as Stefano Allegrezza notes, 'in the digital context there is no trace of the thinning operation. In order to be able to carry out the thinning operation in the digital context, it is necessary that the documents belonging to the same file have different retention times. This ensures that only the documents that are supposed to be thinned out are deleted while the others remain untouched'; Allegrezza, 'Primavera archivistica 2022. La selezione in ambiente digitale'.

[21] The choice made at the time for EU sites was the hybrid method: three captures per year for 69 European institutions/agencies/bodies; for important events, such as the European Parliament elections, two ad hoc captures before and after elections. Monica Steletti, 'Archiviazione dei siti delle istituzioni europee. Il progetto pilota degli Archivi storici dell'Unione Europea tra principi e realizzazione', presentation at Corso ANAI, 18–19 May 2015. I am grateful to Monica Steletti for sharing the preparatory material for the HAEU pilot project with me.

[22] As Pennock explains, 'the main issue in establishing scoped collections is the artificial limits they impose, even at a national domain level. The Internet does not respect collection and national boundaries! Sites in these collections will

frequently link to other sites that are not captured as part of a collection and this can be frustrating for users who inevitably then encounter broken links'; Pennock, *Web-Archiving*, 10.

[23] Lorenzana Bracciotti, 'Web Archiving. Conservazione e uso di una nuova fonte', *Officina della storia*, 10 (2019): https://www.officinadellastoria.eu/it/2019/01/10/il-web-archiving-conservazione-e-uso-di-una-nuova-fonte/.
Very often, especially in the case of state-led website archiving initiatives, a 'mixed' selection method is adopted, i.e., in addition to national domains (.be, .nl, .co.uk, .fr, .pt, etc.), websites of particular interest or which are considered relevant to that country are manually collected; Musiani et al., *Qu'est-ce qu'une archive du web?*, 18–20. As already noted, this is also the criterion adopted for archiving EU websites, or websites linked to it in some way.

[24] Musiani et al., *Qu'est-ce qu'une archive du web?*, 25–26.

[25] According to Gilda Nicolai, selection should be approached 'as a question of preservation rather than as a control of proliferation, especially in the public sector'; Gilda Nicolai, 'Dagli archivi tradizionali all'ambiente digitale: la valutazione e selezione nel contesto internazionale', *Archivi*, XII/1 (2017), 31.

[26] Whatever its typology, it is a dead-end operation. In the digital environment, making a definitive disposal is somewhat more complicated. Digital documents are vulnerable but also extremely persistent, as their content, structure and form exist separately in the system. Therefore, they must be decisively destroyed in order to prevent them from lingering in the system. For more details on this, you can refer to the InterPARES project, which was directed by Luciana Duranti. The first phase of the project, which took place from 1999 to 2001, focused on preserving the authenticity of electronic documents that were no longer needed by the body that created them to fulfill its mandate, mission or purpose. Among the results produced were precisely 'conceptual requirements for authenticity and methods for the selection and preservation of authentic electronic documents'; http://www.interpares.org.

[27] Nicolai, 'Dagli archivi tradizionali', 32.

[28] 'Two types of information are produced from the evaluation process: information on the decision itself and information on the electronic documents selected for preservation, transferred by the producer to the preservation

organisation together with the documents themselves. The latter represents the information required in order to keep the documents in authentic form and includes the terms and conditions of transfer to be referred to in order to determine whether a transfer actually contains the intended documents'; Nicolai, 'Dagli archivi tradizionali', 40.

[29] Nicolai, 'Dagli archivi tradizionali all'ambiente digitale', 40. In this regard, Stefano Allegrezza illustrates the advantages of adopting the WARC (Web ARChive) format for long-term web archiving: it is a non-proprietary, open standard format that promotes transparency and reduces risk of obsolescence; hence, it is highly compatible with a long-term digital preservation process; S. Allegrezza, 'Nuove prospettive per il Web archiving: gli standard ISO 28500 (formato WARC) e ISO/TR 14873 sulla qualità del Web archiving', *DigItalia*, 10/1–2 (2015), 46–91. Retrieved from https://digitalia.cultura.gov.it/article/view/1473.

[30] These are just some of the objectives pursued by WAM. For further details, consult the guidelines: Dooley & Bowers, *Descriptive Metadata*, 7.

[31] WAM proposes fourteen description metadata: Collector, Contributor, Creator, Date, Description, Extent, Genre/Form, Language, Relation, Rights, Source of description, Subject, Title, URL.

[32] Bracciotti, 'Web Archiving', 4.

[33] https://archive-it.org/learn-more/publications/web-archiving-life-cycle-model/.

[34] Jillian Lohndorf, 'Building a Sustainable Web Archiving Program', https://support.archive-it.org/hc/en-us/articles/4402736898068-Building-a-Sustainable-Web-Archiving-Program#h_01F92J9T15JVEADT1RYSD03CES.

[35] Kristin R. Eschenfelder et al., 'What are we talking about when we talk about sustainability of digital archives, repositories and libraries?', in *Proceedings of the Association for Information Science and Technology* (2016): https://doi.org/10.1002/pra2.2016.14505301148.

[36] Kristin R. Eschenfelder & Kalpana Shankar, 'Designing Sustainable Data Archives: Comparing Sustainability Frameworks', iConference 2016, Philadelphia: https://minds.wisconsin.edu/handle/1793/74285.

[37] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, La Jolla, Calif.: Blue Ribbon Task Force on Sustainable Digital Preservation and Access (Francine Berman and Brian Lavoie, co-chairs), 2010, 73 ff., https://discovery.ucl.ac.uk/id/eprint/19116/1/19116.pdf.

[38] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, 1.

[39] See Keith L. Pendergrass et al., 'Toward Environmentally Sustainable Digital Preservation', *The American Archivist*, 82/1 (2019), 165–209.

[40] Gilda Nicolai, 'Dagli archivi tradizionali all'ambiente digitale: la valutazione e selezione nel contesto internazionale', *Archivi*, 12/1 (2017), 31.

[41] Bracciotti, 'Web Archiving', 6.

[42] Masanès quotes a study by Boufkhad and Viennot (2003), who 'have shown using the logs and file server of a large academic website that 5% of pages were only accesses by robots, and 25% of them were never accessed at all. This provisions the indeterminacy of future reader's interests'; J. Masanès, 'Web Archiving: Issues and Methods', 4.