fhiso

CFPS 69
(Call for Papers Submission number 69)

A Transcription Notation for Genealogy

Submitted by: Fitzpatrick, Jerry

Created: 2013-05-05

URL: Most recent version: http://fhiso.org/files/cfp/cfps69.pdf
This version: http://fhiso.org/files/cfp/cfps69_v3-0.pdf

Description: A proposed notation for improving the transcription and translation of genealogical sources.

Keywords: transcription, translation, notation, annotation, note

Contents

1 Abstract

Transcribers often use notes enclosed in brackets to provide supplemental information. Unfortunately such notes can be ambiguous, context-sensitive or verbose.

This paper proposes a more elaborate notation that can overcome these limitations. The notation is defined by formal grammatical rules and is suited for both transcriptions and translations.

2 Introduction

A transcript is intended to be an exact copy of a source. Source information is transcribed as written even if it seems incorrect to the transcriber.

Unfortunately, genealogical sources often have irregularities that complicate transcription. Handwriting can often be difficult to interpret. Sources often contain smudged or faded characters. Engravings can be worn, damaged or have low contrast.

Annotations can be used to describe these irregularities. For example, a transcriber might note that some of the original text was smudged, crossed-out or written in the page margin.

Traditionally, transcriber notes are set off using brackets to indicate that they are not part of the original text. This approach is flexible, but the placement and wording of the note can make its meaning ambiguous. Consider the following transcription:

> 2[?] Jan 1913

Does the “[?]” mean that the ‘2’ character is unclear? Does it mean that the ‘2’ is followed by a missing or unclear character? Does it mean that the validity of the ‘2’ is questionable? Or does the text “[?]” appear literally in the source?

Without a clear definition of the transcriber’s notation there is no correct answer.

3 A Better Notation

Any transcription notation must allow the reader to distinguish between the transcribed text and the transcriber’s annotations. A “universal” transcription notation would:

- Impart information in a concise and unambiguous fashion.
- Not rely on special fonts or formatting.
- Use a small number of special characters that are rarely used in source text.

The proposed notation assigns special meaning to the following characters:

Symbol	Meaning
? (question mark)	An unclear character.
_ (underscore)	A missing character.
@ (at sign)	A source defect that may or not be a character.
^ (caret)	Prefix to a literal character or expression.
/ (forward slash)	Alternative expression separator.
{ } (curly braces)	A group of text.
( ) (parentheses)	Alternative characters or expressions.
| | (vertical bar)	Preprinted text.
[ ] (square brackets)	A postfix note.
< > (angle brackets)	An inline note. Three predefined notes are:
	<blank> data not recorded (blank entry)
	<na> Latin non adicio ("not applicable")

These characters are not common in genealogical sources (at least in English). They belong to ASCII character set ("plain text") and do not depend on special fonts or styling.

In addition to the special characters, the notation follows rules defined by an LALR grammar (Appendix A). The notation is described in more detail in the following sections.

4 Postfix-Style Notes

A conventional transcription note consists of literal text enclosed by a matched pair of brackets. It normally describes the text that precedes it (i.e. it is a postfix notation), although exceptions are common.

Unfortunately, the note's meaning can be ambiguous to the reader due to uncertainty about the span of source writing that it applies to.

4.1 Additional Rules

In the proposed notation, a "postfix-style note" is a conventional note that uses additional rules to reduce ambiguity. The additional rules are:

1. The note *always* applies to the preceding text. Some transcribers use brackets to denote inline notes or preprinted text, resulting in ambiguity.
2. Provide the means to explicitly identify the source text being annotated. The technique for this is described in the next section.

3. When the span of source text is not explicitly identified, the note applies to the entire, contiguous group of characters that precede it (not more or less). A “contiguous group of characters” will normally be a word. In some instances, though, it may be a number or an abbreviation.

If the note cannot conform to the third rule, the notation for explicit grouping or alternatives should be used instead.

Examples:

Text	Description
inatrix [matrix?]	An educated guess about the preceding word.
Harold Spiltz [?]	The transcriber is unsure about the surname. The given name is *not* involved in the annotation.
143 [148?]	The transcriber is unsure about the last digit of the number. The notation for alternatives would be preferable here.
143 [8?]	Invalid – “8?” is not a valid note about “143”.
2 [3]	Invalid – ‘3’ is not a valid note about ‘2’.

Ambiguity can be always be eliminated by using explicit grouping.

4.2 Explicit Grouping

The purpose of a group is to explicitly identify a span of source writing to be annotated. The postfix-style note that follows a group applies to the entire group.

A matched pair of curly braces is used to denote a group. A group can consist of any contiguous span of text, whether or not it includes whitespace. Groups may not be nested.

Examples:

Text	Description
{Jahn Phittshugh} [sic]	Verbatim spelling of entire name emphasized.
borrow a {quid} [pound]	An explanation/translation of “quid”.
{In pace requiescat} [rest in peace]	A translation of a Latin phrase.

