Developing high-end reusable tools and resources for Irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies

Brian Ó Raghallaigh
Dublin City University
brian.oraghallaigh@dcu.ie

Michal Boleslav Měchura
Dublin City University
michal.boleslav.mechura@dcu.ie

Abstract

Irish, a low-resourced lesser-used language, is striving to punch above its weight when it comes to some of the digital language tools and resources available to its users. High-tech language tools and resources for Irish are being developed in a number of universities in Ireland and elsewhere, in language technology areas relating to search, parsing, proofing, speech, translation, etc. (Judge at al., 2012). This paper aims to highlight work done by researchers at Fiontar, Dublin City University (DCU), to make a number of valuable Irish-language terminological, lexicographical, onomastic, and folkloristic data stocks more readily accessible, usable, and manageable using web and database technologies. Tools built with these technologies have facilitated the re-organisation, distributed development, and more widespread dissemination of these data stocks, as well as the creation of new data stocks. These language tools, which are on a par with tools that are available to users of well-resourced languages (take for example the online interface of the multilingual terminology database of the European Union, *IATE*: http://iate.europa.eu/), are now enabling Irish language users, language professionals, and linguists operate in an environment similar to that of their major language counterparts. The public interfaces of all Irish-language tools and resources developed by Fiontar are made available at http://www.gaois.ie/.

1 Introduction

Although Irish is a low-resourced language, the Irish Government's *20 Year Strategy for the Irish Language*, which prioritises the "promotion and protection" of the language (Government of Ireland, 2010), has brought about investment in the creation of digital language tools and resources. Linguistic resources, such as printed dictionaries, are now being made available electronically through retro-digitisation, or being created digitally, and then enhanced with search engines powered by language technologies, such as spelling error detection.

This paper highlights the work done by researchers at Fiontar, Dublin City University (DCU) in the identification of valuable non-digital language resources, the digitisation of these resources where necessary, and the application of web, database, and language technology to these resources to widen access and availability, and to increase effectiveness and usability.

Fiontar's tools and resources include public websites that provide easy, user-friendly access to Irish-language terminological, lexicographical, onomastic, and folkloristic data stocks, as well as web-based tools for managing and developing this data. User-friendliness is seen by Fiontar as key in the promotion of the language on the Internet (Měchura and Ó Raghallaigh, 2009). Single query, all-in-one Google-like search, is also a priority, with sophisticated quick search being a feature on all Fiontar websites. All of Fiontar's digital language tools and resources are made available at or linked to from http://www.gaois.ie/ (*gaois* 'wisdom').

2 Terminology and lexicography

In 2005, in partnership with Foras na Gaeilge, the body responsible for the promotion of the Irish lan-

Proceedings of the First Celtic Language Technology Workshop, pages 66–70,
Dublin, Ireland, August 23 2014.

guage throughout the whole island of Ireland, researchers at Fiontar began development of the National Terminology Database for Irish, *focal.ie* (*focal* 'word'). Retro-digitisation (where a work that was previously published on paper is converted into a digital, computer-readable format) was carried out on 54 different dictionaries and term lists supplied by the Terminology Committee of Foras na Gaeilge (Bhreathnach, 2007), and the dataset was imported into a purpose-built relational database for terminological and lexicographical data (Měchura, 2006). In addition, two web-based interfaces to the new database were developed. The first, a password-protected web application, provided a geographically dispersed group of authorised terminologists with access to the data as well as a set of web-based tools for editing and developing the data. The second, a public website, gave public access to the data via a set of linguistically sophisticated (e.g. inflection awareness, misspelling detection, language selection) search tools (Měchura, 2008; Měchura and Ó Raghallaigh, 2010). This meant that for the first time, Irish-language users, most notably language professionals, had free and searchable worldwide electronic access to this valuable data stock.

The focal.ie system continues to be maintained and developed today. The database currently contains over 342,000 terms, mostly in Irish and English. The technology has gone through a number of major overhauls. Most notably, the database and (private) editorial interface were replaced in 2012 with a new system called *Léacslann* (Měchura, 2012b). In Léacslann, terminological data is now stored as XML. Léacslann also incorporates additional features such as user permission management, a power search feature which allows users to interrogate the data in complex ways, and an extranet application to gather input from external subject and language experts. And in 2013, the public search algorithm was optimised for speed and enhanced with better spelling-error detection.

One of the advantages of the Léacslann system is that multiple data stocks can be stored and managed in the same database. This allows the editorial tools to be reused across multiple terminology and lexicography projects. The system now hosts multiple lexical databases being maintained and developed by Fiontar language experts. It also has the potential to be used to host terminology and lexicography projects for other institutions and languages, as it is flexible and customisable. It can be used to work with various kinds of stocks such as monolingual and bilingual dictionaries, terminology databases or indeed any sort of reference work. Léacslann stocks can accommodate any language and any combination of languages, as long as text in those languages can be encoded in Unicode (Měchura, 2012b). This might prove to be an economical way to develop such resources for other low-resourced languages such as Scottish Gaelic, for example.

