A Quantitative Analysis of Biographical Data from *Ainm*, the Irish-language Biographical Database

Úna Bhreathnach, Cathal Burke, Jeaic Mag Fhinn, Gearóid Ó Cleircín, Brian Ó Raghallaigh

Fiontar & Scoil na Gaeilge, Dublin City University
Dublin, Ireland
E-mail: Una.Bhreathnach@dcu.ie, Cathal.Burke@dcu.ie, Jeaic.MagFhinn@dcu.ie, Gearoid.OCleircin@dcu.ie, Brian.ORaghallaigh@dcu.ie

Abstract

This paper looks at some trends identifiable in the biographical data contained in the *Ainm* collection of Irish-language related biographies. The data structure is described and the reasons for its particular structure are outlined. The structured data is then analysed to identify some notable patterns and significant gaps in the Ainm biographical collection. These features and omissions are discussed in the context of the creation of both the original print biographical dictionary (the *Beathaisnéis* series) and the more recent digital version (www.ainm.ie).

Keywords: Mining biographies for structured information; quantitative analysis; biographical dictionaries; digitizing biographical data; Irish-language biography; Irish biography

1. The Ainm project: background

The Ainm (the Irish word for 'name') project is an online biographical database focused on people, mainly (although not exclusively) Irish, who had a connection to the Irish language. It is written in Irish and has been available online since 2011.

The database evolved from, and now significantly expands, the *Beathaisnéis* ('biography') series of published biographies (Breathnach & Ní Mhurchú 1986-2007). The authors of the *Beathaisnéis* series, Diarmuid Breathnach and Máire Ní Mhurchú, intended to create a dictionary of biography, using relevance to the Irish-language world as the main yardstick for inclusion, and with a strong focus on lives associated with the Gaelic Revival and the period 1882-1982, which are covered in five volumes (Breathnach & Ní Mhurchú 1986, 1990, 1992, 1994 & 1997). The scope was subsequently expanded, in three further volumes, to the previous periods, 1782-1881 and 1560-1881 (Breathnach & Ní Mhurchú 1999 & 2001) and to the subsequent period 1983-2002 (Breathnach & Ní Mhurchú 2002), with a further volume of supplements, amendments and indexes (Breathnach & Ní Mhurchú 2007).

The process by which the *Beathaisnéis* volumes were digitised, tagged and edited by Fiontar & Scoil na Gaeilge, DCU, is described in detail elsewhere (Ó Raghallaigh & Ó Cleircín 2015). Over 600 of the previously published biographies were updated to reflect corrections and additional information which came to light after publication. After the retirement of the *Beathaisnéis* authors from the project, and the digitsation of the print material, DCU and Cló Iar-Chonnacht (the publishers of the series) secured a small amount of funding to update and expand the biographies. A panel of contributors was established in 2013 to continue writing biographies, mainly of recently deceased subjects but also of overlooked lives, with between 10 and 15 new lives added annually.

In addition to new biographies, additional content has been added to the website. Thematic essays are added annually to provide an introduction to different categories of biography (e.g. participants in the 1916 rising; traditional singers; folklore collectors). Visualisation features have been developed too, in particular an interactive map displaying placenames tagged in the various lives. A feature for displaying the social networks of individuals is also in development.

The result to date is a collection of 1,756 biographies, with an average length of 1,223 words and 37 tags or cross-references in each. 1,652 of these biographies are from the original series and 104 have been added since 2010. These biographies overlap with the much larger English-language *Dictionary of Irish Biography*[1] (*c.*420 also feature there). The Ainm database is used widely, with an average of 1,143 searches per day (14/10/2018 - 05/03/19).

