Blog Contribution

How to face the challenges of Web archiving? The experiences of a small library on the edge.

***

National Library of Ireland
Della Murphy, dmurphy@nli.ie
Catherine Ryan, cryan@nli.ie

Internet Memory
Chloé Martin, chloe.martin@internetmemory.net
France Lasfargues, france.lasfargues@internetmemory.net

***

Table of content

1 Context .......................................................................................................................... 2
1.1 National Library of Ireland ....................................................................................... 2
1.2 Internet Memory ........................................................................................................ 2
2 Web archiving project ................................................................................................. 3
  2.1 Project Origins ......................................................................................................... 3
    2.1.1 Current Organisational Requirements; building a 21st century library ............ 3
    2.1.2 The Hand of History ......................................................................................... 3
    2.1.3 'Just do it' .......................................................................................................... 3
    2.1.4 Project scoping; first NLI campaign, the Irish General Election ..................... 4
2.2 Websites selection and Permissions gathering ....................................................... 4
  2.2.1 Selection and Permissions .................................................................................... 4
  2.2.2 Seeds validation .................................................................................................... 5
2.3 Crawl ......................................................................................................................... 5
2.4 Quality Assurance ..................................................................................................... 5
  2.4.1 NLI requirements ................................................................................................ 5
  2.4.2 IM Operations and Services ................................................................................. 6
2.5 Access ....................................................................................................................... 6
  2.5.1 NLI Requirements ................................................................................................ 6
  2.5.2 IM Solution .......................................................................................................... 6
2.6 Collection Publication and Promotion ...................................................................... 7
2.7 Advantages to web archiving for the NLI ............................................................... 7
3 Benefits of a collaborative approach ......................................................................... 8
  3.1 From NLI point of view .......................................................................................... 8
  3.2 From IM point of view ............................................................................................ 8
4 Conclusion .................................................................................................................... 9
1 Context

1.1 National Library of Ireland

The National Library of Ireland was established by the Dublin Science and Art Museum Act, 1877. The Library’s holdings constitute the most comprehensive collection of Irish documentary material in the world and offer an invaluable representation of Ireland’s history and heritage. Material acquired through Legal Deposit, donations and purchases is subsequently processed for storage and access. Providing appropriate storage and care for our collections is a vital part of our work and ensures its preservation for future generations.

The Born Digital programme of the National Library of Ireland had its origins in 2010. It is one of a number of programmes designed to achieve the Library’s strategic objectives in relation to digital material, by delivering solutions encompassing data, processes, people and technology. Our mission statement states that our purpose is to ‘collect, preserve, promote and make accessible the documentary and intellectual record of the life of Ireland’. That documentary and intellectual record is now part digital and increasingly so. The Library collects all manner of publications in print form such as newspapers, books, journals as well as manuscripts, prints, political ephemera and photographs. The big question for the Library is how do we do this in a digital age and ensure that our mission statement is still relevant?

The National Library of Ireland recognises the intrinsic cultural value of born digital materials, and the importance of preserving this material for current and future generations. We also recognise this opportunity of working with Born digital collections as a way to enable users to work with the NLI and NJU collections, in new and innovative ways. A staff member was put in post as Born Digital Programme Manager in October 2010. The overall aim of the Born Digital programme is ‘to identify the role of the National Library of Ireland in relation to the collection of born digital material and to identify, collect and make accessible born digital material as part of day to day NLI activities’. This programme can be broken down into a number of high level policy activities and everyday practical activities, and in its initial phase is operating through a series of iterative pilot projects, in the areas of web archiving, e-publications and e-manuscripts.

In 2011, in conjunction with Internet Memory, we carried out two web archiving pilot projects in relation to the General Election and the Presidential Election. This material is now available on our web site. In relation to 2012, our priority pilot project relates to the area of e-publications. We are conducting a 6-month pilot project to ingest and make accessible e-government publications from a selected group of Government Departments, in the absence of legal deposit legislation i.e. on a voluntary basis. In 2013 we hope to conduct a pilot project in the area of born digital archival material.

