Landscapes, languages and data structures

Isses in building the Placenames Database of Ireland

Michal Boleslav Měchura
Fiontar, Dublin City University
mechrm@dcu.ie

Abstract

This paper reviews some of the data-structural issues that arose during the construction of the Placenames Database of Ireland (www.logainm.ie). The issues covered include linguistic issues such as the internal struture and grammatical properties of placenames; issues related to the fact that the database is bilingual, such as cross-language placename borrowing and anglicisation; and finally issues caused by administrative hierarchies, such as the problem of multiple overlapping hierarchies and the usability problem that occurs when multiple geographical units at multiple levels of hierachy carry the same name.

1. Introduction

The Placenames Database of Ireland is a large bilingual database which records the official names in Irish and English of geographical objects on the island of Ireland, as far down as street level in some areas. The database contains over 100,000 entries, is accessible to the public for free (www.logainm.ie) and serves around 150,000 searches per month. The database and its accompanying website are the result of cooperation between the Placenames Branch in the Irish government's Department of Arts, Heritage and the Gaeltacht, which provides the data, and *Fiontar*, the Irish-language school in Dublin City University, which provides the technology. The technical architecture consists of a back-end database and a public-facing website, as well as a password-protected editorial interface used by the editorial team to input and edit the data.

The public-facing website was launched in 2008 in the context of recently enacted legislation which gave more prominence to Irish-language placenames in the Republic of Ireland; for details on the legal and sociolinguistic situation of placenames in Ireland see Mac Giolla Easpaig (2008 and 2009). The database was created primarily to serve the practical day-to-day needs of translators and public administrators but, in addition to this, has attracted an unexpected amount of interest from historians, folklorists, genealogists and generally anybody with an interest in Irish studies. The reason for this appears to be that in Ireland, the discipline of *dinnseanchas* 'lore of placenames' features prominently in studies of history, culture and language, arguably more so than in other European countries. For more insights into the history and role of placenames in Ireland, see for example Mac Mathúna (1990) and again Mac Giolla Easpaig (2009).

Technically speaking, the Placenames Database of Ireland is a conventional relational database (implemented in Microsoft SQL Server) consisting of tables for the various data categories which include: places, their names, sound recordings of those names, place categories (such as county, townland, mountain, river), position in the country's administrative hierarchy, geographical coordinates, historical citations linked to a bibliography, as well as scanned images of archival index cards used before computerization.

We will not attempt to describe the structure of the Placenames Database of Ireland holisticly in this paper. Instead, we will concentrate on three sub-areas where interesting challenges have arisen, namely: issues relating to the structure and properties of placenames as linguistic objects (in Section 2), issues arising from the fact that the database is bilingual (in Section 3) and issues caused by administrative hierarchies (in Sections 4 and 5). In many cases, the issues presented here are issues that have not been solved satisfactorily in our database yet, or a solution has not been implemented yet. In this sense, the paper is merely a discussion of issues rather than a catalogue of solutions.

2. Linguistic properties of placenames

Placenames databases differ in how much attention they pay to the linguistic properties of the names themselves, such as their gender, grammatical number, inflection and so on. Multilingual databases such as Geonames usually ignore such aspects completely; this is understandable given that their focus is mainly on geography rather than on language and given that it would be unfeasible to accommodate many language-specific annotation schemes for many different languages in a single database. Smaller databases, however, such as the Placenames Database of Ireland, often focus on only one or two languages and part of their mission is language promotion. The Placenames Database of Ireland has it as part of its mission to make Irish-language placenames more prominent and to encourage their use. In such contexts, it may be advisable to think of the database as mainly a lexical database rather than a geographic one, that is, a database that records a lot of linguistic facts about the names. Linguistic properties relevant to placenames can be broadly subdivided into the following three areas.

Internal structure. Practically all placenames that consist of more than one word have an implicit internal structure, just like other linguistic expressions. Some aspects of this structure may need to be made explicit in a database, in the form of XML annotation or as separate database tables. These include: (1) The presence of definite articles at the start of the name, or other textual phenomena that get in the way of alphabetical sorting. In Irish and other Celtic languages this also includes initial consonant mutations that are triggered by definite articles. In some cases the definite article can be more or less optional, or only used in some grammatical contexts. For example, some Irish-language placenames are used without an article in the nominative case but with one in the genitive case: *Gaillimh* ‘Galway’ but *muintir na Gaillimhe* ‘people of Galway’. (2) Composite names such as *Ballaghgowla and Froghan*. Ideally, one wants such names to appear in an alphabetical list under both components, and to be findable in a search engine by typing even just one of the components. (3) Names with disambiguators such as *Black Lough (South)*. These often obtain when a traditional geographical unit has been subdivided into two administrative units. One might want the disambiguator (the component in brackets) to be formatted differently on screen (or in printed output) and the placename to be findable in a search engine by typing even just the un-disambiguated name.

