A data model for sources and citations DRAFT v. 0.4
A proposal for discussion

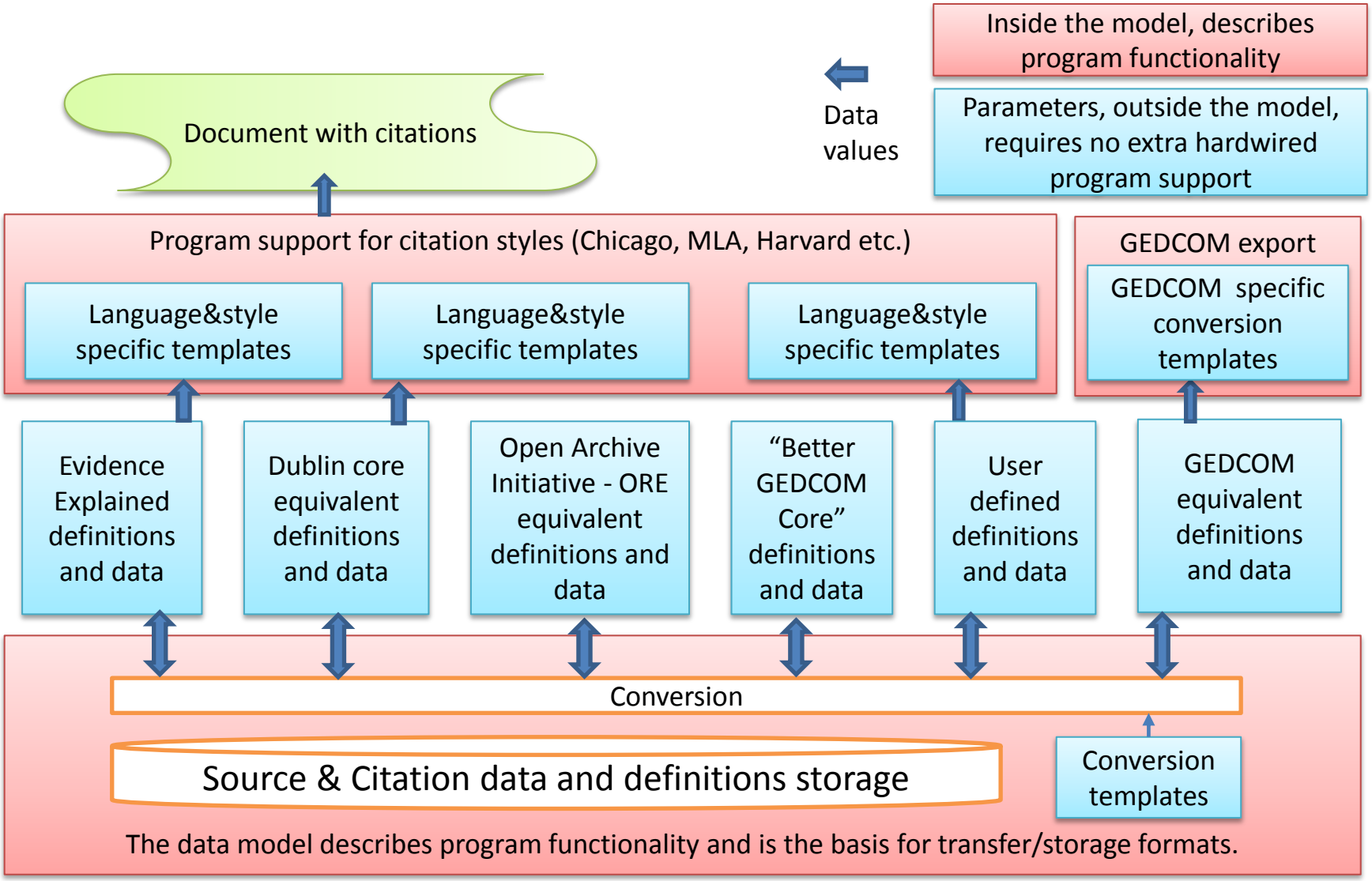
- This model suggests a way to structure data needed to record and generate citations, when these data are transferred form one genealogy program/service to another.
- The model is limited to information relevant to production of citations (notes and bibliographies), other source related issues are not described.
- It can be mapped to a structure in a file (BG-file), API, database or communication protocol.
- It is based on what the author consider best practices in some existing genealogy programs, but extended to provide
  - Independence of citation style, to the extent possible without unreasonable changes in programs. Total independence in every detail will require use of something like Citation Style Language.
  - Support for transfer of definitions of «master source types», «source detail», their «citation elements» and «templates».
  - Translation of such definitions into other languages.
  - Transfer of data in more than one language where appropriate, and generation of citations in a language different from the one (those) used when the data was entered.
  - Conversion of data from one «Master Source Type Set» («meta data set» e.g. EE, Dublin Core) to another.
  - Means for selection of the «Master Source Type» when data about a source is recorded.
  - Sharing of element definitions between master source types based on different «meta data sets».
  - Backwards compatibility with GEDCOM, to the extent possible.
- Practical use of the model requires development of definitions of the data in master source types and templates for presentation of citations and conversion (see next page).
- The model is currently not complete wrt template syntax , data type descriptions, conversion method and GEDCOM incorporation. The focus is on citations only, QUAY and media is not included. More work is required to develop a complete specification, incl. testing on real cases.
- More background material can be found here (and in documents referred to by the documents) http://bettergedcom.wikispaces.com/An+architecture+for+sources%2C+reference+notes+and+bibliographies and here http://bettergedcom.wikispaces.com/EE+%26+GPS+Support but note that there are terminology differences.