1.2 Internet Memory

Internet Memory Foundation, a non-profit institution, was established in 2004 in Amsterdam under the name European Archive Foundation, to support and develop Internet Archive in its mission in Europe. In 2010, it changed its name to Internet Memory Foundation to express its interest in preserving web content as a new media for current and future generations. Also, cultural or heritage institutions are not always in a position to build a full internal project for their web archives for resource and expertise reasons, so many of them need recourse to an outsourced solution, such as that provided by Internet Memory. It enables them to keep complete control on their main focus (selection, crawling parameters, preservation and usage) without dealing with operational workflow issues (qualified human resources, crawl indexing technologies, access issues...)

Currently Internet Memory Foundation hosts hundreds of terabytes of archived websites on open access including its own collection and collections from partner institutions. This includes organisations such as The UK National Archives, the National Library of Ireland and CERN (where the Web originated 20 years ago).

To operate the workflows that facilitate web archiving, Internet Memory Foundation closely collaborates with a young organisation operating in the same field. Internet Memory Research (IMR) was established in 2011 in Paris as a spin off from Internet Memory Foundation. Internet Memory Research works with many institutions in Europe: Institute for Sound and Vision in the Netherlands and several German audio-visual archives, the National Library of Ireland, Research Projects (Inside Installations, enpolitique.com,...), CERN, etc.

Both of these organisations actively support the preservation of the Web for heritage and cultural purposes. To fulfill their goals and because the Web is dynamic, innovative, unlimited, and ephemeral both need to engage in state of the art web archiving activities to help develop technologies in this field. So, Internet Memory Foundation and Internet Memory Research are engaged in several research projects to support the growth and use of web archives. The first
organization is more focused on preservation and web-scale questions, while the second focuses on web archives usage.

2 Web archiving project

2.1 Project Origins

2.1.1 Current Organisational Requirements; building a 21st century library

The NLI embarked on a number of integrated work programmes to build a 21st century library. These work programmes covered areas such as Born Digital collections, digitisation of our holdings, producing a single integrated catalogue, developing the technical hardware and software necessary for a 21st century library, including the development of a digital repository and an organisational review of priorities which evolved into the ‘OSCAIL’ (the Irish word for Open) programme. The identification at a high level of the additional staff and financial resource requirements to build a 21st century library came at a time of well documented and unprecedented financial difficulty for our country and therefore for the Library. At a time when we had identified additional resources required to move forward, we soon realised that not only would these additional resources not be an option, but as an organisation we were likely to suffer a reduction in staff and budget. It is in this environment that we are trying to address the digital challenges as outlined above.

It was envisaged that the Born Digital Programme manager’s job was to look at requirements and policy at a high level and devote 2011 to producing a policy document outlining the next steps in relation to born digital collecting for the NLI. But external events and pressures came to the fore in the shape of a General Election called at short notice, at a time of huge political upheaval. The Library quickly realised that the time for thinking about policy direction, staffing requirements, financial requirements for the Born Digital programme had to be put on hold. We had to adopt a ‘just do it’ approach.

2.1.2 The Hand of History

Externally, the historic nature of the election, coupled with the increasing use of social media by political candidates and our knowledge that no other organisation was in a position to collect and archive websites at the time, pushed us towards investigating whether or not we as an organisation could. Internally, one of our collecting strengths had always been in the political area, with strong political ephemera and manuscript collections as well as a near complete newspaper collection providing a robust print collection in this area. On one level the decision to collect digital material in relation to the General Election for the first time was a ‘no brainer’. It would complement our existing holdings and also fill an ever increasing digital gap in our collections. On another level however, given the structure and position of our organisation within the public sector, the difficulties embarking on this new collection area were still considerable, particularly as we had only a five week window in which to work. As we had never done this type of work before, we realised very quickly that we had to operate on a collaborative basis if we wanted to get the collection organised in the required timeframe.

We needed to find a partner that best suited our requirements at that point in time, and also had experience of working with organisations in our sector, the cultural sector. We knew that we had a small number of staff in the Library with some of the technical skills necessary to engage in work of this nature, but we also knew that given our own shortage of resources, their skills were required elsewhere within the organisation. We wanted to leverage our experience as an organisation in collecting material in relation to politics in any partnership we embarked on. We had already established informal contacts with IM as part of the usual networking process, so we conducted an exploratory conversation via e-mail in relation to timeframes, cost and requirements for this project, and parallel to this conversation we put in process the internal business case required to proceed. We could not afford to wait for one side of the process to conclude before embarking on the next, so the next fortnight was a somewhat chaotic, out of sequence timeline, that resulted in a green light for a collaborative project in a completely new area for the NLI.

