AI & SOCIETY (2022) 37:975–984
https://doi.org/10.1007/s00146-021-01364-0

OPEN FORUM

Managing and accessing web archives: Irish practitioners' perspectives

Maria Ryan[1] · Della Keating[1] · Joanna Finegan[1]

Received: 29 April 2021 / Accepted: 23 November 2021 / Published online: 14 February 2022

Abstract

This article provides practitioners' perspectives on preservation of the Irish web space by the National Library of Ireland (the NLI). The context of this work is outlined including the history of Ireland's national library, its role, resources and place in library, archive, cultural and digital preservation networks. The development of the NLI Web Archive is discussed within the wider context of the Library's mission and digital collecting and preservation policies, as well as international approaches to preserving the web. The article looks at how the NLI has developed its selective web archive over the past decade, and has grown the content and access to it as a way to mitigate against the absence of at-scale solutions. The unusual legal context in Ireland regarding legislative barriers to archiving the Irish web space at scale and the NLI's work for over a decade to change this situation are discussed as are the significant implications of the current legal situation for data loss and long-term access to Ireland's contemporary record. Distinctive Irish aspects of digital cultural heritage preserved in the NLI Web Archive collections are highlighted. The opportunities and challenges in developing outreach and access for the Web Archive are considered together with its relationship with the collecting activities of the Library's Born Digital Pilot Projects. This article will also discuss types of usage and user groups in relation to archived Irish web data. Potential for creative and imaginative uses of Irish web archive collections and data are also considered in relation to the Library's broader public learning and outreach programmes.

Keywords Web archives · Born digital · Cultural heritage · National Library of Ireland · Legal deposit · Big data

1 Introduction

The purpose of this article is to discuss how the Web Archive programme of the National Library of Ireland (the NLI), since its inception in 2011, has adapted and developed to rise to the challenge of preserving the Irish web space. It is written from the point of view of practitioners, the archival and library professionals working in the development and management of digital cultural heritage collections who comprise the NLI Web Archiving team, and Born Digital Pilots team and is benchmarked against literature where appropriate.

The article situates the work of the team nationally and internationally. Nationally, it locates the programme in the context of the importance of the role of the Library's Web Archiving programme in the collection of Irish digital cultural heritage. The programme is also framed in terms of how it has leveraged international practice, particularly in relation to selective web archiving and networks in this arena, to rise to the challenges faced.

The challenges faced by the Web Archiving programme include legal challenges unique to the Irish environment that preclude routine full-domain web archiving and the more ubiquitous ethical and preservation challenges associated with selective thematic collecting and the challenge of collecting the Irish web space at scale. The evolution of web archiving at the NLI over the past decade is seen as a response to those challenges and the article describes various strategies employed by the practitioners by way of mitigation and response to challenges faced. The opportunity offered by AI is viewed as an implicit and much wanted activity in an environment where substantial challenges exist

✉ Maria Ryan
mryan@nli.ie

Della Keating
dkeating@nli.ie

Joanna Finegan
jfinegan@nli.ie

[1] Digital Collections Department, National Library of Ireland, Dublin, Ireland

to collecting at scale as a precursor to interrogating data at scale with techniques such as AI.

The article frames the web archiving programme in the context of the history of collecting Irish cultural heritage at the NLI and more recent developments in relation to digital collecting at the NLI. It then describes different national and international approaches to web archiving in general and proceeds to discuss the NLI interpretation of these approaches to adapt to its own environment with particular reference to the Library's strategic approach to selective web archiving. The impact of the specific challenges posed by the lack of legislation relating to full-domain web archiving in Ireland is addressed, and the evolution of access and outreach strategies to the NLI web archive is described and analysed in the context of potential for creative and imaginative use of content.

2 Context: NLI history and collecting

The NLI is the library of record for Ireland. Our unique collections are permanent, accessible by all and constitute the most outstanding collection of Irish literary and documentary heritage in the world. The mission of the NLI is to collect, protect and make available that recorded memory on behalf of the State for the people of Ireland (National Library of Ireland 2016, p.[6]). This ever-growing record comprises sources which are textual and non-textual, published and archival and in physical and digital formats. Each year the collections of the Library expand through donations, purchases and legal deposit, for the benefit of the Irish public, both at home and the diaspora across the world.

Today NLI has three collecting strands, with published and special collections comprising over ten million physical items including printed books, serials, newspapers and official publications together with photographs, manuscripts, maps, prints, drawings and ephemera. The Library's other collecting strand is digital and includes both digitised and born digital content composed of unique literary and organisational records together with that of the archived Irish web. Currently there are almost 127,000 items and 59 TB of digitised content.

Established in 1877, the Library is now an autonomous National Cultural Institution, operating under the aegis of the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media. In 2019, the NLI had over 280,000 research or exhibition visits and 20 million combined online interactions, with a staff of 87 FTE and funding allocation of just under €8.5 million (National Library of Ireland 2020, pp.36–39).

The NLI is part of national and international library, archive, cultural and digital preservation networks. It is one of Ireland's National Cultural Institutions and a member of the Consortium of National and University Libraries (CONUL), the Library Association of Ireland (LAI) and the Archives and Records Association Ireland (ARAI). The Library actively collaborates with both public and academic libraries although there is no formal statutory relationship because in Ireland, public libraries operate under the aegis of local government and the academic libraries of the higher education institutions largely operate under the Department of Education and Skills (Collins 2019, p.177). Internationally, the NLI is member of the Conference of European National Librarians (CENL), the International Federation of Library Associations (IFLA), the Digital Preservation Coalition (DPC) as well as the International Internet Preservation Consortium (IIPC). Within Ireland, digital preservation networks are largely emergent.[1]

3 Digital collecting at the NLI

In 2015, the NLI established the Digital Collections Department in recognition of the changing external environment and the need to make explicit our commitment to digital collecting and preservation activities and examine how we could provide access to our stakeholders. The department has a mix of IT professionals, librarians, archivists and digital preservation specialists. It is responsible for web archiving, born digital archives, digitisation and digital preservation as well as providing and managing the IT infrastructure for the Library including the Library Management System, Public Catalogue and Digital Repository.