---

[1] dib.cambridge.org

2. The Ainm database: data structure

The biographical entries are stored as XML data (using the SQL Server XML Data Type) in a relational database. This allows us to store and modify the XML data in an efficient and transacted way. It also allows us to conveniently log changes and store versions of the biographical entries. Each entry in the biographies table comprises a unique ID, a text TITLE, and an XML document. The ID field is a permanent unique identifier within the database and can be used to access the biography over the web in HTML[2] or XML[3] format. The HTML is generated from the underlying XML using an XSL transformation.

```xml
<biography id="454" xml:space="preserve">
  <header>
    <fileDesc>
      <titleStmt>
        <title xml:space="preserve">MAC PIARAIS, Pádraig (1879-1916)</title>
        <titleNote xml:space="preserve"/>
        <persName>
          <forename xml:space="preserve">Pádraig</forename>
          <surname xml:space="preserve">MAC PIARAIS</surname>
        </persName>
        <addName type="pseudonym" xml:space="preserve">Colm Ó Conaire</addName>
        <addName type="pseudonym" xml:space="preserve">Siubhán Pháidín Sheagháinín</addName>
        <addName type="pseudonym" xml:space="preserve">Cuimín Ó Cualáin</addName>
        <addName type="pseudonym" xml:space="preserve">Laegh Mac Riangabhra</addName>
        <birth date="1879-11-10" xml:space="preserve">1879</birth>
        <death date="1916-05-03" xml:space="preserve">1916</death>
        <floruit xml:space="preserve"/>
        <birthPlace id="1383282" xml:space="preserve">
          27 Sráid an Phiarsaigh (Sráid Mhór Brunswick), Baile Átha Cliath
        </birthPlace>
        <sex xml:space="preserve">M</sex>
        <school xml:space="preserve">
          Scoil na mBráithre Críostaí, Rae an Iarthair, Baile Átha Cliath
        </school>
        <university xml:space="preserve">An Coláiste Ollscoile, Baile Átha Cliath</university>
        <occupation-b9 xml:space="preserve">rannpháirtí in Éirí Amach 1916 [B9]</occupation-b9>
        <occupation xml:space="preserve">múinteoir</occupation>
        <occupation xml:space="preserve">oideachasóir</occupation>
        <occupation xml:space="preserve">scríbhneoir</occupation>
        <occupation-b9 xml:space="preserve">óglach 1913-1922 [B9]</occupation-b9>
        <faith xml:space="preserve"/>
        <author xml:space="preserve">Diarmuid Breathnach</author>
        <author xml:space="preserve">Máire Ní Mhurchú</author>
```

Figure 1: TEI elements used in the biography of Pádraig Mac Piarais

The original biographical entries digitised from *Beathaisnéis* Volumes 1–9 (Breathnach & Ní Mhurchú, 1986–2007) have ID numbers in the range 1–1999. Stubs (short entries) from the original volumes are in the range 2000–2999. New biographical entries written and added between 2010 and 2016 are in the range 3000–3999, and entries added since 2017 are numbered 5000+.

Figure 2: Web view of the biography of Pádraig Mac Piarais

3. The Ainm database: metadata

The XML data follows the *TEI: Text Encoding Initiative* guidelines for the most part. In addition to the required <title/>, <forename/>, <surname/>, <birth/>, <death/> and <sex/> elements, <addName/>, <school/>, <university/> and <occupation/> are included as metadata in the <header/> element, where known. These metadata are displayed in the biography title and infobox on the public website.[4] The dates are also used on the timeline[5] and thematic tag cloud[6] tools.

In addition to (biographical) entry level metadata contained in the header, certain entity types have been tagged inline in the biography <text/> element. These include placenames (<placeName/>), publications (<opus/>), Gaelic League branches (<conradh/>), educational institutions (<eduInst/>) and political parties (<party/>). This information is used to create the aforementioned tag clouds. Placename tags include a reference to the Placename Database of Ireland[7] where the place is in Ireland and a reference to GeoNames[8] where the place is outside of Ireland. People (<persName>) are also tagged in the <text/> element. People tags include a cross-reference where the person is within the database. A element is included after the <text/> element in some of the newer biographies.

