IDEAS ON THE STRUCTURE AND METHODOLOGIES OF A PUBLIC MEDIEVAL GENEALOGICAL DATABASE – PART 1

by Joseph A Edwards, MEng, ACGI[1]

ABSTRACT

The purpose of this article is to set forth some principles and design specifications by which a public medieval genealogical database could be constructed. This will be reinforced through practical and theoretical examples with explanation and justification. It is hoped that this article will act as a platform from which a wider debate could emerge. The FMG welcomes all correspondence on this subject. This first instalment of the article will deal with more generic database structure. The next instalment will detail the possible presentational formats of a public database and show how it could be effectively administered.

Foundations (2003) ***1*** *(1): 22-30* *© Copyright FMG*

Introduction

Firstly I need to define what is meant by a 'public medieval genealogical database'. It is a database that can be read and contributed to by people from across the world, but which, unlike databases such as RootsWeb[2], is genuinely a single database rather than an interlinked collection of many smaller databases. Such 'collection' databases allow users to submit their own personal database, which is then incorporated into the master index for searching purposes. The major draw back with this approach is that the same person could occur dozens, or maybe thousands of times in different guises.

What I am proposing here is a single database where data is contributed on a per person basis, rather than as a pre-interlinked set of persons. This is very hard to achieve in post-medieval genealogy, as there is very often considerable doubt whether a person in one database is the same as a person with similar details in another database. Medieval genealogy however does not, in the vast majority of cases, suffer this problem. The researcher almost always knows 'who' [3] they are talking about. The questions are more based around what is known of that person and their relationships.

Sources and references

To clarify I define a 'source' as a definable entity containing data pertaining to the subject, such as a book, or an article in a periodical. A 'reference' is defined as an individual block of data within a source such as a paragraph or a numbered person (though the reference might, in the case of very short sources, be the whole source). A

---

[1] The author is a trustee and co-director of the Foundation for Medieval Genealogy, and works professionally in a database design role for a major engineering company. Besides numerous data systems for his employers, he is the designer of the FMG's Medieval Genealogical Registry, and of their internal financial, membership, and library catalogue systems.

Address for correspondence: c/o FMG (see inside cover of this newsletter)

[2] http://www.rootsweb.com/

[3] By 'who' I mean that they are a distinct person in history. The only records that can realistically be built into medieval genealogical trees are those that reference relatively significant historical figures, and where many records point to the same individual. The vast majority of names mentioned in primary medieval sources will never be linked to an actual person.

'primary source' is almost always contemporary as it is not based on any other source. A 'secondary source' is one that obtains its data from primary sources or other secondary sources[4].

A possibly controversial condition needs to be placed on the sources used in this proposed database: the only ones directly cited should be fully accessible to the public. This means in the vast majority of cases that they will be published in one form or another, and that they will almost certainly be secondary sources. To understand the reasoning behind this condition, one must understand what the database is trying to do. I propose that it is the purpose of a large-scale public medieval genealogical database to pull together research from all corners of the globe and present it to everyone. Its purpose is not to support detailed primary research[5]. To do justice to any piece of primary research done in conjunction with the database it should be removed from the main structure and made a separate entity which I shall call an 'appendix'. If made available using the same website as that used for the database's output[6] then each 'appendix' can be considered 'published' and therefore qualify for citation in the database[7].

Private papers such as unpublished articles can also be cited by providing scanned images or transcripts[8] as 'appendices'. In primary research appendices it would clearly also be valuable to have scans or transcripts of the primary sources, but for the reasons discussed above these would still rarely be suitable for direct citation in the database as they require too much interpretation before they become easily useable.

Data Structure and Format

Before proposing how to handle the data in the database one must consider the nature of the database (described in the introduction), its structure and layering, and distinguish the different classes of data. One must then consider the possibilities for a practical format for each data class. The levels of complexity could always be increased, and probably should be, but the following should be considered a minimum.

Data classes

I propose that there are three main classes of 'person'[9] data (with a few subtle variants): relationships, names, and life events. To complicate matters slightly a 'relationship' can

---

[4] It is possible that a source could be both primary and secondary, if for instance a contemporary chronicler were to cite some manorial records, however it is usually safe to class these as primary due to their contemporary nature. Common sense, as always, should be exercised, taking into account how similar sources have been handled previously.

[5] Primary research done well requires a huge amount of ancillary data and discussion to piece together the evidence and this would overload the output of any database, making it unreadable.

[6] It is assumed that the only viable method of output to a wide audience is through the Internet.

