Open Source at the National Library of Ireland

Catholic Parish Registers at the NLI

Project Overview

The Catholic Parish Registers at the NLI project (registers.nli.ie) makes the National Library of Ireland's (NLI) collection of parish register microfilms freely available to everyone interested in their Irish family history, and in particular the Irish diaspora. The registers, which were microfilmed by the NLI during the 1950s and 1960s, contain records of baptisms and marriages from almost all Catholic parishes in Ireland and Northern Ireland from the 1740s to the 1880s. Previously these records were available only to researchers by using manual microfilm-reader machines at the NLI's premises in Dublin. The importance of these records arise from the fact that they are the single most important source for Irish family history research prior to the 1901 Census. In many registers, the period before and during the Great Famine is documented, providing a unique resource for a period when there was huge emigration from Ireland and for which there are almost no other records of most individuals and families. With a view to preservation and the potential for online access, an earlier project (2010-2011) had outsourced the conversion of 550 NLI microfilm reels relating to 3500 registers into approximately 373,000 digital images (over 6 Terabytes) and metadata.

Why was the project started. What problems or issues prompted the project?

The project was started to deliver dramatic service improvement for the large number of researchers seeking to access the Parish Register records through the NLI's Genealogy Advisory Service (GAS). The existing service at the NLI's Kildare Street premises was exceptionally popular but was in considerable need of updating as the records were only available to researchers by using manual microfilm-reader machines. In 2013 researchers, many of whom had travelled from overseas, made over 26,900 in-person visits to use the microfilms; in 2014, this figure rose to over 27,500 visits. The nature of the microfilm itself presented a number of difficulties for researchers. Although the reels were available on a self-service basis, microfilm is an old technology and is not user friendly. The constant use of the microfilm caused wear and tear on the films, and they had to be replaced frequently. The microfilm readers themselves required constant servicing and repair.

The project sought to dramatically update this in-demand service to meet the expectations of modern researchers and to democratise access by facilitating research regardless of location or financial means.

Why Open Source?

Since 2009 the NLI has adopted and contributed back to open source software projects as a key means of delivering on our core remit to collect, preserve and make accessible collections (physical and digital) of national significance. As a result, the NLI has benefited from the resources and expertise of countless developers and domain experts worldwide, and has been able to contribute back in return. For example, the NLI played a key leadership role in the development of the VuFind Discovery Interface which is now used by thousands of libraries in around the world. This early success was recognised in the NLI receiving the first ever eGovernment Open Source Award in 2011. Since then, the strategic decision to participate in

Open Source has proven extremely successful, enabling the NLI to put in place the core components of a modern digital library infrastructure despite a small technical. In particular, the approach has enabled in the implementation of a robust digital repository system based on the Open Source Hydra/Fedora Commons technology stack, which allows the NLI to ensure the long-term management and preservation of over 40 Terabytes of digitised material and to tackle the complex task of collecting “born digital” collection material.

While it may have been possible to deliver some of these requirements with an alternative non-FOSS solution, it was critical from a sustainability point of view to avoid the creation of a large data silo, distinct from the even larger set of digital assets already handled by core NLI workflows and systems. However, it was just as important that the back-end preservation and management systems didn’t restrict our ability to create an entirely bespoke User Interface, designed solely with the Parish Registers data and audience in mind; we needed complete control over the design and navigation of the front-end experience. The flexibility to mix and match loosely-coupled software packages through standard interfaces is at the heart of Open Source software development. Furthermore, enabling reuse of digital collections in different contexts is a key driver for library technologists: for example, the Hydra Repository project used by the NLI, employs the metaphor of the “Hydra” to express the idea of a solid, shared preservation solution (the Hydra’s body) enabling a multiplicity of views and representations (the Hydra’s many heads). Given these complex systems integration concerns and the relatively tight deadline, using Open Source technologies to build upon the NLI’s existing Open Source stack was identified as the only viable solution. As a result, the NLI was able to deliver a compelling, tailored UI, while at the same time ensuring sustainability and minimal maintenance costs (code, data, and metadata) by leveraging and extending established processes and systems.

Open source project solved the problem or issue at hand efficiently.