In terms of policy, the need to collect digital content is clear in NLI's current strategic plan covering 2016–2021 (National Library of Ireland 2016). In tandem with this strategy, NLI was one of the first Irish national cultural institutions to publish a Diversity and Inclusion policy (National Library of Ireland 2018). In 2017, the Minister for Arts, Heritage, Regional, Rural and Gaeltacht Affairs asked all cultural institutions to put in place gender inclusive policies (MerrionStreet 2017). The NLI went a step further and launched a diversity and inclusion policy which has played a part in evolving our collecting practices to include new, inclusive and diverse forms of content both in terms of the Irish experience represented and the variety of formats in our physical and digital collections (National Library of Ireland 2018, pp. [1–2]).

Today, the web archive is an established NLI digital collecting programme and the Library is currently piloting a

[1] In 2018 Ireland's Department of the Taoiseach launched a public consultation for the second National Digital Strategy. Submissions were to be published online. A new strategy is currently being drafted.

second-born digital collecting strand. These pilot projects aim to extend collecting to unique digital content and archives, such as photographs, emails, documents, electronic drafts of literary works and organisational records, including those of campaigns. Collections acquired to date include the Marian Keyes *Mystery of Mercy Close* collection, the Yes Equality Digital Photographic Archive and the Waking the Feminists collection.

These two programmes provide interesting counterpoints between the published, online record and the unpublished, archival digital record. It has been said that newspapers are the first draft of history, today the web is often that first draft. Archiving the web needs to be carried out in real time due to its ephemeral nature and is frequently carried out in a reactive manner in response to unfolding events. This has often flagged major, emergent issues, e.g. diversity, the far right, social justice campaigns and this in turn has informed our work on the Born Digital Pilots. The NLI's web archiving programme has acted as a weathervane for other digital collecting activities, where we have identified issues at an earlier stage and are already able to draw on our learning from archiving the Irish web. As well as being at the forefront in terms of events and issues, web archiving has also brought the NLI in contact with communities we otherwise would not have encountered, or encountered at the much later stage, whether that is content creators or those seeking to develop research and discovery opportunities. The websites associated with the three Born Digital Pilot Projects were all archived long before decisions on their selection as pilot project collections. The Web Archive also provided the pathway for the NLI's involvement in the AURA network.

4 Archiving the web: national and international approaches

National Libraries worldwide are grappling with born digital objects, not only in terms of appraisal, but also preservation and provision of access (Milligan 2019a, p.30). The web, having recently celebrated its 30th birthday, is still a relatively new format when compared with traditional collections and is continuously developing. A website is less stable than a physical format such as paper or parchment although these are also vulnerable if not cared for appropriately (Brugger 2018, p.77). The changing nature of a website means it requires "active maintenance" (Milligan 2019b, p.54) to survive. In the mid-1990s, professionals across the world became concerned with both the increased amount of records and publications being created in digital-only formats and also the volume of information on the web. In 1996, several key memory institutions including most prominently the Internet Archive, but also the National Library of Australia, the Smithsonian Institution, the University of Texas and the Royal Library Sweden began an attempt to combat the looming digital dark age by preserving the web (Milligan 2019a, p.72–75).

The International Internet Preservation Consortium (IIPC) defines web archiving as the process of "collecting portions of the World Wide Web, preserving the collections in an archival format and then serving the archives for access and use" (IIPC n.d.). There are a number of approaches to web archiving that are carried out by national libraries. Brugger categorises these as Macro- and Micro-level web archiving. (Brugger 2005, 9–11). Macro-web archiving is concerned with collecting large portions of the web. This is usually done through domain web archiving, where a country's top-level domain is archived. Domain web archiving is often underpinned by legislation; such is the case in the United Kingdom where the UK Web Archive carries out broad crawls of the top-level domain on an annual basis (Bingham, Byrne 2021, pp 1–2).[2] Micro-web archiving is more selective and limited in nature and while Brugger argues it is usually carried out by those who have an immediate need to preserve a limited web page or object (Brugger 2005, p. 10), it is often used by national libraries to build selective and thematic-based web archives. These web archives are usually created around a topic or theme or could be used to collect the online representation of an event, such as an election. Many national libraries adopt these two strands to achieve a more holistic representation of the online life of their country. Selective web archiving also allows national libraries to work with subject specialists and communities to build unique special web archive collections. Domain web archiving allows a broader, but often shallower, snapshot of a country's web sphere. Each approach to web archiving produces a different web archive, with neither being a substitute for the other. As will be discussed later, web archiving is an imperfect approach to archiving the web. Milligan (2019a) argues that the web is unequal, with a large digital divide evident in the world, between those who have access to the web and those who do not. Domain crawls ensure wider representation, efficiently sweeping up large portions of a top-level domain. But selective collections "help balance out the forces within the broad crawls and to ensure the widest possible representation on the historical record" (Milligan 2019a, p. 88). Approaches to web archiving in national libraries vary depending on factors, such as legislation and resourcing. Web archiving is often carried out by a third party supplier on behalf of the national library

[2] The UK Web Archive is a partnership of the British Library, National Library of Scotland, National Library of Wales, Bodleian Libraries Oxford, Cambridge University Library and Trinity College Dublin.

or can be facilitated in house, such is the case in the British Library.