Combinatorial properties. These are linguistic properties relevant to how the placename interacts with the text in which it is being used and include language-specific features such as gender, grammatical number and inflection paradigms. A placenames database with a linguistic focus, such as Placenames Database of Ireland, should record some or all of these facts explicitly, like a dictionary would. In addition, the following combinatorial properties are relevant to placenames in particular: (1) Whether and how the name can be combined with categorizers such as ‘town’ or ‘county’ when these are not part of the placename proper. Example from English as it is used in Ireland: *Donegal Town* (categorizer follows name) but *County Donegal* (categorizer precedes name). (2) Which locative prepositions are combined with the placename. Many languages display a duality between *in* and *on*: some placenames use the preposition *in* while others use the proposition *on;* their distribution is largely arbitrary and only partially motivated by semantics or geographical features. Irish: *i nGaillimh* ‘in Galway’ but *ar an gCeathrú Rua* ‘in Carrarow’ (literally ‘on Carrarow’). Czech: *v Čechách* ‘in Bohemia’ but *na Moravě* ‘in Moravia’ (literally ‘on Moravia’).

Lexical relations to other words in the language and to other placenames. The former (relations to other words in the language) include demonyms (terms for ‘people from’) and derived adjectives which are sometimes irregular; English: *Galway → Galwegian* [demonym and adjective], Russian: *Москва* [noun] → *москвич* [demonym], *московский* [adjective]. The latter (relations to other placenames) include cases where one geographical object has been named after another, such as *Ballybeg Road* or *Lismore Terrace;* in these cases it may be advisable to make the link between the two geographical objects explicit in the database so that, if the name needs to be translated into another language, it is known *which* Ballybeg and *which* Lismore is the source of the name.

3. Cross-linguistic issues

A separate cluster of issues stems from the fact that the Placenames Database of Ireland is bilingual. This does not seem to pose a challenge at first; the general principle is that every place has two names, one in each language, and this seems to call for a simple data structure: all we need is two text fields. However, that approach would fail to account for the following phenomena.

Borrowing and gaps. Sometimes, a place has the same name in both languages. An example is an area of Dublin called *Dún Laoghaire*. This is an Irish name which is also used in English with unchanged spelling (but with anglicised pronunciation). An obvious solution would be to simply record *Dún Laoghaire* twice, once as an Irish name and once as an English name. However, this is unsatisfactory as it fails to account for the fact that this name is not really *in* English, it is merely *used* in English.

On the other hand, in some strongly Irish-speaking areas, minor features such as crossroads, fields and wells only have Irish names and there are no known English names. One can only assume that if somebody needed to refer to such a place while speaking English, one would briefly code-switch into Irish to utter the name.

A fairly granular data structure is called for here, one that allows us to capture facts as to whether a name in a given language exists or not, whether the name *used* in a given language is also a name *in* that language, and whether the name has been borrowed from another language.

Anglicisation, translation and re-interpretation. Many English placenames in Ireland have been obtained from the original Irish names by a process of writing down an approximated pronunciation (*Gaoth Dobhair → Gweedore*). In other cases, when the English name was created first, the method used to coin the Irish name has often been translation, such as *Butler's Bridge → Droichead an Bhuitléaraigh*. In other cases still, the Irish and English names are independent coinages (example: *Loch Garman/Wexford*).

Translation is sometimes accompanied by re-interpretation. An example of this is an area of Dublin called *Barra an Teampaill/Temple Bar.* The placename originated in English from the personal name *Temple* but was later re-interpreted as the common noun *temple* and hence the Irish name (literally 'the bar of the temple'). Although based on a misunderstanding, it was decided to keep the Irish name as official because it is commonly used.

An ideal data structure would allow us to record these and other etymological relations between names of the same place in different languages. If such relations are explicated and annotated in the database, then we can not only provide better information to users but also extract interesting statistical observations about the relative proportion of these phenomena in the country's body of placenames.

4. Dealing with overlapping hierarchies

Most countries are subdivided into administrative units such as districts, provinces, counties or similar. These units form a hierarchy and such hierarchies are deliberately designed to disallow overlapping, so that each unit always only has one parent: each unit is wholly contained in another. Such hierarchies are called *nested hierarchies* in mathematical terms, are easy to model computationally and facilitate logical reasoning (if object A is in object B and object B is in object C, it follows that object A is also in object C).

