The Curious Case of Archiving .eu

Helen Hockx-Yu, Ditte Laursen, Daniel Gomes

Introduction

Cultural heritage is our legacy from the past, what we want to keep and pass on to future generations. Libraries, archives and museums have been playing a crucial role in preserving national and cultural resources, with their collections serving as collective memories for given societies or mankind. They are therefore often referred to as “cultural institutions” or “memory institutions”. Dempsey described the role of memory institutions in 2000:

“They organise ... cultural and intellectual record. Their collections contain the memory of peoples, communities, institutions and individuals, the scientific and cultural heritage, and the products throughout time of our imagination, craft and learning. They join us to our ancestors and are our legacy to future generations.” (Dempsey, 2000).

Today, memory institutions have extended this role beyond tangible and physical objects by putting in place strategies for ensuring 'born-digital' documents and artefacts become integrated into the cultural record. Content on the World Wide Web is an important category of born-digital documents that is considered national intellectual record and being actively collected by many national libraries and archives. Countries such as Denmark, France, and United Kingdom have included the Web within the scope of Legal

Deposit legislation, to allow systematic collection and preservation of this material for the use of future generations[i].

When it comes to archiving the World Wide Web, memory institutions have in general adopted the strategy of “divide and conquer”, by assuming responsibility for the respective national domain. The National Library of France for example archives the French web, the Royal Danish Library archives the Danish web, and the British Library and other UK Legal Deposit libraries are responsible for preserving UK websites. Territoriality is often an important element of Legal Deposit law or regulations which enable national libraries to archive the Web at scale. These activities are typically based on archiving the country's Top-Level domain (i.e. .fr, .dk or .uk).

While the domain registration system provides an easy way to help determine scope for web archiving, it does not reflect how content is distributed on the web. The country code Top-Level Domain (ccTLD) alone is not the whole of a national web, as parts of a nation's web can reside on other top-level domains. This model breaks even more when it comes to the ccTLD .eu, which covers multiple nations and does not naturally fall into any national memory institution’s remit. Despite being one of the largest top-level domains in Europe and the 6th most popular (domaintyper.com, The .eu domain), a comprehensive web archive for .eu does not exist, as there is not a central heritage organisation for all countries in the European Union that has the formal, ongoing responsibility of whole-domain archiving for .eu.

In this chapter we present an overview of web archiving activities related to .eu, including the only known comprehensive effort to date that dedicated to archiving the entire .eu domain. It was conducted in 2014-2017 by the Portuguese web archive,

Arquivo.pt, and led by one of the co-authors. We also discuss and propose a number of options towards sustainable, long-term archiving arrangements for .eu.

Archiving .eu

The ccTLD for the European Union is .eu, created to promote the European single market on the internet. The .eu ccTLD was approved by ICANN on 22 March 2005 and added to the Internet root zone on 2 May 2005. Since its launch in December 2005, .eu has become one of the largest ccTLD domains (The European Commission, 2016). The .eu extension in the Cyrillic script (.ею) was launched in June 2016. EURid, the registrar of the.eu and .ею domains, reported 3,760,695 registrations by the end of 2016 (EURid annual report 2016). The following year a positive growth for the .eu domain was registered (EURid, 2018). The number of registrations make .eu one of the biggest top-level domain. In comparison, .fr in France has approximately 2.5 million registrations (domaintyper.com, The .fr domain).

Although ccTLDs are designed for a particular country and the European Union (EU) is a super-national union, .eu is still considered a ccTLD as it is region specific, as opposed to topic specific generic Top Level Domains (gTLDs) such as .com[ii]. Like many other ccTLDs, such as .fr, .jp, .it, .ca and .de, the use of the .eu domain is not open for all or without conditions. It is restricted to organisations and individuals within the European Union. The eligibility criteria for .eu registration require businesses to either have registered offices or be established within the EU, Norway, Iceland or Liechtenstein, and

individuals to reside within the EU, Norway, Iceland or Liechtenstein (Domain Name Registration Policy, v. 8.0).