[7] For a good example of a different type of database, one that is used to store and interpret primary sources, see the COEL project at Linacre College, Oxford, run by Katherine Keats-Rohan (http://www.linacre.ox.ac.uk/research/prosop/home.stm). Their database could in turn be cited by our proposed database as a secondary source.

[8] Transcripts should, wherever possible, clearly indicate the bibliographic details of the original source and the details of the transcriber, and/or translator if applicable.

[9] I do not propose to discuss 'source' data, such as referencing standards, as it is fair to assume that any reasonably logical and detailed reference system will leave the reader in no doubt as to the reference in question.

have certain 'life events' associated with it, and a 'life event' can have secondary or tertiary persons linked to it (however I shall be treating these linked persons as part of the life event and not as a separate relationship sub-class). This detailed data must be supported by a range of isolated fields linked directly to the person.

To break the three classes into variant sub-classes we have:

- *Relationships* comprising 'husband-wife' [10] and 'parent-child'
- *Names* comprising 'standard names' and 'nicknames'
- *Life events* comprising 'single events' [11], 'standard events', and 'full life events' [12]

Obligatory supporting fields include: a unique identifier, a version[13], the person's sex (male/female/unknown), and issue/marriage[14] (tri-state toggle). Useful ancillary fields include: illegitimate (Boolean), medieval issue[15] (Boolean), household (arbitrary group identifier), and general biographical notes.

From this point forward a single member of a sub-class will be referred to as an 'event'. Within each event (as detailed below) there are a varying number of 'fields', and I propose that it is at this 'field' level that direct source references should be made.

Breakdown of data layers

Each of the above classes comprises various layers and groupings of data. The layering can cascade within an event or from one event to other events of a similar or associated sub-class. Cascading within a single event goes beyond what could reasonably be considered the 'field' level and is only discussed here for completeness[16]. Fig 1 below illustrates the breakdown of the sub-classes[17].

---

[10] Or more correctly 'male partner – female partner' as will be discussed later.

[11] Being an 'event' that can only occur once during a single person's life, such as birth and death.

[12] This third sub-class covers data which is not strictly an 'event', but is associated with the life of the person, for example heraldic achievements, residences, inheritances, etc.

[13] A more detailed study of version control within genealogical databases will be the subject of a future article. Suffice it to say that although the version can be recorded against the person (as above), this denies the use of a version history, which has distinct benefits. In this article we shall assume the former simpler case.

[14] As with 'husband-wife' this is more correctly 'issue/partnerships'. It indicates a negative condition only, i.e. 0 = may or may not have issue and/or partnerships; 1 = may or may not have partnerships, but does not have issue; 2 = has neither issue nor partnerships. (Conditions 1 and 2 must be supported by references; condition 0 may be supported, but only to negate a reference purporting an alternate condition.)

[15] To indicate whether any issue of the person would be outside the pre-defined medieval period.

[16] 'Dates' (as shown in Fig 1) can be used to reference an exact point in time from 45 BC onwards, if used carefully. The fields required are day, month, year, calendar, and AD/BC (the latter is excluded here as it is not required for the medieval period). From this combination the day of the week can easily be calculated. For further detail see: Cheney, C R (*editor*) (1996) *Handbook of Dates for Students of English History*. Cambridge: University Press. Before 45 BC, when the Julian calendar was introduced, calendars were too variable (certainly in the western world) to pinpoint accurately an exact day in history.

[17] In theory each sub-class should have a separate main data table within the database, but in practice some of the sub-classes are so similar that they can be combined in one table for convenience. It is assumed that each table will have a unique identifier, in whatever form, which can be cross-referenced from other tables.