The Parish Registers infrastructure is open-source all the way down, built on

- Linux (https://github.com/torvalds/linux);
- Puppet (https://github.com/puppetlabs/puppet);
- Apache (https://httpd.apache.org/) and Lighttpd (https://github.com/lighttpd/lighttpd1.4) web servers;
- Ruby on Rails (http://rubyonrails.org/);
- Apache Solr (http://lucene.apache.org/solr/);
- Fedora Commons (http://fedorarepository.org/);
- Hydra Digital Repository (http://projecthydra.org/);
- Traject indexer (https://github.com/traject/traject);
- Blacklight Discovery Layer (http://projectblacklight.org/);
- IIPImage streaming image server (http://iipimage.sourceforge.net/);
- Memcached (http://memcached.org/);
- The Internet Archive BookReader (https://github.com/openlibrary/bookreader);
- typeahead.js (https://twitter.github.io/typeahead.js/);
- Bootstrap (http://getbootstrap.com/);
- Leaflet.js (http://leafletjs.com/), and many more open source components.

The flexibility to combine this wealth of high-quality software products directly addressed the broad architectural considerations and project constraints outlined above. Descriptive Metadata for each Diocese, Parish, and Register was programmatically generated and loaded into the

NLI's main Library Management System, allowing continual improvement in line with existing cataloguing workflows. This also enables the sharing of metadata and previews of the digitised content with Europeana and other aggregators with no additional development or special processing. The 6 Terabytes of image data was processed through the NLI's existing digital asset validation pipeline. This automated process allowed for the identification and correction of a small percentage of corrupt image data. Validated images were ingested to long-term repository storage where they are subject to regular preservation actions such as fixity checks. The NLI team also expanded its use of Hydra to automatically generate the derivative, web-optimised versions of images used in the Parish Registers image viewers (contributing several bug fixes along the way).

Following an initial process of requirements gathering and wire-framing to map the expected flow of users through the site, the use of Open Source allowed rapid, proof-of-concept prototyping of the core front-end elements: an intuitive search interface; a detailed, interactive map interface; dynamic, high-resolution image viewers (see supporting documentation for the original wireframe of navigational flow). Apache Solr and the Blacklight Discovery Interface were selected as a solid search framework upon which to build our customised interface. The limitations of the data available (in particular the absence of full-text transcriptions) placed constraints on the interface. The navigational flow of the site is geared towards getting users quickly to the Parish(es) of interest to them, before directing them to the corresponding register viewers, and pages. This hierarchical entity relationship (Parish -> Register -> Page) is not supported out-of-the-box by Blacklight's flat document store. However, following some discussion with core Blacklight developers, NLI Developers were able to quickly extend Blacklight's default Model-View-Controller design to seamlessly accommodate the Parish Registers data-model. In order to make the process of identifying the correct Parish as simple and intuitive as possible, NLI developers combined Apache Solr and typeahead.js to provide a powerful autocomplete feature which makes suggestions based on known variant forms of Parish names (e.g. Drumlane is also known as Staghall), as well as on-the-fly spelling corrections. The implementation of the other key features, the interactive map and dynamic images viewers, included particularly innovative use of Open Source and are discussed in more detail in the section below.

Contributions by NLI Developers

Since beginning to use Open Source in 2009, the NLI has adopted a policy of engaging with community developers and contributing local modifications back to the community code-base. This approach ensures improvements are immediately available to other users and, from the NLI's perspective, simplifies the process of managing and migrating code. In the course of developing the Parish Registers infrastructure, NLI developers contributed in a wide variety of ways to the many Open Source projects used including: contributing pull-requests/patches against core components of Hydra (the NLI has been accepted as a Licensed Contributor to the project); submitting feature requests; answering queries and solving problems for other users on community mailing lists; correcting documentation and commenting on open bug/issue tracker tickets; completing community surveys to inform future development; advocating for the use of Blacklight and Hydra at Irish technology seminars; releasing new Open Source software for use by others; releasing GIS data in standard formats following Open Data principles. Some public examples of this, include:

- https://github.com/projecthydra/hydra-derivatives/pull/91 (Merged Pull Request)
- https://github.com/projecthydra/hydra-derivatives/pull/42 (Merged Pull Request)
- https://github.com/projecthydra/hydra-derivatives/pull/41 (Merged Pull Request)