While a domain name is a paid-for and owned section of the internet namespace, a website is a host that serves one or more webpages and in general receives one or more links from a webpage on a different domain. It is important to note that the number of registered domain names of a certain ccTLD does not map to the number of websites on the live web or in a national web archive. Multiple websites can be hosted on a single registered domain name. The Internet Archive offers definitions of the key components of web archives that can help understand the structure of a web archive (Goel, 2016).

Systematic web archiving activities related to .eu include the Web Archive (of) EU Institutions (EUWA) (the European University Institute, 2017), containing archival copies of the websites of (~80) EU institutions and agencies. This is the fruit of collaboration between the Historical Archives of the EU and the Publications Office of the European Union. However, the EUWA is selective of nature and only constitutes a very small subset of what is hosted on .eu and therefore consulting it will not form a broad impression of the .eu domain.

Wayback Machine, the largest and oldest web archive created by the Internet Archive also contains content from the .eu domain, collected as part of the effort to preserve the global web. There are approximately 1 billion URLs with a .eu domain name in the Wayback Machine, with the earliest dating back to 2001[iii]. This content is however collected without the focused attention of crawling .eu as a self-contained part of the web, and therefore has introduced inconsistencies or gaps in terms of content coverage. This also applies to other (national) web archives, that collaterally picked up .eu content while

archiving their intended portion of the web. In 2014, a preliminary investigation by one of the authors showed that while 3.7 million .eu domains were registered that year, the national web archives in Denmark, Portugal and UK had all together captured webpages from only a few thousand .eu domains (Bicho et al., 2014). There is no easy way to search or study the .eu content in these web archives as a distinct dataset, as this is not a use case supported by the current user interfaces.

Arquivo.pt, a research infrastructure managed by the Portuguese governmental institution Fundação para a Ciência e Tecnologia, is another active player in archiving .eu content. The main objective of Arquivo.pt is to preserve online information for scientific and academic purposes. Using metadata published by the European Union Open Data portal, the Arquivo.pt team identified over 20,000 URLs of FP7 projects and consequently archived the projects' websites, most of which were hosted on the .eu domain. Arquivo.pt also conducted three whole-domain crawls between 2014-2017. This is the only known comprehensive effort to date that focussed on archiving the entire .eu domain. This is described with more detail in the next section.

Pilot project on whole domain crawls of .eu

RESAW[iv] is a self-financing and self-initiated collaborative network of researchers and web archive providers, coordinated by the NetLab and the Centre for Internet Studies at the University of Aarhus. RESAW has the ambition to establish a European research infrastructure for the study of historical web materials, which would be not be complete

with little or no content from the .eu domain. RESAW actively seeks funding from the EU and coordinates activities that advance the agenda for web-based scholarship.

Arquivo.pt is a RESAW partner who conducted three comprehensive exploratory crawls of the .eu domain in 2014-2017, with the goal to understand the size, the processes and issues of archiving .eu. The first crawl was conducted in 2014 and took 24 days. Considerable effort was put into constructing the initial seeds list (web addresses to be visited by a web crawler), by gathering URLs from various sources, which were then merged and cleaned[v]. The crawl had to be paused after a few days when spam sites overloaded the crawler by queuing up abnormally large amounts of URLs to be visited. These spam sites included poorly designed online shops (e.g. autobazar.eu), link farm sites (e.g. in-links.eu) or a large number of subdomains that referenced multilingual versions of the same site (e.g. en.myface4u.eu, pt.myface4u.eu, dk.myface4u.eu). When the crawl completed, the number of seeds increased to 51,164, from an initial pool of 34,138. After removing spam sites, 135,907 unique URLs were extracted and used as seeds for the 2nd crawl. A similar process took place to build a representative seeds list for the consequent crawl, identifying newly added resources to the .eu domain, merging these into the previous list and removing spam sites.