In some situations, inline notes can be clearer than postfix notes.

5 Inline Notes

An inline note is an annotation that is meant to be read in line with the transcribed text.

A matched pair of angle brackets is used to delimit an inline note. All of the characters enclosed by the angle brackets are considered to be literal.

As with the postfix-style note, an inline note contains supplemental information from the transcriber, not from the source.

Examples:

Text	Description
one <nation> under	A faded/missing word was probably "nation".
Birthplace: <blank>	Missing data in a form-based record.
Freedom <Kansas, USA>	A clarification about the location of Freedom. A postfix note would work equally well here.

Two special cases of inline notes are defined for form-based records:

- <blank> – The data field is blank. Nothing was written for this field in the source record and no additional information is known.
- <na> – The data field is blank, but not applicable in the current context. For example, a death record field denoting "spouse" would not be applicable to an unmarried individual.

Inline notes should be used sparingly.

6 Missing and Unclear Characters

Source writing may have missing or unclear characters. On occasion, the transcriber may be unsure whether a mark is a character or a simply a spot.

Inline notes can be used to address these problems, but "placeholder" characters are more concise. The proposed notation uses three special characters ('?', '_' and '@') to denote missing or unclear characters. These characters are essentially a shorthand form of inline notes.

Examples:

Text	Description
County Alt?inc?m	Two unclear characters within the name.
Jo_n Hillford	A missing character in the given name.
Wilford Jo<n>es	Missing/unclear character was probably 'n'. Might be clearer as "Wilford Jo_es [Jones?]".
Alexander J. Smit@son	A mark in that might or might not be a character.

The ‘@’ (blot) character should only be used to denote a mark that resembles, or seems to be concealing, a character. It is not intended to identify all of the imperfections in the source.

A missing character is usually a character that has become worn or faded to the point of (near) invisibility. Occasionally, a character may be missing due to a printing or writing error. If the source writing is very faint, the exact number of missing characters may be indeterminate.

An unclear character is a character that is visible but illegible. The character may be illegible due to fading or other defects in the source medium. An unclear character (or word) may be clear enough to have multiple interpretations. In this case, the notation for alternatives should be used.

7 Alternatives

Alternatives can be used when characters or phrases in the source are unclear, but have limited possibilities.

Alternative characters are denoted by enclosing the choices in parentheses. Alternative phrases are denoted by enclosing the choices in parentheses, but separating each one using a forward slash.

The alternative characters or phrases should put in order from most likely to least likely, based on normal reading order. For example, since English reads left-to-right, the leftmost alternative is the one believed most likely to be correct.

Examples:

Text	Description
All(ea)n	Most likely "Allen", but possibly "Allan"
March 1(83), 1912	Most likely “March 18”, but possibly “March 13”.
(Maria / Myra)	Most likely “Maria”, but possibly “Myra”
(CK)ath(ae)rin(ea)	Valid notation, but excessive use may be confusing

To minimize confusion, the number of alternatives should be limited to two or three. If there are truly more than three alternatives, an inline or postfix note may be clearer.

Choosing between the notations for an unclear character or alternatives is a matter of accuracy. When the original text is very unclear the ‘?’ character should be used. If the original text has only two or three interpretations, alternatives are more suitable.

8 Other Transcription Issues

A transcription can rarely capture every nuance of source writing. For example, the source may use superscripts, subscripts, underlines, strike-throughs, italics, boldface, footnotes, and writing in margins or between lines.

A source may contain handwriting, type writing or a combination of both. It may also have signatures, seals, emblems or other “decorative” artifacts of importance.

If transcript can use a style that matches the original writing (e.g. underline), the artifact can be captured more directly. If the transcript uses plain text, then notes can be used to capture the additional information.

Examples:

Text	Description
<three illegible words>	Inline note about unclear words.
{ his only child } [in margin]	A note about writing in the margin.
Henry{VIII} [superscript]	A note about superscript writing.

9 Literal Characters

Ambiguity could result if the source contains any of the special annotation characters. For example, if square brackets were transcribed directly from the source, a reader might not be able to determine whether the text is part of the source or whether it is a postfix note.

In the proposed notation, a caret character (‘^’) preceding a special character denotes that the character appears literally in the original text.

Examples:

Text	Description
bob^@gmail.com	Transcript of the literal text "bob@gmail.com"
and Rooby ^[sic^] walked	Transcript of the literal text “and Rooby [sic] walked”
Sarah ^(Walker^) Jones	Transcript of the literal text “Sarah (Walker) Jones”

Literal characters should rarely be needed in genealogical transcripts.

10 Form-Based Transcription

Many genealogical sources are form-based. When transcribing these documents, it is helpful to denote which text was preprinted and which was entered manually.

There is no universal convention for differentiating preprinted text from manually entered text. Postfix and inline notes are not suited for this purpose because they contain information from the transcriber, not the source.

The vertical bar character is uncommon in source writing, so the proposed notation uses a pair of vertical bars to denote preprinted text. An example of form-based transcription using this convention is shown in Appendix C.