27 Nov 2011 G. Thorud

Thanks to GeneJ and Adrian for many contributions, although they do not necessarily agree.

Sources and Citations data model

What is covered by the model, what must be defined by other «documents»



The main entities in the model and their function

The model supports

1. Information about **Master Sources** and **Citations** for production of reference notes and bibliographies. Citation entities may be referenced by persons, events, groups, places etc. in e.g. a GEDCOM file.
2. All the information is recorded in **Citation Elements (CE)**, so named because of their primary function of holding information rendered in a citations, usually entered in data fields in the user interface. Their definitions are not part of the model, rather implementations of the model can be used to store definitions of **Citation Element Types (CET)**. The data recorded for a CET is contained in a **Citation Element Value (CEV).** Additional work will most likely show that some CETs (e.g. notes in GEDCOM, unstructured source excerpts) must be supported by all programs, and may require special processing by programs - this solution is chosen so that these CETs may have the “attributes” of a CEV, e.g. language and be accessed by templates.
3. Definitions of **Master Source Types (MST)** with CEs for the Masters Source itself and «Source Detail». The latter usually identifies the location of the cited information in the Master Source and possibly «reasoning» about the information, and optionally a variant of the information itself (extract, summary). The definitions are contained in **Master Source Type Sets (MSTS)** identifying and describing the set, including issuer etc.
4. The definitions of Master Source Type Sets and Citation Element Types may be translated into various languages.
5. Definitions of **Citation Templates** controlling the information and layout in reference notes, bibliographies etc. Template definitions are contained in **Template Sets (TS)** identifying and describing the set. There may be many TSs per MSTS, for example supporting different languages or citation styles. Citation Templates may be overruled, providing a way to tailor the citation text on a per citation basis, possibly (but not recommended for consistency reasons) ignoring the stored CE values.
6. Recording of CE values in several languages. Many CEs in a Citation will be language independent, it will be inappropriate to translate others, and yet others may be translated using language independent codes, but some fields may be recorded in several languages, for example CEs containing “reasoning” or “summary”. Programs supporting only one language must define the priority of languages to be imported. Although many programs will only support one language per value, it is important that a standard does not prevent more than one.

