20 DHO: Discovery – Stargazing from the Ground Up

Niall O'Leary[1]

Abstract

In 2008, the Digital Humanities Observatory was charged with creating an all-island gateway to Irish digital collections and resources. A key factor in the achievement of this goal was the development of the web application, *DHO: Discovery*. This chapter will describe what *DHO: Discovery* was intended to be and to what extent it achieved this end. In particular it will detail the infrastructure upon which the system is based and explain how external and internal factors affected the choices made in its development. It will describe the approach taken to the data harvested for the system, recounting how it was gathered from a variety of national repositories and showing how the responses of providers to the project shaped its development. *DHO: Discovery* stresses the importance of strong metadata and highlights the importance of best practice and standards-compliant development. Through the use of internationally recognised standards, *DHO: Discovery* can be used as a basis for further development. The development of several innovative data visualisations demonstrates how some of this potential has already been realised. The design of these tools will also be described. In charting the development of *DHO: Discovery*, it is hoped that lessons learnt might contribute to the future development of digital humanities in Ireland.

Keywords: digital humanities, repository, Fedora Commons, SOLR, metadata, Europeana, sustainability, visualisations.

[1] Digital Humanities Observatory, Dublin, Ireland; n.oleary@dho.ie

How to cite this chapter: O'Leary, N. (2013). DHO: Discovery – Stargazing from the Ground Up. In C. Fowley, C. English, & S. Thouësny (Eds.), *Internet Research, Theory, and Practice: Perspectives from Ireland* (pp. 405-426). Dublin: © Research-publishing.net.

1. Introduction

The Digital Humanities Observatory (DHO)[1] was established in 2008 as part of the Humanities Serving Irish Society (HSIS)[2] consortium. The DHO was established to

- identify standards and advise on best practice in the area of digital humanities;
- store, preserve and provide access to the increasingly complex range of e-resources available to the humanities in Ireland;
- enable the fullest exploitation of existing national research collections and data repositories.

Reporting to the Royal Irish Academy, the unit (initially comprised of nine people with skills ranging from data encoding to project management) encourages the innovative use of technology in what has traditionally been a pen and paper domain, promoting internationally accepted standards and best practice.

Key to the DHO's work is its website (http://www.dho.ie) which it uses to disseminate news and promote its activities. This website is also the access point for many of its online developments. The DHO has collaborated with other projects to build several high-profile websites such as the Saint Patrick's Confessio Hyperstack[3] and the Doegen Records Web Archive[4], but in pursuing its own remit it has striven to create its own suite of services. For instance, in an effort to promote existing digital humanities projects across Ireland, one of the first online services the unit developed was *DHO: Drapier*[5]. This interactive database, available from the main DHO website, surveys current activity in Irish digital humanities and documents the methods, formats and standards that are being used. New projects can learn to use similar techniques, or established

[1] Digital Humanities Observatory, http://dho.ie/

[2] Humanities Serving Irish Society, http://hsis.ie/. The HSIS consortium is comprised of the Royal Irish Academy, six of the seven Irish universities, Queen's University, Belfast and the University of Ulster.

[3] Saint Patrick's Confessio Hyperstack, http://www.confessio.ie

[4] The Doegen Records Web Archive, http://www.dho.ie/doegen

[5] Drapier: Digital Research And Projects in Ireland, http://dho.ie/drapier/

projects can locate related work in their field. However, *DHO: Drapier* does not provide access to existing data repositories or make collections of digital material available for analysis. To achieve this, the Digital Humanities Observatory set about developing another online service, *DHO: Discovery*.

The development of *DHO: Discovery* was initiated when the Observatory had a full contingent of staff and a number of years to run. In 2010, I joined the DHO as IT Projects Manager and became involved in the project. In many ways it provided a practical opportunity for me to further my understanding of digital humanities and contribute to the area. However in 2011, the fortunes of the Observatory itself changed radically. Initially given a three-year lifespan, it was hoped that the DHO might acquire additional funding (and maybe even staff) over time. This was not to be. By June 2011, all DHO staff had reached the end of their contracts with little prospect of renewal. Funding stopped in August 2011, but fortunately enough was left to maintain a staff of two. In this reduced form, the unit will continue until August 2013. This has created obvious pressures and has forced the DHO to be creative and extremely cost-effective in developing *DHO: Discovery*.

In this chapter, I would like to describe *DHO: Discovery*, chart its development and share some of the lessons learnt from my experience of the project. In charting its development, I will address the creation of the infrastructure, the gathering of data, and the subsequent innovations the system made possible.

2. What is DHO: Discovery?

On a very basic level, *DHO: Discovery* is a search engine. It allows a user to find related items from many different online sources using one integrated interface. However, it has many differences to traditional search engines such as *Google* or *Bing*. Concentrating on data from national cultural and academic sources, it only deals with online digital objects suitable for primary research. Such objects include texts, images, video and audio. It uses accepted standards, such as Dublin Core, to allow faceted searching, i.e., searching on the basis of fields such as creator, subject matter, date of creation, etc., not always available on the original websites. *DHO: Discovery* provides several ways to navigate content, but always provides links back to the original websites. This makes it easier to create new connections between heterogeneous content from multiple sources. For instance, rather than having to go to the individual websites of Trinity College Dublin (TCD), University College Cork, etc., for information on Michael Collins[1], a search through *DHO: Discovery* will find all relevant digital objects from the principal research repositories of these institutions, and a variety of cultural institutions, displaying these results with descriptions, thumbnails, rights details, spatial data, and other information. Users can discover Ireland's many online treasures by visiting this one online service. In addition *DHO: Discovery* offers researchers new ways to visualise data and hopefully gain new insights into old content (Figure 1). Indeed this is one of the key functions of the system. *DHO: Discovery* tries to demonstrate some of the exciting possibilities available when internationally accepted standards are applied to digital content and explore new methods of processing humanities data.