In the twenty-first century, the web informs most facets of our life. Its invention has changed how the world works, communicates, socialises, learns and seeks information. Milligan argues that web archives offer the researcher an opportunity "to know far more about human culture and activity than ever before" (Milligan 2019a, p.88). However, with the increase in the volume of data on the web comes the challenge of how to preserve, provide access and disseminate it. There are also ethical concerns when archiving websites that must be taken into consideration. When content is published online, the fact that it may be preserved for future study and dissemination is often not taken into account. Furthermore, do web content creators consider the "ramifications of sharing personal data" (Lomborg 2019, p. 105). It has been our experience that sometimes people do not consider their online content as published. The issue is further complicated when considering content related to or created by young or vulnerable people online. Mackinnon (2021) describes the impact the age of the internet has had on the volume of information available on the ordinary individual. What happens if material posted of or by a child is archived? They may be delighted to re-discover it in a web archive or they may simply not want it preserved. Many archives engage in selective web archiving surrounding events, such as protests, political campaigns and referendums. These are often traumatic, emotive events. What are the consequences of archiving personal accounts of such events and making them available for researchers? Although web archives have the potential to make a positive impact on representation and research, they also have the power to cause harm to those included in the archive (Mackinnon 2021, p. 443). Much research is being carried out into the effects of "growing up online" and the impact that, as Eichhorn (2019) puts it, "The End of Forgetting" will have on a generation of young people. They have grown up documenting their lives on line, often without knowledge or expectation that this content may be online forever.

Pamela Graham argues that the practice of transparency is central to mitigating some ethical issues (Graham 2017, p. 107). It is imperative that web archivists document their collection strategies and decisions for researchers in the future. These will show why each website in a selective collection was included and who made that decision. It is also important that web archivists engage with communities and be open to conversations with those who created and featured in websites that are now part of the archives (Graham 2017). Archivists must be mindful of the power of the archive and the ability of its content to re-traumatise those who feature in it. The ethical questions of web archiving will continue to be discussed in the international professional communities as web archiving develops.

5 Selective archiving of the Irish web at NLI: from pilot to permanent

In 2011, the NLI initiated a pilot project on web archiving to collect and preserve the online record of that year's General Election and Presidential Election. The General Election was the first election in Ireland where political parties invested significantly in an online presence (Little 2011, p.6).This is probably in no small part due to the rise in Irish people using the internet. In 2007, the year of the previous general election, 57% of homes in Ireland had access to the internet (Central Statistics Office 2008, p.327). In 2011, 78% of homes were connected (Central Statistics Office 2012, p. 293). Collecting for the General Election took place in February and March 2011 and resulted in the capture of 100 websites. Since then the Selective Web Archive has evolved from this pilot project phase to become the Library's main born digital collecting programme. Staffing for the programme was initially on a part time basis, typical of many web archives at that time.[3] In 2016, there were many commemorative activities occurring online around the centenary of the 1916 Easter Rising, an event of national historic significance, and other major national commemorations relating to the Irish experience of the First World War. The record of these was in danger of loss and in recognition of the importance of preserving these commemorations temporary staffing was secured. This enabled the Library to undertake digital collecting around these centenaries and major national commemorations that year and eventually secure a permanent web archivist post. Currently, 1.4 full-time equivalents (FTE) work on the web archive. Today, there are over 2300 sites, many captured multiple times and 44 TB raw data in the openly available NLI Selective Web Archive. The Web Archive is thematic and event-based with collecting focused on major unfolding events in Ireland. The selective approach provides open, responsive, curated collections, averaging about 600 crawls a year. The NLI works with a technical partner, Internet Archive, to carry out web archiving.

5.1 Approaches to selective web archiving at the NLI

Selection is carried out in house or in collaboration with a subject specialist and the three approaches can be broadly defined as comprehensive within narrow categories,[4]

[3] A 2016 Harvard Library Report on web archiving showed that more than half of the organisations surveyed had no dedicated full-time staff (Truman G 2016, p. 9).

[4] Collections developed through a comprehensive approach are those which can be very narrowly defined, eg. all Irish Higher Education organisations in receipt of HEA funding, all registered Irish Trade

selection by the web archivist and collaborative. The comprehensive approach aims to capture an annual archive of government departments, higher education institutes and local authorities. Websites chosen by the web archivist are usually regarding events such as elections or topics such as literature or music. Numerous methods of scoping are carried out and each decision is documented for future reference.

Collections include those on politics, such as referenda and elections, social, cultural and creative life, central and local government, parliament and public sector organisations.[5] They also cover themes and events unique to Ireland ranging from life on the Irish coast to farming life to Irish literature. Collections also represent global events, such as COVID-19 and climate change, and how they impact Ireland. Milligan argues that "web archives offer the prospect of incorporating more voices and more people" (Milligan 2019a, p.245) and this has been the case in the NLI. Collections are being developed focusing on younger and older people, LGBTQI + and the Irish diaspora. Many sites have both Irish and English language content. The Irish language, as the national language, is the first official language. Public bodies in Ireland are required by legislation to ensure that information, services and certain publications are provided to the public in Irish only or bilingually (An Coimisinéir Teanga 2003, pp.1–2). The archive also develops Irish language content through preserving sites on literature, publishing, social and cultural life.

The NLI works with a range of subject specialists and partners to broaden and diversify the content of the web archive and also connect to communities representing unique aspects of Irish culture. These types of working relationships require different levels of staff input. NLI would like to develop further project partnerships as our existing ones have proved very fruitful, but it is necessary to acknowledge both the value of these partnerships and the staff resources needed to carry them out successfully for all concerned and to factor these aspects into any future plans.