The model supports

7. Use of (uniquely identified) Citation Element Types defined in other Master Source Type Sets, thus allowing several MSTSs to be based on a common set of CETs and simplifying conversion of Citation data from one MSTS to another.
8. Recording of only Citation Element Types in an MSTS, no MSTs, for publication of CETs independent of MSTs.
9. Master Source Types can have "**Selection criteria**" that classifies a MST according to defined criteria that provides information that assists the user in selecting a MST. It is envisaged that such criteria could be standardized, but this information may also be used by vendors to define selection criteria for an MST that fits with the selection functionality in a program. For less frequently used MSTS or recently defined MSTS, the recorded Selection properties can be used by a program until the vendor supplies optimal information.
10. **Conversion templates** that may be used to convert the Citation Element Values of one MST into those of another MST, an MST usually defined by a different MSTS, possibly defined according to a different style or source classification scheme. This will make it easier to collect source meta data from various databases using different meta data schemes, and to convert data into MSTSs supporting a single citation style. Several factors determine how successful such a conversion can be, and it may be necessary to manually edit the result of a template controlled conversion. Enhancements of the Template functionality may improve the quality. Vendors will have to decide how templates are made available to the user.
11. An alternative conversion mechanism is also provided for use when Conversion templates are not available. It is based the establishment of a hierarchy of General and Specialized CEs and MSTs, a concept similar to the "**Dumbed down**" concept used by the Dublin Core scheme for source meta data. A Specialized CE may be converted to a more General (Dumbed down) CE. This functionality may not be needed if Conversion templates are accepted as the way to go – templates are more powerful.
12. Backwards compatibility with GEDCOM will be provided by definition of a special MSTS and Conversion templates and/or conversion using "Dumbed down". Since GEDCOM provides a limited number of CEs, data loss may occur, but it depends on the MSTS defining the data and the conversion rules for conversion from a specific MSTS to GEDCOM. These templates could also be used to extract CEs used to administer the MSIs stored by a program, e.g. search on Author (unless some harmonization is achieved, there could be several CETs holding this info in various MSTSs), or there could be special templates.
13. **Data Types** for CEs defines certain functions that a program can perform on the data in a CE. These are: Control of data on input, selection or manipulation of parts of a CE value to be output by a template, translation of recognized terms in the CE, provision of possible values (code lists) that can be entered for the CE, use of the CE to access webpages/e-mail, qualification (TypeValuePair) of CEs and definition of culturally independent values (for code list, date, name). Some data types reduce the number of CEs that are needed.

Sources and citations data model – DRAFT 0.4

27 Nov 2011 G. Thorud

Master Source Type Set (MSTS)
- Common administrative info (include box content here, see right)
- Source type selection criteria * 1)
  - Selection Criteria ID
  - Is top-1 level criterion
  - Choice sequence #?
  - Subordinate Criterion ID?
  - MST UUID* ➔
  - Prompt/choice: Criterion/Choice description
- Country code 2)
- Supporting MSTSs (with CETs) (Pair: MSTS UUID ➔, URI)*

Entity also used for Citation Element Type Sets only (no MSTs).

Master Source Type (MST)
- UUID 1)
- Master Source Type Set UUID
- Full Name
- Alternative name? 2)
- Short name? (MSI selection) 3)
- Extensive description?
- Short description incl. registration instructions
- Document refs? (e.g. to EE)
- Master Source CEs (Sequence of CET UUID)
- Citation CEs (Source detail) (Sequence of CET UUID)
- ??Dumbed down MST (pair: MSTS UUID, MST UUID)*

Master Source Instance (MSI)
- ID
- Master Source Type UUID
- Citation Element Value+ (include box content here)
- Repository (triplet: Repository ID ➔, Note ID or Note?, Call#?)*

Citation (Instance) (Source detail)
- ID
- Master Source Instance ID?
- Citation Element Value+ (include box content here, see right)
- Overruling Templates (Pair: Template UUID ➔, Language code)* 1)
- Exclude citation from output 2)

Citation Element Type (CET)
- UUID 1)
- Master Source Type Set UUID
- Name (prompt)
- Name in templates 2)
- Full description (help)?
- Short description (hint)?
- Data type ID (see box)
- Code list? (triplet: code, Long description, short desc.)+ 3)
- ??Dumbed down CET (pair: MSTS UUID, CET UUID)*

Template Set (TS)
- Common administrative info (include box content here)
- ("From") Master Source Type Set UUID ➔
- "To" MSTS UUID? (for conv.) ➔
- Templates language code? 1)
- (From) MSTS definition URI?
- "To" MSTS definition URI?

