fhiso

CFPS 74
(Call for Papers Submission number 74)

Family History Lexicon

Submitted by: Tychonievich, Luther

Created: 2013-05-10

URL: Most recent version: http://fhiso.org/files/cfp/cfps74.pdf
This version: http://fhiso.org/files/cfp/cfps74_v1-0.pdf

Description: FHISO should maintain a lexicon of terms and their several meanings to help mitigate confusions caused by terms with multiple distinct usages. And example lexicon is included

Keywords: lexicon, dictionary, terms, definitions, vocabulary

Family History Lexicon

Luther A. Tychonievich

10 May 2013

Abstract: *Terms are not used consistently by every individual and that developers often use terms differently than do researchers. I propose FHISO maintain a lexicon of terms, and I provide a suggested initial lexicon to jump-start that process.*

It is my personal observation that confusion over terminology is a common source of headaches in genealogy and family history, particularly as it relates to data, information, and software. This confusion might be alleviated by the introduction and maintenance of an official FHISO-approved Lexicon.

1 Proposal

I propose that FHISO create and maintain a family history terminology lexicon. This lexicon will consist of both an effort to document how terms are used "in the wild" as well as an identification of a FHISO preferred vocabulary (see CFPS 23 [1]). I propose that an individual or committee be appointed to maintain this lexicon and that the contents and definitions be refined until they can be accepted by a vote of all FHISO members.

The maintenance of a descriptive lexicon of terms as they are used, in addition to a preferred vocabulary, provides at least three advantages.

- People are more likely to discover and adopt preferred vocabulary if it is referenced from an entry describing their current vocabulary usage.
- A lexicon of uses will help people understand materials produced before the lexicon was adopted or produced by people not consulting the preferred vocabulary.
- Having a lexicon of existing uses will help catalyse conversation about the relative merits and potential difficulties of candidate preferred terms.

2 Considerations

One important decision in creating a lexicon is the flavour or tone of the document. Is it to be an authority in itself, or is it to be a reference citing other authorities to defend its definitions? At least three different lexicon styles are possible:

Authoritative Contains terms and definitions without defending those definitions. An authoritative lexicon is thus positioning itself as a source in itself. Most desk reference dictionaries are written in an authoritative style.

The authoritative style offers the most flexibility to the lexicon writers.

Because I propose that the FHISO lexicon be accepted by vote of the FHISO membership, the authoritative style seems to me to be appropriate; the example lexicon in Section 3 is presented in the authoritative style.

Academic Defends each definition of a term by citing or quoting multiple uses of the term with that definition in extant publications. The Oxford English Dictionary is probably the best-known example in an academic style.

The academic style is the most defensible, but also the most work to create. With a small living community there is the additional challenge that many uses may appear primarily in emails, conversations, and other hard-to-cite communications; and that there is a potential for offence in who is or isn't cited for a particular definition.

Aggregated Defends each definition by citing another lexicon, glossary, or other published definition. The only significant aggregated lexicons of which I am aware are maintained by online search engines.

The aggregated style in some sense "passes the buck," deferring the task of creating accurate definitions to others. Although it might make sense for a single author lexicon (such as the example lexicon in the next section), it does not seem in keeping with a lexicon produced by a standards organisation.

In addition to presentation style, FHISO will need to decide on the criteria by which a term is deemed significant enough to warrant inclusion in a lexicon. Is a single use enough? How about a single use within a widely-circulated publication? What about many uses all within a single narrow community? Etc.

The correct selection criterion is far from clear to me. The example lexicon in Section 3 is based on words that I've heard from multiple speakers and/or that I have observed cause confusion in a conversation. Such loose criteria may be sufficient given the relatively small size of the family history lexicon, but a more formal criterion might be preferable if the lexicon grows too large or if its contents becomes contentious.

3 Example Lexicon

My initial effort at populating a family history lexicon is included below. It is presented in the authoritative style because I believe that to be the most appropriate style for the official FHISO lexicon that I am proposing be created. That said, its contents and definitions are based almost exclusively on my own personal observations; it almost certainly includes omissions and mistakes.

