CFPS 98
(Call for Papers Submission number 98)

Thoughts on Genealogy Standards

Submitted by: Hill, Chris

Created: 2013-12-12

URL: Most recent version: http://fhiso.org/files/cfp/cfps98.pdf
This version: http://fhiso.org/files/cfp/cfps98_vV1.pdf

Description: This document has ideas on standards for genealogy systems, how to build them, and on information held within genealogy systems.

Keywords: Standards, Modelling, Data Formats, Structures

Thoughts on Genealogy Standards

Abstract

I have been considering the problems of transferring information between multiple systems over the last few years, having my tree located on my PC and two online services. Along with that, I get contacted by other users, including distant members of the family, and quite often have to send them information, frequently as written notes.

This has always been problem and the need for an improved way of dealing with this because more of an issue. Alongside that, it is necessary to involve the software and service providers, so that they can develop systems that are of more use to us and will enable them to be able to gain more customers.

In this report, I discuss both the need for standards and some ideas on the information that we hold and need to transfer.

Need for Standards

Why do we need standards?

If everyone was working in totally separate systems, with no need to exchange any data with other users, there would be no need for the transfer of information. Luckily, we are not in the position and we do need to transfer information to and from other users or systems. Clearly in the past, and even now, it is possible to do that using a written process, as we do when reading an old document.

In transferring information there is a need to be able to understand the structure of the data that is being transferred, since otherwise we do not know what it is other than being some text, and we need a format into which we can structure the data so that we can identify individual fields by both its identity and its value. This, in itself, is only half of the problem since all transfers relate to a flow of information between two different systems. There is therefore a need for both partners to be able to understand both the format in which the data is transferred and the way that data is structured. This is, of course, the basis for any transfer of information, whether it be by writing or electronically, and is the way that all computing systems must work.

However, delivering a format for the transfer of information between different systems does not create a standardized format, since it can be privately owned. The need for a standardized format becomes necessary when there is a requirement for information to be transferred between multiple, potentially different, systems and users. This is to ensure that all users can understand and interpret data that is transferred, ideally with no loss of information.

What must the standards cover

The biggest problem in exchanging information between users and systems is caused by incompatible interfaces. This is because the interfaces are either through the current GEDCOM 'standard' or through an attempt to read another semi-standard format. In addition, there are differences between systems regarding the information that they hold and can transfer, and how they process that information.

As end-users we need to be able to transfer information, with no loss of content, between multiple systems. We need standards that will allow conversions to permit the following.

1. An extracted file that is later imported back into the same system will result in an identical file.
2. An extracted file that is later imported into the different system will result in a file with the same content, but possibly restructured to fit the new system.
3. An extracted file that is later imported into the different system, and then later extracted and reimported back into the original system will result in a file with the same content. Ideally it will be identical to the original file, but it is possible that it may have been restructured though the intervening system.

4. Local, i.e. non-online, systems may have references to external source repositories. Where these are online the system must be able to locate, view and create suitable source information. This information must be capable of transferred between different systems.
5. Where online based systems have sources to links in repositories it must be possible for these links to be transferred to other systems.
6. It must be possible to transfer media files between different systems, with integration into the format process.
7. Modifications made to a transferred file will be capable of being returned to the original system.
8. When merging multiple files, such as the return of a file previously sent, the merge process must be capable of allowing mismatched fields to create multiple sets of information, with source information.
9. It should be possible for non-standard records to be received and held as text, and to be resent out later.

What standards do we need

Clearly, we are looking at both the structure of the information that is being transferred and the format within it which it is transferred. One can argue that the structure of the information, being dependent on both the sending and receiving systems, is not, and must not be, within our responsibilities. Equally, for us to be able to define the formats for transferring that information it is necessary for us to be able to define a structure for that information, into which different systems can map their own internal structures.

We therefore need to be able to define both a model for the structure to be transferred and a format under which this can be done.