There are some precedents to *DHO: Discovery*, the most obvious being Europeana[2], a huge online database of European art and culture. With over twenty million objects, including a lot of Irish content, this allows faceted searching on a grand scale. However, despite a small amount of work on visualisation (Europeana4D[3], for instance), Europeana has predominantly concentrated on its search tool. Its Irish content is also somewhat limited, coming mainly from large Irish providers such as the National Archives and the National Library, but ignoring many universities and smaller institutions.

From an American perspective, some moves are afoot to replicate the Europeana service with the Digital Public Library of America[4], but this is still in development. The principal competitor to such initiatives is of course *Google*. *Google* has its

[1] Michael Collins (16/10/1890 – 22/08/1922) was an Irish revolutionary leader and politician. Collins was shot and killed during the Irish Civil War.

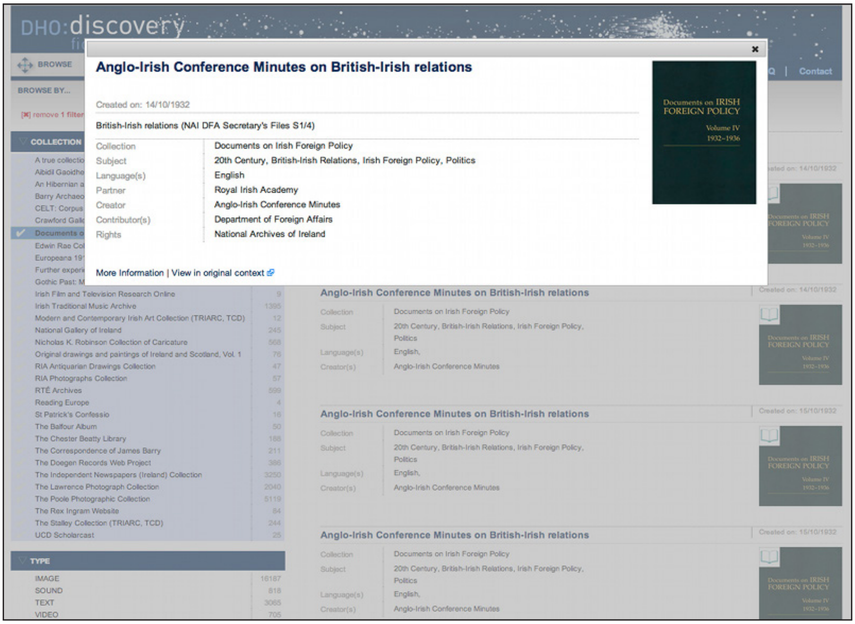
[2] Europeana, http://www.europeana.eu/portal/

[3] Europeana4D, http://wp1187670.wp212.webpack.hosteurope.de/e4d/

[4] Digital Public Library of America, http://dp.la/

own Google Books and Google Art projects[1], each of which covers large and important collections of works. Again though, it has tended to concentrate on the search and display of its content, generally ignoring the potential of metadata. It also seems to be dealing with each type of media in isolation; books and artworks are treated differently[2].

Figure 1. Detail on a document from the Documents on Irish Foreign Policy collection, displayed in DHO: Discovery



Anglo-Irish Conference Minutes on British-Irish relations

Created on: 14/10/1932

British-Irish relations (NAI DFA Secretary's Files S1/4)

Collection	Documents on Irish Foreign Policy
Subject	20th Century, British-Irish Relations, Irish Foreign Policy, Politics
Language(s)	English
Partner	Royal Irish Academy
Creator	Anglo-Irish Conference Minutes
Contributor(s)	Department of Foreign Affairs
Rights	National Archives of Ireland

More Information | View in original context

Documents on IRISH FOREIGN POLICY Volume IV 1932–1936

COLLECTION

A true collectio	
Aibidil Gaoidhe	
An Hibernian a	
Barry Archaeo	
CELT: Corpus	
Crawford Gall	
✔ Documents o	
Edwin Rae Col	
Europeana 19	
Further experi	
Gothic Past: M	
Irish Film and Television Research Online	9
Irish Traditional Music Archive	1395
Modern and Contemporary Irish Art Collection (TRIARC, TCD)	12
National Gallery of Ireland	245
Nicholas K. Robinson Collection of Caricature	568
Original drawings and paintings of Ireland and Scotland, Vol. 1	76
RIA Antiquarian Drawings Collection	47
RIA Photographs Collection	57
RTÉ Archives	599
Reading Europe	4
St Patrick's Confessio	16
The Balfour Album	50
The Chester Beatty Library	188
The Correspondence of James Barry	211
The Doegen Records Web Project	386
The Independent Newspapers (Ireland) Collection	3250
The Lawrence Photograph Collection	2040
The Poole Photographic Collection	5119
The Rex Ingram Website	84
The Stalley Collection (TRIARC, TCD)	244
UCD Scholarcast	25