It is my hope and proposal that this example lexicon be replaced by an official FHISO lexicon approved by vote of the FHISO membership. Until such an official lexicon exists, I welcome additions, clarifications, and citations to this example lexicon either through a comment paper submitted to the FHISO call for papers or by private email to ltychonievich at fhiso.org.

Abstract A description of the contents of a document, artefact, or other **Source**.

Abstracts are one type of **Derivative** of an artefact.

Antecedent The antecedents of an **Inference Rule** are the set of conditions required for the rule to apply. These are usually general in form, such as "a birth event with a person in the mother role and a date."

The antecedents of an **Inference** are the particular claims used to match the antecedents of inference rule being applied. These are always concrete in form: in the above example, it would be a particular person, date, and birth with the necessary interrelationship.

Approximate Having low **Precision**. Less commonly used to mean "having low **Reliability**" instead.

Argument A presentation of **Reasoning**. Also called a **proof argument**. Sometimes the term "argument" is limited to human-readable documents describing reasoning; other times it is used to refer to the reasoning itself, whatever its presentation.

See also **Top-Down** and **Bottom-Up**, which describe processes and data models that support building arguments in two different ways.

Assertion Often synonymous with **Claim**. Sometimes restricted to claims contained in a **Conclusion** or **Proof**. Sometimes used to mean things stated without **Evidence**.

Attachment "Attachment" has at least two distinct uses:

1. A real-world document or artefact that is associated with a **Conclusion**, **Persona**, **Intermediate Conclusion**, or the like, but that is not a **Source**. Depending on the definition of "source" used, all documents might be considered sources and thus "attachment" might not be used at all.
2. A real-world document or artefact a version of which is bundled with, rather than only referenced by, a family history data file.

Because of the overloaded meaning, I suggest that the term "attachment" be avoided wherever an alternative word may suffice.

Bottom-Up Of research: starts with individual **Source**s and **Claim**s and assembles them to discover likely truths.

Of data: structured based on combining claims to build a conclusion.

See also **Top-Down**.

Calender A system for identifying dates. Converting dates between calenders is often complicated by the lack of a shared **Epoch** or by being based on non-synchronized information (e.g., lunar vs solar calenders).

Citation An acknowledgement or indication of a **Source**. There are several kinds of citations:

1. A description of a real-world document, conversation, artefact, or other potential source. This description typically identifies publication details or origin rather than content or meaning.
2. A cross-reference or pointer to a citation (definition 1).
3. A cross-reference or pointer to another element of research, often an **Attachment** (definition 2), **Inference** or **Claim**, that is considered to be part of the **Evidence** of the citing element.

Which of these kinds is intended is usually clear from context.

Claim Some putative fact that a **Source**, **Inference**, **Argument**, or **Conclusion** states, supports, or contains. Also called an **assertion**.

Often, though not always, a claim is assumed to be atomic; that is, it cannot be split into two or more distinct claims without losing some significant part of its meaning.

Comment Within the context of family history software, a "comment" is generally a piece of free-form text associated with another data element.

Some written messages posted by individuals on Internet sites, including family history sites, are also called "comments."

Conclusion A view of how the world was within the context of the genealogical or family historical research being performed. Often, but not always, a conclusion is assumed to include exactly one **Consistent** world-view.

The top-level claims of an **Argument** are sometimes called that argument's conclusion.

"Conclusion-" is sometimes used as a prefix to refer to the final aggregate view a researcher holds of something after applying all available reasoning. Thus a "conclusion-person" is a person as a researcher believes they were; a "real-world person" is a person as they actually were; and an "**Evidence**-person" or "**Persona**" is a person as they appear to be through the lens of a single **Source**.

Consequent The consequents of an **Inference Rule** are the set of **Claim**s that may be supported by the rule if the antecedents of the rule hold. These are often expressed in terms of the **Antecedent**s, such as "the mother in the antecedent was born at least ten years before the date in the antecedent."

The consequents of an **Inference** are the set of claims whose **Source**s include that inference.

Consistent Not containing any contradictions.

Consistent **Conclusion**s or sets of **Claim**s do not contain impossible assertions like "John is his own father."