For the original collection of biographies, all inline tagging (i.e. markup of named entities in the body text of the biographical entry) was done automatically using a purpose-built tagger written in Python. The tagger searched for and tagged named entities (i.e. names, placenames, publications, institutions, and political parties). The tagger included a custom NLP function to deal with initial mutation (e.g. *gC*orcaigh) of entities in the Irish-language text. This function tagged the mutated entity and inserted the base form (e.g. Corcaigh) as an attribute of the entity. All inline tagging was subsequently manually checked. Newer biographies are tagged manually. Older biographies are gradually being re-checked as they are prepared for use as "biography of the week" on Twitter, Facebook and in the project newsletter. User feedback is also considered.

The *Beathaisnéis* collection (on which the Ainm database is primarily based) is the result of the passionate work of two committed amateur biographers. One of the challenges this poses for the creation of standardised metadata is that the original authors did not complete profile sheets or index

---

2 e.g. www.ainm.ie/Bio.aspx?ID=454
3 e.g. www.ainm.ie/Bio.aspx?ID=454&xml=true
4 e.g. www.ainm.ie/Bio.aspx?ID=454
5 www.ainm.ie/Timeline.aspx

6 www.ainm.ie/Tags.aspx
7 www.logainm.ie
8 www.geonames.org

cards such as those commonly used by other dictionaries of national biography (Warren, 2018; Reinert et al, 2015) and which can make the creation of entry-level metadata (e.g. where the person was born) relatively straightforward. While Breathnach and Ní Mhurchú did follow a typical formula in constructing their biographies, they did not list key elements such as profession, religion, gender or place of birth/death independently of the text. In order to retrospectively register such information (i.e. the entry-level metadata as opposed to the named-entity recognition previously described) it has been necessary to manually extract the relevant details – a slow process that is still ongoing in the case of certain elements. The image below from a typical entry displays the typical categories for which data has been extracted to date. It includes date of birth, date of death, place of birth, gender, school, third-level education and occupation. Other common categories such as religion and place of death have yet to be extracted.

dáta breithe	5 Aibreán 1846
dáta báis	4 Samhain 1916
áit bhreithe	Cnoc na Coilleadh, Co. Mhaigh Eo
inscne	fear
scoil	Scoil San Uilfrid, Manchain, Sasana
ardoideachas	Coláiste Oiliúna Naomh Mhuire, Hammersmith, Londain
slí bheatha	scoláire

Figure 3: Typical categories used for data extraction (example: Tomás Ó Flannghaile)

4. Quantitative analysis

4.1 Background

The digitisation of biographical collections offers the opportunity to examine collections at a scale not previously possible when only utilizing biographical text. With the creation and linking of standardised datasets from unstructured text, the overall contents of the collection can be revealed and trends can be analysed. The Finnish BiographySampo[9] (Tamper et al, 2018), for example, illustrates the potential for such examination, interrogating the biographies as a collection of linked data, as does the Netherlands' BiographyNet[10] (Fokkens et al, 2017). Not only do such datasets illustrate the overall makeup of a national biography, and therein the history of the nation,

but also the history of the collection itself, as Warren (2018) suggests in his examination of the *Oxford Dictionary of National Biography* (ODNB); the digitisation of national biographies can be used, he says, to analyse dually the history of both the nation and of the dictionary itself. In doing so, Warren illustrates not only what the ODNB contains, but also how it came to be:

> "...investigating the *ODNB* (1.) *in its entirety* and (2.) *as an historically contingent digital artifact* offers wider purchase on the historical knowledge it makes available and the historical knowledge-making it constrains..." (2018)

Searches by occupation, birthplace or year tell less about their individual importance in the history of the nation than they do about the imagination and biases of the various contributors. Warren notes that mothers of women in ODNB are quite often queens, and mothers in general tend to be actresses, teachers and noblewomen, while fathers are frequently landowners, army officers, clergymen or merchants. National biographies do not generally attempt to capture the typical member of the nation, but rather the atypical, the exceptional names deemed important by the biography's contributors. It might seem odd that "naval officer" is the third most frequent profession in the ODNB and that *Britain and the Defeat of Napoleon* (1996) is the most referenced monograph in the entirety of the ODNB, but the context of the ODNB's construction casts light on these seeming oddities: a prolific naval historian, Sir John Laughton (1830-1915), was responsible for 1,000 biographies of naval figures, roughly 1 out of every 38 *Dictionary of National Biography* entries, all of which, it was decided, would be added to the subsequent ODNB. This required additional research which often referenced *Britain and the Defeat of Napoleon* (1996).

