Ainm.ie: Breathing new life into a canonical collection of Irish-language biographies

Brian Ó Raghallaigh, Gearóid Ó Cleircín
Dublin City University
Dublin, Ireland
brian.oraghallaigh@dcu.ie, gearoid.ocleircin@dcu.ie

Abstract

In this paper we present the Ainm.ie online collection of Irish-language biographies. This collection is a product of a project to retro-digitise the *Beathaisnéis* series, published between 1986 and 2007, as well as ongoing biographical work to expand and enrich the collection. The Beathaisnéis series comprised biographical accounts of 1,650 lives and an additional 520 amendments and supplementary articles, written in Irish. Persons were chosen for inclusion in this series according to their relevance to the Irish-language world. This canonical collection is an invaluable research tool for Irish-language scholars, historians and others but the print volumes risked becoming obsolete and inaccessible. As well as producing a digital version of the original texts, a key aim of the Ainm.ie project has been to ensure the continuity of biographical research in Irish by providing an online platform for publication. This paper introduces the *Beathaisnéis* collection and explains its context. It goes on to describe the digitisation process and the editorial work carried out to enrich the digital version, as well as the motivation for this version. Finally, it discusses how the project has facilitated contemporary biographical research in Irish.

Keywords: Irish language, digitisation, biographical dictionary

1. Introduction

In this paper we present the Ainm.ie online collection of Irish-language biographies. This collection is a product of a project to retro-digitise the *Beathaisnéis* series (Breathnach & Ní Mhurchú, 1986-2007), written and published between 1986 and 2007, as well as ongoing biographical work to expand and enrich the collection. With additions since initial digitisation and publication online, the collection now comprises 1,720 biographies, with a further 10-15 being added annually.

In addition to presenting the digitisation project and the resulting digital resource, we will motivate the creation of this digital resource by looking at the advantages of the digital version of the collection over the original print version.

While the Ainm.ie website is bilingual, the biographical accounts are available in Irish only. The site contains a number of browse and search facilities, which draw on a limited set of metadata stored for each biography. This metadata was added as part of the Ainm.ie project, as described in Section 4.

The Ainm.ie project is one of multiple research projects being carried out by Fiontar that involve the identification of valuable non-digital language resources, their digitisation 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 (Ó Raghallaigh & Měchura, 2014).

2. *Beathaisnéis*: some context

The original *Beathaisnéis* series comprises biographical accounts of 1,650 lives and an additional 520 amendments and supplementary articles, written in Irish. Persons were chosen for inclusion in this series according to their relevance to the Irish-language world. Some of the persons are nationally renowned and also appear in other national biographical resources but most are not widely known outside of the small Irish-language community. The *Beathaisnéis* project can therefore be seen as an alternative dictionary of national(ist) biography, using connection with the Irish language as a yardstick for inclusion. The timeline covered, from the 17th century to the present day, encompasses a period during which Irish went from being the dominant language of the country to the language of a socially and geographically disparate minority. The persons included reflect this, with 17th century chieftains, scribes and theologians rubbing shoulders with 19th century revivalists and revolutionaries as well as modern day folk singers, academics and language activists.

The original authors of the nine volume print series, Diarmuid Breathnach and Máire Ní Mhurchú, were colleagues in the archives of the Irish state broadcaster *RTÉ*. They were often asked to provide biographical information on relatively well-known figures in the Irish-language community, particularly for obituaries, and they became increasingly aware of the lack of an authoritative biographical dictionary. In order to fill in the blanks they regularly had to do basic biographical research themselves, contacting relatives and tracking down birth and death records. Eventually, in 1979, they decided to start work on a biographical dictionary themselves with the intention of covering the 100 year period from 1882 to 1982 (Breathnach & Ní Mhurchú, 2001:17). This was published in five volumes between 1986 and 1997. Breathnach and Ní Mhurchú subsequently went on to expand the scope of the project to take in the

periods from 1560 to 1881 and from 1983 to 2007. It was while working on the final published volume in the early 00's, that they began to think about passing on the responsibility to a new generation of biographers. They were extremely enthusiastic about the potential of a digital version and provided a significant amount of information and support during the early years of the project while continuing to draft new biographies.

