21 May 2011 G.Thorud

From Repository Source Meta Data to BetterGEDCOM and reports – A future scenario

This is an attempt to describe a possible future, inspired by input from GeneJ, Mark Tucker and Tamura Jones.

Genealogists will exchange genealogy data from many types of sources, and also data about them that produce source citations (= e.g. footnote or bibliography). The sources will relate to many countries and will be stored in archives, libraries, and public offices etc, possibly made available on the Internet. There will be many “style guides” telling users how to structure citations, and which data to include in them. Programs supporting BetterGEDCOM will have to support all these situations unless they are limited to a particular user community.

Data about sources and citation (S/C Meta Data) have traditionally been typed into the genealogy program by users, but there are already some Internet services that allow the users to automatically download these data into programs (primarily for other sciences, e.g. the programs that can be downloaded from www.zotero.org accessing e.g. amazon.com). If such solutions could be implemented in genealogy programs or supporting programs, and if databases containing S/C Meta Data for genealogy sources could make the data available for download via the Internet in a structured form (and optionally also the information in the source), it would be much easier to create citations and the quality of genealogy data and documents would in many cases be improved.

Figure 1 describes a situation where a user accesses a database via the Internet, and possibly downloads info from sources by using a browser. A browser extension or supplementary program also allows downloading of source meta data and stores it in a file. A genealogy program can pick up these data and convert them into its internal database containing C/S Meta Data, and it could also store info about the downloaded data in the source (e.g. the local file name). (The genealogy program could also pull these data directly from the database). The meta data can then be printed in reports or charts and be exported to Gedcom or BetterGEDCOM files. See Mark Tuckers video here http://www.thinkgenealogy.com/better-online-citations/ for a demo. Another source of source meta data is incoming Gedcom/BG files.

The big problem is all the various ways to structure source meta data, what semantic info they contain and the various styles used to present/write citations. If a program is to operate in such a heterogeneous environment, it must be possible to convert data between the various data structures, or even better, to convert the data to and from one independent general data structure. This general data structure could be used internally in genealogy programs and in BetterGEDCOM. The pieces of info in the structure are (in BG) called “citation elements” (e.g. author and title). In contrast to the solution in the book Evidence Explained where an extremely large set of element types have specific names tailored to each specific source type, they should be general so they can apply to many source types, and thus we need fewer element types. (But an instance of a general element could in addition to its type identifier have a label (a qualifier) carrying a specific meaning in the context of the source type, this label could appear in user interfaces (e.g. as a prompt) and could thus carry the element type’s name according to Evidence Explained.) An issue would also be if it is possible to define general source types in order to limit the number of types, making it simpler for users to select a source type.

The presentation (rendering format) of the citation elements for a particular source type in a citation in e.g. a report depends on the source type. A source type would define the type of content in the source, medium, its jurisdiction and other characteristics). Whether the use of rules is realistic, and whether there is a need, will depend on the formats of S/C Meta Data in the various repository databases. It is envisaged that the presentation in in foot/end/inline-notes and bibliographies can be described by a source type specific “citation template”, one for each source type, describing which elements to present, in what order, punctuation etc.

Another type of “templates” could also be envisaged. One use of these could be to describe how the general citation elements for a source type shall be carried in (or converted into) the few source/citation TAGs in current Gedcom. I will call these “conversion rules”. One could also envisage conversion rules for conversion of the specific citation elements defined by a citation style into the general citation elements, and you could also “convert/map” source types.

But let us go back to the flow of data in figure 1. The various functions have been further detailed in Figure 2, together with the parameters (citation templates, source type definitions, conversion rules) controlling how the functions should operate on meta data for a particular source type. The figure also shows where standard specifications would be required to describe how descriptions of templates, source types and rules would be encoded. I have included a complete description of Figure 2 in Appendix 1.