The Irish Traditional Music Archive (ITMA) is a national public reference archive and resource centre for the traditional song, instrumental music and dance of Ireland and a new project partner for the web archive which will enable the development of connections and collections on Irish music. Collaborating with the Irish Community Archive Network (iCAN), an initiative of our sister cultural institution, the National Museum of Ireland, working with volunteers and local Heritage Officers nationwide, will facilitate greater representation and preservation of community archives in digital form. Our longest running partnership, which is with The 100 Archive, provides an opportunity to develop a collection of websites of Irish design excellence which also connects us with a community of creators of digital content; this is of particular value as dialogue with content creators is a vital part of any preservation work (Digital Preservation Coalition 2015, p.6).

In terms of working relationships, the NLI also continues to work with government in preserving its published record in online form just as the library's print collection contains thousands of physical format official and government titles. Since 2018, there has had a particularly busy period of activity archiving the individual websites of each Irish government department prior to the migration of their online presence to a new centralised gov.ie location.

When websites are selected for the web archive, the notification process begins. Each website owner is notified in advance of archiving. This allows the owner to explore the concept of web archiving, ask questions and find out more about the web archive. A takedown policy ensures that website owners can engage with NLI to remove material from public view at any point. However, the requirement to notify each website owner is problematic when a website does not have contact information listed. If there is no contact information listed on the website, it cannot be included in the web archive regardless of value. This issue can be particularly prevalent around political events, such as elections or referendums, where the origin of the website is often unclear. In other countries, this content would be collected anyway through a crawl of the top-level domain (TLD), a process which does not have the same notification requirement as that for inclusion in a selective, open web archive. However, this is not currently possible in Ireland, resulting in the loss of websites and the resultant future historical record. The NLI documents the steps taken when creating a web archive collection. Decisions on what web sites are included or why some are excluded are documented. Efforts are also being made to work with communities to ensure wider representation in the web archive.

The NLI uses the Archive-it service to carry out archiving. The web archivist directs the web crawler to a target website or "seed". From the homepage of the website, the crawler visits each page on the website as directed. It downloads the web content it finds along its way, including PDFs, jpegs, and audio–visual material. The result is a copy of the website that can be navigated, so far as is possible, like a live website. Websites go through an extensive testing process, where modifications are made to the crawler's parameters

Footnote 4 (continued)

Unions, all registered political parties, local authorities and government departments.

[5] NLI is the only national cultural institution in Ireland carrying out web archiving for the benefit of the public. In some countries multiple institutions engage in web archiving for public use. For example, in the UK, the National Archives (TNA), National Records Scotland, the UK Web Archive and the UK Parliament all carry out web archiving and make collections publicly available.

to ensure the highest fidelity capture. When the final archiving is complete, each website undergoes a quality assurance process where emphasis is placed on replay ability but also on the capture of documents such as PDFs. At this point, metadata are added to each archived website and made available through the NLI web archive portal. Web archiving is an imperfect activity, constantly grappling with changes in technology and web design (Milligan 2019a,p. 118). However complex, it is the best tool we have to preserve online content for the future.

6 Domain-level archiving at the NLI: an Irish perspective on the challenges of loss and legislation

National libraries are essential for the development of a knowledge society through ensuring universal and equitable access to information and have various legal instruments to support this work. Digital collections offer challenges of scale, formats and legislation (Hosker R 2020) and in the Irish context, the latter is particularly true. The NLI currently has one approach to preserving the Irish web, through selective web archiving; the other approach we need and want to take is domain-level web archiving.

Legal Deposit is the statutory provision which obliges publishers to deposit their publications in named libraries, one of which is usually the national library. This is to ensure the development of a national collection of published material and guarantees citizens and researchers within the country and abroad, access to a research collection of the country's published material in various formats (Larivière 2000, pp.vii-5). NLI has been a legal deposit library since 1927, building our published collections through this provision for public benefit. However, there is an aspect of preserving the national record which unfortunately distinguishes Ireland from nearly 60% of other EU countries: in Ireland, legal deposit legislation does not cover archiving of the web (Fitzpatrick 2021, p.29). Currently, it is not possible to carry out domain web archiving to preserve the national web space at scale by archiving Irish sites including those with the national country code in the top-level domain. In Ireland, this is the ".ie" domain, and by the end of 2020, there were 309,853.ie domains, almost a 50% increase in five years (IE Domain Registry 2021, p.2).

It is important to emphasise however that selective and domain web archiving are not substitutes for each other, they each have unique characteristics. To truly meet its mandate, the NLI needs to be able to undertake both approaches. Focused, event-based archiving provides a flexibility and responsiveness to enable capture of unfolding events which is not possible with domain crawling but, as Milligan observes, it "misses the content that does not belong to an event …Comprehensive crawls are needed, similar to those carried out by national libraries in countries like France and the United Kingdom" (Milligan 2019b, p.59). In 2011, a copyright review process was initiated by the then Department of Jobs, Enterprise and Innovation and since that time NLI has actively participated through a variety of channels in the process to update legal deposit legislation (Ireland. Department of Jobs, Enterprise and Innovation 2011). NLI made submissions to the Review Committee individually in 2011 and 2012 (National Library of Ireland 2011) (National Library of Ireland 2012) and also as part of the wider library community through CONUL (2012), (CONUL 2016). In 2013, the committee's report "Modernising Copyright" was issued (Copyright Review Committee 2013), and in 2017, the Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs, on behalf of NLI, published a consultation paper on Legal Deposit of published digital material in the twenty-first century in the context of Copyright legislation (Ireland. Department of Culture, Heritage and the Gaeltacht, 2017). In 2019, the legislation was amended to extend legal deposit to e-books and journals which certainly improves the NLI's ability to collect and preserve publications; but the continuing absence of legislation to enable legal deposit domain crawls seriously impairs our ability to preserve the contemporary record of our country. The absence of legislation to support at scale archiving of the Irish web space means a serious and irretrievable loss of the record, both of and for the Irish public. The NLI continues to work with our parent department, the Department of Tourism, Culture, Arts, Gaeltacht, Sports and Media, and do everything possible to work towards the amendment of the legislation in relation to web archiving needed to allow us to archive the contemporary record of our country.