Citation Templates
- UUID
- Template Set UUID
- Master Source Type UUID ➔
- Templates (pair: Template type, Template)+ 1)

Conversion Template
- Template Set UUID
- "From" MST UUID ➔
- "To" MST UUID ➔
- Conversion rule (pair: "To" CET UUID, From Template)+ 1)
- Manual control most likely needed! ? 2)

Only selected arrows representing relations are shown in full, the others are indicated by ➔ The model could have been broken up into more entities.

? = 0 or 1, * = 0 or more,
+ = 1 or more, otherwise 1

Citation Element Value (CEV)
- Citation Element Type UUID ➔
- Value (triplet: long value, short value?, language?)+ 1)
- List separator? 2)
- Type? (if type-value pair)

Common Administrative info
- Document UUID
- Title
- Document description
- Issuing organization/company /user
- Issuer contact details ID?
- Document URI?
- Document date
- Document version
- Document language code 1)
- If this is a translation
  - -Original UUID
  - -Original title
  - -Original version
  - -Original issuer
  - -Original issuer Contact Details ID

Data Types
- String
- String with coded terms
- Date
- Person name
- Place
- URI/E-mail
- Coded (code, single choice)
- List (comma, semicolon etc separated)
- Numeric
- TVP - Type (coded) and value pair, multiple data types

Repository
- ID
- Repository Type? 1)
- Name?
- Abbr. name?
- Contact info ID
- Latitude?
- Longitude? 2)
- Coordinate format 1)
- Note ID *

Contact info
- ID
- Name? (Addr line1)
- Addr line 2?
- Addr line 3?
- Addr line 4?
- Place/city?
- Postal code?
- Country subordinate area?
- Country?
- ??Postal address template? 1)
- Electronic Communic. addresses (Pair: Type, Value)* 2)

For notes, see the following pages.

Abbreviations/definitions

CE – Citation Element
Citation – Reference note or
Bibliography entry
Reference note – Inline/Intext note,
Footnote or Endnote.

Model Entity Notes

Master Source Type Set

1) Specifies selection criteria, for the MSTs in the set, their choices and result, see page later in this doc.
2) Countries where the source types are located (or created, TBD). This is the primary selection criteria. Possibly limit to only one country. There should preferably be only one country per MSTS. A code is needed for "universal".

Master Source Type (MST)

1) Same UUID for all translations of the MST
2) Some source types have a very long formal name, but is often referred to by a shorter informal name.
3) Some programs use the source type to group types for selection

Repository

1) Standard codes TBD
2) The location of the coordinates must be reviewed, move them to contact info or later a place record.

Citation

1) Some programs allow free text reference notes (FTM, Legacy) . This raises some problems wrt consistency of data and language. It could be solved by a simp0lified (user friendly) template with syntax for [CE ids] , formatting and free text only - or a normal template. A special case is templates containing just text, no special syntax. Re. several languages, see CEV note 1).
2) This is at least useful for one user transferring between different programs. It is independent of QUAY.

Citation Element Type (CET)

1) Same UUID for all translations of the CET
2) Language dependent name used in templates to identify the CET in the user interface. Must be unique within the MSTS definition, and programs must internally ensure uniqueness across MSTSs (possibly by prefixing this name) – UUIDs are used for identification of CEs in templates transferred between programs. .
3) Lists code values and their-description for the Coded data type, the same for the type field of TLV data type and String with codes. Code values should be language independent – e.g. a number, but may also be strings. Values should not be limited to those in the code list, but other values should be the exception.
   Codes may have along and short description, e.g. for "ed."

Template Set (TS)

1) Shall be present for citation templates. The code identifies a language and optionally a country ("dialect" of the language). Conversion templates are language independent.

Citation Templates

1) Template types:
a. ??Title template??? Use?
b. Reference note template?
c. Short note template?
d. Bibliography template?
e. Inline reference template?
f. ??Source label template?
g. Others?

Conversion Template

1) Conversion to GEDCOM fields for backwards compatibility are done by defining an MSTS and a set of CEs for GEDCOM.
2) Correct automatic conversion of most CEs in the MSI is better than no automatic conversion.