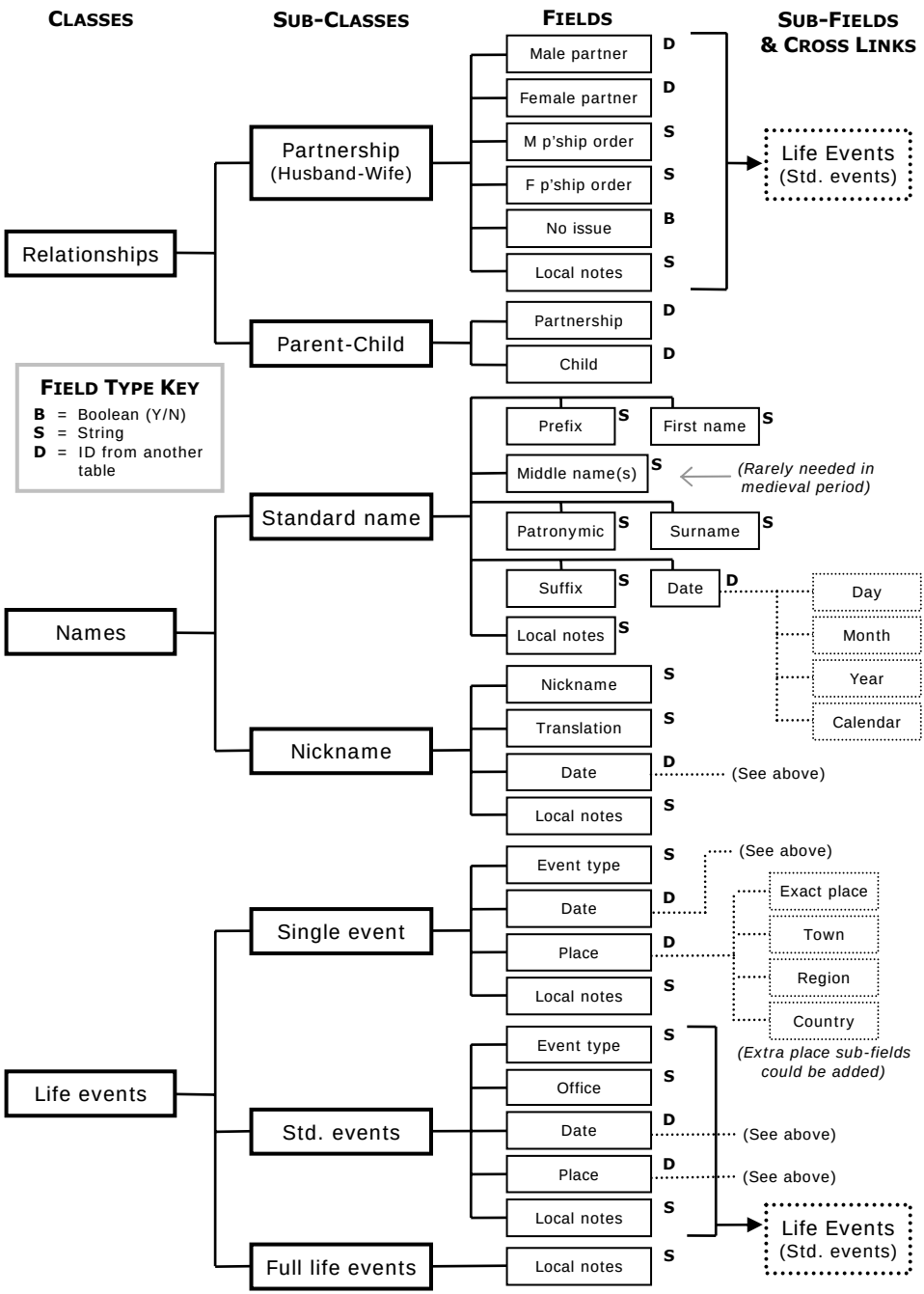


Fig 1. Breakdown of sub-classes into fields and sub-fields

As can be seen from the figure if target events in any cross-linking are limited to members of the 'standard life events' sub-class then the possible configurations of the network are drastically reduced from the theoretical maximum. This turns out to be a very reasonable simplification as standard events make up the majority of the complex detail associated with a person[18].

Components of a reference / event-data link

As stated above, it is not the purpose of this article to address the referencing of sources, and the many and various standards in existence. However we do need to address how a reference is linked to an event. In actual fact to do the job properly one must reference the individual field and not the event as a whole. Almost every database currently in existence uses the latter rather inadequate method, and the failings of this are clear. Take for example the simple case of a birth. Source A says "b. 1260, London" and source B says "b. 1265, Windsor". If the real answer is "b. 1260, Windsor", the only way to indicate the validity of the sources is with a narrative note. This destroys all the potential of a database as a logically organised set of fields each containing a single pre-defined and simple piece of data; a principle well illustrated by the emerging standards for XML[19]. If the date is referenced separately from the place then neat logical analyses can be performed by the database. This is why it is imperative that the data is referenced at the field level (as illustrated in Fig 1) if the database is to be as useful and detailed as possible.