Likewise, most common years of death in ODNB not only reflect periods of illness or bloodshed, but also the inherent biases contributors brought in with their selection of subjects:

> "The local peaks in 1883 and 1908... once again remind us to attend to the data infrastructure. Rather than marking some hitherto unknown plague afflicting the Victorian aristocracy, 1883 marks the point at which contemporaneous deaths ceased to be meaningful to Stephen, his deputy Sidney Lee, and their collaborators." (Warren, 2018)

What biases, designed or unintended, can therefore be found within Ainm? Inspection of the most common years

of birth or death reveal the clearest influence the selection criteria and original aim of the biographical dictionary had on the dictionary itself. The original impetus for *Beathaisnéis* was to produce a biographical dictionary based on 100 years of the Gaelic Revival, covering those who had died between 1882 and 1982. The scope of the project gradually changed as the authors decided to include lives from both before and after that arbitrary period. Nonetheless, the fact that the first five volumes focused exclusively on that 100 year period and that only two of the nine volumes cover the period from 1560 to 1881 means that the collection is inevitably biased towards the period from the mid 19th century onwards. This is shown clearly in figures below (Figure 4).

The image of the nation captured in the pages of the ODNB is less a reflection of the nation's history than of the making of the dictionary itself (Warren, 2018). Between 1450 and 2000, France, the Netherlands, and the United States of America all apparently supersede England, Scotland, Ireland, and Wales in importance in the ODNB, Warren claims, because of the presumed Englishness of each biographical subject. Relevant countries besides England were mentioned specifically, while any reference to the dictionary's own nation was left assumed, and therefore left out. Likewise, there is a presumed continuity among the Ainm biographies: each person played a particular role in the Irish-language world. Their relevance to this world is fundamental to their inclusion in this collection and remains the principal criterion for evaluating suitability. The first volume of the *Beathaisnéis* series outlined criteria for inclusion:

> "...Irish speakers who did something remarkable or who achieved a level of excellence in their lives. Undoubtedly there are also Irish speakers who wouldn't earn a place in the national pantheon but who are still of importance or who are well-known in the context of the Revival period. Both types are included in this volume and will be included in future collections."[11] (Breathnach & Ní Mhurchú, 1986, 11)

Although it might not seem noteworthy to specify the first official language of the nation in a national biography, the case of Irish is somewhat exceptional, in that it exists simultaneously as official and minoritised, essential to the establishment and imagination of the nation while only being spoken by a minority of the same nation. These biographies, therefore, aim specifically to capture, and write into being, the Irish-language nation not previously recorded in biography. In highlighting the important lives of the nation which are relevant to the Irish language, the writing of these biographies inherently asserts the importance of the Irish language itself in both the basis and the imagination of the nation.