Citation Element Value (CEV)

1) Zero or one language means the value will have to do for all language s when imported.
2) Programs that supports only one language, or do not specify the language, should let the user select preferred language(s) (prioritized) for import. CEVs with one or no language shall then be imported, while the value with highest prioritized language shall be imported for CEVs having values in more than one language.

Contact info

1) Some programs record where address fields are located in a printed postal address., this can be recorded in a template. It reduces requirements for duplicate input.
2) For phone, mobile phone, fax, e-mail, home page, Twitter, Facebook etc. Codes for types TBD.

Common Administrative info

1) All MSTs and CETs within a MSTS, and all templates within a TS, must be defined using the same language. Language does not apply to conversion templates. The code identifies a language and optionally a country ("dialect" of the language).

Continued on the next page

General Notes

1. The information in the model is restricted to that needed for citations. Several entities will have additional data for other purposes.
2. All records should also have a last changed time
3. There are additional structures in GEDCOM that need to be included, or mapped to CETs.
4. Structures for multimedia are not included, they will depend on the realization of the model in e.g. an extended GEDCOM. The same may apply to data in sources.
5. "Surety"/QUAY will need a solution, but is not covered by this model.
6. It is expected that one or more websites will provide a registry for MSTSs and TSs.

Data types

1. String
2. String with coded terms. Predefined terms within the string (page/p./pages/pp./volume/ vol./chapter/line/folio/issue etc.) are translated on output, if properly entered. Code list applies. Syntax for identification of codes in the transferred value are needed, also indicating if the long or short description matched. An exporting program translates recognized terms to codes (matching the longest code descriptions in a code list), an importing program translates the codes to the terms in the appropriate language (using a translated code list). A code description may comprise several words. This data type is not intended for translation of general text, but rather terms frequently used in the CE.
3. Date, a syntax separating day, month, year must be defined. Problems: ranges, uncertainty (ca.), seasons.
4. Person name - Given names, surnames, surname prefix/suffix that can be contained in ONE CEI must be defined. Delimiter between names etc. Sorting rules? A solution for multiple names is needed, also rules for sorting. Functions for initials, extract parts, "reverse" TBD. The definition of this data type depends on the solution for author etc.
5. Place, comma separated list, most significant last. The difference between place and list is that a link? to a place may be present.
6. URI/E-mail – could allow text prefixing/suffixing the URI/E-mail. E-mail may not be needed.
7. Coded, one code from a code list is recorded as the value. See note for CET. Templates may display ?code? or description. Use: Language, media type, recording format (Geir's ref: cf. EP GJ 22.11.11)
8. List (of strings), use: list of entities in an archive hierarchy separated by e.g. a comma, geographical hierarchies, most significant or least significant first to be decided. Can be reversed on output, or (most/)least significant extracted..
9. Numeric, use: Presented in Arabic or Roman number, ordinal (2nd or second) - language independent. A range must be supported. Problem: These numbers are often surrounded by additional text, may require splitting of CEs into several CEs, cf. examples in String with coded terms.
10. TypeValuePair, use: e.g. ISBN/ISSN, Author/Editor/Transcriber/Contributor etc., page/section/volume /issue/folio/chapter may reduce the number of CEs in a MST. TVP will in practice give rise to several data types, depending on the data type of the value field (e.g. person name, string, string with codes).

More work is needed on the detailed specification of the data types, and they must be tested on a MSTS.

Some data types are adopted from RM, but extended. The data types that can have short values must be determined (cf. || in RM).

Entity presence – usage scenarios

A BG file may contain any of the top level entities (MSTS, TS, MSI) with subordinate entities in the model, or only

- A TS with "ordinary" citation templates
- A TS with only conversion templates
- An MSTS with CETs, and a TS, without MSIs.
- An MSTS with only subordinate CETs (in this case the term MSTS is a bit misleading).
- An MSI may be present without any Citations linking to it.
- An MSI and a Citation with no definitions, assuming the definitions are known
- More?

A TS must be supported by an MSTS containing at least the CETs used in the TS, in the same language as the TS.