TYPE

IMAGE	16187
SOUND	818
TEXT	3065
VIDEO	705

Anglo-Irish Conference Minutes on British-Irish relations — Created on: 14/10/1932

Collection	Documents on Irish Foreign Policy
Subject	20th Century, British-Irish Relations, Irish Foreign Policy, Politics
Language(s)	English,
Creator(s)	Anglo-Irish Conference Minutes

Anglo-Irish Conference Minutes on British-Irish relations — Created on: 15/10/1932

Collection	Documents on Irish Foreign Policy
Subject	20th Century, British-Irish Relations, Irish Foreign Policy, Politics
Language(s)	English,
Creator(s)	Anglo-Irish Conference Minutes

Anglo-Irish Conference Minutes on British-Irish relations — Created on: 15/10/1932

Collection	Documents on Irish Foreign Policy
Subject	20th Century, British-Irish Relations, Irish Foreign Policy, Politics
Language(s)	English,
Creator(s)	Anglo-Irish Conference Minutes

In Ireland, there are also some notable developments that deal with some of the issues *DHO: Discovery* set out to address. Trinity College Dublin has used the TARA institutional repository to store much of its digital resources[3]. Effectively

[1] Google Books (http://books.google.com) and Google Art (http://www.googleartproject.com).

[2] In October 2012 *Google* launched its Cultural Institute (http://www.google.com/culturalinstitute/#!home) which addresses this shortcoming to some extent.

[3] Trinity's Access to Research Archive (TARA), http://www.tara.tcd.ie/. More recently the Digital Resources and Imaging Services Department at Trinity College Library Dublin launched its own digital repository, http://digitalcollections.tcd.ie/home/

this is a database of images, texts, and other media, stored with appropriate metadata. There is sufficient metadata associated with each object to satisfy the requirements of several recognised digital library formats, for instance Dublin Core or METS, and the objects can be found using a search engine responsive to this information. University College Dublin (UCD) takes a similar approach with its repository, the UCD Digital Library[1]. Repositories such as these also act as hosting mechanisms, making the data they hold available on the Internet. However in each case these services are restricted to the university's own collections (or even subsets of those collections). In addition, there has been little work done on alternative approaches to navigation and visualisation of the data held in these repositories.

In 2008, the DHO set out to address the shortcomings of the systems mentioned above, or at least indicate how they might be addressed. To do this it decided to develop a system that would

- allow users to find content of research interest in an easily navigated way;
- provide a unified interface for national digital resources of digital humanities interest;
- demonstrate the potential of technologies available for digital humanities;
- allow for the serendipitous discovery of objects in disparate collections, creating new connections;
- use the platform as the basis for the development of visualisations that reveal more than any one site can do by itself.

In revealing the many data collections available throughout Ireland it was hoped that it would increase the visibility of these research resources and, as a by-product, drive the growth of digital humanities scholarship.

Given the DHO's national remit, and taking a cue from the work at UCD and TCD, it was decided to develop a national digital repository that would hold the nation's digital objects and associated metadata. In tandem with the development and population of this repository, a website, *DHO: Discovery*, would be

[1] UCD Digital Library, http://digital.ucd.ie/

developed, accessible from the DHO main website. This website would allow a user to navigate and make sense of the repository's contents. Once the main site had been created, a series of visualisations would be developed to showcase what was possible using the data.

3. Building the infrastructure

A digital repository stores, manages and provides access to digital content. In 2008, two of the leading open-source repository solutions were Fedora Commons and DSpace[1]. The limited resources available to the DHO called for a development framework that was flexible, powerful, standards compliant, and allowed for rapid development. It should also be cheap (preferably free). While Fedora Commons and DSpace, both satisfied many of these needs, Fedora Commons provided more flexibility (Gourley & Viterbo, 2010) and was adopted as the technology set for this repository.

Fedora Commons is not a repository in its own right, but is instead a digital asset management (DAM) architecture. It provides the foundation upon which many types of digital library might be built. Unlike a package such as Microsoft Word, it is not a complete application itself, but is instead built from distinct modules. A developer must choose these modules and put them together himself. In this way, the final product reflects the needs of the individual system. For instance, out-of-the-box searching is achieved using a search engine module developed by a third party, the company Lucene, but this can be swapped with other more powerful modules. In our case, we swapped it for Lucene's own enterprise solution, Apache SOLR[2].

The decision to use the SOLR module was not an arbitrary one. SOLR provides a powerful full-text search tool, allowing faceted searches, scalability, and a lot of opportunities for integration with other systems. Crucially though, the

[1] In 2009, the Fedora Commons organisation and the DSpace Foundation, providers of these software products, amalgamated to create Duraspace.

[2] Apache SOLR, http://lucene.apache.org/solr/

website that would give users access to Fedora was being developed using a programming language called JavaScript, specifically a library of JavaScript code called JQuery. JavaScript is a programming language used extensively for the web. It is predominantly a client-side language (works on a user's own machine), but can be used in other ways. Many JavaScript libraries, such as JQuery, are available that make development easier, providing ready-made functions for common tasks. The JQuery library supported the use of SOLR, which made the search engine key to the website. In essence SOLR connected the website to the repository.

