IDEAS ON THE STRUCTURE AND METHODOLOGIES OF A PUBLIC MEDIEVAL GENEALOGICAL DATABASE – PART 2

by Joseph A Edwards[1]

[Continued from vol. 1, no. 1, pp. 22-30]

ABSTRACT

This two part article sets forth some principles and design specifications by which a public medieval genealogical database could be constructed. This is reinforced through practical and theoretical examples with explanation and justification. It is hoped that this article will act as a platform to stimulate wider debate. The FMG welcomes all correspondence on this subject. This instalment details the possible presentational formats of a public database and shows how it could be effectively administered.

*Foundations (2005) **1** (5): 338-343*

Since the publication of the first part of this article (Edwards, 2003) I have received very useful constructive criticism from a number of people, for which many thanks. Taking this into account, I have developed my theories to take the project in a slightly different direction. In this second part I shall wrap up some outstanding points from part one, based on the original premise, and then explain how I see it progressing. I aim to tackle the new ideas in future articles.

Recapitulation

In part 1 of the article[2], I addressed the technical architecture of the underlying database (expanded below with further detail on reference-data links). I defined classes, sub-classes, fields, and, in a general sense, sub-fields. I also looked at the structure of the tables, their interlinking, and the obligatory and optional fields involved. Simple examples were used to demonstrate how the system would work. A pilot version of a public medieval genealogical database was mounted on the FMG website, however it is still under development[3] and the full functionality is not currently available. Anyone wishing to contribute ideas, talents, or assistance is encouraged to contact the author at the FMG.

Branches and Households

This idea was mentioned briefly in part 1, but I should like to expand on the concept of an arbitrary group identifier for a person, as this could help with the apportionment of editing and moderation of submitted data (see later).

Broadly speaking, the set of people to be included in the proposed database is limited to those where lineage and inheritance are of prime importance (Edwards, 2003,

---

[1] The author is a trustee of the FMG, and works professionally in database design and analysis. *Address for correspondence:* c/o FMG (see inside cover of this journal)

[2] Part 1 was originally published in *Foundations* (January 2003) **1**(1): 22-30; it is now available on open access from our website at http://fmg.ac/

[3] The Medieval Genealogical Registry is at present offline pending fundamental changes. The Source element (presently equating to the FMG library catalogue) is still available at http://fmg.ac/MGR/. See that page for any future information regarding the MGR.

p.22, note 3). As such, in almost all cases it is possible to identify an arbitrary 'head' of each family, from whom all members descend. Let us describe this group as a 'household' and subgroups within it as 'branches' (each of which must have its own branch head from whom all branch members are directly descended). These categorisations are purely arbitrary and reflect the choice of the compiler, but their uses are threefold:

- They enable users to browse to find a person very simply by narrowing down the search
- They allow useful grouping for both indexing and statistical purposes, which will be very useful to prosopographers
- Editors and moderators of incoming data can be allocated a series of branches, so that new persons in the database are automatically assigned

A simple set of rules can be set to govern and control the choice of branch/household: a person can only be a member of one branch/household[4]; one (or both) of their parents must also be a member, or the person in question must be the head. Using this simple logic it will be seen that the branch structure actually corresponds to a genealogical tree, as demonstrated in Fig 1[5]:

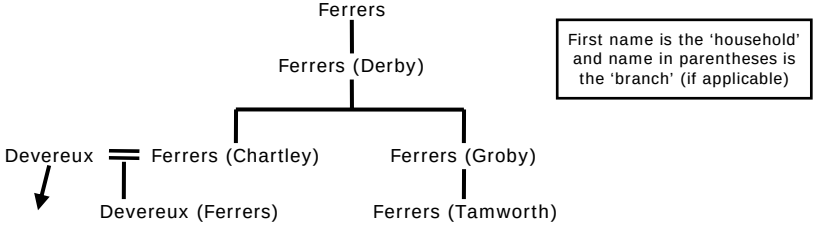

Fig 1. ***Example of a Household/Branch Tree***

There is clearly a need for a facility to merge branches, as the database grows and as people from the same theoretical branch are added initially in unconnected blocks. It would also be useful to have a holding system for individual unconnected people, such as extra spouses, rather than have to create a branch for them all on their own. When (or if) their branch develops they can be added to a real branch.