Consistent data obeys the assumptions of the underlying data model, having no broken links, cyclic dependencies, or missing elements. In some data models it is possible to have consistent data that expresses an inconsistent conclusion or set of claims.

Derivative Of a document or artefact making **Claim**s, "not **Original**" with all of the associated variations that "**Original**" contains.

Common kinds of derivatives include **Abstract**s, **Extract**s, **Indices**, and **Transcript**s.

Direct The **Claim**s explicitly stated or intentionally made by a **Source** are direct. Non-claim **Information** (definition 1) could probably be direct, though I have never heard it so called.

Direct **Evidence** contains only direct claims.

Epoch A moment in time from which a **Calender** system is (or may be) measured. Converting between different calenders requires expressing a single epoch in both calenders.

The word "epoch" can also mean an era or period of time.

Event "Event" has several different uses

1. Something occurring at a relatively **Precise** moment in time, often but not always one that effects a change in the studied people; e.g., a birth, a death, a marriage, an eclipse, etc.
2. Something occurring over an extended window of time, often but not always one of broad historical impact; e.g., a war, a drought, a time in school, etc.
3. An action or occurrence detected by a program that may be handled by the program, often but not always resulting from user input; e.g., a key press event, a mouse move event, a timeout event, etc.

The first two definitions are sometimes confused for one another, but I have seen no alternative terms in common use. I suggest users explicitly clarify which kind of event they mean. The third definition is used in a distinct technical context and I have not observed any confusion resulting from it.

Some people also distinguish between **personal** events, which impact only a few people; and **historical** events, which impact entire regions or cultures. This distinction is orthogonal to the moment versus time window distinction noted in the first two definitions above.

Evidence The evidence of a **Claim** is the **Information** (definition 1) that supports the claim, either **Direct**ly or **Indirect**ly (as the **Antecedent**s of an **Inference**).

It is not uncommon to hear the word "evidence" used to refer to information (definition 1) or to claims without them used as being evidence *of* anything. This usage is particularly common with developers because "information" has many other uses in software and because evidence and information are often stored using identical data structures.

"Evidence-" is sometimes used as a prefix to refer to the the view of something that can be defended using only a single **Source**. See the discussion under **Conclusion** for more on this prefix usage.

Explanation Within the context of family history software, an "explanation" is generally a **Comment** that presents a human-centric summary of the reasoning used to support the data element to which it is attached.

"Explanation" is also sometimes used as a synonym for "**Argument**" or "**Reasoning**".

Extract A partial copy or **Transcript**ion of a document or artefact. Also called a quotation or excerpt.

Extracts are one type of **Derivative** of an artefact.

Fact "Fact" has at least three distinct uses:

1. What actually transpired; the truth that a researcher is trying to discover.
2. A **Consistent** assertion or **Claim**; one plausible or putative truth.
3. A strongly held belief in any context, including observations about tools, algorithms, processes, researchers, people, events, etc.

Because of the overloaded meaning, I suggest that the term "fact" be avoided wherever an alternative word may suffice.

Family History The history of a family, including but not limited to parent-child, spouse, and other familial relationships; births, marriages, deaths, and other events; anecdotes, images, and other "colour"; etc. Family history often defines "family" broadly, including cousins, in-laws, ex-in-laws, close friends, neighbours, co-workers, etc.

Family Tree Often the set of direct-line ancestors of a single person; sometimes also including descendants and/or collateral lines. Sometimes referring instead to a particular graphical representation of these people and their relationships. Sometimes used to refer all of the people in a family history.

"Family tree" sometimes refers to **Conclusion**-people and sometimes to real-world people. It is common for a single person to use both of these definitions in different contexts.

Fuzzy Having low **Precision**.

Genealogy Literally "line of descent or pedigree." Sometimes restricted to mean direct-line parent-child relationships; sometimes to the vital statistics of people in such a direct line; and sometimes used more broadly as a synonym for "**Family History**."