Returning to the actual link between the field and the reference, I propose that there are three components necessary to document the link. Firstly the reference must be quoted verbatim, using the original script (such as the Greek or Russian alphabet) without any corrections[20]. Secondly the reference should be rated as to its accuracy. Thirdly there should be notes, where necessary, to explain the choice of accuracy and how the reference interacts with other references with regard to the data in question. Obviously each of the three components is increasingly subjective from the previous component.

The second, 'accuracy', needs further elucidation. The way that accuracy is judged is open to debate. It can range between anything from 'Correct' or 'Incorrect', to a scale of say 0 to 100 of 'percentage correctness' [21]. There are however a couple of other cases which should be considered: 'Inconclusive' which is subtly different to 'Unknown Accuracy' (see footnote[21]) in that the reference has not tried to draw a conclusion; and 'Implied' which is where the conclusion is not specifically stated but is clearly meant (such as marriage orders where, say, the previous wife is discussed but the wife that is

---

[18] 'Names' and 'Single events' (such as births and deaths) are always isolated and would not cross-link. 'Full life events' are more biographical and textual in nature, not being fixed to points in history. 'Relationships' are very different in nature and cannot link to themselves, excepting that a parent-child event requires a partnership ID.

[19] For a good overview of the principals of XML visit http://www.xml.com/

[20] Nevertheless 'snipping' out irrelevant phrases from the middle of sentences is acceptable, so long as the meaning of the sentence is in no way altered by that action.

[21] In the Medieval Genealogical Registry (http://medgenreg.co.uk/) the FMG has chosen to use 'Correct', 'Unknown Accuracy', and 'Incorrect', as it was felt that a single compiler would either have a polarised view or be totally undecided. Various other compilers could then give their views and the majority vote could be taken, so long as the minority does not have a particularly convincing proof. Editors' discretion will always be required.

of interest is not specifically stated as the second wife)[22]. Most researchers would be able to use intuition and experience in the majority of cases to make a judgement on this component without necessarily needing to be fully familiar with the detail of the evidence (though it is always recommended). It is in the third 'notes' component that the real original research should be detailed.

Figure 2 below shows how 'events' within the sub-classes interlink with each other and with the person and source/reference data. As can be seen the three life event sub-classes have been combined into one table, as have the two name sub-classes. Nevertheless at first glance it appears that an unnecessary extra table ('Partners[23]') has been created where the data could have been combined with the data in Partnerships[24]. The reasons for this are threefold:

- It saves data redundancy if one of the partners in Partnerships is unknown (i.e. if Partners were not used there would often be an empty 'Person' and 'Order' field in Partnerships)
- It enables much simpler searching and querying as all like data is in the same field (e.g. 'Father ID' and 'Mother ID' are combined in 'Person ID' in Partners)
- The number of fields that can referenced is reduced by two, as above

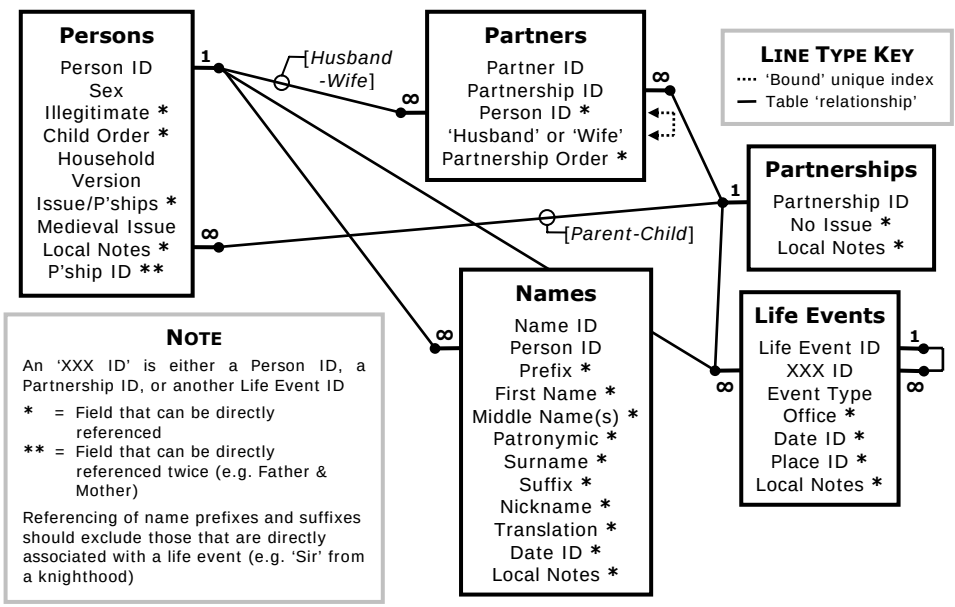


Fig 2. Data tables and fields to support each sub-class

---

[22] Experience of the researcher is the only way that this condition can be used as it could never be a guaranteed solution to any logical construct, though if used well it clearly has a great many applications.