Maximum interoperability will be ensured at all times if the MSTS s, CETs for all MSTs and CEs having instances in the file, and at least one TS , are always included in the file.
When MSIs are present, the supporting MSTSs with subordinates, and at least one TS, should be present if the file is used for archival purposes. Several MSTSs and TSs may be present in the same file.
An application could download only one MSI (and optionally one Citation) , possibly accompanied by source data/image, from an Internet repository or source meta-data database.
Details TBD.

Continued on the next page

Template issues

1. The template syntax and functionality must be defined.
2. Templates will not be able to control all formatting of citations, e.g. sorting of bibliography entries, handling of “et. al” – and more, see e.g. Citations Style Language. Some general rules or recommendations may be needed, e.g. sort on first CE in template in bibliographies.
3. Template syntax depends on the data types, i.e. the syntax must support functionality associated with the types.
4. RM contains a good candidate for a template language, and there are also some other good ideas in other programs.
5. A simplified syntax should be considered for overruling templates, see Citation.
6. Templates used for conversion will have very limited capabilities wrt structural conversion, and will not be able to extract pieces of a CEV (unless functionality similar to regular expressions are built into templates, not very user friendly). There may be problems if meta data schemes bundle a lot of different info into a CE.
7. CEs are identified by their UUID in templates in order to prevent clashes between MSTS. A better solution that ensures uniqueness is welcome. See also Template name in CET.
8. If it should turn out that a CET can have more than one occurrence in a Citation (including the MSI CEs), there is a need to identify the correct occurrence in templates.
9. Templates must be able to select TVP type CEs with a specific Type value.
10. Default values for CEs (no page etc.) are handled in templates.
11. Conversion templates can be useful in a transition from the various existing source type implementations to a standard set.

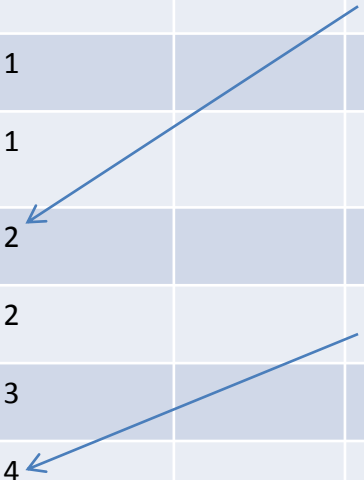
Other issues to be decided

1. Multiple master sources per citation is allowed by several citation styles, incl. EE. It is not supported by the current version of this model.
2. Current programs does not seem to allow a CE to occur more than once in a MST. This has consequences for Author, Transcriber, etc. A discussion about how to handle repeating creators (authors, editors etc.) is needed, few (none) programs handle this well although almost many sources have several creators of mixed type. A standard should have a proper solution and it should be the vendors task to map it to the user interface, the standard should not be the limiting factor.
3. A few style guides have lists of abbreviations, e.g. NARA, NST. A good solution will require new separate entities (or a special template type).
4. Hints for CEs per ST. Would allow the hint to be context sensitive, i.e. be specific for the MST without limiting the info that can be recorded in the CE.
5. Exclusion of CEs from output, for e.g. notes, extracts/summaries, or alternatively let the user control it with a overruling template.
6. Requirements to programs for support of the functionality must be defined.

Source Type selection criteria

The criteria comprise a multi level three structure starting at the top level. The top level criterion is the countries defined by the MSTS, possibly qualified by the MSTS Title if there is more than one MSTS for the country - although programs may in addition have a preferred MSTS. At each subordinate level there is a list of Choice Descriptions, headed by a Criterion Prompt (identified by Choice sequence # = 0). When a choice is made for a criterion, a set of Master Source Types (Full name, and optionally alternative name) are listed, or a subordinate criterion is displayed. There may be many criteria "paths" leading to the same MST. There shall be only one criterion below the top level (Top-1). Most of this is modeled after Legacy. See the MSTS definition.

Criterion ID	Is Top-1 level	Subordinate Criteria ID	Choice sequence #	MST UUIDs	Criterion prompt / Choice description
1	y		0		Select the main type of source
1		2	1		Book
1		3	2		Archival material
1			3	123456789, 456789123	Gravestone
2			0		Select type of author
2		4	1		Unknown author
3					
4			0		Select medium
4			1	234567891	Paper
4			2	345678912	Scanned



The maximum number of levels is TBD.