Towards the end of 2009, the website was progressing well and Fedora was installed. DHO staff began to turn their attention to content. As well as looking for digital content for the repository, they began to devise a data model to reflect the objects that would populate the system. As important as the technologies used, the model by which one maps one's data is crucial. In order to properly index a digital object within a repository, one must enumerate the various fields, or facets, by which a user might categorise that object. There are obvious metadata elements that should be recorded: the title of a painting for instance, or the author of a particular book. However, a richer model is required when allowing users to search and browse resources from across different collections. Though it is tempting to create a bespoke model, this can be dangerous. A facet might be left out at the outset that proves necessary as the project grows (what about an alternative title for the object?). Some terms might be too ambiguous, not allowing enough specificity. Without the right data model, a project cannot be integrated with similar projects. Adopting an arbitrary model might doom your project from the outset by restricting its scalability or flexibility. A recognised model, by contrast, ensures that your data can be understood internationally. In the case of *DHO: Discovery*, it was important that the system could be developed by other parties and integrated with other systems.

As has been mentioned, other initiatives had already addressed these issues, for instance Europeana. As it happens, the DHO was working with the Irish Manuscripts Commission to aggregate content for Europeana and it was highly likely that content harvested for *DHO: Discovery* might eventually find its way into Europeana as well[1]. Using a variation of the Europeana data model would allow us to benefit from their expertise, future-proof *DHO: Discovery* and allow for future interoperability between the two systems[2]. In addition, the Europeana data model adopts many terms from Dublin Core, a model established by the Dublin Core Metadata Initiative (DCMI) for the very purpose of describing resources for discovery[3]. By using Europeana's model we ensured future compatibility with other Dublin Core compliant systems. For all these reasons, with very little tweaking, Europeana's data model was adopted as our own.

4. Gathering content

In 2010, as the data model was being finalised, I began work at the DHO as IT Projects Manager. It soon became apparent that getting actual data for the system was proving problematic. There were many reasons for this. For instance, some content providers did not have the resources to contribute digital objects to the repository. For some of the other providers the notion of an external organisation hosting their images, video and audio, even one with which they were a partner, seemed to represent too much of a loss of control. Overall, providers were more disposed to giving information about their objects rather than the objects themselves. Without those objects though Fedora could not be populated and *DHO: Discovery* could not become the rich platform we envisaged. Another approach had to be found.

Fortunately the modular nature of Fedora Commons proved useful in tackling this problem. It was SOLR, not Fedora, that the website depended upon, and, as a self-contained module, it could be used by itself. Indeed we could supply it with data directly as spreadsheets. This presented a new model. Rather than store the objects in a repository like Fedora, we could simply store the metadata

[1] Irish Manuscripts Commission, http://www.irishmanuscripts.ie/

[2] Europeana Semantic Elements v3.2 XML Schema. Europeana's current data model, backwards compatible with version 3.2, is Europeana Data Model Definition v5.2.3.

[3] Dublin Core Metadata Initiative, http://dublincore.org/

associated with those objects in SOLR itself, as a search index. With Fedora Commons and digital objects no longer required, the DHO: Discovery website was back on track. However, we still needed content, even if only in the form of spreadsheets.

There are many ways to gather content from contributors. An automated approach might use the Open Archives Initiative Protocol for Metadata Harvesting (or OAI-PMH) to harvest content on a scheduled basis, ensuring data is always up to date[1]. However, to use this protocol the data provider must also use their own OAI-PMH compliant system and few institutions have the resources to do this. The flip-side of this is that few institutions have the resources to make frequent changes once the initial effort to publish a collection has been made. More realistic is a manual approach. Metadata for indexing can be supplied via manually generated text files. The format for these text files is XML, a language that allows free text to be encoded in a structured way and is used to build customised markup languages like the Text Encoding Initiative (TEI)[2] language. A spreadsheet can be converted into an XML file very successfully using the column headings to label the content. A suite of web forms was developed to take spreadsheets and other files, transform them into the files SOLR needed and submit them directly to the index. A collection could be added to DHO: Discovery within a matter of minutes. It was not entirely automated, but it wasn't too far from it.

Content providers find spreadsheets a lot easier to create and supply. Where databases are used, it is often possible, with relatively little difficulty, to create a report of data directly as a spreadsheet. We decided to give each of our contributors template spreadsheets listing all the fields that could be populated and asked them to complete it. Many contributors were happy with this approach. What made them especially happy was the fact that we did not need their digital objects.

[1] OAI-PMH, http://www.openarchives.org/OAI/openarchivesprotocol.html

[2] "Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable" (XML, 2013, para. 1). TEI is used to encode a wide variety of primary texts, such as novels, letters, journals, etc.

Why did we not need the digital objects after all? The Internet allows websites to display files hosted on other servers; the files don't have to belong to the website. Now that we no longer needed to populate Fedora, we could use this property to our advantage. By including the web addresses of the contributors' objects in the spreadsheets, *DHO: Discovery* could display their images etc. without actually hosting them.

Even with this new approach collecting content proved difficult. As part of the Humanities Serving Irish Society project, most Irish universities were obliged to submit content to *DHO: Discovery*. However, not all were in a position to do this. One institution was re-organising its own repository and did not feel their metadata was of a sufficient quality to be submitted. Others had no content or else did not have enough resources to prepare it. Nevertheless we approached all our consortium partners in an attempt to get at least one representative collection from each and largely succeeded. We also approached cultural institutions. From our work aggregating content for Europeana, we had already forged many relationships with these, and given the similarity of the two data models, had metadata in a format that could be harvested quickly with very little manipulation required. We contacted these institutions and got their approval for content to be included in *DHO: Discovery*.