Multi-Compiler Data Referencing

The only way to undertake a project as massive in scope as this is with the help of numerous people from across the world, doing as much as they felt able. This is not just by contributing data to the database but also in moderating the submitted data.

---

[4] In principle any one individual could be allocated to the same branch (or household) as either of their parents, but not both. This gives the database compiler the opportunity to select the most relevant branch, whether maternal or paternal, for the person concerned.

[5] An interesting example of different methods of relationship representation, for both groups and individuals, can be seen at http://eclectic.ss.uci.edu/~drwhite/, particularly their PGraph project. This is definitely more prosopographical, demonstrating any type of relationship rather than specifically lineages, but the ideas have a wide applicability. This type of system could clearly be used to interrogate our proposed database for onomastic, occupational, demographic, and other sorts of data, in order to perform statistical analyses.

See also the various articles in Keats-Rohan (2002) for related material.

Approved moderators/editors need to be assigned to receive data submitted for persons in their allocated branches. The data received should then be checked, as far as possible, edited and added to the master database.

Resolving Different Opinions

Obviously there will be times when contradictory opinions are submitted on the same reference-data combination, and the most accurate outcome is not obvious to the primary moderator[6]. This will affect the Accuracy and/or Notes fields. There are a number of different ways to tackle this. In cases where the submitted conclusions are equally valid and the moderator(s) cannot confidently decide the more accurate outcome, then the data from all submissions should be presented, with the edited Accuracy field (A_x in Fig 2) set to 'Inconclusive'.

Where a decision should be possible, but the moderator is not confident in deciding alone, then the data from all submissions plus the current entry in the master database should be put up for discussion amongst the primary and secondary moderators for the branch. This discussion could take the form of anything from a simple majority vote to a detailed online discussion in a purpose built forum. If an outcome is agreed upon by a sizable majority then it should be carried forward as the current accepted outcome on the master database, but all new material that passes the initial scrutiny should initially be posted on the master database as footnotes to the current data, to allow others to check the work of the moderators and to see all competing theories.

This system is equally valid for disagreements between submissions that cite different references, however in this case it might first be useful to inform the contributors about the alternative opinion with the evidence cited, to see if the disagreement can be resolved among the contributors.

Quality Checking of Submissions

In order to maintain the quality of data in the master database the moderators will need to vet submissions to a certain extent. I feel that limiting submissions to recognised or pre-approved contributors would deter newcomers who might have valuable contributions to make. A better system would be one of internal acceptance ranking, purely amongst the moderators. This could be used both to fast track contributions from those with an established track record, and to flag warnings against contributors whose previous submissions were dubious or poorly researched. The latter could be checked more thoroughly, and if necessary views sought from other moderators.

No one will ever be able to produce the definitive answer on any given topic, as new research constantly leads to changes in accepted ideas. However if all submissions are clearly linked to sources, with accurate transcripts and a clear explanation of the conclusion drawn, then others can easily question the interpretation and make their own submissions. This is how the database would evolve to become progressively more accurate.

---

[6] The primary moderator is the one to whom the Branch is allocated, and of which there can only be one per branch, though of course they can moderate multiple branches. If they require further consultation they could consult a series of secondary moderators, of which there could be many for each branch.