2.1.3 'Just do it'

The project broke down into a number of phases and on the NLI side involved a range of staff at different times during the project. Day to day project management required two staff who were only available part time. Our duties included:

- Project scoping and finalisation of contract with Internet Memory
- Site selection
- Permissions gathering
• QA (look and feel)
• Publication and promotion

At various times along the process we needed to consult with our own colleagues in print and visual collecting areas, as well as our own Information Systems team and with Internet Memory through e-mail and conference calls. Our Keeper of Collections provided overall guidance for the project. Some parts of the project are worth highlighting and are outlined below.

2.1.4 Project scoping; first NLI campaign, the Irish General Election

The National Library of Ireland contacted Internet Memory 5 weeks before the Irish General Election (February 2011). Time pressure was a factor in this project, as the event date was fixed, so there was a sense of urgency around getting this project up and running.

By week 2, IM put a proposal to the NLI, outlining 3 different types of Web archiving campaign. Each of them included one snapshot before Elections and one snapshot after:

• Domain crawl (ie): large snapshot with a crawl period of 3-4 weeks each, without manual Quality Assurance.
• Selective crawl: deep crawls (up to 500 units, according to our machine availability at that time), with manual QA
• Thematic crawl: from a small number of websites (eg.5 like political hubs), IM will have collected all discovered links in pyramidal type during a crawl period of 2 weeks, without manual QA

The selective approach was chosen, which was more relevant for General Election purposes. The number of units (100 seeds) was defined by NLI and were determined by IM and NLI needs and costs. In the current financial context, both institutions had to settle on which cost basis they could collaborate. By the end of 2 weeks, the budget and scope of the project were clarified, and formal approval from the National Library of Ireland was received to proceed with the project. In the meantime, on both sides, teams were already working on the project: NLI on the URL list and IM on the contract and on putting in place infrastructures for this campaign (human resources and machines). This meant the NLI was ready to proceed by the time the project received formal approval.

2.2 Websites selection and Permissions gathering

2.2.1 Selection and Permissions

We had a number of different selection criteria to help us determine what sites to include as part of the archiving process, ranging from general to the specific. Websites themselves were chosen to present a rounded picture of activity at election time with more formal information communicated on the official government and candidate websites and more reactive comment on the blogs/ commentary websites. General selection criteria included the following;

• Website presence. Websites were evaluated on the basis of use. Only those sites that received regular and recent updates were included. Token websites were in the main excluded.
• Technical reasons. Websites that were ‘clunky’ or that contained substantial moving image (YouTube) or flash were more likely to be excluded as there was no certainty around final capture.
• Cut-off date. Sites had to be selected in a very short time frame which meant that some sites may not have been up and running in a meaningful way by that time and were therefore not included.
• Women Candidates. We consciously tried to include women candidates.
• Twitter has a collecting arrangement with the Library of Congress (USA) so where decisions between sites for inclusion were tight our preference was for the non-Twitter sites.

Site specific criteria was determined by the category of site in question. For example with political party sites and political candidate sites we wanted to ensure that all parties were represented and that there was full geographic coverage. We also engaged in a very rewarding consultation process with a number of academics with a known interest in the field, and asked them to recommend sites for inclusion. Sites in the Irish language were also archived, and some sites were chosen specifically because they complemented our paper based collections such as newspaper and our visual material. There were a number of sites that contained creative visual political ephemera which was for
us a clear example of how campaigning had moved online in an innovative way. These sites proved trickier to archive from a technical point of view.

We contacted all participants in the project, which surprisingly sometimes required postal contact, and briefed them on the project. Many respondents were delighted to be included, and only one site declined to participate, and that was due to the fact that they were closing down their site. The permissions gathering phase was a very pressurised but necessary part of the project.

2.2.2 Seeds validation

When IM received URLs lists from NLI, the IM team had to comment and highlight specific points to improve crawl results. IM checked:

- URLs: are they correctly written?
- Duplicated websites, eg. when one domain AND one subdomain were specified
- Redirects, which would not be followed by the crawler considering them as out-of-scope
- Website parts which will not be captured because they will be out-of-scope due to external links
- Dynamic websites, which are difficult to crawl
- Links to defined social Web (Facebook, Twitter and YouTube) to include them in the crawl
- Server hosting multimedia content (on the same domain, or not)

All comments and questions were then discussed during conference call to finalise changes.