3. Advantages of the digital version

The Ainm.ie project was inspired by the digitisation of other canonical biographical resources like the *Oxford Dictionary of National Biography*[1] and the *Australian Dictionary of Biography*[2] which proved that the move from print to digital could make such collections more accessible and potentially increase their user base. These projects also highlighted the kind of added value that the digital edition could bring such as quicker updates, regular thematic features and the potential for using the biographical data in new and interesting ways.

The initial application for funding for the Ainm.ie project in 2009 coincided with the publication of the nine volume *Dictionary of Irish Biography* (McGuire & Quinn, 1999) which was made available concurrently in an online version.[3] Advice was sought from researchers in the Royal Irish Academy, where the Dictionary of Irish Biography (DIB) project is based, regarding best practice in creating an online biographical collection.

3.1 Accessibility

It became clear from examining similar online collections and from talking to colleagues in the field that a digital version of *Beathaisnéis* could offer some substantial benefits. The most obvious of these was the possibility of making the material more accessible. Breathnach and Ní Mhurchú had deliberately published the collection in relatively small volumes with the intention of setting themselves achievable targets. Another benefit for the authors was that publication of subsequent volumes in the series allowed them to include corrections and additions to previous editions (2001:25). However, the reality of a nine volume collection published over a twenty year period, was that volumes regularly went out of print, a fact that was exacerbated by the limited size of the Irish-language publishing market. A freely accessible digital edition would make the entire collection available to all.

3.2 New possibilities

Creating a digital version of the collection would enhance its usability. Full text search and clickable cross-references would undoubtedly allow users to drill down into the collection more quickly than before.

Other possibilities such as named entity extraction and network analysis would also be opened up by the creation of a digital version of the collection.

Making the digital version available online would open up the potential to link the biographies with equivalents in other online collections, such as the DIB, and to link metadata to other online resources.

3.3 Public interaction

Putting the collection online would open up the editorial process allowing members of the public to suggest inclusions, to highlight errors and to provide various other types of feedback. This has been encouraged and facilitated by the use of Twitter and Facebook accounts to share news and features such as the *Biography of the week*.

4. Retro-digitisation

The first stage of the Ainm.ie project involved the retro-digitisation of the nine volumes of the *Beathaisnéis* series. Volumes 5, 6, 8 and 9 were made available by the publishers in a QuarkXPress publishing format that could be exported to Microsoft DOC format. These volumes were exported in this way, before being checked, exported to text, cleaned and processed for publication online.

Volumes 1, 2, 3, 4 and 7, which were not available in any digital format from which text could be extracted, were scanned and converted to Microsoft DOCX format using OCR, before being checked, exported to text, cleaned and processed. Scanning and OCR was carried out by outside contractors. Checking the texts that were created using OCR involved the reinstating of characters lost or misinterpreted during the automatic recognition stage.

Before exporting the volumes to text, bold and Italics text formatting in the DOC and DOCX documents was converted to a form of markdown, that could be retained after exporting to text. Markdown is a plain text formatting syntax.[4] Our version of markdown involved enclosing bold formatted text between asterisks (e.g. *this is a bold example*), and enclosing Italics formatted text between plusses (e.g. +this is an Italics example+).

The volumes were then exported to text, and some programmatic cleaning was carried out, e.g. spurious line breaks and superfluous white space were removed. Once cleaned, individual biographies were extracted from the volumes in text format, and saved as individual text files. The individual biographies were then processed.

4.1 Pre-processing

Before the biographies were added to the Ainm.ie database, a number of pre-processing tasks were carried out. These tasks included the extraction of basic metadata

[1] http://www.oxforddnb.com/ Accessed on 24 June 2015.
[2] http://adb.anu.edu.au/ Accessed on 24 June 2015.
[3] http://dib.cambridge.org/ Accessed on 24 June 2015.
[4] http://daringfireball.net/projects/markdown/ Accessed on 24 June 2015.

and the insertion of cross-references.

Firstly, each file was assigned a unique identifier. Each file was then converted to a simple XML format which comprised a header containing metadata relating to the article and a body containing the biography text.

Basic metadata was then added to each file. Firstly, global metadata regarding the volume and collection was inserted. Secondly, each person's first name, surname, date of birth, and date of death, where given, were parsed and extracted from the first line of each source file, and inserted into the metadata header of the XML file.