In Ireland, the situation is far from this computational ideal. The basic units (*counties → baronies → civil parishes → townlands*) may have originally been designed as a nested hierarchy but are no longer so because of boundary changes that have not been propagated up and down the hierarchy. Several more recent systems have been designed to overcome this problem, such as the system of electoral divisions. While these new systems are nesting and computationally tractable when taken in isolation, they often overlap with the traditional system in a way that cross-sects a particular townland into several electoral divisions and so on. Because the traditional system of townlands and civil parishes is well known and inspires local patriotism, one cannot simply ignore it in a placenames database. In many cases, a townland is a unit of naming and a target of loyalty, while an electoral division is nothing more than a utilitarian unit of administration. Townlands mostly have long-established names while electoral divisions often have awkward, recently coined names with disambiguators (*Ashtown A*, *Ashtown B*) or composite names (*Terenure-Cherryfield*).

The consequence when building a placenames database is that we must work with an *overlapping hierarchy* in mathematical terms, one where a child may have more than one parent. This complicates things; for example, logical reasoning is no longer always possible. If we know that A is in B and that B is in C_1 and C_2 simultaneously, we can no longer infer whether A is in C_1 or C_2 or both. In fact, there is no way to know this other than by deriving it from a dataset of geographical boundaries, or by recording it explicitly. The former (geographical boundaries) would be preferred, but the latter (explicit recording) is the approach taken in our database. Even though it introduces a potential for inconsistencies, we have had no other choice as we do not have unimpeded access to accurate boundary data.

5. Place as an abstract concept

Consider the placename *Dún na nGall/Donegal* which can be found in the north-west of Ireland. The question to ask is: how many places called *Dún na nGall/Donegal* are there in this corner of Ireland? Depending on one's perspective, the answer may range from one to five. An outsider will see only one, a county of that name. A local inhabitant will probably see two, the county and its capital town of the same name. A placenames researcher will see a townland of that name contained in a civil parish of the same name, contained in a county of the same name. A local politician will probably see an electoral division of that name and also a town of that name with its town council. In total, there are five units called *Dún na nGall/Donegal* in that part of Ireland.

In our database, these are treated as separate objects which, as far as the database knows, only *happen* to have the same name. That, however, is unsatisfactory as it fails to distinguish a case like this from cases such as the 19 places called *An Baile Mór* (literally 'the large town') which can be found all over Ireland and which genuinely *happen* to have the same name. The fact that the five objects called *Dún na nGall/Donegal* share the same name is not a coincidence.

Another reason why this is unsatisfactory is that it introduces a potential for inconsistency because data such as the Irish name's grammatical information need to be recorded five times instead of once. Also, there is an uncertainty as to which of the five records we should attach historical citations to: the town, the county, the townland...? Last but not least, a casual user searching for *Dún na nGall/Donegal* in the database may be confused by a listing of five places where he or she intuitively only expects one or two.

An ideal data structure would provide a way to connect several concrete places (such as the five Donegals) to a single “abstract place”. The abstract place would contain all information common to the concrete places, such as names and historical citations, and these would then be inherited by the concrete places. This would introduce a guarantee of consistency into the database and would also make it possible to present such cases in a more user-friendly way to non-specialist users, for example by grouping the five Donegals under one heading.

6. Conclusion

This paper has reviewed some of the data-structural issues that have arisen during the construction of the Placenames Database of Ireland. Many of the issues are inherently linguistic and stem from the fact that we conceive of our database as primarily a lexical database and only secondarily as a geographical database.

A second cluster of issues is caused by administrative hierarchies. Our placenames database must accommodate a heritage of conflicting and overlapping administrative hierarchies, which complicates the task of database design and impedes automated reasoning. But the mere existence of an administrative hierarchy, even a mathematically ideal one, still creates issues for the database designer as he or she has to deal with the multitude of perspectives different users have. Where a casual user perceives a single geographical unit and a single placename, the specialist perceives several units at several levels of hierarchy. The casual user’s view, however, cannot be merely dismissed as uninformed or incorrect because the hierarchical units the specialist perceives are in fact manifestations of a shared abstract concept.

References

A. Placenames databases

Geonames: http://www.geonames.org/

Placenames Database of Ireland: http://www.logainm.ie/

B. Literature

Mac Giolla Easpaig, Dónall (2008) ‘Placenames Policy and its Implementation’ in Caoilfhionn Nic Pháidín, Seán Ó Cearnaigh (eds) *A New View of the Irish Language*, Dublin: Cois Life, pp. 164-177

Mac Giolla Easpaig, Dónall (2009) ‘Ireland’s heritage of geographical names’ in *Wiener Schriften zur Geographie und Kartographie,* vol. 18, pp. 79-85

Mac Mathúna, Liam (1990) *Ár dTimpeallacht Logainmneacha: Inniu agus Amárach* [our placenames environment: today and tomorrow], Dublin: Coiscéim