'Factoids'

The concept of 'factoids' is introduced by Bradley and Short (2002), pp.12-15[7]. In our system a factoid equates to the event data – reference link. It provides clarity and simplifies presentation. Developing my discussion of enhanced handling of factoids, using multiple compilers, sources, and opinions (Edwards, 2003, p.29, note 29), I have expanded the table structure, as shown in Fig 2. This is the main aspect where I have developed the initial concepts from part 1 of the article (see also the final paragraph of this article):

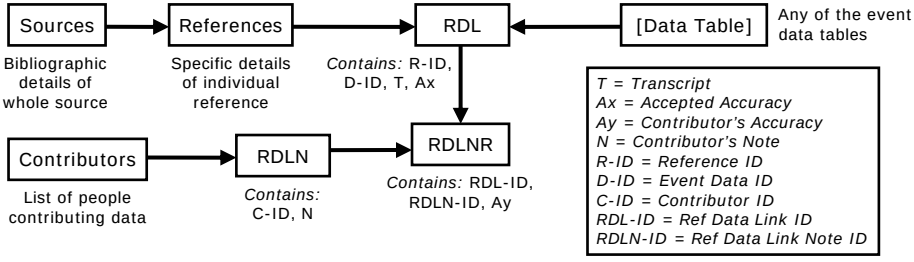

Fig 2. ***Event Data – Reference Link Tables***

Graphical Presentation

It is necessary to present the database in a variety of formats, to cater for different users with diverse aims. As well as a range of text formats, graphical representations such as genealogical trees provide a simple summary for the casual browser, and help in navigation between persons and the understanding of complex relationships.

Tabular Data

The beauty of an online project is that data does not all need to be crammed into the same space, but can be spread out over a large number of sub-pages. The tiered structure described here fits this ideally. I propose that at the highest level against the person there is just a summary of the accepted facts and nothing more. This provides an accessible 'front page' for each person that both catches the interest by being streamlined and not too wordy, and gives the facts in an easily digestible form. This level alone will be sufficient for the casual browser and those who just want to learn a little about the person concerned.

From this summary one can then drill-down to the full 'accepted' detail for the person, *i.e.* all the event data fields. This could be structured in a collapsible tree format building on a more detailed summary. At this level the genealogist or prosopographer can access much more detail, can reliably cite the database (quoting the person's unique identifier and the version of the page) and be reasonably confident that what is presented has been through a rigorous approval and editorial process by the moderators. A very useful addition at this level would be a full source/reference list, and hyperlinks to each stated piece of data, though not the detail of the factoids.

---

[7] They define a factoid as "*an assertion in the form: This source at this point says this about this person*".

At the final level of detail all the data is seen, including everything in the factoid tables, and how each piece of data has been interpreted by each contributor, with their notes (see augmented table structure in Fig 2). This final stage would by necessity be fairly dense, so good filtering systems would be needed to home in on the data being sought. This level would be of interest only to those doing in-depth research, or wanting to make new contributions to the database.

Obviously, major historical figures will be cited by a very large number of sources that give essentially the same data. To cite them all would quickly overwhelm the system. This is the role of the moderators, to weed out sources and contributions that add nothing to the body of knowledge or to the discussion on the data item. It would also be prudent to have periodic pruning sessions as the database grows, conducted jointly by a group of moderators, to remove sources and contributions that are no longer useful. This should be done with great care, particularly where usually reliable sources disagree with the currently accepted data.

Genealogical Trees

The logical nature of genealogical relationships makes it relatively easy to represent them graphically, though there is a definite problem of exponential increase with number of generations (usually worse in descent trees due to the expanding population size). As such it is much more efficient to represent and easier to decipher ten generations with 17 six generation charts (with a two generation overlap[8]) than to use one ten generation chart[9]. Another problem with graphical trees is the presentation of implexes[10] and intricate relationships of incest and illegitimacy. This is such a complex topic that it deserves a study all of its own and I hope to address this in a future article.