An overview of the three .eu domain crawls is presented below:

Time of crawl	# of Files collected	Data volume (TB)
21/11/2014 to 2014/12/16	129 793 987	5.8
2016/01/07 to 2016/01/26	61 863 684	3.1
02/06/2017 to 10/07/2017	105 823 552	11

Table 1. Comprehensive .eu crawls 2014-2017

The first crawl has been indexed, is searchable and publicly available. Arquivo.pt is processing the remaining two crawls, with the goal to make them available to researchers in the same manner. Technical documents related the first crawl, including the crawl log, crawl configuration, and report generated by the crawler software Heritrix are also available (Bicho et al. ibid.).

A key finding from the first crawl is that it's fairly common for .eu pages to redirect to other top-level domains like .com. Analysis shows that 22% of the URLs include redirects point to other non-eu ccTLDs.

Without an authoritative list of registered .eu domains, Arquivo.pt had to construct the initial seeds list and argument it for all crawls. This time-consuming and not necessarily faultless process could be spared if the .eu registrar EURID would support the initiative by providing an up-to-date list of registered domains. The list could be in the form of a (partial) “Zone File”, which contains just enough information to support web archiving but without revealing any personal information of the registrants. There are

precedents elsewhere, such as Denmark and France, where the non-print legal deposit legislation obliges the registrars to provide the national libraries with such lists, which ensures comprehensive and timely collection of the national web.

The main issue related to archiving .eu, however, is the lack of an organisation which takes the ongoing responsibility of its preservation. Arquivo.pt had funded all the crawls described in this section but could not put in place additional resources for further analysis of the crawl data or comparison with other web archives, which could give us more insight into the .eu domain. Although Portugal is a member state of the European Union, and the web content collected by Arquivo.pt is useful for the research communities served by them, it is unreasonable to assume that the ongoing archiving responsibility of .eu lies permanently with Arquivo.pt. In the next section, we propose and discuss a number of options towards sustainable, long term archiving arrangements for .eu.

Possible models for archiving .eu

An archiving solution for .eu first of all requires an owner or a custodian, an organization or a group of organizations who cares enough about the longevity of content on .eu, and is willing to take the long term responsibility of collecting, preserving and providing access to the archived European web content. The custodian organisation’s own longevity is also important. It should be expected to exist for long term, have the appropriate mission and stable sources of finance. A number of possible models are proposed and discussed below.

1. Direct funding and involvement of the European Commission

Direct funding and involvement from the European Commission, at least seeding the establishment of an .eu archive as a project, and ideally funding a web archiving operation on an ongoing basis, is the most effective and desirable option. This is in direct alignment with EC's role "to assist and complement the actions of the Member States in preserving and promoting Europe's cultural heritage" (The European Commission, May 2018). It provides not just the required financial support but also the perceived longevity – no member state is in a better position than the EC to lead pan-European effort in preserving Europe's digital heritage. The advantage of this model also includes the possibility of obtaining authoritative list of registered .eu domain names from EURid, who is currently not in a position to provide such information and stated that the ultimate ownership of such database is in the hands of the European Commission. It is also possible for the EC to mandate all the information related to their funded projects is archived by a designated archiving service provider, setting this as a condition of the grant. A foreseen difficulty is the identification of the right authority within the EC, which would regard the European web as part of cultural heritage and understand the necessity and benefits of archiving it.

2. Collaboration

Self-initiated networks such as RESAW could, through membership fee or other forms of financial contribution, form a consortium and support a pan-European archive including .eu for research. This would cover wider activities than collecting content hosted on .eu, but also linking to existing web archives. The actual operation of the .eu archive could

either be undertaken by one of the partners or outsourced to web archiving service providers.

The main issue related to this model would be legal restriction which would prevent RESAW from collecting .eu content and offering public access. As RESAW's goal is to support scholarship, not commercial exploitation, an "opt-out" or "notice and take-down" model could be considered. In this model content would be collected from the web and removed from the web archive when being notified by copyright owners.