[23] For clarity all further references to tables will be double underlined.

[24] The alternative is to create four extra fields in Partnerships: 'Father ID', 'Mother ID', 'Father Partnership Order', and 'Mother Partnership Order'.

The tables in Fig 2, plus a 'Source' table, a 'Reference' table, and a number of 'Reference-Data link' tables equal to the number of tables in Fig 2, are all that is required to create the most basic form of genealogical database[25] that meets the needs of the highest standards of scholarship. There are of course other routes to the same goal, but the above method has been developed as that most suitable for the proposed public medieval genealogical database. The referencing of data, as described above, can now be mapped out as in Fig 3:

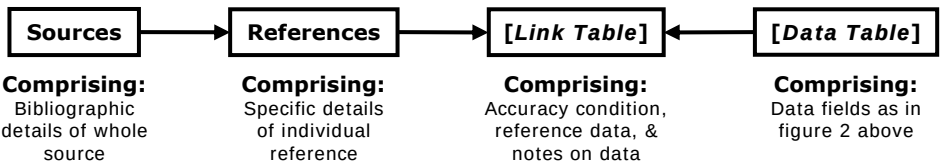


Fig 3. Referencing of data

Practical Examples

The above description of data structure was by necessity a little dry, and to follow it will require at least a basic understanding of the principles of systems analysis for relational databases. In order to clarify the flow of logic, the following examples may be useful:

Life Event

Let us take an example from three sources describing a 'birth' event. The detail from the sources is as follows[26]:

- Reference A – "Person X was born 1234, in London"
- Reference B – "Person X was born 14 Jan 1234, in England"
- Reference C – "Person X was born 16 Jan, 18 Hen III, in France"

If the true answer is that person X was born 16 Jan 1234, in London, then clearly before we try to analyse the references we must break the data into its fields, in this case 'date' and 'place'. The division here is obvious, but for more unusual life events a degree of sensible judgement will be required, the over-riding principle of all databases being: whatever you choose to do, always do it that way (which in turn means think carefully about data entry formats, etc. before building the database).

Taking first 'date' it can be said that references A and C are correct, and B is incorrect. To clarify this: reference A gives incomplete data but is still correct; reference B has sub-

---

[25] The Medieval Genealogical Registry (http://medgenreg.co.uk/) being developed by the FMG uses the above principles, but currently has 98 tables, so it is clear that the structure suggested in this article should only be a foundation.

[26] For the present we shall ignore the sources that are referenced by secondary sources cited by the database. The problems of citation of a source's source and the additional problems of reference loops (where a 'live' source such as a database cites a source that in turn cites an earlier version of the same database – this is not a problem with 'static' sources, such as books, where the source takes a snapshot in time and even an earlier edition of the book cannot create a reference loop as it can be distinguished as a separate source) will be discussed in a future article.

fields which are correct but the overall result of the field in incorrect[27]; reference C gives the full answer but in a 'cryptic' form[28]. Using the accuracy conditions defined earlier I could have said that reference A was 'inconclusive' but this is not the case, unless the context had given further information to indicate this, as it is just limited in detail. One cannot assume to know the author's intention in putting a year only without the reference going on specifically to discuss the point. If the only data given is that as quoted above then the only conclusion that can be drawn, assuming that the author is intending to convey the most accurate picture possible, is that the data is, to the knowledge of the author, correct and as detailed as possible from the information that they had available. There are of course many sources that only give a summary of data and might for instance only show the year of birth. This should be obvious from the source and the data should be treated accordingly using a common sense approach.

Turning to 'place' references A and B are correct, and C is incorrect. Again reference B does not give all the information but is still correct. In this case it could be said that reference A also gives incomplete data, although the example of 'London' is obvious. If the reference had simply said 'Newtown' then there are dozens of places with this name in Great Britain alone. Unless the context of the reference made it absolutely clear which 'Newtown' was meant, the accuracy condition would have to be 'inconclusive'. If a further reference then said just 'Montgomeryshire' (which would tie the place to a single town), and was judged to be correct, then the notes field of the relevant reference link table[29] should discuss the two references and show that together they provide the whole answer even though each one alone is insufficient.