Legislation in relation to web archiving of national domains does impose artificial boundaries given the global nature of the web (Winters and Prescott 2019, p.398). The situation is further complicated on the island of Ireland as it comprises two political units with different country domains (ccTLD[6].ie and ccTLD.uk). In itself, this presents a rich area of potential research as shown by Webster's study of the Northern Ireland religious web and the limitations of ccTLD (2019). However, large .ie domain datasets are unavailable to researchers including Webster who notes that there was no Irish equivalent to the British Library's JISC UK Web Domain Dataset available with which to investigate the same patterns (2019 p.14). Without a legal instrument to enable at scale archiving and preservation, an integral part of Ireland's heritage and key historical and research source is being lost. National domain archives are essential to comprehending a country's development and use of the web,

[6] ccTLD is the country code top-level domain.

while also providing the opportunity to develop tools for data analysis (Brügger N, Laursen D 2019, pp.2–3).When we talk about web archives, we cannot provide access or develop the potential of AI for unique Irish content if we cannot collect at scale in the first place.

6.1 Web archiving opportunities from an Irish perspective: discovery and access

The NLI Selective Archive is continually developing and its content is available for discovery and re-use. Access to the Web Archive is provided on the NLI's Archive-It portal[7] which is linked to on the homepage of the Library's website. Websites are also included in the NLI catalogue. In terms of preservation, the WARCs themselves are preserved in the NLI's preservation storage system.[8] However, there is a challenge for institutions such as the NLI, with a mandate to serve the general public, in terms of identifying the Designated Communities required by the OAIS digital preservation model. Talboom and Underdown (2019) discuss the usefulness of identifying communities of users by their different types of usage of digital archives. They describe three different user groups: readers, data users and the digitally curious. In terms of discovery and access to NLI Selective Web, the Archive-It portal currently serves 'readers' viewing content in a more traditional way through browsing and search of sites at collection and individual unit, ie. website level. A full-text search with options to narrow by file format, date, etc. is also provided.

However, both traditional search and discovery methods and unstructured free-text search have considerable limitations when dealing with the extent of digital collections. What seem unscalable mountains of data, particularly web archives, cannot be searched in the same way as the live web (Winters J, Prescott A 2019, p.397).The 'data users' identified by The National Archives UK (TNA) are using computational analysis on large scale digital collections. The NLI has begun working with academic partners and Irish data users, using network analysis to explore our 2016 web archive collections and demonstrate how it can be used to produce ego networks for parliament and government departments (Greene D, Ryan M 2019).We will look to increase this aspect of discovery services through developing further collaborative partnerships with the research community in relation to our cultural heritage data.

6.2 Web archiving opportunities from an Irish perspective: different types of usage

Of course, many users do not have advanced data analysis skills but are aware of the potential they offer. The presence of these 'digitally curious' users lacking more highly developed computational skills but with an interest in web archive data (Talboom and Underdown 2019), together with the known difficulties of working with WARCs (Milligan I 2020) and the need to lower the barriers to usage have been identified by digital archivists, historians of the web and web archiving service providers. The Mellon Funded collaboration between Archives Unleashed and Archive-It includes integration of the Archives Unleashed toolkit and Archive-it to provide easy to use scalable tools and datasets, such as Web Archive Transformation (WAT) files, Web Archive Named Entities (WANE) files and CDX index files, to enable computational queries and analysis of web data (Archive-It 2020). It aims to give archivists and librarians tools for making web collections available as data and supporting computations analysis. As a client of the Archive-It service, the NLI anticipates, this will provide new opportunities for discovery of archived Irish web content for both the digitally curious as well as experienced data users. On the user side, researchers will also have to consider technical skills going forward. Schafer raises the issue of "the training tomorrow's historians will need as part of their University courses" to deal with web archive data (2019, p.155). Winters and Prescott go further by saying that "Digital Historians will have to adapt to meet new theoretical and methodological questions posed by working with these and other data formats yet to be imagined" (2019, p.285). Interest in web archives is emerging among Irish researchers. In 2020, the NLI supported and participated in Engaging with Web Archives (EWA) which was the first researcher led web archives conference in Ireland, hosted by the Maynooth University Arts and Humanities Institute (EWA 2020).

6.3 Web archiving opportunities from an Irish perspective: connecting data and people

Enabling users to discover, connect and reuse Irish web archive data within the web archive itself and across the richness of our collections is an ongoing part of the Library's activity. The NLI also works to develop external engagement with all collections. One method of achieving this is through Open Data which seek to make data held by public bodies freely available and easily accessible online for reuse, redistribution and connection with other datasets. The drivers for this in Ireland are at international, European, national and local level (Ireland. Department of Public Enterprise and

[7] The NLI Selective Web Archive is hosted on Archive-It at https://archive-it.org/home/nli.

[8] The WARC format (Web ARChive) is the internationally accepted (ISO 28500:2017) file format for storing web crawls and archived websites.

