Problems with databases and the XML solution

John G. Keating, Denis Clancy, Thomas O'Connor and Marian Lyons

Databases of biographical material are now a common feature of historical research.[1] Many historians have converted archive-derived data into electronic format to facilitate the storage, querying and presentation of the material.[2] In some cases, these electronic databases have been made available online.[3] However, as biographical databases have proliferated on the web, problems have emerged, many of them linked to the type of technology used by historians. Most history databases are implemented using unsophisticated tools like Microsoft Access. While these systems are adequate as data-storage facilities they impose certain limitations. In particular, they oblige historians to shoehorn historical data into pre-existing categories that are not custom built for the material in question. Nor do they easily accommodate variation in the density of information between individual records. This results in the generation of unfilled 'white space' so that database size and inflexibility increase even though many of the records remain sparsely populated. While commercially produced databases have query facilities these become more difficult to operate as the individual database becomes more complex. Merging is also problematic. This is because most researchers design their databases primarily as data stores: they do not construct them in formats that facilitate merging with other databases. This causes unnecessary splintering of the research enterprise and encourages duplication. Traditional databases also suffer from accessibility problems. Because individual researchers are typically concerned with organizing data for a specific research purpose (a PhD, for example, or a funded project), their databases are not designed to facilitate access or querying by other research interests, for instance by geographers or sociologists. Web hosting is also difficult. While it is straightforward to create personal databases it can be difficult to make these available online. The result is that many valuable prosopographical databases remain inaccessible to the larger research community and to the general public. Lastly, traditional databases are not 'future proofed'. This problem has two aspects. First, traditional databases are designed to respond to research needs at a particular time. However, research needs inevitably change over time and there is growing concern that rigidly constructed, format-filled databases will be insufficiently flexible to evolve appropriately. Second, as information technology continues to progress, new document delivery systems will become available. Again researchers fear that traditional databases will prove incompatible with these. Thus data painstakingly collected and entered by historians in one generation is seriously at risk of becoming inaccessible to the next.

[1] See, for example, David Bracken, 'Irish migrants in Paris hospitals, 1702-1730' in *Archivium Hibernicum* lv (2001), pp 7-47 and Virginia Davis, 'Material relating to Irish clergy in England in the late middle ages' in *Archivium Hibernicum* lvi (2002), pp 7-50.

[2] Virgina Davis, 'Medieval English Clergy database' in *History and Computing* ii, 2 (1990), pp 75-87.

[3] For example, Steve Murdoch's, Scotland, Scandanavia and Northern Europe Project database. Visit http://www.abdn.ac.uk/ssne/. Visit http://www.linacre.ox.ac.uk/research/prosop/prosopon.stm for medieval prosopography. For the Clergy of the Church of England Database 1540-1835 visit http://ww.personal.rdg.ac.uk/~lhstalrs/cced.htm

There is not a general consensus that the solution to these problems lies in a closer working relationship between database professionals and history researchers to produce databases that are appropriate to the historical data, flexible enough to facilitate querying and robust enough to enable web hosting and the creation of an integrated electronic research environment. The starting point is the historical data themselves. Together with database professionals, historians must take due cognisance of the unique nature of their raw material. This involves developing, at the outset, an electronic language, which is appropriate to the unique nature of the data and sufficiently flexible to permit storage, querying, management and web hosting. In the case of biographical data in particular, designers must take into account that data density and type will vary according to sources, epoch and milieu and will rarely be consistently rich over the full range of records. Furthermore, allowance must be made for the fact that new data will inevitably become available for existing records. This poses the problem of integrating new material into the existing body of data and also of assuring its accessibility through modified query and reporting facilities. It is also important to consider how data will be used in the future. Without adequate technical input at the beginning, databases can prove unequal to the changing demands and expectations of the research community.

A modern approach to biographical database creation

To illustrate some of the problems currently facing researchers working with computerised data, consider the following example. Let us say that the following personal record exists in an archive: James Byrne, born in 1755 in Maynooth: entered a seminary (Paris) in 1772; ordained in 1780. In a traditional, relational database, these personal data are entered into a single (simple, limited detail) table with the following fixed fields: first name, last name, address, date of birth, seminary, entry date and departure date.