- https://github.com/traject/traject/pull/91 (Merged Pull Request)
- https://github.com/puppetlabs/puppetlabs-passenger/pull/86 (Pending Pull Request)
- https://wiki.duraspace.org/display/hydra/Hydra+Licensed+Contributors (Hydra Contributor registry; requires sign-up)
- http://sourceforge.net/p/iipimage/discussion/299493/thread/bebbf543/ (Accepted feature request)
- http://sourceforge.net/p/iipimage/discussion/299493/thread/9e80d8c1/?limit=25#df73 (Providing community support/solutions)
- https://groups.google.com/d/msg/hydra-tech/_L2ysL1A0M8/2agSkvWVFAAJ (Providing community support/solutions)
- https://github.com/projecthydra/hydra-derivatives/issues/81 (Providing community support/solutions)
- https://groups.google.com/d/msg/blacklight-development/NQDw1d9pKgQ/ejJzrayYDwAJ (Providing community support/solutions)
- https://groups.google.com/d/msg/blacklight-development/ubsUasZD7Xc/cpI1F9QdHUcJ (Providing community support/solutions)
- https://cwiki.apache.org/confluence/display/solr/Spell+Checking?focusedCommentId=51811299#comment-51811299 (Documentation correction)
- https://github.com/jbeard6/jbeard-nfs/pull/11#issuecomment-109581640 (Accepted feature request)
- https://docs.google.com/spreadsheets/d/1uosMP08OPoohR-VWM7zbufbcr2tcbSlAlxx4rIdFUGM/edit?usp=sharing (Blacklight Community Survey)
- http://www.interleaf.ie/news/Open-Source-in-Libraries-Seminar (Open Source Advocacy, July 2015)
- https://bitbucket.org/nlireland/ (Original Source releases)

Having built a highly successful technology strategy around Open Source software, the NLI is committed to the long-term sustainability of Open Source projects upon which we depend. Building on our achievements with VuFind, Hydra and the recent success of the Parish Registers implementation, the NLI intends to re-double our participation in Open Source communities. For example, in the last month we have joined dozens of other libraries as members of the Duraspace foundation, making a direct financial contribution to the Fedora Commons digital repository project upon which Hydra and the NLI's repository depends.

Innovations to solve specific problems during the project

Interactive Map: Open Data via Open Source

Location is one of the key ways genealogical researchers track their Irish roots; often a county or parish name is their only clue. In the case of Catholic Parishes, this task is made considerably more difficult due to the fact that Catholic parish boundaries continually shifted over time with the result that detailed maps such as are found for Civil Parishes were not available for use in the site. We evaluated various mapping options including traditional HTML Image Maps, as well as SVG coupled Javascript visualisation libraries. However, the requirement to show complex labelling and boundary divisions and the need for first-rate mobile/touch support, indicated the need for better raw data and the use of a modern, mature web-mapping library. The availability of high-quality Open Source GIS solutions allowed rapid prototyping during this evaluation phase, where commercial GIS software would have been prohibitively expensive. Working with genealogist, John Grenham, who has painstakingly researched historical parish boundaries, the NLI used Open Source software to convert his low-resolution maps (the most complete available) to fully geo-referenced ESRI Shapefiles. Specifically, this involved:

- Geo-referencing raster images using MapWarper (https://github.com/timwaters/mapwarper);
- Using QGis (https://github.com/qgis/QGIS) to overlay the geo-referenced images on existing open boundary data from the OSI in order to sub-divide into parishes;
- Applying custom styles and labelling using Tilemill (https://github.com/mapbox/tilemill), GeoJson (http://geojson.org/) and CartoCss (https://github.com/mapbox/carto);
- Deployment and testing using TileStream (https://github.com/mapbox/tilestream) and Mapbox.com;
- Integration and custom interactions using UTFGrids (https://github.com/mapbox/utfgrid-spec) and Leaflet.js (https://github.com/Leaflet/Leaflet);

This original dataset shows the Parish boundaries as they existed in the mid-19th century, allowing users to understand the topographical relationship between parishes, diocese and counties and find the Parish of interest to them. The simple visual design of the map is intentional: while the GIS data allows for overlaying other map layers etc., the uncluttered interface used in registers.nli.ie: a) reflects the purpose of the map as a navigational aid and b) reflects the fact that, due to incomplete historical record, drawing 19th century parish boundaries is often an art rather than an exact science. Coupled with tools such as Leaflet.js, the data allowed NLI developers to build an entirely responsive “slippy map”, providing a familiar user-experience similar to Google Maps or Openstreetmap. The result is an entirely unique, modern web-based map, which works seamlessly across all devices, greatly enhancing the navigation of the overall site.