2.3 Crawl

Seeds were gathered in batches using the following scope parameters: domain, host and path. Specific attention was paid to social web content in order to crawl home page of Twitter, Facebook and Youtube pages. The General Election crawls ran twice, before and after the elections with a similar URL list. The crawls were launched according to NLI technical parameters. Robots.txt files were not followed and a high level of politeness was setup. A Webmaster page (http://www.nli.ie/crawl-ge2011.html) was designed and published on NLI website. It gives some information on the crawl’s context, technical details and contact information.

Specific incidents occurred during the crawl period and required some technical changes on the fly:

- Modification of scope due to offline domain which redirects to another site
- Pending crawls due to maintenance operation on website
- Adaptation of the politeness due to Webmaster complaints
- Inclusion specific rules for robots.txt to get more deep content

Feedback from the crawl and QA time were used to improve the quality of the post election crawls by adding:

- Associated URL to help to catch the target content
- Politeness adjustment related to webmaster feedback

2.4 Quality Assurance

2.4.1 NLI requirements

It was very important for NLI that Internet Memory carried out as complete QA as possible, for resource and expertise reasons. We felt that the nature of the project suited manual QA. We were therefore in a position to carry out basic 'look and feel' QA, using software supplied by Internet Memory, which though it was new to us proved relatively easy to use. Our input was necessary because we had a clearer picture as to what aspects of the collection were most important to us, unlike Internet Memory which was not in a position to assign value of that nature. We also carried out QA using different browsers. As with any other NLI collection purchase where money is expended we felt it necessary to perform these activities to ensure that the product delivered was the product required. At times as part of our QA process it was necessary to communicate with website owners, which helped build a relationship with them that ultimately helped in the wider promotion of our web archiving activity. This was an unintended but very welcome consequence of our QA activities, and we view it as opening up a new set of relationships with potential contributors/stakeholders in future digital collecting activity at the Library.
2.4.2 IM Operations and Services

Capturing Web content is full of challenges and harvesting tools used have clear technical limitations. Because of these limitations and the now obvious incompleteness of Web archives, most European institutions are using Quality Assurance methodologies and tools that can be applied to Web archives.

The most used Quality Assurance methodology is the visual method, followed by a capture of the identified missing resources. It applies mainly to selective harvesting and consists in visually checking pairs of Web Pages: the live version versus the archived version. It implies the use of the live version as a reference when checking quality of captured content and therefore means that it should always be done as fast as possible after the crawl, due to the ephemerality of Web content. IM offers the option of a methodological human driven QA, which includes repairing operations. The QA team takes prompt action to resolve as many issues identified as possible. Levels of quality assurance, automatic, manual (light/1 or deep/2) were defined to match NLI's requirements:

- **QA level 1**: includes both QA and reparation action on the main entry points of the sites (homepage+1).
- **Chosen by NLI: QA level 2**: deep manual QA relying on methodological checking for each crawl made (homepage+2).

QA reports were operated through a bug tracking tool Jira: [http://www.atlassian.com/software/jira/overview](http://www.atlassian.com/software/jira/overview).

In this particular web archiving campaign, temporal coherence was an important criteria because during the election period, web content of many of the selected sites changed fast. To guarantee a high quality of its web archive, NLI had chosen a QA Level II, which meant that the content of snapshots were checked from the primary URL to the level II (homepage + 2 hops) up to 2 hours per snapshot. The calculation is simple: 100 x 2h = 200 hours of QA, based on an 8 hour working day. IM needed at least 25 man-days to carry out QA to agreed levels. So, for the first snapshot (pre election) NLI had to prioritize some sites over others for the QA process before the date of the election because all crawls were either not finished or just finished at this date. The snapshot approach follows a full QA process before being validated. The NLI could either validate QA tickets by clicking on "Accept instance", ask for clarification or add some QA requests of their own by clicking on "Reject instance": Upon NLI's request, validated crawls were "published" in the main NLI index for public access after validation. The NLI preference was to complete all QA prior to publication of the collection.