Relationship

Relationship examples are a little more complicated as the principal behind them is very different. They are not distinct tangible pieces of data, as such, but rather connectors between collections of tangible data (i.e. 'Persons'). They do of course have a certain degree of intrinsic data value, otherwise they would be pointless, but it is far harder to quantify, in the way that I am defining them, than say a birth date. The important thing to remember is that all detail data about any person in the relationship, such as their

---

[27] Sources which go down to the level of the sub-fields defined earlier are sadly very rare. If a source does discuss the component elements of, say, a date then each description of the whole date field can be considered a separate reference within the source, so a conflict of detail will never arise. Alternatively the data can be discussed in the reference-data link note.

[28] To back up this example of a Regnal date a further reference would be required, for instance: 'Cheney, C R (editor), "Handbook of Dates" Cambridge: University Press (1996)', linking the Regnal year to the real year.

[29] As discussed above there are three components of a Reference-Data Links table: transcript, accuracy, and notes. If put as fields in a single table (the simplest form) then each record becomes tied to a specific data-reference combination. A better solution from the users point of view would be retain the first two components in the link table and move the notes to a separate table (Reference-Data Link Notes) that links between the specific data (as before) and a new Compilers table that stores the list of people working on the database. There then should be a tertiary level table (Reference-Data Link Note Refs) that links Reference-Data Link Notes back to individual references, and contains the fields: RDLNR ID, RDLN ID, Reference ID, and Text. This way a Compiler's note can span multiple references, which clearly would be required for the above 'place' example. The alternative with the simpler single table form is to write an individual note for each reference-data combination in question and cross-reference textually, something not to be recommended as the cross-reference must therefore be uncontrolled. Further complexity can be added by allowing multiple Person ID links to be embedded in the text, but that is too complex for the current study.

name, is **not** held in the relationship. It is detailed in a separate name event or life event. A relationship merely serves to link two arbitrary unique identifiers[30].

For my example I will take the following references:

- Reference A – "Person X was son of person Y and person Z[31]"
- Reference B – "Person Y married person Z and had issue"
- Reference C – "Person X was son of person Y"

The above references have been made 'correct' for simplicity, but accuracy conditions would apply as before. First the data must be split into separate relationships and fields. References A and C both contain a **single** 'parent-child' relationship. The fact that reference A has two parents and reference C has only one is irrelevant. That the parents had issue, as stated in reference B, does not constitute a relationship, though it could be used as counter-proof to "no issue" stated by another reference, in the person specific 'Issue/Partnerships' field. References A and B both contain a 'male partner – female partner' relationship, though counter intuitively so does reference C, it is just that the second partner's ID is 'null'. A relationship still exists because a child can only be a child of a partnership, not a person, and a partnership requires a 'male partner – female partner' relationship, even though in this case it only contains one person[32].

When referencing the above it is important to remove all possible personal detail from the transcript component of the reference, such as names. A reference transcript should be of the form "The father was almost certainly [*ID 1*] but possibly [*ID 2*]". In the output of the database these IDs should be replaced by hyperlinked arbitrary names of the persons in question, where the content of such names is itself of no intrinsic data value[33]. By far the most common reference for a relationship can therefore be seen to be "[*ID 1*]", as is shown, for instance, in a simple diagrammatic family tree (unless tentative links are shown).

In the second part of this article I shall discuss the principles behind the operation of the hypothetical public medieval genealogical database and the infrastructure required. This will include possible output formats and how they could be generated in a version-controlled environment. I shall also discuss what is currently being done in this field and the levels of technology available to the current generation of Internet users. The problems of regulating and reviewing the data are very complex and these articles can only offer suggestions for a wider debate. All comments are welcomed.

---

30 A relationship as defined here can only contain two identifiers: partnership and child, or male partner and female partner. There is no actual father-child or mother-child link, although these two theoretical links are referenced separately as the vast majority of sources do not define partnerships. The detail of this mechanism will not be discussed here.

31 Names have been excluded deliberately to avoid confusing name events and relationships.

32 A much-simplified variant of this system is used in the GEDCOM format, where the dataset is broken up into 'Persons' which hold all the life event and name data, and 'Families' where the people are linked together. The 'Family' relationship in GEDCOM effectively combines partners, partnerships, and children in a single block.

33 This 'Short Person Name' field could be an extra field in Persons, and should be as helpful as possible in identifying the person as a distinct figure in history (a birth-death year date range after the name is a sensible addition). As the name is arbitrary it must be uncontrolled.