Many national libraries in Europe are actively preserving their portion of the web, enabled either by legal deposit, or through smaller scale web archiving programmes. A variation of the collaborative model is to take advantage of existing pan-European initiatives such as the European Digital Library, which has the union catalogue of European national libraries and provides access to collections of 47 national libraries of Europe and leading European Research Libraries (The European Library). The open portions of the participating libraries' web archives could become part of this platform. Metadata records of web archives could be also added to the catalogue and become discoverable as part of the European Digital Library.

3. Crowd sourcing

Crowd sourcing is a "social archiving" model which relies on content owners and other individuals to undertake part of the archiving responsibilities. Networks such as RESAW could support them by developing or maintaining personal archiving tools and packaging them up and distributing them in the desired format. Content archived by the crowd could then be uploaded to a central .eu archive for access. A certain level of funding would still

be required to maintain the infrastructure for this. The main disadvantage of this model would be the patchy coverage of the .eu domain as the model relies on the good will of the crowd. It does, however, address the permission issue to some extent if IPR holders submit their own content to the archive. The best perhaps is to use this model as an alternative in combination of option one or two. The crowd sourcing model is also more applicable for social network content such as Tweets than general websites.

4. Commercial venture

It is possible to set up a commercial venture to offer professional web archiving services for content owners on the .eu domain. This would require start-up and operational funds and a viable business model. It will again not immediately deliver a web domain-level archiving solution. As it is likely going to be a paid-for service, it may result in even more patchy coverage than option three.

5. Aggregate what has been archived

Many existing web archives, both within and outside Europe, already contain .eu content. It makes sense to bring all this content together for it to serve as an aggregated .eu web archive. This builds on existing efforts and is the best way to pull together historical webpages which have already disappeared from the live web. Even if the aggregation is only possible at metadata-level, due to copyright restrictions, it would still provide some insight into the development of the .eu web domain.

Conclusion

Efforts around the world to archive the World Wide Web since mid-1990s have led to the existence of large-scale web archive collections, including the Wayback Machine created by the Internet Archive, and national web archives by memorial institutions. These web archives have become valuable resources for understanding contemporary events and history, offering scholars new opportunities and challenges to conduct research and study the web. National web archives also serve as a historical record of the respective country's web presence and cultural heritage. Despite sharing the same class of internet domain names, a comprehensive web archive for .eu does not exist, because no organisation so far has taken formal, ongoing responsibility of whole-domain archiving for .eu.

In this chapter we have presented an overview of the web archiving effort related to the .eu domain, including the three pilot crawls of the entire .eu domain conducted by the Portuguese Web Archive. Further analysis of these crawls could shed light on interesting details and lead to better understanding of .eu domain. Such analysis could, for example, investigate the use of natural languages: is it mainly English? What are the other top languages? It would also be interesting to see what topics the webpages cover and whether it would possible to cluster them in broad categories based on semantics. We could also compare the pilot crawls with .eu content held in the Wayback Machine and national web archives, to understand how comprehensive the pilot crawls were, whether any content was missed. This would also reveal how .eu content is distributed in various web archives and whether there is any overlap. These analyses were outside the

scope of the pilots but demonstrate the usefulness and potential of web archives across borders. Arquivo.pt and RESAW are committed to providing assistance to researchers who are interested in studying and analysing the crawls.

Not much progress has been made since the pilots. In the meantime, content has disappeared from the EU web. We have proposed a number of options towards sustainable, long-term archiving arrangements for .eu. Involvement and funding from the European Commission on an ongoing basis is the most effective and desirable option.

The value of a web archive dedicated for .eu is indisputable, especially when we take into account some of the recent events which could significantly re-shape Europe, as well as the European web. In a referendum on 23 June 2016, majority of the participating UK electorate voted to leave the EU. As a result, the British government invoked Article 50 of the Treaty on the European Union on 29 March 2017 (BBC News, 2017). The UK is thus on course to leave the EU on 29 March 2019. This means over 300,000 UK registrants of the .eu domain names no longer meet the eligibility criteria. The EC announced on 27 March 2018:

“As of the withdrawal date, undertakings and organisations that are established in the United Kingdom but not in the EU and natural persons who reside in the United Kingdom will no longer be eligible to register .eu domain names or, if they are .eu registrants, to renew .eu domain names registered before the withdrawal date.” (The European Commission, March 2018).