This solution may seem complex. An initial implementation of BG would most likely choose to implement only some of the features, the most important and fundamental one being the use of general citation elements (possibly with qualifiers), and probably general source types. Other features, e.g. citation templates could be hardwired into the program, although user defined citation templates, and source types, would be very desirable if the program is to be used in many countries. But, when developing this initial functionality it would be wise to try to see how they would fit into a larger long term scenario, especially because it will affect the definition of general citation elements and source types. The purpose of this document has been to describe one such possible scenario.

Further reading:

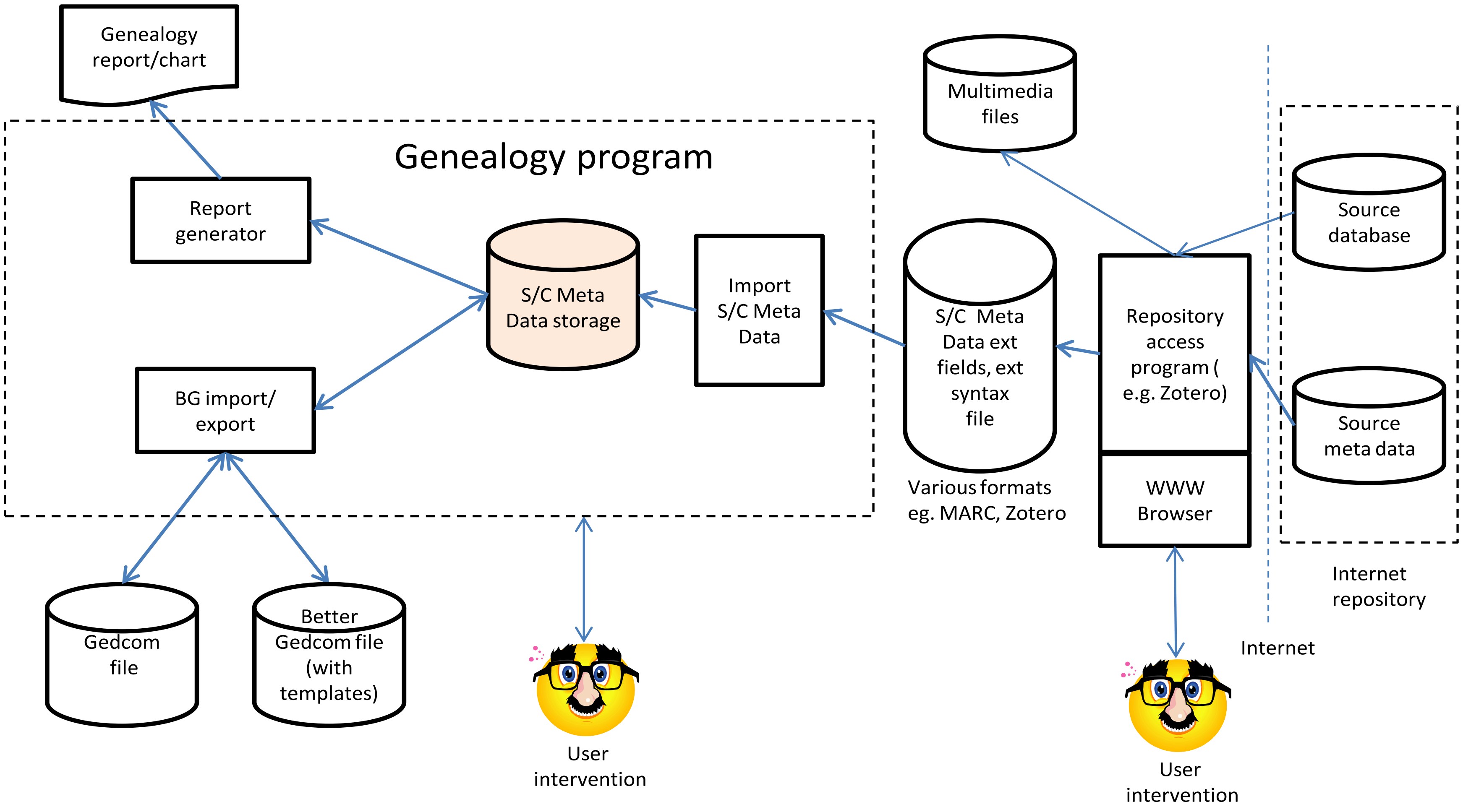
Better Gedcom EE and GPS support

Citations in BetterGEDCOM – Some high level considerations

Summary of the Citation Style Language 1.0 specification

Another document is in preparation suggesting technical solutions for some of the data described in this document.