Now we learn from another archival source that Byrne was appointed to Athy parish in 1780 and served in several other parishes before taking up his final appointment in Maynooth in 1795. It is at this point that the database becomes more complex. How can these new data be integrated into this record? Data modelling studies suggest that efficient, easily searchable data should be ‘normalized’ to minimize data redundancy.[4] This requires the allocation of a unique identification number, also called a key field, to every individual’s name in the database, in this case, Byrne. This key may be used when referring to the unique record in other tables. This is how *relational* databases work. Thus, in our case, a ‘parish’ table is created and includes both the list of parishes and the positions held by the cleric. However, as there will be unique information other than its address, relating to each parish, this will also need to be recorded. A new ‘parish’ table is therefore required to hold this information and a unique key assigned to every parish record. In order to associate key fields of the ‘personal’ and ‘parish’ tables, a *third* table is now required. Its purpose is to indicate which clerics served in which parish. As clergy move around parishes on new assignments, it will be necessary to record the date of the assignment in the ‘parish’ table or elsewhere. This requires a modification. Variable-length biographical-data, associated with a record, always require additional tables, as data recorded in a database table field must be atomic. For example, it is not possible to

[4] For example, see C.J. Date, *An introduction to database systems*, (Reading, Mass., 1995).

have a ‘date’ field associated with a record that holds multiple dates. In this case, if we want to record Byrne’s, and his colleagues’ various assignments, it will be necessary to create another table of records containing a key field and an associated date. Thus it can be seen that as more biographical information is recorded, the number of tables and relationships, and the consequent complexity of the database, increase. It becomes more difficult to query the database and users require a sophisticated understanding of the table and relationship structure in order to upload, query and report on collected data.

To avoid these problems we can now consider an alternative mechanism for data encoding. Let us again consider the data record for James Byrne, described earlier in Access-type relational database language. Instead of sundering our painstakingly collected data into a number of related tables, we can describe or encode the data in a markup-language (data description language) based on standard XML (extensible markup language). XML is a data description language derived from SGML (as is HTML).[5] It is an already well-recognized standard, used successfully in other domains for data modeling, exchange and storage. Using XML we can describe or encode the data as follows:

```
<record>
  <personal>
    <name>
      <firstname>James</firstname>
      <lastname>Byrne</lastname>
    </name>
    <born>Maynooth</born>
    <birthdate>1775</birthdate>
  </personal>
  <seminary location=”Paris”>
    <entered>1772</entered>
    <ordained>1780</ordained>
  </seminary>
</record>
```

This XML record contains the information in a single record organized into two major sections that are ‘tagged’ using the labels <personal> and <seminary>. It is easy to add further, variable-length data to the record by creating additional tagged sections. For example, to encode the parish assignments, the following section could be easily inserted into Byrne’s record:

```
<parishes>
  <parish name=”Athy” date=”1780” position=”Curate”>
  <parish name=”Maynooth” date=”1795” position=”Parish Priest”>
</parishes>
```

For biographical databases, this system is far superior to the relational-model described earlier. First, the encoding language describes the data as it really is. XML permits us to write a specific language to describe our own unique data in a way that is universally

[5] For details, background and other resources on XML, and information on the W3C working group on XML, visit http://www.w3.org/XML/. SGML is the acronym for ‘standard generalised markup language’. HTML is the acronym for ‘hypertext markup language’.

accessible. Second, in this encoded form, our data is susceptible to a broad range of manipulations. Third, new data sections can be added with ease. Fourth, it is possible to query the data with standard software packages. Last, it is easy to migrate from one encoding scheme to another. Database merging is simple and, like so many other functions, can be executed automatically.