4.2 Timespan

Magnus Ó Domhnaill (c.1490-1563) is the earliest-born in the collection, with 2017 the most recent year of death. Spanning seven different centuries, there are 306 different years of birth. An interesting demographic is revealed when we analyse the years of birth and death on a broader range. The most significant is the fact that 815 people were born in the 19th century, 46% of the lives. If we add in those who were born in the 20th century, 514 people, we reach 1329 lives, 75% of the total collection. Furthermore, 1064 people died in the 20th century, 62% of the total. Therefore, having been born in the 19th century, and having died in the 20th century, the majority of lives lived through the revival period of Irish, something which comes in line with the understanding that the *Beathaisnéis* project initially centred around those most active in reviving the language during the late 19th and early 20th century. There are however 110 people with no year of birth or year of death stored as metadata. (In most cases, these lives were in the form of short 'stub' articles by the *Beathaisnéis* authors, where birth and death dates were not included in the title and therefore not automatically extracted. This is an area for future improvement. Having identified this gap, we will begin manually extracting available missing data in order to store it accordingly.)
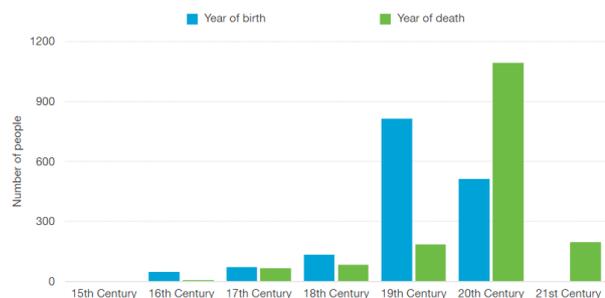

Figure 4: The predominance of the 19th Century (year of birth) and the 20th Century (year of death) in the above graph helps to further illustrate the original aims of the *Beathaisnéis* project.

4.3 Gender

Men account for almost 90% (1,580) of the biographies. There are only 176 biographical accounts of women. These figures highlight a major gender imbalance in the collection, and surely represent the greatest area for future research opportunities, but they are not out of sync with

---

[11] Authors' translation.

other international biographical databases (Farr, 2012). Of the women included in the collection 86% were born from the year 1847 onwards and 76% died in the twentieth century. Of the biographies written since 2013, 26% are of women, showing a significantly increased representation.

4.4 Birthplace: country and county

A connection to the Irish language is the primary condition for inclusion in the collection, yet 20 different countries are represented in the database. Ireland (including Northern Ireland) is the top represented country with 1,217 people. England is next on the list with 63. Germany, Scotland and the United States are next, each with 15. The other countries represented in the database are India, Norway, Switzerland, Sweden, Italy, Wales, France, the Netherlands, Australia, Denmark, Malta, Belgium, China, the Czech Republic and Japan. The total number of people born outside of Ireland is 140, or 8% of the collection.

399 people have no recorded place of birth stored as metadata. There are a number of very short biographies (134) which lack key biographical information and require further research. Filling this gap represents an area of future improvement for the project. It was not possible for the original authors to find records of a place of birth for some 103 lives from the 16th, 17th and 18th century.

Each county in Ireland is represented in the collection (Figure 6), with Cork (the largest county by size) being the highest represented county with 197 people, or 16% of those born in Ireland, and Fermanagh and Leitrim (both small counties) being the lowest represented counties with 3 lives each. The top six represented counties are Cork, Dublin, Galway, Kerry, Donegal and Waterford; all except Dublin contain an Irish language speaking area, or 'Gaeltacht'. These counties represent 42% of the collection. The number of people born in England, 63, is higher than any of the other 26 counties.
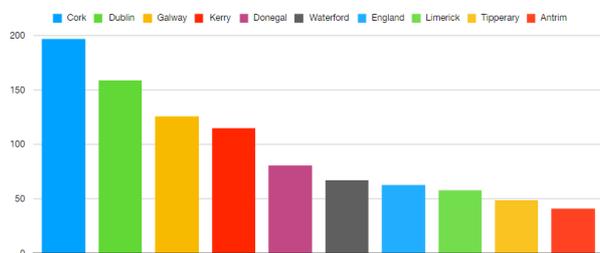
While each province is widely represented in the collection, Munster is the highest represented province with 534 lives (44% of those born in Ireland), almost double that of Leinster (280), followed by Connacht (210) and Ulster (193). The noticeable difference is something which has been previously alluded to by the original authors of the biographies. Although they made an effort not to neglect other areas (referring to Connacht and northern Leinster particularly), it is clear that there were more people of interest to them in Munster, due mainly to the historic strength of Irish in the province, particularly around the time of the revival period: '…there is a sort of nucleas or kernel of literacy, as you'd say, in Munster, and especially in Cork, and maybe part of Kerry as well. But, I saw figures from the time of the revival… seventy or eighty percent of the people reading the language were in Cork.'[12] Diarmuid Breathnach also states his belief that 'they had very good Irish, especially those from Cork and a lot of Munster people particularly.'[13] Of the biographies published since 2013, 17% are from Munster. This reduced proportion may be attributed to the fact that the Irish-speaking community is no longer Munster-dominated, or to the bias of the *Beathaisnéis* authors.
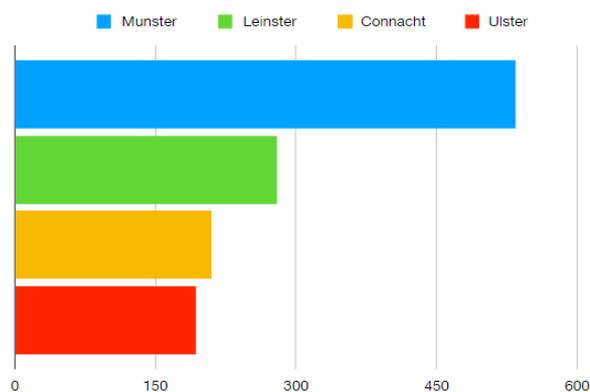