Granularity A type of **Precision**. When applied to things presented in a discrete hierarchy, such as time (under most **Calender**s) or location (under most political naming schemes), "granularity" indicates how many levels of the hierarchy are provided. For example, "2013" is courser-grained than "2013-05"; "USA, Virginia, Richmond" is finer-grained than "USA, Florida."

Index "Index" has at least three distinct uses. The first is common among genealogists; the second (from databases) and the third (from programming languages) and are common among software developers.

1. A partial **Transcript**ion of a document intended to enable content-based searches of the document.

   This definition of "index" is one type of **Derivative** of an artefact.
2. An data structure designed to accelerate common database queries.
3. A numeric positional key into a list or array.

Because each meaning is relatively narrow I have not seen signifcant confusion arrise from using all three together.

Indirect In general, anything that is not **Direct**. **Claim**s derived via **Inference** or contained "between the lines" of a **Source** are indirect. Indirect **Evidence** is evidence that makes use of one or more indirect claims.

Inference Inferring is the process of deriving new **Claim**s from existing claims and other **Information** (definition 1). The derivation used is the **Inference Rule**; the pre-existing claims and information are the **Antecedent**s and the new claims are the **Consequent**s of the inference.

"Inference" is also used to refer to the process of extracting information from a **Source** and any other part of the research process where judgement calls need to be made by the researcher.

Not all inferences are identified as such and not all identified inferences identify their inference rules.

Inference Rule The context under which an **Inference** may be made. A description of an inference rule identifies the **Antecedent**s and **Consequent**s of the rule.

Some inference rules represent logical deductions, such as the relationship between age and birth date. Some represent constraints of natural law, such as ages of fertility and life expectancy. Some represent trends within a particular culture, such as the relationship between given names and gender. Some people restrict their usage of "inference" to a subset of these kinds of rules, but I have not been able to identify trends in which subsets are most common.

Information "Information" has at least two distinct uses. The first is common among professional genealogists; the second is from information theory [2] and is common among professional software developers.

1. That which may be learned from a document, meaning primarily the set of **Claim**s in a **Source**, but also any other intelligence that may be extracted from the document's existence, format, contents, etc.
2. The "meaning" contained in a signal, often measured in bits of entropy. Because of information's importance in computing, many computing fields and positions are called, e.g., "information systems", "information technology", etc.

Because of the overloaded meaning, I suggest that the term "information" be avoided wherever an alternative word may suffice.

Many developers use the word "**Evidence**" in place of "information" (definition 1) in every instance.

Intermediate Conclusion Like a **Conclusion** in most particulars, except that it is used as a step along the way toward building or supporting another conclusion.

Negative Of **Evidence**: based on the (apparent) lack of something rather than its presence. For example, a census not including an expected child might provide negative evidence supporting that child's earlier death. Negative evidence is a kind of **Indirect** evidence.

Of **Claim**s or **Conclusion**s: asserting that something is not true. A record asserting someone is "single (never married)" is a negative claim about a marriage event. A negative conclusion typically refutes an otherwise-reasonable assumption, such as "this John Doe is not my John Doe."

In my experience, few researchers distinguish between positive and negative conclusions.

Original Of a document or artefact making claims, the version of the document created by the originator of the claims. Sometimes including official and/or automated copies or replicas.

Sometimes used as a direction on the **Provenance** line, as in phrases "more original" and "less original."

Persona Refers to a particular element of a particular class of data models; the plural in this context is "personas" not "personae." The approximate meaning of a persona is "the view of a person supported by a single **Source**; or the view of a person derived by combining two or more single-source personas."

"Persona" is also used to describe tools and data models that are based on having a data structure for each single-source set of **Claim**s and combining those single-source structures into larger **Conclusion**s.

"Persona" is sometimes used to imply that a system or data model uses **Bottom-Up** data, whether or not that data is organized as personas.

"Persona" is sometimes used as synonymous with "**Evidence**-person"; see the discussion under **Conclusion** for more on this usage.

Precision The degree to which the described element pinpoints a single alternative from the universe of discourse. For example, "turned 27 today" is a more precise age than is "in her 40s;" "Luther A. Tychonievich" is a more precise name than is "John;" etc.

Precision is logically distinct from, but sometimes used interchangeably with, **Reliability**.