5. Going live and going forward

By 2011, we had enough content to populate *DHO: Discovery* and on March 30th, Deputy Seán Sherlock, Minister of State for Research and Innovation, launched the site. The site as launched provided a faceted browse and search service for around 6,000 digital objects and three basic visualisations, i.e., a treemap, a basic tag cloud, and a node-link graph. The DHO's efforts could now concentrate on acquiring more metadata and developing new visualisations. Unfortunately 2011 also saw the departure of much of the DHO's staff and by July 2011, only two remained; one for outreach, one for IT.

Content destined for Europeana had provided metadata before. Now it seemed feasible to reverse the process and take data from Europeana. In September 2012, all metadata held by Europeana was made publicly available under a creative commons license, but it had been available for non-commercial purposes for some time beforehand. In essence, this meant that a developer could use any of the metadata stored by Europeana, provided any rights associated with the original object (e.g., use of the thumbnail or original image) were respected. With this in mind any Europeana Irish data not already in *DHO: Discovery* was imported, helped by the fact that our common data model meant that this data was already in the format needed. A program was developed making it possible to select and harvest collections contained within Europeana directly from that system's index. A wealth of content became available and at time of writing *DHO: Discovery* has indexed 20,701 objects from 30 collections.

When dealing with many collections from diverse sources, one quickly becomes aware of the varied vocabularies used by different curators to describe similar objects. For instance, a collection of architectural drawings might use the term 'Gothic Vaults (Architecture)' while another might use the term 'Gothic Architecture'. Developing a number of online tools for submitting data to SOLR allowed me to incorporate some filtering of these terms. Without removing anything, common terms were added to some objects with similar subject terms (e.g., 'Gothic'). However, this involved a huge manual effort and the issue could only be addressed at the most basic level. Ultimately the use of diverse vocabularies by different providers highlighted the importance of using common standardised ontologies[1]. This can only happen if curators work together.

Despite these inconsistencies, the value of good metadata, particularly subject terms, was clear. One of the earliest collections to be indexed was correspondence from the Documents on Irish Foreign Policy project[2]. Using key phrases such as 'Anglo-Irish Treaty' or 'British-Ireland Relations', it was possible to collate all documents pertaining to the Anglo-Irish Treaty of 1921. This formed the basis

[1] Organisations such as the Getty Research Institute are actively developing such ontologies.

[2] Documents on Irish Foreign Policy, http://www.difp.ie/

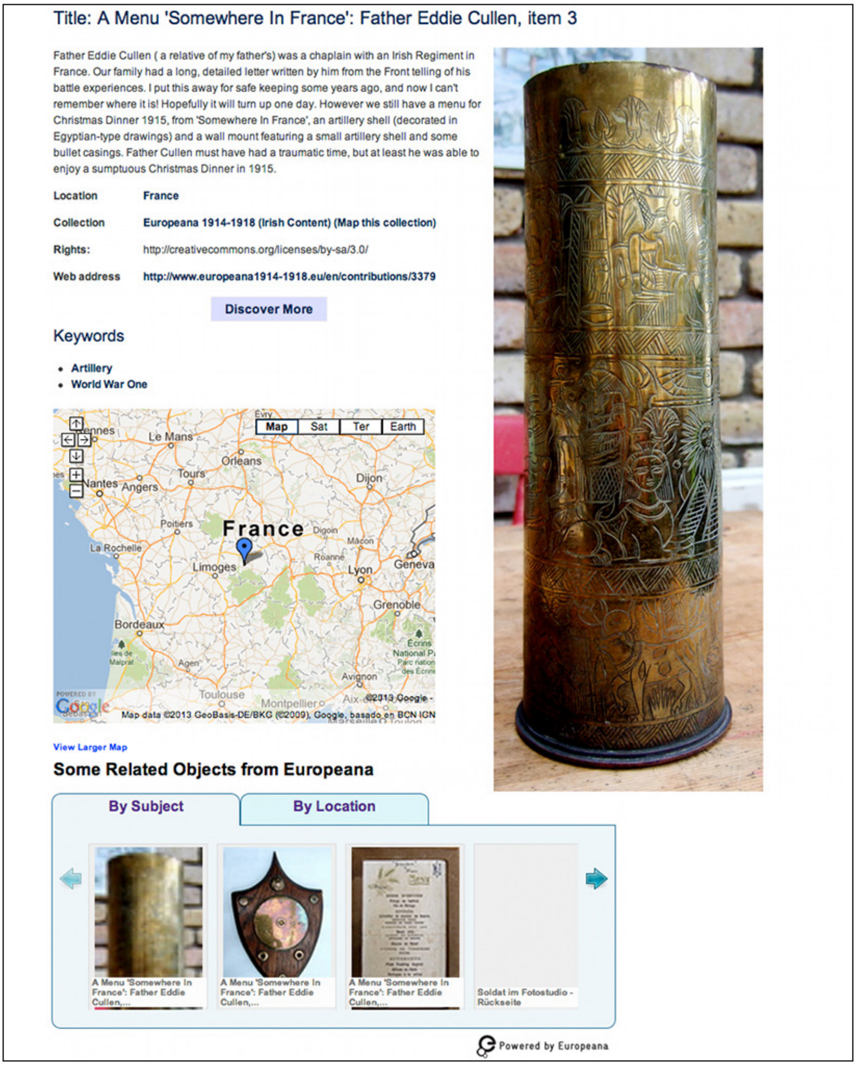
for an e-book produced by the DHO in collaboration with the Documents on Irish Foreign Policy and the National Archives of Ireland (Crowe, Fanning, Kennedy, Keogh, & O'Halpin, 2012).

