Originally published at http://www.gaenovium.com/presentations2014.html

Copyright © Louis Kessler 2014

GÆNOVIUM LEIDEN 2014-10-07
THE GENEALOGY TECHNOLOGY CONFERENCE

Reading Wrong GEDCOM Right

Louis Kessler, Author of Behold, GenSoftReviews

www.beholdgenealogy.com www.gensoftreviews.com

Special thanks to Tamura Jones for his suggestions and reviews of this talk.

3rd party photos and illustrations in this presentation are all royalty-free Office.com clip art.

How do we read GEDCOM “Right”?

- GEDCOM 5.2 and earlier
  - Specifications don’t exist.
  - But we can reverse engineer the specs.

- GEDCOM 5.3 and later
  - Specifications exist.
  - They are imperfect, but do provide rules.

We can and should develop best practices.

Outline

1. Reading the Header
   a. GEDCOM Version Number
   b. Program Name and Version Number
   c. Character Set
2. Structural Problems
3. Level 0 Records
4. The CONC Tag
5. User Defined Tags
6. Odds and Ends

Reading GEDCOM in Behold

- A flexible, forgiving GEDCOM reader
- “Understanding” of GEDCOM grammar
- Generalized data structures
- A list of valid tags, by GEDCOM version
- Handling of special cases

- My goal: Try to read everything

GEDCOM 101

Gedcom_line :=

Level + [xref_id] + tag + [line_value]

```
0 @1234@ INDI
1 NAME Will /Rogers/
1 CHIL @1234@
```

Finding Sample GEDCOMs

- Google search (> 500)
  “0 HEAD” filetype:ged
  about 20,300 results
  - most are older
  - only 140 are from the past 10 years

- User files (>150)

- Size
  - < 1 KB (very small files)
  - 324,738 KB – Good-Engle-Hanks (prpletr.com)
    - largest file of people (741,968 individuals)
    - *Formerly at: http://prpletr.com/Gedcoms.htm – but now removed*
  - 650,134 KB – CoL2010.ged (catalog of life – Paul Pruitt)
    - largest file in use (about 2,100,000 individuals)
    - *See: http://famousfamilytrees.blogspot.ca/2008/07/species-family-trees.html*
  - > 73 GB – GedFan 28 (Tamura Jones)
    - largest test file (268,435,455 individuals)
    - *See: http://www.tamurajones.net/GedFan.xhtml – GedFan*

1. Reading the Header

Hello World

```
0 HEAD
1 SOUR 0
1 SUBM @U@
1 GEDC
2 VERS 5.5.1
2 FORM LINEAGE-LINKED
1 CHAR ASCII
0 @U@ SUBM
1 NAME X
0 TRLR
```

Every valid file requires:

- A HEAD(er) record
- A SOUR(ce) line
- A SUBM(itter)
- A GEDC(OM) spec
- A CHAR spec
- A TRLR line

Source: *http://www.tamurajones.net/TheSmallestGEDCOMFile.xhtml*

1. Reading the Header

a. GEDCOM Version Number

GEDCOM Version

(1 GEDC 2 VERS xxx)

- 5.3 (14 – FTW 1.01 to 3.40, Family Origins 1 to 3)
- 5.4 (4 – Family Origins 4.0)
- 5.5 (much less than 60%) – many are 5.5.1
- 5.5.1 (15% plus those 5.5's that are 5.5.1's)
- 5.5 EL (3 – PCAhnen 2004 – 2006)
- 5.6 (1 – Tim Forsythe - timforsythe.com)

The Version Number may Lie

Of 413 files claiming GEDCOM 5.5:

- 71 have CHAR UTF-8 (mostly PAF)

GEDCOM 5.5.1 added these tags:

EMAIL, FAX, FACT, FONE, ROMN, WWW, MAP, LATI, LONG

- MyHeritage, FTB (5.5) uses EMAIL, FAX, WWW
- FTM, Pro-gen, PhpGedView, BK, PAF, … (5.5): EMAIL
- RootsMagic Vers 2 & 3 (5.5) uses MAP, LATI, LONG