Figure 6: Province of birth for those born in Ireland (where metadata is available)

4.5 Education and profession

A university education is recorded for 767 subjects; another 989 do not have metadata stored regarding university education and some of these were also university educated. Of those with available metadata, 40% of women (71) attained some form of university education, in comparison

---

[12] Translation of extract from unpublished interview with Diarmuid Breathnach and Máire Ní Mhurchú, 2010.

[13] Translation of extract from interview with Diarmuid Breathnach and Máire Ní Mhurchú, 2010, www.ainm.ie/Info.aspx?Topic=resources.en

to 44% of men (696). University College Dublin (173) was the most commonly attended university of the database, followed by Trinity College Dublin (111), St. Patrick's College, Drumcondra, Dublin (71), National University of Ireland, Galway (69), St. Patrick's College, Maynooth (61), and National University of Ireland, Cork (56). 45 attended either Oxford University, Cambridge or Harvard, but only 6 of these were born in Ireland. There are accounts of people attending university all across Europe, most notably universities in England, Germany, France, Italy, Spain and Belgium. Of these, the Irish Colleges of Rome and Paris, and Leuven University, Belgium, appear more frequently than others. The preponderance of these religious institutions can be attributed to the large number of clergymen who travelled abroad for education during the period from the 16th to 18th centuries when this was prohibited to Catholics in Ireland.

Of the 1,756 lives, 1,690 have at least one occupation recorded, with 'teacher' being the most common profession, among both men and women, with a representation of 20% of men and 24% of women (21% total). This makes sense in the context of the central role the Irish language played, and continues to play, in the education system, however the original aims of the collection certainly influence this propensity towards education, given the necessarily central role of teachers in the revival of any language. Many of those involved with the Gaelic Revival spent time teaching Irish to others.

There is a very high proportion of clergymen. There are 239 Catholic priests (bishops, archbishops, Christian brothers, Franciscans, Jesuits) and 42 Protestant ministers, which represents around 18% of men documented. In comparison, there are only two nuns, Mary Bonaventure Browne and Máire Treasa Ó Murchú, recorded in the collection.