Regardless of improvements and new content, site usage statistics suggested the service was not being used. Admittedly given the reduction in staff at the DHO, promotion of the site was limited, but there was a more fundamental problem. It was not appearing in search engines. Initially *DHO: Discovery* was developed using HTML and JavaScript (specifically the AJAX-Solr and JQuery libraries), which meant that all the processing of search queries was made within the user's browser. The content of a page was created when a user made selected facets or entered a search term, but the address of the page remained the same. Effectively the site was made from one page and without human interaction this page was blank. Search engines, such as *Google*, follow the links on unique pages to index content. Despite the thousands of objects indexed by *DHO: Discovery*, there were no links to follow and no unique pages, at least not without human interaction. All *Google* could 'see' was a blank page. *DHO: Discovery* was in danger of never being discovered.

Search engines need pages to index, preferably ones with appropriate keywords. The DHO had those keywords – the objects' metadata – it just needed to expose them. Obviously pages for every single object could not be created manually, but a program to automatically generate those pages when a link was followed could be developed, and following a link was the very way search engines interact with a site. One link on the main page – 'Collections' – led to this program. This generated a 'Collections' page linking to an individual 'Collection' page that in turn listed links to all the objects in that collection. Each object now had its own page detailing all its related metadata (Figure 2).

In this way a website of one page (as far as *Google* or *Bing* were concerned), suddenly became a site of thousands of pages. In addition each object page gave further utility, presenting the object's location on a map and offering customised links to, for instance, *Wikipedia*. Search engines began to index *DHO: Discovery* and in September 2012, there were over 730 visitors from 52 countries.

Figure 2. An object page describing an item of Irish content in Europeana 1914-1918



Title: A Menu 'Somewhere In France': Father Eddie Cullen, item 3

Father Eddie Cullen (a relative of my father's) was a chaplain with an Irish Regiment in France. Our family had a long, detailed letter written by him from the Front telling of his battle experiences. I put this away for safe keeping some years ago, and now I can't remember where it is! Hopefully it will turn up one day. However we still have a menu for Christmas Dinner 1915, from 'Somewhere In France', an artillery shell (decorated in Egyptian-type drawings) and a wall mount featuring a small artillery shell and some bullet casings. Father Cullen must have had a traumatic time, but at least he was able to enjoy a sumptuous Christmas Dinner in 1915.

Location	France
Collection	Europeana 1914-1918 (Irish Content) (Map this collection)
Rights:	http://creativecommons.org/licenses/by-sa/3.0/
Web address	http://www.europeana1914-1918.eu/en/contributions/3379

Keywords

- Artillery
- World War One

Some Related Objects from Europeana

By Subject
By Location
A Menu 'Somewhere In France': Father Eddie Cullen,...
A Menu 'Somewhere In France': Father Eddie Cullen,...
A Menu 'Somewhere In France': Father Eddie Cullen,...
Soldat im Fotostudio - Rückseite
Powered by Europeana

Metadata proved useful in creating pages suited for search engines, but it also provided the data needed to create search queries to external websites. Just as

DHO: Discovery built queries for SOLR automatically, with a little customisation these queries could be tailored for Europeana[1]. For instance, the value in the ‘Creator’ field, i.e., who wrote, painted or generated the object, could be used to explicitly query Europeana on the basis of ‘Creator’, safe in the knowledge that their use of the field corresponded with our’s. The results they returned, formatted as XML, could be readily transformed into HTML and integrated into each object page to show related objects from around the world[2].

6. Exploiting the data

DHO: Discovery was intended to provide insights into humanities data that could only be done because of:

- the large number of aggregated objects available;
- the standardised nature of the metadata collated;
- the specialised skillset available in the DHO.

Above all, the value of a standards-based approach to data needed to be highlighted. The different ways of navigating the collections, and the added value provided by the individual object pages, achieved this in part. Using a common data model meant a common approach to programming could be adopted. Throughout 2012, this concept was further exploited in the development of a wide range of innovative visualisations employing open source libraries and software. From a personal perspective, these developments gave me an opportunity to explore the possibilities of structured data and experiment with new approaches to data analysis. I would like now to discuss some of them in more depth.

Scholars in the digital humanities field will be familiar with a variety of tools for visualising large datasets, whether those datasets are the words in a text,

[1] Europeana uses OpenSearch protocol to pass queries to their system. OpenSearch, http://www.opensearch.org/Home

[2] More recently a widget returning results from the Digital Public Library of America was also developed using this approach.

or the complete oeuvre of an individual author. There are, however, relatively few online applications that process content on-the-fly[1]. The huge amount of metadata within *DHO: Discovery* provided an excellent dataset to explore this approach.

