fhiso

CFPS 75
(Call for Papers Submission number 75)

Functional requirement for recording searches

Submitted by: Smith, Richard

Created: 2013-05-22

URL: Most recent version: http://fhiso.org/files/cfp/cfps75.pdf
This version: http://fhiso.org/files/cfp/cfps75_v1-0.pdf

Description: This paper discusses the requirement for recording searches, whether proposed, successful or unsuccessful, in addition to the results of the search.

Keywords: search, inference, negative inference

Abstract

This paper discusses the requirement for recording searches, whether proposed, successful or unsuccessful, in addition to the results of the search.

1 Introduction

Knowing what data was found is only part of the story of genealogical research: it is also important to know what was searched, and what *wasn't* found. It is one thing to know of a baptism in a particular parish; but it can be much more powerful to know if this was the *only* surviving baptism record for an individual of that name in the whole county. Many genealogical data models, and certainly the present GEDCOM 5.5 standard [1], fail to provide a means for recording this information. The GENTECH data model is one of the few that does: it provides `SEARCH`, `ACTIVITY` and `RESEARCH-OBJECTIVE` objects for that purpose [2].

2 Requirements

The FHISO's data model should include some form of search object. It should be possible to record searches that are proposed, that are underway, and those that are complete. An application may wish to use the list of proposed searches as a form of 'to do' list to help the researcher plan future archive visits effectively.

It should be possible to record when, where and by whom the search was made, including when a search extended over several archive visits, where multiple archives were involved, or where several researches participated. This might perhaps be achieved using a sub-search restricted to the activity of one researcher on one day at a single location. An application targetting the requirements of professional genealogists might use this as the way of recording chargeable work.

It should be possible to associate records with one or more search, allowing for the possibility of more than one search finding the same record, as might happen when searching for all marriages for a several given surnames; and it should be possible for a completed search to have no associated records if it was unsuccessful. The term record is used here loosely to mean any genealogical information including images and transcriptions of records, as well as machine-readable genealogical data extracted from them.

Researchers should be able to ignore the mechanism for recording searches if they so wish. At the opposite extreme, applications may choose to provide a option whereby users can require that results only be entered as the result of a search.

3 Search specification

This paper does not specify how the detail of the search should be recorded. At one level, a text field written by the researcher might be sufficient. But the FHISO may wish to consider whether a machine-readable specification of the search is feasible. In particular, an application that processes negative inference rules, as discussed in CFPS 5 [5], can only do so with a machine-readable way of determining the believed completeness of its data, for example by the scope of a search. Without knowing the scope of the search, an application cannot systematically distinguish between absence of evidence and evidence of absence.

If an application can infer how complete its data is believed to be, it can determine how complete its answers to other queries are. For example, a family history society might sell a database of all baptisms in a county. If this is is loaded into an application and the application has a machine-readable way of determining that this dataset is the result of a complete search of surviving baptism records for the whole county, then the application can determine which queries it can answer exhaustively and which are being answered with incomplete data. This becomes particularly powerful in conjunction with inference rules. With a suitably powerful inference engine, a user might instruct the application to look of baptisms for individuals in a family where there is a single suitable baptism within, say, 5 miles of the expected place and no more within 25 miles.

Depending on the form of the data model, it may be feasible to use a standard query language such as XQuery [3] or, perhaps better, SPARQL [4] to specify what is being searched for in terms of the output that a positive search would yield.

References

[1] Church of Jesus Christ of Latter-day Saints, 1996, *The GEDCOM Standard (Release 5.5)*,
`https://devnet.familysearch.org/docs/gedcom/gedcom55.pdf`

[2] GenTech Lexicon Working Group, 2000, *A Comprehensive Data Model for Genealogical Research and Analysis*,
`http://www.ngsgenealogy.org/cs/GenTech_Projects`

[3] World Wide Web Consortium, 2010, *XQuery 1.0: An XML Query Language*,
`http://www.w3.org/TR/xquery/`

[4] World Wide Web Consortium, 2013, *SPARQL 1.1 Overview*,
`http://www.w3.org/TR/sparql11-overview/`

[5] Luther Tychonievich, 2013, *Inference Rules* (CFPS 5),
`http://fhiso.org/files/cfp/cfps5.pdf`