Data Standard

Here we need to able define the basic entities in the structure – people, families, media, events and sources – with individual fields relating to them - Facts. We need to be able to relate entities to each other – parent/child, event/person(s) – and the role between the entities. We also need to allow for a Fact to be changeable as a replacement or as an alternative, and perhaps at a specific time or event – thus a change from unmarried to married after a marriage. We must always be aware that some fields may be missing or that the content may be only partially known, such as a date to the nearest year.

We also need to be aware of the differences between a genealogy system, which is purely tracing the history of ones ancestors, and a family history system tracing how our ancestors lived and worked. We therefore need to deal with relating events to entities, changes of state, adopted children etc.

Format Standard

The format that we structure our information must be both flexible and capable of handling the structure of the data model.

In defining the format structure we need to recognise that under an entity we may have many sub-structures and/or multiple fields. These can be structured to show the type of sub-structure and its depth. Each sub-structure can be regarded as a record, of different types.

When creating the data format, we have the ability to specify the existence that each record or field must have. We can categorize these as being one of these:

1. Mandatory – this record or field MUST be present. If it is not present, then the record must not exist.
2. Required – this record or field SHOULD be present, but it may be omitted. Should we have this option instead of making the record or field Mandatory?
3. Optional – this record or field MAY be present. This will allow us to deal with missing information. For a record it implies that all lower level records and fields are missing.
4. Non-standard – this record (or field?) is not within our standard format but is generated by the originating system. This has been one of the problems within GEDCOM with the _XXXX record type. Ideally, we should not allow these, but we may need to do it if the sender regards it as being needed. I think that we can deal with it by

using to create a non-standard entity or field within the data model, with the ability to resend it out on later transfers with the same identity.

In defining the way that we format the data for each field, dependent of the format chosen, we need to be able to deal with incomplete information, and with the links between different entities.

To a degree, the transmission format is separate from the data format, since a set of data can be formatted in multiple different ways. Examples of these include:

1. A GEDCOM style, structured by a depth value and record id followed by fields. Reasonably easy to read.
2. A structured format, as used by UN/EDIFACT (ISO9735) for business data transfers. This would be more compressed and not easily human readable.
3. A textual style, as shown by XML or JSON in the GEDCOM X style. Easy to read but would use more data for transfer.

We therefore have the ability to define a format for the data that we need to transfer, identified as records, fields and structure, and a separate specification on its external format for transfer.

Who should be involved?

Standards, of any kind, cannot be created independently. Much as we, as users working with genealogical structure, would like to define a standard for formatting data and then hope that the many systems and services would start to use them, we know that will never happen. When setting standards it is essential to include people from all of the businesses and organizations working in this market. [Per se, I was involved, to a degree, for 15 years during the standardization of business transfers within the European Automotive business, working with vehicle manufacturers, suppliers and service providers.]

In order to generate standards that will be used, we must include members from all of the interested parties:

1. Users – ultimately, as we need to transfer information between ourselves and to/from online systems, we are the driver for these standards.
2. Software developers – most of us will be using one of the many software solutions. While we do not want to impact their solutions, we do want them to support any new standards and to provide experience regarding data and format standards.
3. Service providers – we will be using their services to hold online versions of our family trees and to provide us with access to many online databases copied from original paper-based records. Again, we want them to support any new standards for transferring our data and to provide us with experience in this process. We can also look at the possibilities of extracting data, in a standard format, from their databases.
4. Family History groups – Family History and Genealogy groups can provide experience in the way that these systems can be structured and transferred.

This way, we can ensure that all interested parties will be able to help on the design of the standards and will be involved in ensuring that both the standards and their systems are compatible.

We must be aware that we are not wanting to impact on the systems used by the software and service suppliers. Indeed, by involving them in the settings of new standards we are providing them with the means of making it easier to transfer information. This will provide reasons for users to move to their system, since it will be easier for them to transfer information.

Entities

When looking at the data model and the top-level records in the format structure we are looking at the basic entities within our model. Clearly these will include people, families, events, media and sources. Of these it is clear that a person or a family, excluding the question of what one regards as a family, is a single entity. Other entities, such as an event or a source, can be common to multiple, different entities. We must apply the process of singularity to these and only record them once, referencing them by reference from other entities.