Legacy textual cross-references, i.e. "[q.v.]", "[B1]" (i.e. *Beathaisnéis*/Volume 1), "[B2]", etc., were then removed and replaced with cross-references tagged/marked up with a target identifier, i.e. the unique ID number of the target biography. These new cross-references were created programmatically by searching for each name for which there was a biography in the collection. Where a match was found, it was tagged with the target identifier. These cross-references were subsequently manually verified.

Further named entities were then searched for and tagged in the body of each biography. Placenames, as well as a closed set of publications, organisations, educational institutions, professions and political parties, were tagged during this stage of pre-processing. Some of the lists of named entities were based on indexes included in the *Beathaisnéis* series, others were compiled specifically for this purpose.

Placenames found were tagged with target identifiers from Logainm.ie, the Placenames Database of Ireland[5], the authoritative source for Irish toponymic data, and a dataset also developed and hosted by Fiontar, in conjunction with the Placenames Branch of the Government of Ireland. Place objects in Logainm.ie contain toponymic and geographic data, as well as links to other geographical databases, such as GeoNames. Tagging of placenames was done programmatically by searching for each placename in the Logainm.ie database in each of the biographies. Base, mutated and inflected forms of each placename were searched for using a linguistically aware search algorithm. In cases of ambiguity, where multiple places in Logainm.ie had the same name, all possible references were added, and the correct one was selected by hand afterwards.

4.2 Editorial processing

The files were then committed to a central Subversion data repository to which an editorial team was granted access. Editors worked on local working copies of the repository, and committed changes as they worked. Editors worked on the XML files using a locally installed XML editor. A stylesheet was developed to facilitate user-friendly editing.

The editorial team enhanced the collection in a number of ways. Firstly, all automatic pre-processing was checked, and OCR errors were corrected. In addition, a style guide was developed for the digital edition in an attempt to standardise items such as references, quotations, dates and numbers as well as certain spelling and grammatical issues. The original volumes were published over a twenty year period and thus contained a certain amount of inconsistencies that could be cleaned up. The style guide is now circulated to new contributors to ensure consistency.

The most significant editorial enhancement was the integration of supplementary notes to the primary biographies. As mentioned in Section 3, the authors had included new information relating to over 500 biographies in appendices at the end of each volume. This allowed them to include new research that had come to light since the primary biography was published and also to correct any factual inaccuracies. The digital edition provided an opportunity to amend the relevant accounts to reflect this additional information. Careful redrafting was necessary in some instances where the supplementary information was extensive and the original authors were consulted when appropriate. This element of the editorial process continues today as newly-published research is reviewed and new information relevant to a biography is added. The editorial team also accept submissions from the public via email and correct inaccuracies or add minor details when verified.

4.3 Post-processing

Once all biographies had been checked and enhanced by the editorial team, the collection of biographies were prepared for publication online. This stage in the project involved the development of a tool to export the collection from the repository into a purpose built relational database.

For each biography in the collection, the tool extracts the metadata from the article's XML header before inserting it into the database in normalised form. The tool then extracts and cleans the XML body before inserting it into the database. The tool is now run weekly to update the Ainm.ie database.

5. Tools and resources

Drawing on Fiontar's experiences from the Téarma.ie[6] and Logainm.ie projects, web and database technologies were harnessed to publish the biographies online. A web application was built to present the biographies in a user-friendly way to a new audience.

The Ainm.ie web application comprises a home page, an information section, a number of tools for browsing and searching the collection, and a biography viewer. The home page also includes a *Biography of the week* widget which can be embedded on other sites.

[5] http://www.logainm.ie/en/ Accessed on 24 June 2015.

[6] http://www.tearma.ie/Home.aspx Accessed on 24 June 2015.

The first of the browsing tools is the alphabetical list. This tool groups the biographies alphabetically according to the surname, and comprises a paging tool to browse the letters of the alphabet. One of the novel aspects of this tool is that it lists women under both the feminine and masculine forms of their surnames. For example, the biography of *Áine Ní Raghallaigh (1868 - 1942)* will be listed both under "N" and under "O", amongst instances of the masculine form of that surname, i.e. "Ó Raghallaigh". This feature is language specific.