At time of writing there are 10 visualisations, all to be found at http://discovery.dho.ie/discover.php. They include:

- Exhibit Visualisations
- Google Maps
- Google Charts
- Image Galleries
- Wordle-style Word Cloud
- Raphaël Visualisations
- Europeana4D
- Treemap Visualisations
- Tag Cloud
- Node-Link Visualisations

Key to these developments is SOLR. It can provide results in several formats including JSON and CSV[2]. For the DHO's purposes though the most powerful format it delivers is XML. XML is a crucial technology in the world of digital humanities. If one understands the schema of a particular XML format, one can transform it into almost any other text format using Extensible Stylesheet Language Transformations (XSLT). XSLT is a language for transforming XML documents into other XML documents, or other objects, such as web pages. The DHO's approach to visualisations was to take the SOLR results in XML format and convert them into a format that could be used by a particular web application, e.g., Google Maps. To minimise the dependence on client-side programming (browsers can be fickle in their support for this), and to maximise the toolset available, the decision was made to combine client-side

[1] A notable exception is TAPoR, http://www.tapor.ca/

[2] JavaScript Object Notation (JSON), is a text-based open standard designed for representing simple data structures and associative arrays. Comma-separated values (CSV) files are used to store tabular data in plain-text form.

technologies, such as JavaScript, with the power of server-side programming (mostly using a language called PHP).

7. Developing the visualisations

Other DHO projects employed the Exhibit framework to add a rich database-like experience to websites[1]. Developed by MIT, Exhibit is a collection of JavaScript files that can be integrated into web pages to create timelines, tables, maps and faceted searching. When the browser loads the page,

> “the Javascript reads in one or more JSON data files and builds a local database in the memory of the machine running the browser. Data can then be filtered and sorted directly in the browser without having to re-query the server. The design of Exhibit is optimised for browsing faceted data” (Exhibit, 2012, section “Overview”, para. 1).

Exhibit sites are normally created manually, but here was an opportunity to generate customised ‘exhibits’ dynamically based on a user’s selected criteria.

Although SOLR can deliver JSON automatically, its format is not suited to Exhibit. Instead an application was developed that

- created a SOLR query dynamically from user choices selected on a web form;
- submitted the query and received the results as an XML file;
- used XSLT to rewrite the XML results into the required Exhibit-specific JSON format;
- created the Exhibit web pages dynamically from this automatically generated JSON content.

By having an intermediate script take the results and reformat them, the resulting web pages would be assured of getting the data exactly as needed.

[1] For other examples of Exhibit, see http://research.dho.ie. For more information on Exhibit, see http://www.simile-widgets.org/exhibit/

This ‘proxy’ approach proved to be a successful model and formed the basis for many further developments.

Maps are an especially effective way of conveying data. Exhibit is capable of displaying Google-like maps, but only when it has the appropriate geo-spatial coordinates. To explore the possibilities of online maps directly meant developing explicitly for the Google Maps platform[1].

Google Maps uses an XML file format called KML in which place names, geo-spatial co-ordinates (e.g., ‘53.3473,-6.2591’) and even web addresses can all be encoded. Much of this metadata was in *DHO: Discovery*. Indeed some providers had even supplied the geo-spatial co-ordinates needed to pinpoint a place on a map. Unfortunately these providers were the exception rather than the rule. It was necessary to pair a place name with its co-ordinates.

Fortunately much of our content was to be found in Ireland. Co-ordinates for many Irish towns and cities are easily available and so a database of the most common Irish places was set up. This had the added benefit of allowing the DHO to make entries for places and geographical features that were specific to some of our objects, but for which current information was not easily available (e.g., Dun Laoghaire was once called Kingstown and was referred to as such in many 19th century images). Each visualisation that required a map (remember Exhibit has a mapping tool) could use this database. Obviously countries and major world cities were also added to account for collections (such as that of the Chester Beatty Library) that had a more international dimension.

Another benefit of this database was that there was now a resource against which programs could check for place names. Many of our providers simply didn’t populate the location field. However, the location was often referred to in either the title of the object or its subject terms. A filter was developed for the data input tool that checked the title and subject fields of new collections for any words that corresponded with the place names in the database. If there

[1] Google Maps, http://developers.google.com/maps/

was a match the place was added to the location field. While not foolproof, once combined with a brief manual check, this opened a lot of collections up to mapping.

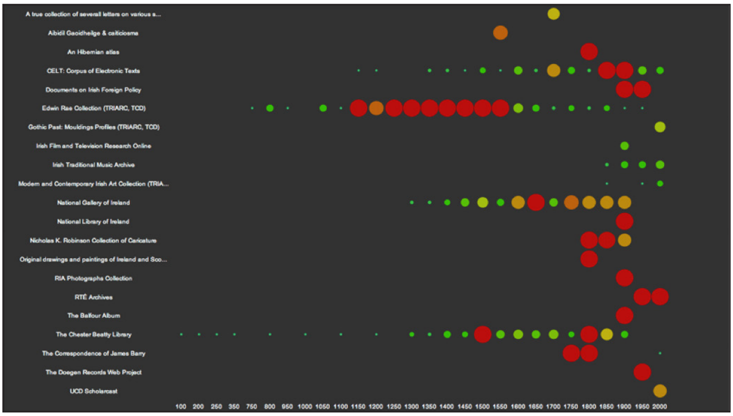
The Exhibit application provided a web form that built up a SOLR query which could now be re-used. Now instead of converting the results into JSON, they were formatted them into KML (the file format recognised by *Google*), and this KML file was supplied directly to Google Maps. In addition, because the pathways to the thumbnails of the objects were available, the 'pins' used to locate places on the map could be customised, or if more than one collection was involved, the objects colour-coded by their collection.