By adopting XML history professionals, working with large amounts of biographical data, finally have a standard to facilitate data storage, interchange and investigation. The concept of a universally accepted data standard is not new. Librarians have been using the MARC (MAchine Readable Cataloging)[6] encoding standard for over thirty years, and it is the *de facto* standard for information interchange by library systems. It is no surprise that the Library of Congress Network Development and MARC Standards Office are developing a framework for working with MARC in an XML environment. They realize that the flexibility of the system allows users to work with data in ways specific to their needs while still retaining the ease of interchange with other users and systems.

Irish in Europe Project XML pilot progamme, 2003-4[7]

During the academic year 2003-4, the Irish in Europe Project (department of modern history) and the computer science department at NUI Maynooth conduced a pilot study to assess the feasibility of using XLM to develop an encoding language for biographical data collected on the Irish migrants to the Continent in the early modern period. The pilot project had three objectives: first, to use XLM to create a markup language appropriate to the biographical data in question; second, to develop software to facilitate the querying of the data encoded; third, to put the resulting database online. The particular dataset used in the study was the biographical set, compiled by Professor L.W.B. Brockliss (professor of early modern French history, Magdalen College, Oxford) and Professor Patrick Ferté (professor of French history at the university of Toulouse) and originally based on archival records concerning Irish students who attended French universities in the seventeenth and eighteenth centuries.[8] We modeled the biographical data entries using XML and produced a 350-element language suitable for general biographical usage.[9] XML allowed us define a 'grammar' to describe our specific data. Because it was supported by a well-defined suite of standards and software, we were in a position to test various tools, with a view to developing a comprehensive querying facility.[10] Lastly, a

[6] For details on MARC, and the MARC 21 XML Schema, visit The Library of Congress' Network Development and MARC Standards Office website at http://www.loc.gov/standards/marcxml/.

[7] Visit http://irishineurope.com.

[8] L.W.B. Brockliss and P. Ferté, 'Irish clerics in France in the seventeenth and eighteenth centuries: a statistical study' in *Proceedings of the Royal Irish Academy*, 87, C, 9 (1987). The original typescript of the prosopography was deposited with the Royal Irish Academy and a copy was made for the Russell Library, St Patrick's College, Maynooth. Professor Ferté's revised version of the prosopography is reproduced in this issue of *Archivium Hibernicum*.

[9] Denis Clancy 'Brockliss and Ferté: an online biographical database' (unpublished undergraduate thesis, department of computer science, NUI Maynooth, 2004).

[10] Ibid. We found that currently available freeware software packages have problems with native XML integration. This is not serious, however, as the problems will be addressed in future releases, and there are perfectly acceptable workarounds in the meantime.

proto-type web page was designed with a view to uploading raw data directly from archives into the database *via* a GUI or web interface.

The pilot project, despite the limitations of time and funding, concluded that the XML encoding option was the best one for web-hosted, biographical databases. Particularly impressive was the multi-purposing of the data allowed by the unique flexibility of the tailor made markup language. From an initial requirement for universal access to the Irish in Europe Project database we produced a prototype, Internet accessible, portable Migration Map generation software using freely available, standard tools and software. The cost of the software required was negligible as it was available on most platforms. We used the Apache Web Server and associated tools.[11] The current programmes are written using the Perl programming language and utilise a number of programming libraries to parse XML documents and generate PNG (GD library) and PDF (PerlMagick software) documents all of which are available from the Comprehensive Perl Archive Network (CPAN).[12] The programmes are Internet accessible and utilize the CGI (Common Gateway Interface) software library to facilitate Internet browser access.

Used in this way, the technology delivers a range of promising tools, including a capacity to generate migration density maps automatically, which we can illustrate here. As mentioned above, the pilot programme prototype software utilizes synthetic data based on the results presented by Brockliss and Ferté. In their report they produced a table summarizing the diocesan origin of Irish clerics at the University of Paris for the periods 1590-1639, 1640-1689, 1690-1739 and 1740-1789. We built a simple mathematical model, a non-linear discrete distribution, which we used to seed a random cleric generator. This generator was used to produce a random XML record for a hypothetical cleric, born in a year band, who migrated from a particular diocese, according to the distribution. It took just twenty seconds to generate a database of 1000 XML records, each containing around ten fields. It took approximately five seconds to parse this file and produce a colour-coded Migration Density Map based on a time period specified by the user (using a Macintosh G4 Powerbook running OSX 10.3.4). Figures 1 and 2 show grey-scale automatically generated maps for the periods 1590-1639 and 1740-1789, respectively.