Figure 1. From repository meta data to BetterGEDCOM and reports

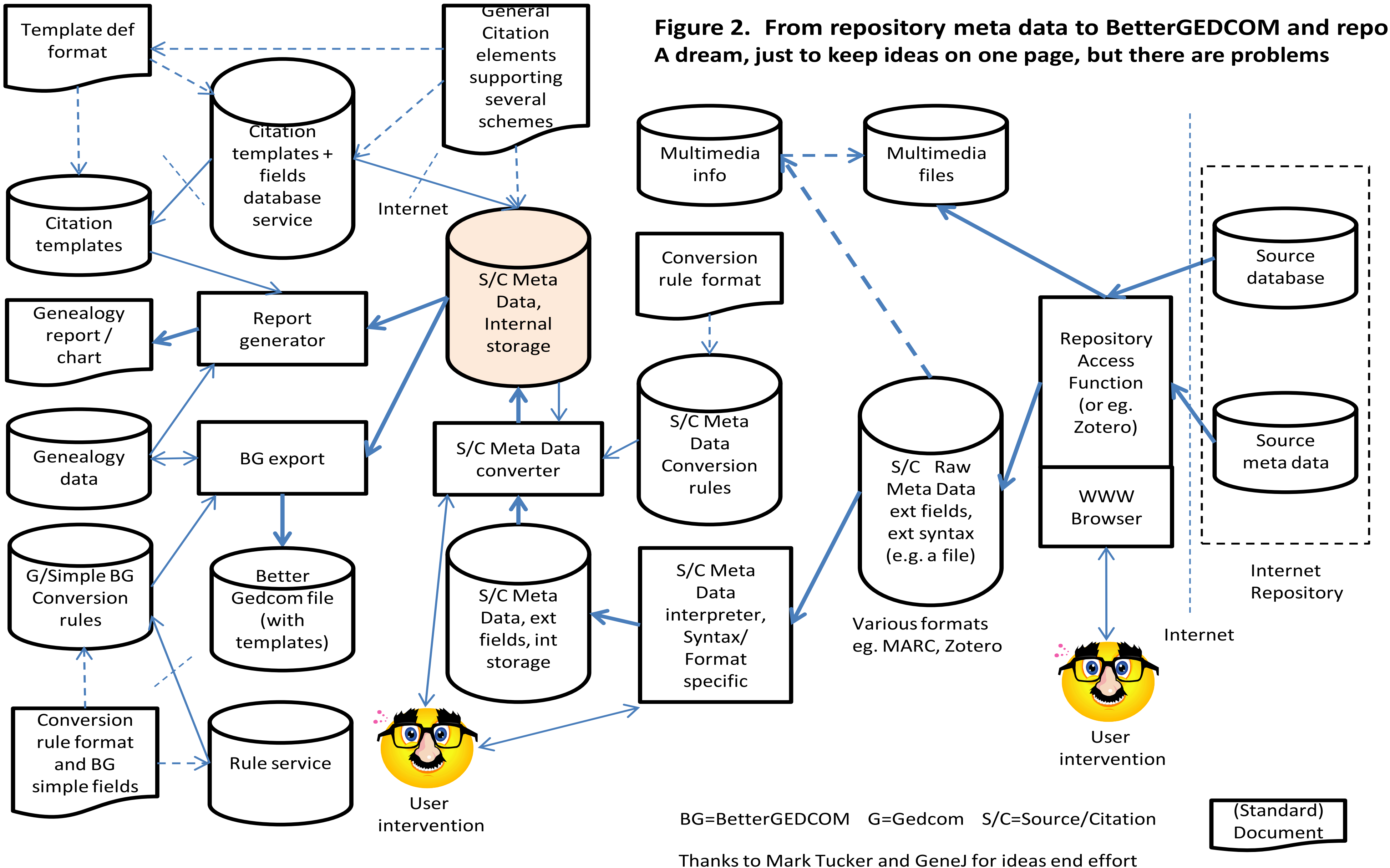


BG=BetterGEDCOM G=Gedcom S/C=Source/Citation

Thanks to Mark Tucker and GeneJ for ideas end effort

10.05.11 G.Thorud

Figure 2. From repository meta data to BetterGEDCOM and reports
A dream, just to keep ideas on one page, but there are problems


Template def format
General Citation elements supporting several schemes
Citation templates + fields database service
Citation templates
Internet
Genealogy report / chart
Report generator
S/C Meta Data, Internal storage
Genealogy data
BG export
S/C Meta Data converter
G/Simple BG Conversion rules
Better Gedcom file (with templates)
S/C Meta Data, ext fields, int storage
Conversion rule format and BG simple fields
Rule service
User intervention
Multimedia info
Multimedia files
Conversion rule format
S/C Meta Data Conversion rules
S/C Meta Data interpreter, Syntax/ Format specific
S/C Raw Meta Data ext fields, ext syntax (e.g. a file)
Various formats eg. MARC, Zotero
Repository Access Function (or eg. Zotero)
WWW Browser
User intervention
Source database
Source meta data
Internet Repository
Internet
(Standard) Document

BG=BetterGEDCOM G=Gedcom S/C=Source/Citation

Thanks to Mark Tucker and GeneJ for ideas end effort

25.03.11 G.Thorud

Appendix 1. Description of Figure 2.

The functional building blocks in a possible future solution for handling of source/citation data as shown in Figure 2 are described below. It is a detailed description of the building blocks in Figure 1.

Terminology: S/C – Source and citation – should have been S&C.

For non-techies: The squares in the figure perform functions on data, they are pieces of program code. The arrows show that data is transferred between these functions, this can in principle be done via a file, internally in a program, via the Internet or a database. Thus, one box does not necessarily represent a program, several functional boxes (pieces of code) can be grouped into a program, but where the boundaries between programs are, is not shown because it is not