Cannot always rely on the GEDCOM Version Number

- GEDCOM Earlier Versions
  - 1.0 (2 – Anstfile)
  - 1.2.3 (1 – a test file called “all gedcom 5.5.ged”)
  - 4+ (1 - RootsIV 1.1)
  - 4.0 (~20 – Ancestry 1.0, FamRoots 4.3, EasyTreeV5.2)
  - 5.0 (1 – Reunion V4.0)
  - 5.01 (5 – Reunion V3.0, 3.0c, V4.0, Ancestory)
  - 5.2 (1 – CFTree 1.0)

These specifications are not available

FTW Text Files

- FTW TEXT (3 – FTW)
- FTW TEXT 5.3 (2 – FTW 1.0, FTW 3.00)
- FTW TEXT 5.5 (9 – FTW 4 to 9, FTM 13 and 16)

if you search Google for:

"0 header" filetype:ged

there are 7 results.

```
0 HEADER
1 SOURCE FTW
1 DESTINATION FTM
1 DATE 1 Mar 1999
1 CHARACTER ANSI
1 FILE C:\PROGRA~1\FTW\FRASER3.GED
0 @I001@ INDIVIDUAL
1 NAME James Edwin /Fraser/,Jr.
1 SEX M
1 BIRTH
2 DATE 30 Aug 1949
2 PLACE Rochester, NY
1 FAMILY_SPOUSE @F01@
1 FAMILY_CHILD @F02@
```

- GEDCOM Missing Version Number (15%)
  Likely GEDCOM 3.0

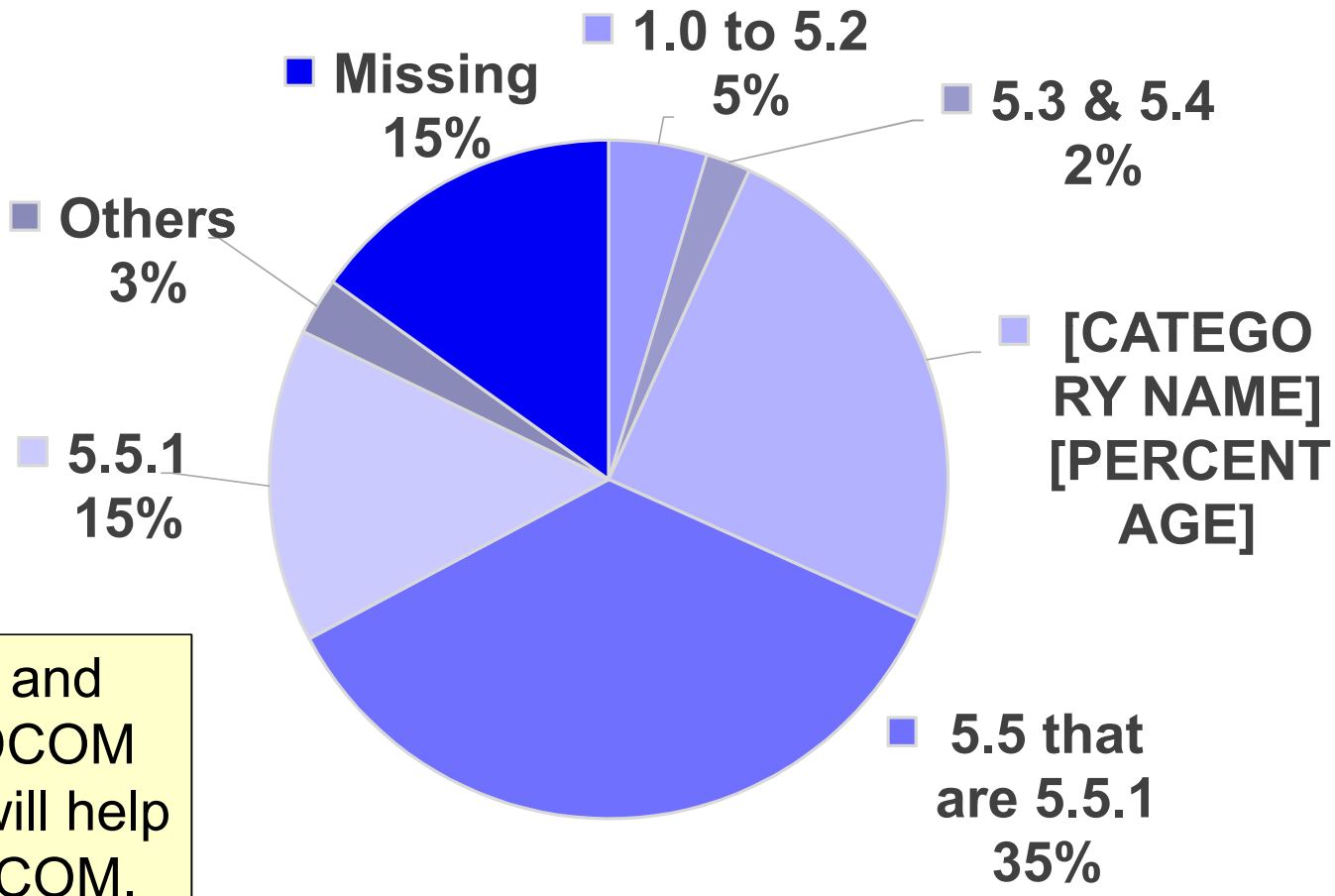
PAF up to 2.3.1, Brothers Keeper up to 5.2, and TMG with DEST (Destination) = DISKETTE and others exclude the GEDC and VERS lines.