2.5 Access

2.5.1 NLI Requirements

With ease of use a key objective, provision of full text search to the collection is very important to us. There are a number of ways to access the collection. From the IM’s website, users can browse by URL, or carry out keyword searches across the full-text of the collection. Additionally, our Information Systems team at the National Library of Ireland in collaboration with Internet Memory developed a widget based on the OpenSearch protocol which displays results from the web-archive collection directly in our catalogue alongside our other resources. We intend to build on this work in the future to make the collection even more accessible through our own interfaces, and indeed the issue of a huge results set from the web archive is one of the factors compelling us as an organisation to look at how best to integrate catalogue results from all our collections on our website.

2.5.2 IM Solution

Access to the archive is provided through a wayback-machine-like approach (index page, navigation with on-the-fly link rewriting, banner on top of pages, etc.) using IM internal access tools. The IM access tool was the first tool to implement a server-side link rewriting 4 years ago and has been constantly improved since then. It enables IM to implement custom rewriting rules for difficult sites or to embed custom video players and other advanced on the fly presentation improvements.

It includes:

- A front-end web server that fetches the closest in time archived content for a specific URL and reference date. Resources of a certain type get rewritten on the fly, such as HTML documents that have their links modified so that they point to the archive, or a banner can be inserted.
A distributed index that allows fast access to a resource given a URL and a date.

IM provides a full-text search service based on Lucene. Currently, all our collections are indexed with Lucene and the web interface access provides a keyword-based search facility. This access tool is already supporting millions of requests per month.

Added to this, IM provides an OpenSearch interface to the full text index to let external systems query it programmatically. This allows, for instance, inclusion of full text search results on a third party website. The XML responses conform to OpenSearch version 1 and contain for each hit, among others, the original URL, the date when the URL was retrieved and a snippet showing the text around the searched keywords.

2.6 Collection Publication and Promotion

IM were in a position to make all sites that had fully gone through the QA process available immediately. However NLI preference was to wait until the process was complete for the collection as a whole as we wanted to alert the public to it in that way. Given that this was a new collecting activity for us, and one that we were anxious to promote, we had to do some thinking around how best to do this. Would we reach our intended (and new) audiences through traditional print media or would we have to do something different? It turns out that the answer is a combination of the two. We were fortunate that at the time we were ready to publish the General Election web archive, the NLI was establishing an online social media presence on Twitter in particular, and had also developed a blog, which we were able to use to promote the collection (http://www.nli.ie/blog/index.php/2011/10/26/generalelection-2011-web-archiving/).

Some of the participants in the project became involved in promoting the project online also. The print media in the area of technology were interested in what we were doing and we noticed that when they wrote about our activities there was a spike in usage. The collaborative nature of the project was a major selling point from our point of view, and of course Internet Memory promoted the collection on their side also. Usage figures have increased over time, however we feel that the real value of this collection will only be realised fully in the medium term (5-10 years) as more and more information is ‘born digital’ and research patterns and expectations adjust accordingly.

2.7 Advantages to web archiving for the NLI

Collection of digital material opens up new opportunities in terms of delivery of material to our users. It allows the National Library of Ireland explore the possibility of being able to reach new audiences of people who cannot for whatever reason travel to Dublin and also continue to work with many of our users who quite rightly expect to be able to conduct their research online.

We felt that this collection in particular had the following potential uses;
To enable researchers to compare online content before and after the election using pre- and post- election sites for comparison.

To facilitate research into how ‘online’ this election was.

To assess the impact of technological developments such as social media and web 2.0 in relation to the communication of campaign information.

To act as a record of campaign information.

3 Benefits of a collaborative approach

3.1 From NLI point of view

For the National Library of Ireland there were numerous advantages in undertaking a collaborative approach to web archiving. We knew at the outset that this was an activity that we wanted to engage in on a long term basis and very much viewed our first collaborative project with Internet Memory as a pilot project, which;

- Provided us with a mechanism to enter into a new collecting area despite our lack of resources and expertise
- Facilitated collection of material that was of high historical importance and that no one else in our country was in a position to collect at that time
- Enabled us to collect this material relatively quickly in time for a specific event of historical importance
- Allowed us to maximise and leverage our own curatorial skills and institutional knowledge in the selection process without having to worry about the technical aspects
- Allowed us gain new technical skills with good supervision and back up from our partner organisation in a collecting area that we hope to continue to build on

3.2 From IM point of view

For IM, the benefits of collaborating with heritage institutions, such as the National Library of Ireland include helping

- To operate rapid deployment of Web archives and to develop Web archiving initiatives
- To address new challenges in this area: for example, our partners (including NLI) really appreciate specific effort on social media content.