Most of those teachers and clergymen had a second occupation for which they were more recognised; clergymen were often professors. For both men and women, writers, scholars and poets complete the top five professions: being a published writer was one of the suggested criteria for inclusion in the collection.[14] This preponderance of writers, and initial suggestion for their inclusion, also corresponds with the original focus on the Gaelic Revival, in which the construction of a modern, written literature in Irish played an important role. Other occupations to feature highly on the list include civil servants, musicians, singers and folklore collectors, politicians, lecturers, translators and editors; there are also

lawyers, doctors, astronomers, actors, journalists, artists, engineers, miners, broadcasters, soldiers, and publishers.
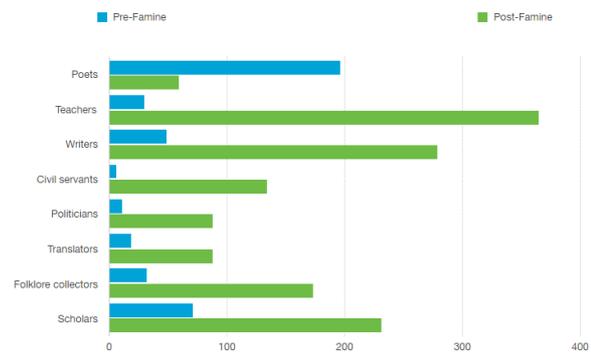

Figure 7: Professional demographics for subjects born pre and post the Great Famine
(where metadata is available)

Priests, poets, and writers dominate the professions early in the 16th and 17th centuries. The 19th century (see Figure 7) sees a decline in poets, 70% of whom were born before the start of the Great Famine (1845); this can be attributed to the decline of the bardic poet tradition in Irish. The numbers of teachers, civil servants, politicians and translators begin to rise around the same time. The end of the 19th century and beginning of the 20th century also sees a rise in folklore, music, and song collectors, no doubt due to the desire to recuperate all that was lost in the previous century of famine, emigration and political unrest.

5. Conclusion

The Ainm example highlights some issues which confront digitisers of biographical dictionaries: omissions or unstructured data in original material, and text which is not easily tagged. These issues are still being addressed by the editorial team.

The preponderance of 19th and 20th century lives in Ainm is a reflection of the original editorial aims, rather than of the most important era for the Irish language, which had begun to decline as a literary and administrative language long before then. Quantitative analysis can be used to confirm the authors' acknowledged bias towards certain regions (Munster) and professions (writers), as well as the usual gender disparity. As Warren (2018) found for the ODNB, so too for Ainm: it tells the history of both the nation and of itself.

---

[14] Unpublished interview with Diarmuid Breathnach and Máire Ní Mhurchú, 2010.

6. Acknowledgements

7. References

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

Farr, M. (2012). Review of *Online Dictionaries of National Biography*, (review no. 1259), https://reviews.history.ac.uk/review/1259, (accessed 09.05.2019).

Fokkens. A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V. (2017) BiographyNet: Extracting Relations Between People and Events. In: Bernád, Á. Z., Gruber, C., & Kaiser, M. eds., *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*. Wien: New Academic Press. pp. 193--224.

Muir, R. (1996). *Britain and the Defeat of Napoleon, 1807-1815*. New Haven: Yale University Press.

Ó Raghallaigh, B., Ó Cleircín, G. (2015). Ainm.ie: Breathing new life into a canonical collection of Irish-language biographies. In *Biographical Data in a Digital World (BD)*. Amsterdam: CEUR-WS.org, pp. 20--23, http://ceur-ws.org/Vol-1399/paper4.pdf, (accessed 15.05.2019).

Reinert, M., Schrott, M. and Ebneth, B. (2015). From Biographies to Data Curation - The Making of www.deutsche-biographie.de. In *Biographical Data in a Digital World (BD)*. Amsterdam: CEUR-WS.org, pp. 13--19, http://ceur-ws.org/Vol-1399/paper3.pdf, (accessed 15.05.2019).

Tamper M., Leskinen P., Apajalahti K., Hyvönen E. (2018) Using Biographical Texts as Linked Data for Prosopographical Research and Applications. In: Ioannides M. et al. (eds) *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. EuroMed 2018. Lecture Notes in Computer Science, vol 11196. Springer, Cham.

Warren, C. (2018) Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics*. 22.11.2018, 10.31235/osf.io/rbkdh, (accessed 08.05.2019).