The second browsing tool is the themes tool. This tool allows users to generate lists of biographies that share named entities. This tool uses the tagged named entities in the body of the biographies to build visual tag clouds. The named entities include placenames, publications, organisations, educational institutions, professions and political parties, all of which were tagged and verified in the pre-processing and editing stage. The tag clouds are rebuilt each time the database is updated.

The third browsing tool is the timeline tool. This tool groups the biographies by birth and death dates. Once a year is selected from the timeline, a list of persons born on that year as well as a list of persons who died that year are presented to the user.

Additional browsing tools are incorporated into the biography viewer, in the right hand column. The first is a Wikipedia style infobox. This infobox contains links to other persons that share an occupation with the current person. The second tool lists persons in the collection with the same surname as the person being viewed. This tool is linguistically aware in that it will list both men and women with the same surname. The third tool lists biographies that contain cross-references to the current biography. Finally, all cross-references from the current biography to other biographies in the collections, or to places in the Placenames Database of Ireland (Logainm.ie), are clickable hyperlinks. These links are created during the transformation of the biography from database entry to web page.

Finally, the full text of all biographies in the collection can be searched using the search tool. This tool can be accessed from the home page.

6. Continuity

A central concern of this project from day one was ensuring the continuity of Ainm.ie as an authoritative Irish-language biographical resource. As mentioned in Section 2, the original authors were keen to hand over the responsibility to a younger generation of researchers so it was important to create a sustainable structure. To this end, a panel of 'joint-editors' was established in 2013 to write new biographies and to provide information regarding the update of existing biographies in the collection. This editorial panel produces 10-15 new biographies per year. A shortlist of candidates for biography is agreed upon at an annual meeting between the joint-editors and the publisher with each joint-editor being allocated a number of biographies to work on. The texts are then processed and published by Fiontar.

7. Future plans

Fiontar currently has a limited amount of funding to host and maintain the website which makes it somewhat difficult to plan major developments. We would like to further develop the site in a number of ways, with a view to strengthening links with other projects and resources, and thus enhancing the user experience. We hope to enhance the search and browsing tools by incorporating the *Irish Surnames Index* we are developing as part of the Dúchas.ie project, a collaboration with University College Dublin to digitise the National Folklore Collection of Ireland (Ó Cleircín et al, 2014). This resource would facilitate the suggestion of related biographies based on relationships between different surnames. We also intend to link the entries in this collection, where possible, to related entries in other collections and databases. We undertook a comparable project with Logainm.ie, using linked data to connect places in the Placenames Database of Ireland with places in other datasets such as GeoNames (Lopes et al, 2014). Finally, we plan to redesign the home page of the site to enhance usability. We also plan to enhance the editorial experience by developing web-based editorial tools which would supersede the current setup, which involves offline editing of individual XML files checked out from a repository.

8. Acknowledgements

The project is a partnership between Cló Iar-Chonnacht, an Irish-language specialist publisher that holds the copyright to the material, and Fiontar, Dublin City University, who developed the technical solution described in this paper. Funding for the project was provided by the Irish government.

9. References

Breathnach, D. & Ní Mhurchú, M. (1986-2007). *Beathaisnéis* (9 volumes). Dublin: An Clóchomhar.

Breathnach, D. & Ní Mhurchú, M. (2001). 1882-1982 Beathaisnéis: Fiontar Taighde. *Studia Hibernica*, 31, pp. 17-25.

Lopes, N., Grant, R., Ó Raghallaigh, B., Ó Carragáin, E., Collins, S., & Decker, S. (2014). Linked Logainm: Enhancing Library Metadata using Linked Data of Irish Place Names. *Communications In Computer And Information Science*, 416, Theory and Practice of Digital Libraries, pp. 65-76.

McGuire, J. & Quinn, J. (Eds.) (2009). *Dictionary of Irish Biography*. Cambridge: Cambridge University Press.

Ó Cleircín, G., Bale, A. & Ó Raghallaigh, B. (2014). Dúchas.ie: Ré Nua i Stair Chnuasach Bhéaloideas Éireann. *Béaloideas*, 82, pp. 85-99.

Ó Raghallaigh, B. & Měchura, M. B. (2014). Developing high-end reusable tools and resources for Irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies. *Proceedings of the First Celtic Language Technology Workshop* (CLTW), 23 August 2014, Dublin, pp. 66-70.