As well as its map application programming interface (API), *Google* has provided tools for developing other web features. One of the most powerful is Google Charts[1]. As with Exhibit and Google Maps, the key to developing charts is to supply the data in the format required by the application code. Formatted as a JavaScript array, the data could be rendered as, for instance, pie-charts or barcharts. It is not the only API available though. Other graphics libraries, such as Raphaël, can be used to created alternative displays of large datasets. Using visualisation, a user can tell at a glance the subject matter of a collection, the frequency of topics, the time range covered and much more. Data visualisation is not a replacement to standard methods of study, but it is a powerful addition (Figure 3).

Again and again, the same basic web form could be used to generate a SOLR query and then use XSLT to convert the results to a format recognised by a JavaScript library. This library along with the formatted results were embedded in a dynamically created web page delivered to the user by a server-side program. Even the formats could be re-used. For instance, Europeana4D, an application that charts geographical and spatial data together, uses KML files. There was no reason why this method could not be re-used and further developed over time.

[1] Google Chart, http://developers.google.com/chart/

Figure 3. A graph illustrating a collection's subject matter over time, created using Raphaël



8. Conclusion

The Digital Humanities Observatory has satisfied its remit of exposing Ireland's rich online resources to the digital humanities community. It has done so using best practice and open standards. With *DHO: Discovery* it illustrates how a disciplined, structured approach to primary digital data can open up a vast array of research opportunities. With minimal expenditure, using open-source software, and building on the rich resources Ireland has to offer, it has transformed existing digital data into new material for scholarship . However, what it achieved is only a beginning. At this point, the lessons learnt in its development are as important as the service itself.

Ireland is a small country. There is no need for isolated silos of data or overly proprietary attitudes to its curatorship. Particularly in the case of public bodies, there is an obligation to open Ireland's digital heritage to all. *DHO: Discovery* demonstrates how easily this can be achieved, and, despite the downturn in the economy, how it can be re-purposed cost-effectively. Properly done, Ireland can become a shining light in the world of digital humanities.

In 2011 the Digital Repository of Ireland was created to develop a national digital repository[1]. This presents a valuable opportunity to achieve that to which *DHO: Discovery* aspired. However, all owners of important digital content need to work together to make this happen. With a holistic vision and appropriate funding, such repositories will be essential to the future of digital humanities in Ireland.

Acknowledgments. The Digital Humanities Observatory and the work described in this paper are funded by the Irish Higher Education Authority (HEA) under the Programme for Research in Third Level Institutes Cycle 4 (PRTLI-4). The DHO was established under the auspices of the Royal Irish Academy. Additional support for repository hosting and programming is provided by the Dublin Institute for Advanced Studies (DIAS) and the Irish Centre for High-End Computing (ICHEC). The author would like to acknowledge the contributions of all the staff at the Digital Humanities Observatory, past and present, especially Randall Cream, Emily Cullen, Shawn Day, Don Gourley, Kevin Hawkins, Faith Lawrence, Katie McAdden, Dot Porter, Susan Schreibman, Paolo Battino Viterbo, and Bruno Voisin.

The image of an object page (Figure 2) incorporates images drawn from Europeana (http://www.europeana.eu/portal/) and made available under the Creative Commons Attribution-ShareAlike 2.0 Generic license (CC BY-SA). Arthur Mathews was the original contributor of the photographs used. It also includes a map image supplied by Google Maps (incorporating map data ©2013 GeoBasis-DE/BKG (©2009), Google, basado en BCN IGN).

References

Crowe, C., Fanning, R., Kennedy, M., Keogh, D., & O'Halpin, E. (Eds.). (2012). *The Anglo-Irish Treaty, December 1920 - December 1921* [Adobe Digital Editions version]. Retrieved from http://research.dho.ie/1921treaty.epub

[1] Digital Repository of Ireland, http://www.dri.ie/

Exhibit (web editing tool). (2012). *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Exhibit_%28web_editing_tool%29

Gourley, D., & Viterbo, P. B. (2010). A Sustainable Repository Infrastructure for Digital Humanities: The DHO Experience. In M. Ioannides, D. Fellner, A. Georgopoulos, & D. Hadjimitsis (Eds.), *Digital Heritage: Third International Conference, EUROMED 2010 Lemessos, Cyprus, November 8-13, 2010 Proceedings* (pp. 473-481). New York: Springer.

XML. (2013). *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Xml

Published by Research-publishing.net
Dublin, Ireland; Voillans, France
info@research-publishing.net

Internet Research, Theory, and Practice: Perspectives from Ireland
Edited by Cathy Fowley, Claire English, and Sylvie Thouësny

Research-publishing.net has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate. Moreover, Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before (except in the form of an abstract or as part of a published lecture, or thesis), or that it is not under consideration for publication elsewhere. While the advice and information in this book are believed to be true and accurate on the date of its going to press, neither the authors, the editors, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Typeset by Research-publishing.net
Cover design: © Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-04-9 (Paperback, Print on Demand, Lulu.com)
ISBN13: 978-1-908416-05-6 (Ebook, Kindle Edition, Amazon Media EU S.à r.l.)
ISBN13: 978-1-908416-08-7 (Ebook, PDF file, Open Access, Research-publishing.net)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Bibliothèque Nationale de France - Dépôt légal: juin 2013.