important and gives us flexibility wrt how functions are grouped into programs. The important thing is that the boxes do things with data. In some cases I have indicated that data is (or can be) transferred via the Internet. The "bear cans" represent "data storage", again whether the data is stored in a file, database, Internet server or whatever is not important – they are just stored. The boxes with the "waves" are documents, "BG-specifications" or genealogy reports/charts.

The central point in the figure is the **"S/C Meta Data internal storage"**, typically in a genealogy program, where data about sources and citations are stored. These would preferably be general citation elements. To the right of this storage there are functions for transfer of S/C-data from Internet repositories/services, services possibly also holding source data. Bottom left are the functions for exporting (/importing) data to/from Gedcom and BetterGEDCOM. To the left are functions for report output. The rest are storage of parameters controlling how these functions operates, and specifications required.

The building blocks are:

Import of Source Data and S/C Meta Data from the Internet

Internet Repository – A web site or service holding Source Data and/or Source Meta Data accessed by genealogists. The meta data would be in a structured format, e.g. XML based, MARC etc. The access could be via HTTP (used for e.g. www) or some other protocol. The repository could be operated by companies, organizations, archives, libraries, bookshops etc. Some of them will hold Source Meta Data only (as many library databases currently do, although most of us would not call that a repository because it does not contain source data, e.g. www.WorldCat.org) and some only the Source Data (but these are of less interest here because the focus here is on the meta data, example of a source data only service is www.InternetArchive.org). The most important repositories are those where you can access both the source data and the meta data about them – not many of those around for genealogists at the moment.