Corpora for use in lexicography have also been developed. One such corpus, a parallel Irish-English corpus of Irish and European legal texts, made available to Fiontar by the Irish Government and the European Commission, known as *ParaDocs*, has been made available to the public on gaois.ie.

3 Onomastics

In 2007, in partnership with the Placenames Branch of the Government of Ireland, the body that conducts research into the placenames of Ireland to provide authoritative Irish language versions of those placenames for official and public use, researchers at Fiontar began development of the Placenames Database of Ireland, *logainm.ie* (*logainm* 'placename'). A new relational database for bilingual Irish-English toponymic data was purpose-built for the project, and data already digitised by the Placenames Branch was imported into this database (Mac Giolla Easpaig, 2009). The architecture adopted for the terminology project was reflected in the placenames project in that two web interfaces, one public and one private (editorial), were built on top of the placenames database to allow dissemination as well as distributed editing and development of the data via the web (Měchura and Ó Raghallaigh, 2012).

A mapping interface, which used Google maps, was added to the public website in 2010, and in 2014, the data structure was enhanced with the inclusion of *place clusters*. These so-called clusters better reflect how people think about 'places' such as *Donegal*, for example. People don't normally think about the distinction between the various administrative units called 'Donegal' in County Donegal (i.e. the parish, townland, town, and electoral division), all of which are stored as distinct objects in the placenames database, but rather think of just one place, Donegal. The new data structure allows clustered place objects to be grouped and presented in a more user-friendly way (Měchura, 2012a).

Other developments include a collaboration with the Digital Repository of Ireland to make the dataset available as Linked Data, i.e. as exposed RDF data objects that are linked to equivalent objects in other geodatasets such as GeoNames (Lopes et al., 2013), and a project to match the dataset with Ordnance Survey Ireland so that logainm.ie data can be displayed on OSi maps, and in turn so that those maps can be used in place of Google Maps on the website (Byrne et al., 2013). As of May 2014, the English and Irish versions of the OSi medium-scale *Basemap* are being used on logainm.ie in place of Google Maps (Satellite View).

Data, some of which has to be digitised (originating on maps or on hand-written cards, for example), continues to be added to the placenames database, and development is ongoing. Additional resources such as maps, articles, and educational resources are also added periodically. The database currently contains entries for over 108,000 geographic places on the island of Ireland.

Another onomastic project, which has recently been established aims to produce a surnames database, which will group related Irish and English surnames. The intention is to use the database to enhance the names search interface to the folklore collection described in Section 4, and to make this database freely available to search or to download and reuse. The project is in its infancy and will be fully reported on at a later date.

4 Folkloristics

In 2012, in partnership with the the National Folklore Collection (NFC) at University College Dublin, home to one of the largest collections of oral and ethnological material in the world, researchers at Fiontar began development of *dúchas.ie* (*dúchas* 'heritage'), a new digital version of the NFC. The project was initially funded by the Government Department of Arts, Heritage and the Gaeltacht on a pilot basis for one year (2012-13) and has now been funded from the same source for three more years (2013-16) to digitise, digitally catalogue, and publish online 14% of the NFC. The NFC comprises multiple collections, including a music archive, a map archive, an audio and video archive, a collection of paintings, and a collection of photographs. One collection in particular, a manuscript collection comprising handwritten stories, gathered as part of a Government-sponsored scheme in 1937-39, has been chosen as the first collection to be migrated to dúchas.ie. Known as *The Schools' Collection*, it was chosen primarily due to its popularity (Ó Cléircín et al., forthcoming).

Since The Schools' Collection comprises manuscript only, digitisation in its case involves the scanning of pages to create digital image files. The text written on these pages is not being transcribed, as this would be not be feasible, but a digital catalogue of the pages and the stories written on them is being compiled as part of the project, to make the collection electronically searchable. It is envisaged that 46% of the Schools' Collection, i.e. *c.* 339,000 pages, will be scanned and catalogued by 2016.

As with the terminology/lexicography and the placenames projects, the dúchas.ie project comprises two web applications, one public and one private (editorial), and two databases, one for each web application. The public system is used to present the digitised collections to the world, and provides the user with a number of search interfaces. Currently, The Schools' Collection can be searched by *person* (the names of the people who told or collected the stories) or by *place* (where the stories were collected). The private system is used to manage and edit the digital catalogue. The contents of the private database are transferred to the public database weekly. In this instance, the Léacslann platform was reused, and a customised editorial/management application was added for this data stock.