If we ignore 'fan charts' as just a presentational variation, there are three main types of tree: ancestors, descendants and 'hourglass', the latter being a combination of the first two with a single person at the centre. To treat implexes simply without going into too much detail it is useful to have the facility to show or hide the resulting duplicate persons. If hidden it is sensible to show, for instance, an arrow indicating where else on the chart the first instance of that person can be located. Fig 3 shows an example:

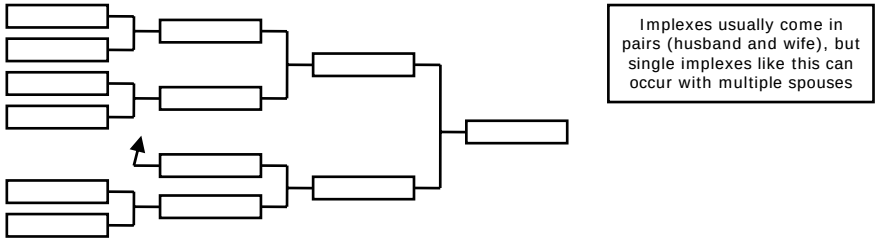

Fig 3. ***Example of an Implex with Hidden Duplicates***

---

[8] Although a single generation overlap can be used, two generations helps the user to get their bearings and offers them a greater confidence that they are following the correct line.

[9] For an example of the former preferable method see http://fmg.ac/Projects/CharlesII/ on the FMG website, which is a graphical index to Neil Thompson and Charles Hansen's ancestry of Charles II series in *The Genealogist*.

[10] A French term for multiple instances of the same person within a descendant or ancestor tree, i.e. intermarriage within a family.

There are a number of programs available which will convert standard GEDCOM files[11] into online narrative and graphical pedigrees. One of the best systems currently available is a program called HTML Pedigree[12] which has a very strong system for generating manipulable and customisable genealogical trees. There will always be limits to what an automated system can achieve, but charts more complicated than a program can generate are usually only needed for specific illustrations so can be created manually.

Future Developments

I started this project wanting to provide a central repository where data could be amassed to provide an overall picture of the current state of knowledge. Researchers could then use this as the backbone from which their detailed researches could stem. The results of their research could then feed back into the database and, through continuous iteration, the overall picture portrayed would become more and more accurate.

Whilst this aim has not changed, the consensus of feedback received from part 1 of the article is that to foster interest and involvement the database would need to have formal detailed research more integrated into the system, *i.e.* providing a clear facility for publication of peer reviewed material directly linked to the database. This could provide an alternative to the detailed multi-source factoid recording, for a specific study topic, by providing a link to a specialised published paper within the database. The paper would present the necessary discussion of the data and any conflicting conclusions in a more traditional, free text format. This would in effect remove the 'automation' of the presentational format, created by having to work through the database and present data in a specific way, but would allow more flexibility to contributors according to their preferred style of working. Clearly this moves closer towards primary research, which I specifically excluded at the beginning of this article. The approach to handling this extra dimension will need careful thought. Nevertheless I feel that the database structure and approaches to data acquisition discussed in this article have considerable merit, and the FMG welcomes all feedback.

References

Bradley, John & Short, Harold (2002). Using Formal Structures to Create Complex Relationships: The Prosopography of the Byzantine Empire – A Case Study. In: Keats-Rohan, K S B (editor), *Resourcing Sources*. Oxford: Unit for Prosopographical Research. pp. 3-21.

Edwards, Joseph A (2003). Ideas on the Structure and Methodologies of a Public Medieval Genealogical Database [part 1]. *Foundations.* **1**: 22-30.

Keats-Rohan, K S B (editor, 2002). *Resourcing Sources*. Oxford: Unit for Prosopographical Research.

---

[11] GEDCOM is an international standard for exchanging genealogical information between systems. Almost all genealogical programs will output in this format. It does however have limitations and does not approach the level of complexity of my proposed database.

[12] More information about HTML Pedigree can be found at http://www.htmlpedigree.com/, or see the advertisement at the end of this journal. It produces a very flexible, but highly cross-browser compatible system, which can be used with little or no knowledge of web design.