fhiso

CFPS 21
(Call for Papers Submission number 21)

Proposal for Handling Personal Names

Submitted by: Proctor, Tony

Type: Technical proposal

Created: 2013-03-19

Last updated: 2013-04-20

URL: Most recent version: http://fhiso.org/files/cfp/cfps21.pdf
This version: http://fhiso.org/files/cfp/cfps21_v1-1.pdf

Description: Proposal for Handling Personal Names

Keywords: Personal-names, Culture-neutral

Contents

1. Abstract

This proposal makes a case for handling personal names in a generic way that accommodates not only worldwide variations but also unstructured names.

Each Person entity must be allowed to have alternative names. These alternatives may be time-dependent, and of different types (e.g. formal, informal), and in different languages.

2. Proposal

2.1 Name Structure

When we think of personal names, we immediately think of given-names and surnames, often using the more informal and less-correct terms forename and last-name. In almost all places where names are stored - in genealogy and in other areas such as government databases - there is an attempt to formalise names and categorise their elements. Hence, the parochial Western knowledge of worldwide names often gets enshrined in data storage and it can adversely affect people from other cultures.

So, for Western type names, we obviously need to account for possible multiple middle names. Then we find there may be academic titles (e.g. Dr. or Prof.), honorific prefixes (e.g. the honourable, or his holiness), honorific titles (e.g. Sir, Lord, Dame, Lady), or post-nominal letters (e.g. VC, OBE, PhD), generational titles (e.g. .Jr, Sr, I, II, III, etc).

As we move out of the English-speaking world, we find cultures with multiple surnames. The surname category itself has to include patronymic or matronymic names which are a different type of inherited name element. In Far Eastern cultures, there is a generational-name concept that we don't have in the West. There is also a general class of name element called a 'name particle', analogous to a grammatical particle. This includes all those small joining words such as: “von”, “van”, “der”, “de [la]”, “d’”, “the”, “[son] of”, “mc”, “mac", "Ó", "Ní", "Nic", "Mhic", "Bean", "Ui", "y", etc.

Even if the storage supports extended-Latin or non-Latin alphabets, we find different rules for capitalisation (sometimes it is not the first letter of an element, and sometimes it is more than one letter), and sorting (e.g. sorting on the first, last, or other name element).

The Native American cultures effectively have unstructured names that defy all attempts at formalisation, so this is a problem even within the US. A name like *Running Deer* has no surname concept. If our representation of names is to include pseudonyms, stage names, and other alternatives - any of which could be a simple mononym - then we have to abandon any formal categorisation of the tokens.

The suggestion in this proposal is therefore to treat a name as a simple sequence of words, or tokens. Any attempt at formalisation of the tokens will fail somewhere so it is best to avoid it.

2.2 Time Dependency

In the West, name changes are mostly associated with a woman changing her name through marriage, although some men taken on this convention too. The next most common type of change is probably through adoption or deed poll. In all these cases, the new name starts at a much later time than the birth of the person.

For professional names then they may run in parallel with their primary names. Hence, the alternatives may be overlapping in the general case.

A Native American might have different names at different periods of their lives, e.g. an infant name like "little rabbit", later changing to a war name when a boy becomes man, and changing again for the later periods of their life.

In summary, each alternative needs a starting date (defaulting to birth) and an ending date (defaulting to death).

2.3 Name Types

Each person will have a number of accepted variations in their name. This may include informal versions, name changes, possible spellings in other languages, Romanised versions, and professional names. They are distinct from unaccepted variations such as typographical errors, transcription errors, and misheard names. The latter group form evidential properties yielded by a particular source and so must be recorded verbatim where the source is referenced.

A possible list of types for accepted name alternatives:

- Alias – General pseudonym, including also-known-as, nom de plume, pen name, and nom de guerre. Some cases may have a specific *type* available for them.
- Married – Name adopted after a marriage ceremony, or other type of union.
- Nickname – Informal alias.
- Personal (default) – Normal personal name.
- Pet name – Hypocorism. A term of endearment used in more intimate circumstances.
- Private – For cases where a personal name is only used within certain circles, as with some Native American tribes.
- Professional – Includes stage name.
- Public – Some Native American tribes distinguish a private name, used within their own tribe, from a public name used outside of it.

2.4 Name Matching

A person will typically have more variations of their name that are accepted as input (both in real life and for digital searching), but fewer standardised forms for presentation. STEMMA handles this by representing the acceptable alternatives as token sequences (i.e. a list of name parts for each alternative), and having a *canonical name* concept for the presentation versions. The canonical name may have variants according to the style of presentation, such as formal, semi-formal, informal, and listing. The latter is for alphabetic listings and is mentioned below.

Each alternative is matched in sequence against a particular name being sought. Note that this is not as rigorous as parsing a computer language since we're not dealing with sentences and clauses. Hence, solutions using tools such as YACC and BNF are inappropriate.