Image Quality

Another considerable challenge was the problem of poor image quality carried over from the original 1950s microfilm. In many cases, the images are under or over exposed, or lack an even exposure with the result that correcting the exposure for one part of the image would render another part entirely illegible. While full-text search of the register content would have been a technically straight-forward solution, the task of transcribing every page was well beyond scope and budget of the current project, so the necessary data simply wasn’t available to NLI developers. The solution implemented by the NLI was to a) provide low-bandwidth, high-performance zoom, b) allow real-time adjustment of image contrast, brightness, and inversion, c) enable users to intuitively filter the potential number of pages to be examined by event type (e.g. Baptism, Marriage) and date (e.g. May 1783).

This was accomplished through the innovative combination and extension of Open Source software, namely the IIPImage High-Performance Imaging Server, the Internet Archive BookReader, and Apache Solr. IIPImage provides standard APIs for manipulating and streaming image tiles, while limiting bandwidth use by sending the minimum amount of image data needed for a given client. The IA BookReader provides a javascript client interface allowing users to intuitively navigate and scroll through digital volumes. Apache Solr provides powerful search and data faceting, allowing very performant filtering across hundreds of thousands of register page metadata.

The NLI already had experience of working with the IIPimage and the BookReader for delivering digitized content through our main catalogue interface (http://catalogue.nli.ie). Building on this work, the team added the crucial ability to manipulate image settings ensuring the Parish Register images are as legible as possible. For genealogists who know an approximate date for their ancestors, the facility to filter can dramatically reduce the amount of time needed to research a given volume. Furthermore, the NLI team upgraded various aspects of the IA

BookReader to provide excellent user experience on tablets and touch-centric devices, including full-screen browsing.

This is an excellent example of how disparate Open Source components can be used in new, innovative combinations to effectively address a particular challenge.

Performance testing and scalability

Given the huge public appetite for genealogy and family history research, the NLI correctly predicted widespread interest in the release of records as significant as the Parish Registers. Considerable work was carried out at a hardware level to upgrade the NLI's bandwidth, network and load-balancing capacity. However, accommodating a high number of concurrent users also required the ability to scale all aspects of the Parish Registers software infrastructure. The NLI achieved this through a) using Open Source performance testing tools to test and improve the various components under heavy load, and b) adopting Puppet and Capistrano as automated deployment tools.

While the NLI had previously used Capistrano for the automated deployment of locally developed code, the Parish Registers project was our first use of Puppet Configuration Management in production. Puppet allows developers and system administrators to encode the exact software configuration needed by a given virtual machine, allowing new versions of a server to be brought online in a matter of minutes. This is extremely useful when testing different performance characteristics or when "scaling out" an application by distributing requests across multiple copies of the same server. The core technology behind Puppet is entirely Open Source and a great many Puppet Modules, or building blocks for configuring particular pieces of software, are freely available. The NLI made extensive use of these existing Open Source Puppet Modules, contributed improvements to them, and also developed and publically released source code for a number of new Puppet Modules, including components for Lighttpd, Solr 5, IIPImage, NewRelic system and ruby agents, and Tilestream.

In order to simulate realistic load, the NLI designed a number of test scenarios using the Apache JMeter testing framework. We then used Blazemeter and New Relic cloud services to run performance tests at scale, gather metrics and identify bottlenecks. The flexibility inherent in Open Source allowed us to address bottlenecks effectively. For example, early testing identified the dynamic image server as a potential bottleneck given the sheer amount of data that could be requested (the facility for dynamic image adjustment, described above, results in an imponderable number of potential image tile combinations). Testing allowed us to make informed decisions about such factors as: switching web servers (Nginx -vs- Lighttpd), using and pre-warming Memcached for RAM-based object caching, fine-tuning the optimal number of IIPImage FCGI processes per instance, fine-tuning the number of load-balanced VMs, etc.