We also modeled a hypothetical destination for the clerics, assuming they migrated to a number of destinations in Europe. In this case we randomly assigned a destination in France, Spain, Italy, England, Scotland and Wales to each randomly generated cleric. The software then built the Migration Density Map as before and created a thumbnail of this map in a larger European map showing the destinations. An example of this map is shown in Figure 3. The maps presented here are low density copies produced for viewing using an Internet browser. However, it is possible to produce high-quality with additional programming and higher-quality base maps.

[11] For a complete description of the open-source software projects of the Apache Software Foundation, a community of software developers and users, visit the foundation website at http://www.apache.org/.

[12] This is probably the largest repository of free software worldwide. The CPAN (Comprehensive Perl Archive Network) is a large collection of Perl software and documentation. Visit http://www.cpan.org/. For information on the Perl, a high-level programming language developed by Larry Wall, visit http://www.perl.org and see Larry Wall et al, *Programming Perl* (O'Reilly, 2002).

Conclusion

From this pilot project we concluded that this facility promotes ease of data interchange and dissemination. Using this technology, we can take records encoded with XML and automatically produce HTML (Hypertext Markup Language) for web rendering. We can also produce PDF (Portable Document Format) for journal printing and PNG (Portable Network Graphics) for graphical presentation. The range of data manipulation permitted by the technology is comprehensive. Further, our prototype grammar may be expanded as the data, and as research needs, require. With XML technology, history and social science professionals can concern themselves with describing their data, using a formal notation. They can then avail of software, some already available, to assist in storage, mining and manipulation. It is our expectation that our XML-based data standard, developed for the universally accessible Irish in Europe database, will, by dint of its performance, recommend itself as the universally recognised standard for history professionals.

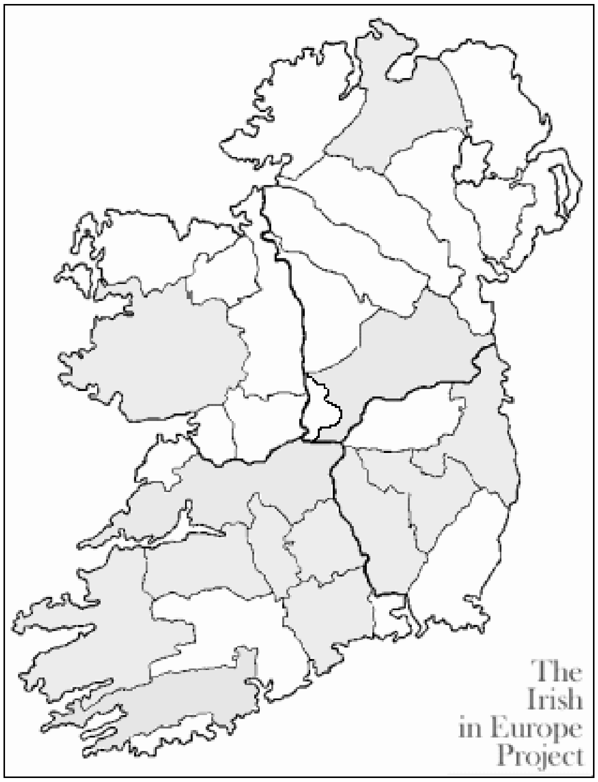

Figure 1 Automatically produced Migration Density Map showing migration from Ireland to France during the period 1590-1639 (after Brockliss and Ferté, 1987).