During name matching, it is recommended that the date ranges are ignored in order to provide a more relaxed operation. A person's name doesn't cease to be valid after their death anyway so that is an intuitive strategy. However, in order to derive a Person's full formal name then they should be honoured and in the order they are written, just in case there's any overlap due to fuzzy dates.

Character matching should be relaxed here. This means treating each pair of tokens in a case-blind and accent-blind fashion, and handling the composed/decomposed equivalences as recommended by the Unicode recommendations: http://www.unicode.org/reports/tr15/.

2.5 Sorting and Collation

Sorting and indexing personal names is a minefield. Assuming we primarily want to sort by surname, a name may have more than one surname or none at all. The surname may be at the beginning, or the end, or part-way through the name.

If the name includes name particles then the rules are culturally dependent, and often down to personal references (of the recorded individual, not of the end-user). Some particles are part of the surname and some are not. Some are subject to sorting and others are effectively transparent.

STEMMA handles this by having a style of “Listing’ that can be applied to a canonical name, e.g. “Proctor, Tony”. This is only a partial solution as it does not yet solve the issue of a surname with a transparent name particle. An example is ‘Willem de Kooning’ whose surname is ‘de Kooning’ but which would be sorted under ‘K’.

Another option might be a ‘SortAs=’ attribute. The essential problem, though, is that no amount of pre-programmed rules will correctly deal with all cases.

3. Not Covered or Not Required

The issue of highlighting a portion of the name – typically the surname or family name – is not covered here. This is primarily because STEMMA V1.0 was still experimenting with various schemes. The STEMMA discussion at http://www.familyhistorydata.parallaxview.co/research-notes/worldwide-fh-data (section 7.1, Capitalisation) explains that this should be done through some type of mark-up and should not constitute a material change to the stored name (e.g. all uppercased or enclosing solidus).

The issue of which name to use for presentation is not directly covered here. Rather than specify a primary name, STEMMA allows each Person entity to have a title property associated with it. This allows the identification of unnamed children, indications of unknown names, and differentiation of persons with similar names (e.g. “John Smith (1908)”). This avoids having to overload the true personal-name data with text that is not a real name.

Some type of mark-up for identifying uncertain transcriptions (e.g. characters that may have more than one interpretation) is covered in a separate proposal.

STEMMA syntax is used for the illustration but is not mandated by this proposal.

4. Illustration

The following STEMMA example involves someone called Grace Ann Murphy who doesn't always use her middle name and sometimes goes as Gracie. However, she's Irish and also has an Irish version of her name. Using an informal notation, this would require the following two token sequences:

{Grace,Gracie} [Ann] Murphy
Gráinne [Ann] Ní Murchú

This would be stored as:

```
<Names>
    <Sequences>
        <Canonical>Grace Ann Murphy</Canonical>
        <Sequence>
            <Tokens>
                <Token>Grace</Token>
                <Token>Gracie</Token>
            </Tokens>
            <Tokens Optional='1'>
                <Token>Ann</Token>
            </Tokens>
            <Tokens>
                <Token>Murphy</Token>
            </Tokens>
        </Sequence>
```

```
            <Sequence Language='gle'>
                <Tokens>
                    <Token>Gráinne</Token>
                </Tokens>
                <Tokens Optional='1'>
                    <Token>Ann</Token>
                </Tokens>
                <Tokens>
                    <Token>Ní</Token>
                </Tokens>
                <Tokens>
                    <Token>Murchú</Token>
                </Tokens>
            </Sequence>
        </Sequences>
</Names>
```

5. Use Cases

The Data Model must be applicable to all names, irrespective of their structure, cultural origin, or type. If this means sacrificing the ability to categorise every single token then so be it – it makes no sense to have a system that works fine for English names and has to be bypassed to accommodate all other names.

The Data Model must also accept that names are not fixed (i.e. they may change for various reasons), and not unique (i.e. there may be multiple concurrent alternatives).

6. References

Useful documentation resources on personal names:

- W3C internationalisation guide. Discusses personal names around the world: ga-personal-names.
- Citation Style Language (CSL) 1.0, deals with sorting of names involving name particles: citationstyles.
- Wikipedia. Personal names: Personal_names.
- ROCIC Law Enforcement Guide to International Names: law-enforcement-guide-to-international-names
- IFLA Universal Bibliographic Control and International MARC Program. National Usages for Entry in Catalogues: NamesOfPersons_1996.
- Wikipedia Manual of Style.
- Name Particles. http://www.grammarphobia.com/blog/2010/07/nobiliary-particle.html.

Useful documentation on Native American names:

- personal-names-among-the-indian-nations-east-of-the-mississippi

- dissertation_lombard_c
- Family Education - Baby Names

STEMMA discussion of personal names.
http://www.familyhistorydata.parallaxview.co/research-notes/persons-places (Section 4 - personal names) and http://www.familyhistorydata.parallaxview.co/research-notes/worldwide-fh-data (sections 6 & 7).