During the first 24 hours following the launch, the Parish Registers site handled a peak of 7,500 concurrent users with minimal performance degradation, and served approximately 128 million requests (a.k.a. "hits"), 111 million of which were made to the IIPImage servers. Furthermore, the Puppet configuration allows us to scale capacity down and up as required with comparative ease and predictability. Since launching the Parish Registers, we have migrated several other staging and production systems to use Puppet.

As Virtualisation, Public/Private Cloud, and "DevOps" continue to make inroads into Enterprise and Public Sector deployments in Ireland, the NLI is an exemplar of a small, agile organisation fully leveraging these new technologies to improve its core public services. This approach goes

far beyond the simple use of a single open source package, fully adopting Open Source as the native technology stack of the web and scalable, cloud-based solutions.

Please outline steps you have taken to ensure your service is accessible and assists visitor ability to use the service.

In developing the Parish Registers interface, the NLI adopted a Behaviour Driven Development (BDD) methodology. Feedback from wireframing and requirement sessions were converted into 16 User Stories/Features and 194 Scenarios using Cucumber, an Open Source Domain Specific Testing Language (https://cucumber.io/). These stories and scenarios are written in plain English following a simple set of syntactic rules. Once created, these stories provide rich documentation about the expected behaviour of a piece of software in language that promotes unambiguous communication between the software developers and non-technical project members. Combined with making regular and early releases and evaluations, this collaborative approach and common language puts the emphasis on the user experience, and ensures the product developed aligns closely to business objectives. Furthermore, the domain specific language used in describing the desired features can also be interpreted by automated testing frameworks (see screencast: https://www.youtube.com/watch?v=4vbJPFdraIs) . Along with unit tests written with RSpec (http://rspec.info/), this ensured that every code change committed by NLI developers could first be automatically validated against the entire test suite of expected behaviour. These continuous integration techniques greatly reduce the risk of introducing bugs and breaking existing behaviour. They also ensure the overall sustainability of the software product, since the self-documenting nature of BDD allows new developers to quickly understand the overall requirements. For final acceptance testing, the NLI was able to supply the exact same User Stories to human software testers to manually review and independently validate the entire set of expectations across a wide range of devices (smartphones, tablets) and browsers. With its strong emphasis on user interactions verified across all sorts of devices, the BDD approach greatly enhanced accessibility.

In terms of front-end design, the Parish Registers interface was developed using the Open Source Bootstrap framework, following mobile-first, responsive design principles. The NLI developers collaborated with graphic designers from New Graphic (http://newgraphic.ie/) to establish a strong visual identity that would scale well across all screen sizes, with the result that every single feature available on the Desktop is equally available on the smallest of smartphone. Even data-rich and sophisticated components such as the Map and Image Viewer work seamlessly on small screens and through touch interaction. The consideration of different devices greatly enhanced visitors ability to use the service. On the launch day over 20% of visits came from tablets and 24% from mobile devices, so many users first impression of the service was through a touch-based mobile device. This mobile trend has remained strong, with just under 20% of all sessions coming from tablets and 12% of all sessions coming from smartphones.

The use of a mature front-end framework like Bootstrap ensured support for the structural and navigation elements needed encouraged by the Web Content Accessibility Guidelines to which NLI developers sought to adhere. Furthermore the NLI took additional measures to ensure key content is available without the use of Javascript: for example, users who are unable to use the Javascript based Image Viewers, have the option to page through and download all register images using standard, static HTML content.

Conscious of the high-level of interest from overseas, the NLI team regularly tested the client/browser-side load times of the site from New York, London and Sydney (using the free http://www.webpagetest.org/ service). These tests allowed various optimisations including the use of geographically distributed Content Delivery Networks and targeted HTTP headers to encourage browser caching where appropriate. This work ensures a great user experience regardless of location, and in the case of Australia halved the initial page-load time for first-time visitors.

Other steps taken to enhance accessibility and usability, included: creating two simple video tutorials to visually explain how to use the site; offering a detailed set of FAQs; providing both English and Irish versions of the interface; providing a dedicated contact form/email which ensures users receive a response within 48 hours of asking a question or leaving feedback.