Reform 2017, p.4).[9] The NLI's first dataset was published in 2020 on Ireland's Open Data portal and while this was geo-data, the opportunities offered by open data are also important to consider in terms of derived web archive datasets and which of these are appropriate for publication.[10]

Traditional methods of search and browsing across the web archive collections and developing new types of data-level access are not the only potential means of discovery and use for the NLI Web Archive. Through the work of its own Education, Learning & Programming Department and Development Department, the Library also shares the story of Ireland with a broad audience in a variety of imaginative and engaging ways. This includes very successful exhibitions such as those on Nobel Laureate Seamus Heaney and at the Museum of Literature Ireland (MoLI) which is a partnership with University College Dublin. In 2019 there were 210,000 onsite visitors to six NLI exhibitions and both the physical and online exhibitions as well as the accompanying book on Nobel Laureate WB Yeats are award-winning (National Library of Ireland 2007, pp.13–16) (National Library of Ireland 2009, p.16). The NLI's membership of Flickr Commons since 2011 has also been a very successful initiative, bringing much wider exposure and user engagement with the photographic collections.[11] How can we use born digital content including web archives in creative ways in both our physical and online exhibition spaces? The work of interpretation, exhibitions, learning and outreach are as relevant approaches to sharing the story of Ireland as library catalogues, online portals and data analytics. Sharing web archive data can enable its reuse in imaginative ways and facilitate users to share their own knowledge and creativity. It is worth bearing in mind that these web archive data are also a rich source of non-textual cultural heritage data on Ireland and that visual data analytics can support new forms of historical analysis when used with web archives (Ben-David A, Amram A, Bekkerman R 2018, pp.95–106). What do born digital collections such as web archives and technologies like AI offer to users and what can users offer in return? We do not have the full picture of this yet, but we do know that, as always, the NLI wants to share the story of Ireland in creative and imaginative ways and develop our services to support a broad range of users in doing the same with our digital cultural heritage.

7 Conclusion

The development of web archiving at NLI was an extension of library practices concerning the national record together with recognition of the role of the web as a key primary source for Ireland's unique and distinctive documentary and creative heritage. A major challenge in Ireland continues to be the absence of policy instruments to support archiving the national web at scale through domain crawling. To realise opportunities offered by AI for research and discovery, the Library must be able to collect at scale. The challenge for NLI is legislative. While archiving the web at national level is imperfect and complex given the global nature of the web (Winters J, Prescott A 2019, p.398), memory institutions such as national libraries are essential in the preservation and provision of long-term public access to the online record.

However, ten years of sustained web archiving for the Selective Web Archive has resulted in unique, distinctive, openly available and ever-growing digital content which forms part of the national collections of Ireland and seeks to reflect the breadth of Irish experience. Over this period, NLI web archiving practices have evolved based on strategies for maximizing capacity through partnerships on the technical side and collaborations on collecting activities with relevant organisations, communities and as part of national commemorative events. These collaborations also provide a valuable opportunity for advocacy in relation to the essential role of digital preservation and the NLI in ensuring future access to the record of Ireland.

Digital collections, including the archived web, present challenges of scale and we need solutions at scale as well in relation to processing, describing and enabling discovery and reuse of these collections. AI will certainly offer the NLI opportunities to innovate in this regard. Given the extent of digital collections data 'automation is no longer a choice but a necessity' (Seles 2020). We are beginning initial collaborations with the higher education community in relation to the potential of AI for processing large volumes of digital material at scale. National libraries including the NLI have a richness of unique and distinctive, immensely valuable data. As practitioners, we need to develop our skills which include not only those around the opportunities AI can bring regarding access, engagement and discovery, but also those around the considerations of which we, as collection managers need to be aware. AI does not just need data, it needs *good* data. Memory institutions, such as libraries and archives, have long been concerned with issues of ethics, inclusivity, transparency, consent and privacy (Jo E, Gebru T 2020).

The Library's strategic aims of innovation and trialling new forms of engagement to connect people with the story of Ireland are certainly relevant to the opportunities offered

[9] The Open Data and Re Use of Public Sector Information (PSI) Directive was transposed into Irish law in July 2021.

[10] The geo-data https://data.gov.ie/dataset/catholic-parish-registers-geodata concerns the NLI collection of Catholic parish register microfilms and is used in the mapping features of the Library's online resource https://registers.nli.ie/.

[11] By end of 2019 there were a total of Flickr 12,368,535 views (National Library of Ireland 2020 p.8).

by AI. Collaboration and research partnerships with Higher Education are essential to development of our skills and capacity as practitioners to deal with our ever-growing digital collections relating to Ireland, including the Irish web. For the NLI participating in the AURA network is an important part of developing these capabilities. Our commitment to inspiring and supporting inclusion, creativity and learning begins at home.

Irrespective of the challenges and problems with web archiving, it will play a vital part in any future research of the late 20th and early twenty-first century (Winters 2019, p.285).

Acknowledgements We would like to thank colleagues Jenny Byrne, Kieran O'Leary and Ger Wilson of the NLI Digital Collections Department for their help and input to this article.

References

An Coimisinéir Teanga (2003) Official Languages Act 2003 Guidebook. An Coimisinéir Teanga, An Spidéal. Version archived on 22 January 2021. Retrieved from the Wayback Machine: https://web.archive.org/web/20210122185933/https://coimisineir.ie/userfiles/files/Official%20Languages%20Act%20Guidebook_eng.pdf Accessed 28 April 2012

Archive-It (2020) Archive-It and Archives Unleashed join forces to scale research use of web archives. Archive-It Blog 2020, July 28. Version archived on 11 March 2020. Retrieved from the Wayback Machine: https://web.archive.org/web/20210311211108/https://archive-it.org/blog/post/archives-unleashed-partnership/ Accessed 28 April 2021.