Primary Of a **Claim** or assertion, made by (1) someone with personal knowledge of the asserted facts (2) at or near the time when those facts became true. Of **Evidence**, supported by primary claims. Of a **Source**, used to supply primary claims.

Some people get upset when a source is called "primary;" others used the phrase "primary source" frequently.

Some people used "primary *X*" to mean "the most **Reliability**Reliable of the available *X*s," though this usage seems to be rare.

Proof A proof **Argument** that demonstrates that the preponderance of evidence supports the **Conclusion** argued and that considers the results of a sufficiently exhaustive and complete search for, and analysis of, **Source**s.

Proofs are not final; they stand until additional evidence is discovered.

Many researchers do not refer to proofs at all. Those who do generally use the term as defined in much greater depth by the Board for Certification of Genealogists [3, 4].

Proof Argument see **Argument**.

Provenance The chronology of ownership, custody, location, and replication of a document or artefact. A fully-identified provenance of a **Source** identifies the source of the source on back to the humans who first created, discovered, or recorded the information.

Reasoning The set of **Inference**s that ties a set of **Source**s to a set of **Conclusion**s.

"Reasoning" is frequently used to refer to the researchers' mental process; the portion of that process that is shared or stored in written, spoken, or digital form is often called an "**Argument**."

"Reasoning" is sometimes used to refer to all parts of a researcher's mental process including those steps that do not make it into the final argument; other times it is restricted to those steps that end up contributing to linking sources to conclusions.

Reliability The likelihood that the element in question accurately reflects what actually happened historically. In general, **Primary Evidence** is more reliable than **Secondary** evidence; multiple agreeing **Source**s are more reliable than a single source; and reliability decreases if there is reason to believe any party involved might lie, mis-record, or not care about the accuracy of the information.

Reliability is usually categorised only in relative terms (e.g., this evidence is more reliable than that evidence) rather than with numeric probabilities; accurate probabilities are not generally known.

Secondary Not **Primary**.

Of a **Claim** or assertion, made by someone either without personal knowledge of the asserted facts, or with some time after the facts became true. Of **Evidence**, supported by secondary claims. Of a **Source**, used to supply secondary claims.

Some people get upset when a source is called "secondary;" others used the phrase "secondary source" frequently.

Some people used "secondary *X*" to mean "not the most **Reliability**Reliable of the available *X*s," though this usage seems to be rare.

I have never heard anyone use "tertiary," "quaternary," "quinary," etc., within a family history context.

Source Some use "source" only as part of the phrase "*Y* is a/the source of *X*" meaning that *Y* is a real-world document whose contents help substantiate **Claim** *X*. Some use it this way but allow *X* to be an intermediate **Conclusion** as well as a real-world document.

Some people use "source" in isolation to mean "a real-world document" whether or not that document is currently the source *of* anything. Within this usage, some people restrict "source" to mean documents that could be used to substantiate a claim while others allow it to refer to any real-world artefact including things like sound recordings and images that contain few if any claims. Those who restrict sources to documents containing claims typically call other documents **Attachment**s.

Top-Down Of research: starts with a goal or research question and seeks out the **Information** (definition 1) necessary to reach the goal or answer the question.

Of data: structures the data based on the final **Conclusion** or belief and fills in details as they are discovered.

See also **Bottom-Up**.

Transcript A representation of an analog artefact such as handwriting or audio files as digital text comprised of a string or sequence of characters.

Transcripts are one type of **Derivative** of an artefact.

References

[1] Tony Proctor. "Proposals to Define a Preferred Vocabulary." *FHISO Open Call for Papers* CFPS 23. http://fhiso.org/files/cfp/cfps23.pdf Retrieved 2013-04-30.

[2] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1949. ISBN 0-252-72548-4

[3] Board for Certification of Genealogists, "The Genealogical Proof Standard." http://www.bcgcertification.org/resources/standard.html Retrieved 2013-04-30.

[4] Board for Certification of Genealogists, *BCG Genealogical Standards Manual*. Ancestry Publishing, 2000. ISBN 978-0-91-648992-2.