The vertical bar character must be used carefully so the reader doesn't confuse it with the characters '1' (one), 'l' (el) or 'I' (eye).

11 Acknowledgements

My thanks to various members of the Association of Professional Genealogists discussion group for their comments on the original notation. Special thanks to Donn Devine and Ray Beere Johnson II for their insights.

12 Bibliography

1. Elizabeth Shown Mills (Editor), *Professional Genealogy*, Genealogical Publishing Company, Baltimore, Maryland, 2001.

2. Church of Jesus Christ of Latter-day Saints, Family History Department, *The GEDCOM Standard, Draft Release 5.5.1*, Church of Jesus Christ of Latter-day Saints, Salt Lake City, Utah, 1987-1999.

3. Colorado State University Writing Guide, *Writing@CSU*, Colorado State University web site, http://writing.colostate.edu/references/sources/working/.

13 Appendix A – Notation LALR(1) Grammar

```
transcript:
   element_list

group:
   { element_list }

element_list:
   element
   element_list element

element:
   ?
   _
   @
   text_fragment
   preprinted_text
   group
   ( alternatives )
   note
   escape

preprinted_text:
   | sequence of any characters except vertical bar |

note:
   [ sequence of any characters except right bracket ]
   < sequence of any characters except right angle bracket >

alternatives:
   text_fragment
   alternatives / text_fragment

text_fragment:
   sequence of non-whitespace characters except | / [ ] { } < > ( ) ? _ @ ^

escape:
   ^ followed by any one of | / [ ] { } < > ( ) ? _ @ ^
```

14 Appendix B – Prose Example

SUYDAM GETS DIVORCE

The Court Did Not Waste Any Time on the Proceedings.

New York, Sept. 28. – It required less than fifteen minutes for Justice Clark t<o> hear the evidence and grant an interl__utory [interlocutory] decree of divorce to Walter L@p(ea)nard Suydam Jr., the Blue Point, L. I., millionaire, from his wife, Louise White, who left him for Frederick Noble, son of a Brooklyn plumber. Suydam himself testified only to his marriage to the defendant. Mrs. Cecilia McMara, employed at the Suydam home, told of seeing Mrs. Suydam and Noble together at the Blue Point house.

15 Appendix C – Form-Based Example

For form-based transcripts, the vertical bar character is used to enclose the preprinted text items. Although preprinted and manually-entered text may be interspersed, a tabular format is recommended for readability.

<table>
<tr><td>| PLACE OF BIRTH |</td><td></td></tr>
<tr><td>| County of |</td><td>Winnebago</td></tr>
<tr><td>| City of |</td><td>Oshkosh</td></tr>
<tr><td>| (No. St. Ward) |</td><td>804 Pearl</td></tr>
<tr><td>| FULL NAME OF CHILD |</td><td>Alice Mary Krantz</td></tr>
<tr><td>| Sex of Child |</td><td>F.</td></tr>
<tr><td>| Color or Race of Child |</td><td>W.</td></tr>
<tr><td>| Twin, Triplet, or other? |</td><td><blank></td></tr>
<tr><td>| Legitimate? |</td><td>yes</td></tr>
<tr><td>| Date of birth |</td><td>Feb. 4 |19| 18</td></tr>
<tr><td>| FATHER |</td><td></td></tr>
<tr><td>| FULL NAME |</td><td>Albert D. Krantz</td></tr>
<tr><td>| RESIDENCE |</td><td>Oshkosh</td></tr>
<tr><td>| AGE AT LAST BIRTHDAY |</td><td>25</td></tr>
<tr><td>| BIRTHPLACE |</td><td>Wis</td></tr>
<tr><td>| OCCUPATION |</td><td>Farming</td></tr>
<tr><td>| MOTHER |</td><td></td></tr>
<tr><td>| FULL MAIDEN NAME |</td><td>Mabel H(ea)rring</td></tr>
<tr><td>| RESIDENCE |</td><td>Oshkosh</td></tr>
<tr><td>| AGE AT LAST BIRTHDAY |</td><td>24</td></tr>
<tr><td>| BIRTHPLACE |</td><td>NY</td></tr>
<tr><td>| OCCUPATION |</td><td>H. W.</td></tr>
<tr><td>| Number of child of this mother? |</td><td>2</td></tr>
</table>

<table>
<tr><td colspan="2">| CERTIFICATE OF ATTENDING PHYSICIAN OR MIDWIFE* |</td></tr>
<tr><td>| (Signature) |</td><td>{O. E. Werner} [?]</td></tr>
<tr><td>| (Physician or Midwife) |</td><td>[Physician; “Midwife” crossed out]</td></tr>
<tr><td>| Address |</td><td><blank> [“Oshkosh” entered on preceding line]</td></tr>
<tr><td>| Filed |</td><td>Feb 11 |19| 18</td></tr>
<tr><td>| Local Registrar |</td><td>{E. H. Bische} [? (signed)]</td></tr>
</table>