Ben-David A, Amram A, Bekkerman R (2018) The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. Int J Digit Libr 19:95–106. https://doi.org/10.1007/s00799-016-0202-6

Bibliothèque nationale de France (2017) The WARC File Format (ISO 28500) - Information, Maintenance, Drafts. http://bibnum.bnf.fr/WARC/. Accessed 28 April 2021.

Bingham NJ, Byrne H (2021) Archival strategies for contemporary collecting in a world of big data: challenges and opportunities with curating the UK web archive. Big Data Soc. https://doi.org/10.1177/2053951721990409

Brügger N (2018) The archived web: doing history in the digital age. MIT Press, Cambridge, MA

Brügger N, Laursen D (2019) Introduction: digital humanities, the web, and national web domains. In: Brügger N, Laursen D (eds) The historical web and digital humanities : the case of national web domains. Routledge, Abingdon, pp 1–10

Brugger N (2005) Archiving Websites. General Considerations and Strategies. The Centre for Internet Research, Arhaus, Denmark.

Central Statistics Office (2008) Statistical Yearbook of Ireland 2008, Central Statistics Office, Dublin, version archived on 25 April 2016. Retrieved from NLI Web Archive: https://wayback.archive-it.org/org-1444/20160425050144/http://www.cso.ie/en/media/csoie/releasespublications/documents/statisticalyearbook/2008/Statistical_Yearbook_2008_for_web_complete.pdf. Accessed 8 April 2021

Central Statistics Office (2012) Statistical Yearbook of Ireland 2012, Central Statistics Office, Dublin, version archived on 19 October 2019. Retrieved from NLI Web Archive: https://wayback.archive-it.org/org-1444/20191019043333/https://www.cso.ie/en/statistics/statisticalyearbookofireland/statisticalyearbookofireland/statisticalyearbookofireland2012edition/. Accessed 8 April 2021

Copyright Review Committee (2013) Modernising Copyright : a report prepared by the Copyright Review Committee for the Department of Jobs, Enterprise and Innovation. CRC, Dublin. https://enterprise.gov.ie/en/Publications/Publication-files/CRC-Report.pdf. Accessed 23 April 2021

Collins S (2019) The National Library of Ireland.Alexandria. 28(3):177–181.https://doi.org/10.1177/0955749019878523

CONUL (2012) Submission by CONUL In response to Copyright and Innovation: A Consultation Paper prepared by the Copyright Review Committee for the Department of Jobs, Enterprise and Innovation. https://enterprise.gov.ie/en/Consultations/Consultations-files/CONUL.pdf Accessed 28 April 2012

CONUL (2016) Copyright and Related Rights Act, 2000: Submission by CONUL to the Department of Jobs, Enterprise and Innovation on Extending Legal Deposit provisions to Digital Content. http://www.conul.ie/wp-content/uploads/2017/10/Submission-to-DJEI-on-digital-Legal-Deposit-for-Ireland-final18May2016.docx Accessed 28 April 2012

Digital Preservation Coalition (2015). Digital Preservation Handbook, 2nd Edition. Version archived 28 November 2019. Retrieved from the Wayback Machine: https://web.archive.org/web/20191128165622/https://www.dpconline.org/handbook. Accessed 28 April 2021

EWA Conference Organisers. (2020, October 7). Book of Abstracts: #EWAVirtual 2020. Presented at the Engaging with Web Archives: 'Opportunities, Challenges and Potentialities' (#EWAVirtual), Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland. 10.5281/zenodo.4058013

Fitzpatrick S, O'Dowd M (2021) Libraries and Archives: 'Guardians of the word hoard'. Royal Irish Academy Culture and Heritage Working Group Paper, 2021. RIA, Dublin. https://www.ria.ie/sites/default/files/libraries_and_archives_news.pdf. Accessed 23 April 2021.

Greene D, Ryan M (2019) Exploring Selective Web Archives via Network Analysis: An Irish Case Study, Derek Greene, School of Computer Science, University College Dublin, Ireland, Maria Ryan, National Library of Ireland, Ireland. Poster at 2019 LIBER Conference. 10.5281/zenodo.3250157

Hosker R (2020) Beautiful Messy Data: Archival Access and Data Protection. AURA Network Workshop 1: AI and Archives: Current Challenges and Prospects of Born-digital archives, 28–29 January 2021. IE Domain Registry (2021) .IE Domain Profile Report 2020. IE Domain Registry, Dublin. https://www.weare.ie/wp-content/uploads/2021/01/IE-Domain-Profile-Report-2020.pdf. Accessed 23 April 2021

IIPC (nd) Web Archiving: Why Archive the Web? version archived on 13 March 2021. Retrieved from the Wayback Machine: https://web.archive.org/web/20210313224804/https://netpreserve.org/web-archiving/ Accessed 26 April 2021

Ireland. Department of Jobs, Enterprise and Innovation (2011) Consultation on the Review of Copyright and Related Rights Act 2000. Ireland. Department of Jobs, Enterprise and Innovation, Dublin.