Individuals

We should also allow differences in the way that family names are created, such as the Icelandic way of creating a last name from the father's (or mother's) first name plus –son or –dóttir, which is also beginning to be used in Scandinavia. Other counties do not necessary format names the same way that we do, generally in countries which have retained their historical naming and have not been converted to a European-styled naming structure. Other countries, such as Irish, may support both native and European systems with both an Irish name and an English name.

Families

Traditionally, I suspect, Family History and Genealogy systems tend to be European and Christian based. That tends to imply the structure of a Family to be a pairing of a male and female pair, with the addition of children. In our modern society, we now have to deal with single-sex partnerships, more adopted children and the ability for a person's sex to be changed.

Facts and Events

Within an entity there will be data fields holding the information that we are collecting – these are Facts. Some of these will be static, such as a proper name, while other may be changeable. In this context I am not referring to manual changes which have a permanent effect; I am referring to changes which effectively modify the status of a field during a period of time, since the modified entity, such as a person, has an existence within which it has lived. Typically, these could apply to status flags, such as

```
unmarried →married →divorced/widowed
           |                |
           |------ ← -------|
```

These changes are the results of an Event, and are constrained to occur at certain times. Other examples could include an alternative name, or nick-name, or a military rank.

An Event records some information that has occurred to an Entity. That may be a simple record, such as a census record, or it might result into a change to the value of a Fact, such as a marriage. An Event can be applicable to multiple Entities and should therefore only be held once.

The relationship between an Event and an Entity becomes a Role which indicates how the Event relates to the Entity. An Event may have different Roles against Entities it is related to. Against each Entity one should be able to record specific information relating to the Event. For example, a census record could list Occupant for the family members but Boarder for non-members of the family, and could have different information related to the Entity.

Multiple values

We need to understand that some Facts may multiple values, and ensure that the model includes support for associating multiple values at different times to a Fact. In some cases this will be because, during the existence in time of the Entity, the value has changed. It could also be because there are different versions, with different values, of the same Fact at the same time. This may be because the information has been collected independently and the users have produced different information on it. The data model has to allow for this and it must be possible for sources to be associated with each version.

Places and Locations

Since the same location (place) can be used in many different entities it would be sensible to make each location into an entity and to refer to it where it is needed. This will reduce the size of the data and format models, and ensure that the content will be consistent everywhere it is used.

We also need to be aware that a location may move over time. My ancestors, 150 years ago, lived in villages in the county of Middlesex; now these are locations within London, which grow from 1890 to 1965 by taking over most of Middlesex, and parts of Surrey, Kent and Essex. We therefore need for our model to be able to specify, against a static location (Latitude and Longitude), the different addresses that was applicable at different times. Similarly, we may need to hold administrative, religious or legal addresses against a single location, since these can be different at the same time.

Dates

When looking at Dates it is important to recognize the difference between the Value of a date and the way into which it is shown or interpreted.

ISO 8601 fixes a reference calendar date to the Gregorian calendar of 20th May 1875, which was the date the Convention du Mètre (Metre Convention) was signed in Paris. However, calendar dates before that date are still compatible with the Gregorian calendar all the way back to the official introduction of the Gregorian calendar on 15th Oct 1582, depending on when the country changed from Julian dates. In addition, the Gregorian calendar can be extended backwards until 1st Jan 1, and could also be extended further back into BCE dates. These values, as a number of days from 1st Jan 1 (= 1), are the true value for a date, and this how they should be held in the data model. As a note, working backwards past day 1 raising questions of whether there was a day 0 or a year 0, but we are unlikely to need to worry about that.

The ISO specification allows for the date to be only partially defined. Thus we can use a full YYYMMDD date, or a month date as YYYYMM(00), or a year date as YYYY(0000). Holding the date in this format, with zeroes for missing days and months, will enable them to be correctly sorted. The use of the durations is allowed within the ISO standard and can be used to create a UK Quarter Date in the format YYYMM00P3M, with MM = 1, 4, 7 or 10. The need for durations is needed in our systems since records are not always accurate and we have to allow for a date range.