Repository Access function – A program, or a function integrated into a **web-browser**- or other program (could also be in a genealogy program), capable of accessing and capturing Source Data and/or Source Meta Data from an Internet Repository, and store it (see below) on a computer in some format. It could be specific to a particular repository or general. The standalone or browser-based functions offered by Zotero (www.zotero.org) is an example.

Multimedia files – These are the Source Data files (not the meta data) downloaded from the service. What is not shown in the figure is that the Source Data could also be structured (e.g. XML or CSV encoded), or could be unstructured text, and could then be imported into the genealogy program (and these data would also be accompanied by meta data).

Multimedia info – These are the data structures in a genealogy program that keeps track of multimedia files, normally stored externally to a genealogy program (e.g. in files on the same computer). The info would for example be the file name, the type of file, a caption for a photo etc. (This is not info intended to be output in a citation, but some Source Meta Data could also be stored in the multimedia info as well as in the S/C Meta Data Internal storage (TBD), but that is not important in this context).

S/C Raw Meta Data, external data fields and syntax – S/C Meta Data, possibly with a structure, syntax and semantics specific to the Internet service or the Repository access program. Alternatively the structure, syntax and/or semantics could be according to some standard.

S/C Meta Data interpreter, syntax/format specific – Interprets the syntax in the S/D Meta Data (there could be one variant of the function for each syntax/format) and converts the data into an internal syntax independent structure holding citation elements. The interpreter could for example be the first step in a S/C Meta Data import function in a genealogy program (the output is then internal to the program), but it would not be an import function if the Repository Access Function was built into the genealogy program. The semantics of the citation elements output from this interpreter function would still be specific to the Internet Repository (or a standard). Note: The feasibility of such an interpreter would be highly dependent on the ability to isolate and identify the individual pieces of information in the specific format/syntax and to identify them, which will be difficult if the specific format has few citation elements with a lot of info compressed into one data field. A complicating factor is also that the data would perhaps not confirm to the specification of the format. Manual intervention might therefore be needed to improve the output of the interpreter, but that could be a simpler task than retyping all the info.

S/C Meta Data converter – this function would operate on a more semantic level, converting the specific citation elements in the S/C Meta Data into (general) citation elements (and source types) to be stored in S/C Meta Data internal storage. Its operation could be directed by **S/C Meta Data conversion rules** specific to the Internet service, specified using a (standard) **Conversion rule format.** Manual intervention may be required. The arrow from S/C Meta Data storage to the converter indicates the possibility of fetching data from the storage for conversion from one citation style to another, the result being put back in the storage.

Gedcom and BetterGEDCOM import/export

The figure is simplified to only show export, and only BG. Depending on the internal structure used in S/C data internal storage, BG **export** will be a more or less complex function. If the internal structure uses general data elements, it should be relatively simple. Note that the BG file could contain Citation template definitions.

G/Simple BG Conversion rules – Gedcom/simple BG Conversion rules– For programs using general citation elements, these rules are used by the BG export function to convert data into the S/C Meta Data fields of current Gedcom, one rule per source type. It is assumed that the same fields could be present in BG files also, for backward compatibility reasons, a “Simple” variant of BG’s citation elements. Depending on the set of general citation elements, this might be overkill. The rules would be specified using a (standard) **Conversion rule format**.

Report/chart generation

Citations (footnotes, endnotes, bibliographies) in reports/charts are produced by the **Report generator** (possibly helped by a user). The format of the citations is specified by **Citation templates**, specified by a (standard) **Template specification format**.

The templates, rules and source types for various Citation styles could be downloaded from a **database service** on the Internet - everything based on a standard set of **General citation elements**. And/or, they could be transferred in a BG file.