5 Digitisation, management, and dissemination

Expertise in digitisation project management, as well as web-based data management and publication has allowed Fiontar to transition other Irish language legacy data stocks to the web. One example is the biographies database, *ainm.ie* (*ainm* 'name'). This project involved the digitisation of nine physical volumes of biographies (*c.* 1,700 lives) written and published between 1986 and 2007. Once again, this resource has been digitised, managed online, and published online with associated electronic browsing, navigation, and search tools, all of which involved the reuse of existing infrastructure, technologies, and expertise. Another example is the legacy research sound archive of the Placenames Branch, which is accessible to researchers at http://www.logainm.ie/phono/.

6 Technologies and hosting

All of the projects described here were built using web and database technologies. The Microsoft .NET Framework and SQL Server platform were used in each case. Hosting for all websites and databases is provided by DCU Information Systems and Services in conjunction with the HEAnet. Binary files created for the dúchas.ie project are hosted by UCD Research IT.

7 Conclusion

This paper described some of the tools and resources for Irish developed and made available online by Fiontar, Dublin City University, as well as the web and database technologies utilised in their deployment. It was highlighted that all of these tools and resources encompass technologically and linguistically sophisticated search interfaces. The use of technology in this way to enhance the resources available to Irish-language users and professionals is serving to place their language-related activities on a more level playing field with their major language counterparts, and goes some way towards the promotion and protection of the language.

Acknowledgements

This research was undertaken with support from Fiontar, Dublin City University. The research described here is being undertaken with financial support from Foras na Gaeilge and from the Department of Arts, Heritage and the Gaeltacht of the Government of Ireland.

References

Úna Bhreathnach. 2007. www.focal.ie – A New Resource for Irish. *Translation Ireland*, 17(2):11-18.

Maria Byrne, Brian Ó Raghallaigh and Mairéad Nic Lochlainn. 2012. Synchronising the Ordnance Survey Ireland (OSi) and Placenames Branch (logainm.ie) bilingual toponymic datasets. In *Placenames Workshop: Management and dissemination of toponymic data online*. Dublin: 153-162.

Government of Ireland. 2010. *20-Year Strategy for the Irish Language 2010-2030*. Online at http://www.ahg.gov.ie/en/20-YearStrategyfortheIrishLanguage2010-2030/ [Retrieved 9 May 2014]

John Judge, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell and Elaine Uí Dhonnchadha. 2012. *The Irish Language in the Digital Age*. Springer, London, UK.

Nuno Lopes, Rebecca Grant, Brian Ó Raghallaigh, Eoghan Ó Carragáin, Sandra Collins and Stefan Decker. 2013. Linked Logainm: Enhancing Library Metadata using Linked Data of Irish Place Names. In *Linking and Contextualizing Publications and Datasets (LCPD 2013)*. September 2013, Malta.

Dónall Mac Giolla Easpaig. 2009. Ireland's heritage of geographical names. *Geographical Names as a Part of the Cultural Heritage, Wiener Schriften zur Geographie und Kartographie*, 18:79-85.

Michal Boleslav Měchura. 2006. Finding the right structure for lexicographical data: experiences from a terminology project. In *Proceedings of the 12th Euralex International Congress*. Torino: 189-198.

Michal Boleslav Měchura. 2008. Giving Them What They Want: Search Strategies for Electronic Dictionaries. In *Proceedings of the 13th Euralex International Congress*. Barcelona: 1295-1299.

Michal Boleslav Měchura and Brian Ó Raghallaigh. 2009. User-Friendliness: the key to promoting a minority language on the Internet. In *International Conference on Minority Languages (ICML 12)*. May 2009, Tartu.

Michal Boleslav Měchura and Brian Ó Raghallaigh. 2010. The Focal.ie National Terminology Database for Irish: software demonstration. In *Proceedings of the 14th Euralex International Congress*. Leewarden: 937-948.

Michal Boleslav Měchura and Brian Ó Raghallaigh. 2012. The logainm.ie Placenames Database of Ireland: software demonstration. In *Placenames Workshop: Management and dissemination of toponymic data online*. Dublin: 115-122.

Michal Boleslav Měchura. 2012a. Landscapes, languages and data structures: Issues in building the Placenames Database of Ireland. In *Digital Humanities Conference (DH 2012)*. July 2012, Hamburg.

Michal Boleslav Měchura. 2012b. Léacslann: a platform for building dictionary writing systems. In *Proceedings of the 15th Euralex International Congress*. Oslo: 855-861.

Gearóid Ó Cléircín, Anna Bale and Brian Ó Raghallaigh. Forthcoming. Dúchas.ie: ré nua i stair Chnuasach Bhéaloideas Éireann. *Béaloideas*.