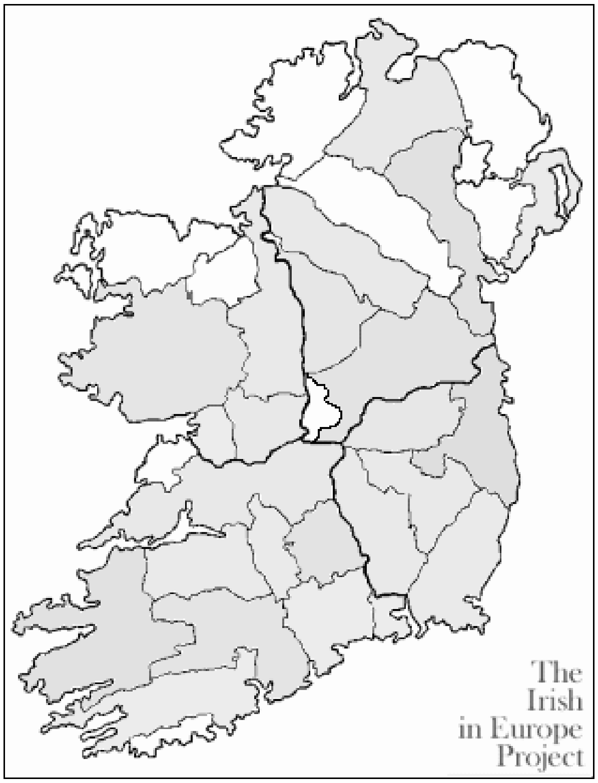

Figure 2 Automatically produced produced Migration Density showing migration from Ireland to France during the period 1740-1789 (after Brockliss and Ferté, 1987).

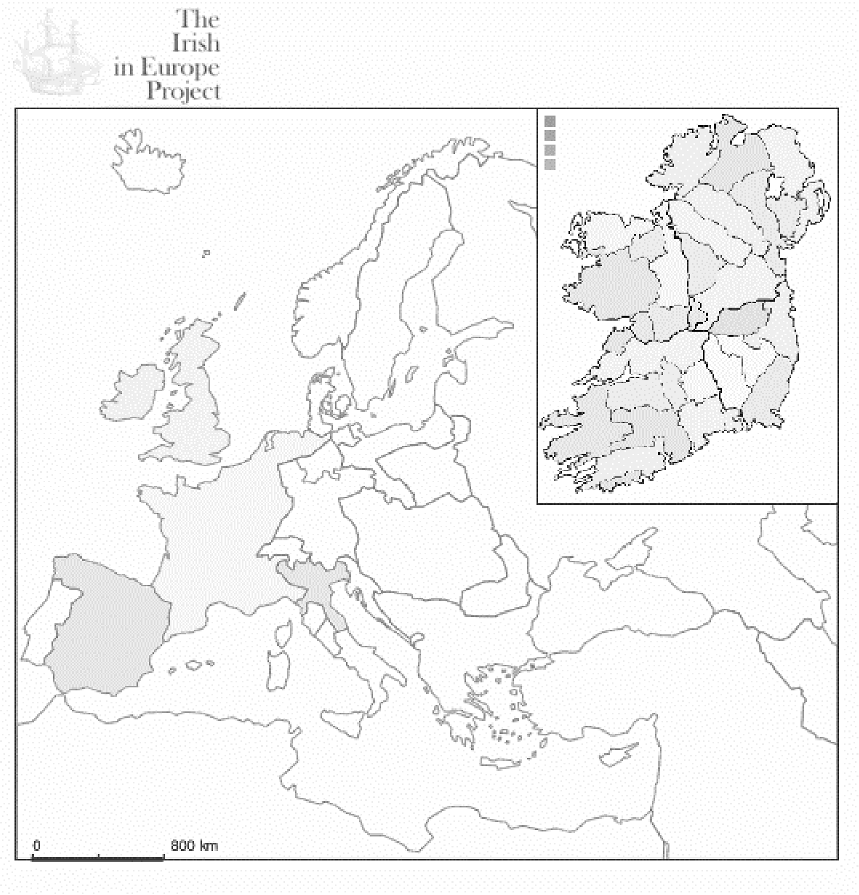

Figure 3 A Migration Density Map showing the origin and destination of migrants.