We should also be aware that the Gregorian system is based on the UTC, or GMT, date and time format, and therefore should have an adjustment to allow for the time zone it which it was recorded. If I write this at 6am in the morning on 10th Dec it will be 23pm on 9th Dec in Arizona. That is probably not an issue that we need to worry about other than to realize that the date of an event will be on the local date and not necessary the same as the UTC date.

When we look at how a date is interpreted we are looking at different formats which can be converted to or from a standard date. Thus we can take a date value and ask for it to be formatted as Gregorian, Julian or any other date formatting that can be defined. Equally, when processing a date we need to know the identity of the date format under which it was created.

Thus today, 10th Dec 2013, can be represented as

• Julian Date	27th Nov 2013
• Hebrew Date	7th Teveth 5774
• Islamic Date	6th Safar 1435
• Indian Civil Date	19th Agrahayana 1935
• French Revolutionary Date	Décadi II Frimaire 222

Regarding Roman structured dates, regarding dates as being ante diem the following Kalends, Nones or Ides of a month can be sensible within the Julian Date system, from 45BCE and allowing for adjustments made around 1CE. Again, looking at Roman dates pre 45BCE is more complex and not an area that we should need to worry about.

For Regnal dates, one needs to know the King, or Queen, and their date of ascension in order to convert a date. To make it more complex, the settings can be different for some of them, such as John whose years are based on the value of Ascension Day rather than the date of ascension.

For legal, civic and ecclesiastical dates this becomes more complex as the start of year was changed, with reference to the standardized calendar date. Until the end of the 13th century the New Year began at Christmas, 25th Dec, and therefore all dates between 25th Dec and the following 24th Dec would be regarded as the calendar year at 25th Dec. From the 14th century until the conversion to Gregorian in 1752 the New Year began on Lady Day on 25th March, and then moved to 1st Jan 1753. This implies that a written date of 1st June 1250 will be referring to a calendar date of 1st June 1251, and a written date of 1st Feb 1750 will be referring to a calendar date of 1st Feb 1751. [Interestingly, the UK financial year is still based on Lady Day converted from Julian 25th March to Gregorian 5th April.]

Presumably, there are equivalent systems for other countries, should we have to deal with them? We could regard any system of that type as a pure textual value, and convert it, as best as we can, into a Gregorian value.

Potentially, we can have a Date held as a Gregorian date, together with a textual equivalent and a default format for viewing it. On exporting it we have the ability to specify both the Gregorian date, and a format and conversion of it to be viewed in a different calendar. On import, if the format is specified, we can use that to convert it back to Gregorian, or we can ask the user to supply the format it is in.

Sources

It is all too easy, as I know from my own family tree, for us to forget to record where we got our information from, and the quality or certainness of it. In some cases this will be from verbal memories or it can be from published information. With the modern web-based world it is most likely that the sources will be links back to an online resource repository. We therefore need the data model to be able to hold these various source definitions in a way that can be transferred and which can be accessed while working on a system, either online or on a local system. We should also be aware that a source may disappear, as a site is modified or removed, and therefore be able to make local copies of the information.

All Sources will relate to one or more Entities in the system. That indicates that a Source must be an Entity in itself, and must not be regarded as part of an Entity. We should be able to record additional information against the Source, as opposed to the Entity it is referring to.

Many Sources will come from the same repository, which will become an Entity in itself as a Master Source. The Repository in itself will also be an Entity, against which we can also add information.

Media

The system must include support for Media records. These should be held separately with links to them held within the system database. When transferring information, it must be able to extract and include Media files as part of the transfer process, and for the import of information with attached media to extract the media either into a standardized location or to ask the user where to put them, potentially for each file. Given the complexities of creating a file structure automatically, it would seem simplest for them to be located into a single location. We will then need to be able to select one or more media files, relating to multiple Entities, and relocate them into a structured file system, with the embedded links to them being updated automatically.