With Brexit looming in the horizon and UK's web presence disappearing from the .eu domain, archiving and preserving .eu has become an urgent matter. It would be a great shame if this historical change is not already captured and preserved, as it would contribute to our understanding of the evolution of the European Commission, the UK/EU relations, and, on a broader canvas, the European web sphere.

References

Bicho, Daniel; João, Miranda & Gomes, Daniel, 2015. "A first attempt to archive the .EU domain Technical report": http://arquivo.pt/crawlreport/Crawling_Domain_EU.pdf.

BBC News, 2017. "Brexit: Article 50 has been triggered – what now?": http://www.bbc.com/news/uk-politics-39143978.

Dempsey, Lorcan, 2000. "Scientific, Industrial, and Cultural Heritage: a shared approach: a research framework for digital libraries, museums and archives", Ariadne Issue 22: http://www.ariadne.ac.uk/issue22/dempsey/.

Domain Name Registration Policy, v. 8.0: https://eurid.eu/media/filer_public/0a/fe/0afef70a-85c0-4075-8052-d2853fbe1dff/registration_policy_en.pdf.

Domaintyper.com. "The .eu domain": https://domaintyper.com/domain-names/top-level-domains/ccTLD/eu-domain.

Domaintyper.com. "The .fr domain": https://domaintyper.com/domain-names/top-level-domains/ccTLD/fr-domain.

EURid, 2018. "EURid's 2017 Annual Report Shows Positive Growth for the .eu Extension": https://eurid.eu/da/nyheder/2017-annual-report/.

EURid annual report 2016: https://eurid.eu/media/filer_public/61/6a/616a9b08-13ca-4379-8e11-0a3580201bb5/annual_report_2016.pdf.

Goel, Vinay, 2016. "Defining Web pages, Web sites and Web captures": https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/.

The European Commission, 2016. "10 Years of .eu!": https://ec.europa.eu/digital-single-market/en/news/10-years-eu.

The European Commission, March 2018. "Notice to stakeholders: withdrawal of the United Kingdom and the EU rules on .eu domain names": https://ec.europa.eu/digital-single-market/en/news/notice-stakeholders-withdrawal-united-kingdom-and-eu-rules-eu-domain-names.

The European Commission, May 2018. "Supporting cultural heritage": https://ec.europa.eu/culture/policy/culture-policies/cultural-heritage_en.

The European Library. "Discover Contributors": http://www.theeuropeanlibrary.org/tel4/discover/contributors.

The European University Institute, 2017. "About the web archive of the EU institutions": https://www.eui.eu/Research/HistoricalArchivesOfEU/WebsitesArchivesofEUInstitutions.

The Internet Archive. "The Wayback Machine": https://archive.org/web/.

[i] See a list of countries which included web content as part of Legal Deposit: http://netpreserve.org/web-archiving/legal-deposit/.

[ii] .eu was first added to the ISO 3166 list when the EUR code for the euro currency was created. ISO gave greenlight to the European Commission on September 7, 1999, for EU being used as a ccTLD identifier. Stéphane Van Gelder included a detailed account of this history in *Dot EU – The first decade*: https://eurid.eu/media/filer_public/d3/85/d38538c1-dac5-4e28-a779-cc31b8259697/boek_dot_eu_v05.pdf.

[iii] See key summary of .eu content in the Wayback Machine at http://web.archive.org/details/eu.

[iv] RESAW, a Research infrastructure for the Study of Archived Web materials see: http://resaw.eu/.

[v] For a list of sources used to build the initial seeds list, see Bicho et al.