On one hand, this material is both ephemeral and highly contextualized, making it increasingly difficult for archivists to decide what to preserve. A recent Library of Congress blog post refers to the fact that every day a quarter billion photographs are uploaded to Facebook, 300 billion e-mails are sent and 340 million tweets are posted to Twitter. 150,000 new websites are added to the web every day. On the other hand, the social web is based on dynamic websites and uses specific technologies to publish content. Thus, collecting the social web requires new tools as well as new strategies.

So Web archiving institutions and consortia address these challenges in a number of ways. They publish and circulate guidelines on crawls parameters. (See for instance guidelines published by the Internet Archive to enhance the quality of social media content: https://webarchive.jira.com/wiki/display/ARIH/Archiving+Social+Networking+Sites+with+Archive-It). Moreover, organisations such as The European Commission funds research projects to address this issue such as the ARCOMEM project (Collect-All ARchives to COmmunity MEMories - http://www.arcomem.eu/) in which Internet Memory is involved. In contrast to classic web archiving crawlers, ARCOMEM will provide facilities for extracting complex objects from the web, by changing the behaviour of a crawl depending on the kind of web application targeted, and by interacting with content analysis modules. An “Application aware crawler” is being developed for this purpose. This module aims at making a crawler aware of the particular kind of web application it is crawling, in terms of its general classification (wiki, social network, blog, Web forum, etc.), its technical implementation (MediaWiki, WordPress, etc.), and even its specific instance (Twitter, CNN, etc.), in order to adapt the crawling strategy to the task at hand. The crawler will rely on a hierarchical knowledge base of web applications, which specifies how to recognize an application and how to crawl it. This could be very efficient for thematic and selective campaign such as Web archives about Elections.
• To improve technologies based on the needs of our partner institutions and to assess needed processes we should automate to reduce processing time and costs: automated Quality Assurance to match with temporal coherence challenges

The “visual” methodology described above, if it provides good results on selective harvesting, obviously requires trained human resources and is time and money consuming. To overcome these difficulties and apply quality review on larger sets of Web content, institutions and companies are now looking into implementing automated or semi-automated QA.

A first option is to use metrics related to crawls as references and comparison tools. This method, that can be automated to detect problematic crawls, is useful when crawling the same list of domains at different frequencies. For instance, one could decide to store metrics from one reference crawl, which would have gone through visual checking and compare them to all future crawls during a set period of time.

Other options consist of developing specific QA tools based on proxy, execution (Selenium - http://seleniumhq.org/) or image comparison tools (see for example the project SCAPE - http://www.scape-project.eu/), to mimic “manual” QA and either detect missing elements to fetch and index them automatically, or evaluate quality of crawls at a high level to minimize human intervention and reduce costs.

4 Conclusion

Also, as hopefully this article has shown, web archiving institutions use the knowledge and experience gained over the years they have spent as part of larger organisations involved in the collecting, preserving and provision of access to material of historical and cultural importance, to inform to help create collections of long term value.

In the framework of this partnership, including the two web archiving Campaigns (General Election and Presidential Election), the National Library of Ireland, with the collaboration of Internet Memory, archived:

• General Election
  o 18,495,771 URLs
  o 1,14 Tb
  o 10,405 ARCs
• Presidential Election
  o 7,333,399 URLs
  o 278,10 Gb
  o 2,513 ARCs

After completion of crawls, Internet Memory has provided an interface through which ARC files containing the crawled resources, with all the associated metadata, were downloadable by NLI. Thus all data are preserved by the National Library of Ireland, which retains control on its preservation policy.

Moreover, and with NLI permission, Internet Memory Foundation also hosts and gives access to a copy of these collections. Currently Internet Memory is moving to a sophisticated distributed Storage engine (HBase on top of HDFS), which automatically manages replication and fault-tolerance. Any resource inserted in the system is automatically replicated, and the failover mechanism ensures that this replication is maintained when a disk failure occurs by creating a new copy. Although this native storage conforms to the Open-source Hadoop file system representation, IM also provides an import/export mechanism based on the standard WARC format.

To learn more on these projects and institutions:

• http://www.nli.ie/blog/index.php/2011/10/
• http://internetmemory.org
• http://collections.europarchive.org/nli/