Ireland. Department of Culture, Heritage and the Gaeltacht (2017). Public Consultation – Legal Deposit of Published Digital Material. Version archived on 5 March 2018. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/10702/20180305220147/https://www.chg.gov.ie/arts/culture/projects-and-programmes/public-consultation-legal-deposit-of-published-digital-material Accessed 28 April 2021

Ireland. Department of Public Enterprise and Reform (2017). Open Data Strategy 2017–2022. Version archived on 24 August 2018. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/org-1444/20180424235225/https://data.gov.ie/uploads/page_ images/2018-03-07-114306.063816Final-Strategy-online-version1.pdf Accessed 28 April 2021

Ireland. Department of the Taoiseach (2018) Government seeks views on Ireland's Digital Strategy. Version archived on 22 October 2019. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/org-1444/20191022175709/https://www.gov.ie/en/press-release/69baa0-government-seeks-views-on-irelands-digital-strategy/. Accessed 28 April 2021

Jo E, Gebru T (2020) Lessons from archives: strategies for collecting sociocultural data in machine learning In: FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency p.306–316. https://doi.org/10.1145/3351095.3372829

Larivière J (2000) Guidelines for Legal Deposit Legislation. UNESCO, Paris. https://www.ifla.org/files/assets/national-libraries/publications/guidelines-for-legal-deposit-legislation-en.pdf Accessed 28 April 2021

Little C (2011) The general election of 2011 in the republic of Ireland: All changed utterly? West Eur Polit 34(6):1304–1313. https://doi.org/10.1080/01402382.2011.616669

Mackinnon, K (2021). Ethical Approaches to Youth Data in Historical Web Archives (Dispatch). Studies in Social Justice, 15(3), 442–449. https://doi.org/10.26522/ssj.v15i3.2541

MerrionStreet (2017, March 22) Minister Humphreys hosts gender policy workshop with Cultural Institutions.

Milligan I (2019a) History in the Age of Abundance? How the Web Is Transforming Historical Research. McGill-Queen's University Press, Montreal

Milligan I (2019b) Studying the web in the shadow of Uncle Sam : The case of the.ca domain. In: Brügger N, Laursen D (eds) The historical web and digital humanities : the case of national web domains. Routledge, Abingdon, pp 45–63

Milligan I (2020) You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure. WARCnet, Aarhus. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf Accessed 28 April 2021

Merrion Street, News Room. Version archived on 6 April 2017. Retrieved from NLI Web Archive: https://wayback.archive-it.org/org-1444/20170406145929/http://www.merrionstreet.ie/en/News-Room/Releases/Minister_Humphreys_hosts_gender_policy_workshop_with_Cultural_Institutions.html Accessed 28 April 2021

National Library of Ireland (2020) Annual Report 2019. National Library of Ireland, Dublin. https://www.nli.ie/getAttachment.aspx?id=e0e03ff7-08f0-4b23-8935-521b74dc1ebe Accessed 28 April 2021

National Library of Ireland (2011) Submission of the National Library of Ireland to the Copyright Review Committee, Department of Jobs, Enterprise and Innovation: Consultation on the Review of Copyright and Related Rights Act 2000. National Library of Ireland, Dublin. https://enterprise.gov.ie/en/Consultations/Consultations-files/National-Library-of-Ireland.pdf Accessed 28 April 2021

National Library of Ireland (2012) Submission by the National Library of Ireland in Response to Copyright and Innovation: A Consultation Paper Prepared for the Department of Jobs, Enterprise and Innovation by the Copyright Review Committee. National Library of Ireland, Dublin. https://enterprise.gov.ie/en/Consultations/Consultations-files/National-Library-of-Ireland1.pdf Accessed 28 April 2021.

National Library of Ireland (2018) Diversity and Inclusion Policy 2018–2021. National Library of Ireland, Dublin. Version archived on 15 July 2020. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/org-1444/20200715171920/http://www.nli.ie/GetAttachment.aspx?id=f13f83cf-582f-4861-84fb-81845037cd8d Accessed 13 April 2021

National Library of Ireland (2016) Strategy 2016–2021. National Library of Ireland, Dublin. Version archived on 15 July 2020. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/org-1444/*/https://www.nli.ie/getAttachment.aspx?id=ceb18c07-f24d-4fc3-a0c6-82aa1ad5a133 Accessed 13 April 2021

National Library of Ireland (2007) Annual Report 2006. National Library of Ireland, Dublin. Version archived on 15 July 2020. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/org-1444/*/https://www.nli.ie/GetAttachment.aspx?id=0c29359c-49ca-4f62-807a-0f76806e0f99 Accessed 28 April 2021

National Library of Ireland (2009) Annual Report 2008. National Library of Ireland, Dublin. Version archived on 15 July 2020. Retrieved from the NLI Web Archive: https://wayback.archive-it.org/org-1444/*/http://www.nli.ie/GetAttachment.aspx?id=f47939e3-f928-4deb-8352-f24faf7f0622 Accessed 28 April 2021

Schafer V (2019) Exploring the "French Web" of the (1990). In: Brügger N, Laursen D (eds) The historical web and digital humanities : the case of national web domains. Routledge, Abingdon, pp 145–160

Seles A (2020). Artificial Intelligence and Archives. Presentation at Emerging Technologies, Big Data and Archives Webinar 9 June 2020. https://www.youtube.com/watch?v=nmE01ZTA4zI . Accessed 23 April 2021

Talboom L, Underdown D (2019) 'Access is What we are Preserving': But for Whom? DPC Blog. https://www.dpconline.org/blog/access-what-we-are-preserving Accessed 28 April 2021.

Truman G (2016) Web Archiving Environmental Scan. Harvard Library Report. http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314 Accessed 23 April 2021.

Webster P (2019) Understanding the limitations of the ccTLD as a proxy for the national web: lessons from cross-border religion in the northern Irish web sphere. In: Brügger N, Laursen D (eds) The historical web and digital humanities : the case of national web domains. Routledge, Abingdon, pp 110–123

Winters J (2019) Digital History. In: Tamm M, Burke P (eds) Debating new approaches to history. Bloomsbury Academic, London, pp 276–300

Winters J, Prescott A (2019) Negotiating the born-digital: a problem of search. Archives and Manuscripts 47(3):391–403. https://doi.org/10.1080/01576895.2019.1640753

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.