Legacy 3.0/4.0 left the GEDCOM version blank

- but Legacy 2.0/2.0.1 says VERS 5.5

```
0 HEAD
1 SOUR Legacy
2 VERS 3.0
…
1 GEDC
2 VERS
```

GEDCOM Version Numbers



When it's there and correct, the GEDCOM Version Number will help to read the GEDCOM.

1. Reading the Header

b. Program Name and Version Number

The Program (1 SOUR xxx 2 VERS xxx)

My test files include:

- ~ 100 different programs
- ~ 200 different program/version combos

Likely > 500 programs that write GEDCOM

You need to use SOUR and VERS
to customize your input action
for certain programs.

Version Number Abuse (1 - 15 chars)

SOUR	VERS
FTW	← VERS tag not included under SOUR (it is optional)
FTW	1.0
FTW	7.00
FTW	11.0
FTW	Family Tree Maker 2005 (12.0.337) July 30, 2004
FTW	Family Tree Maker 2005 (12.0.345 SP1) August 20, 2004
FTW	Family Tree Maker (13.0.281)
FTW	Family Tree Maker (16.0.350)
FTM	Family Tree Maker (17.0.0.440)
FTM	Family Tree Maker (22.0.0.1243)

The NAME of the program can be 90 characters, but not VERS.

See: *http://www.tamurajones.net/EarlyLookAtFTM2008Beta.xhtml*

1. Reading the Header

c. FORM and CHARacter

2 FORM LINEAGE-LINKED

- Every single GEDCOM must include it.
- GenoPro has “LINAGE-LINKED”
- Legacy <V6 and EasyTree V8 have “LINEAGE_LINKED”

FORM only has one valid value.
It can be checked or ignored.

See: http://www.tamurajones.net/GEDCOMForm.xhtml - GEDCOM Form

Valid Character Sets/Encodings

- ASCII (10%) (1 CHAR xxx)
- ANSEL (20%)
- UNICODE (17)
  - UNICODE introduced in GEDCOM 5.3
- UTF-8 (20%)
  - UTF-8 introduced in GEDCOM 5.5.1
  - But half of my examples are UTF-8 with GEDCOM 5.5 (including PAF 5.2, GRAMPS 3, GenoPro 2, Reunion 9, AncestralQuest 12, PhpGedView 3.3)

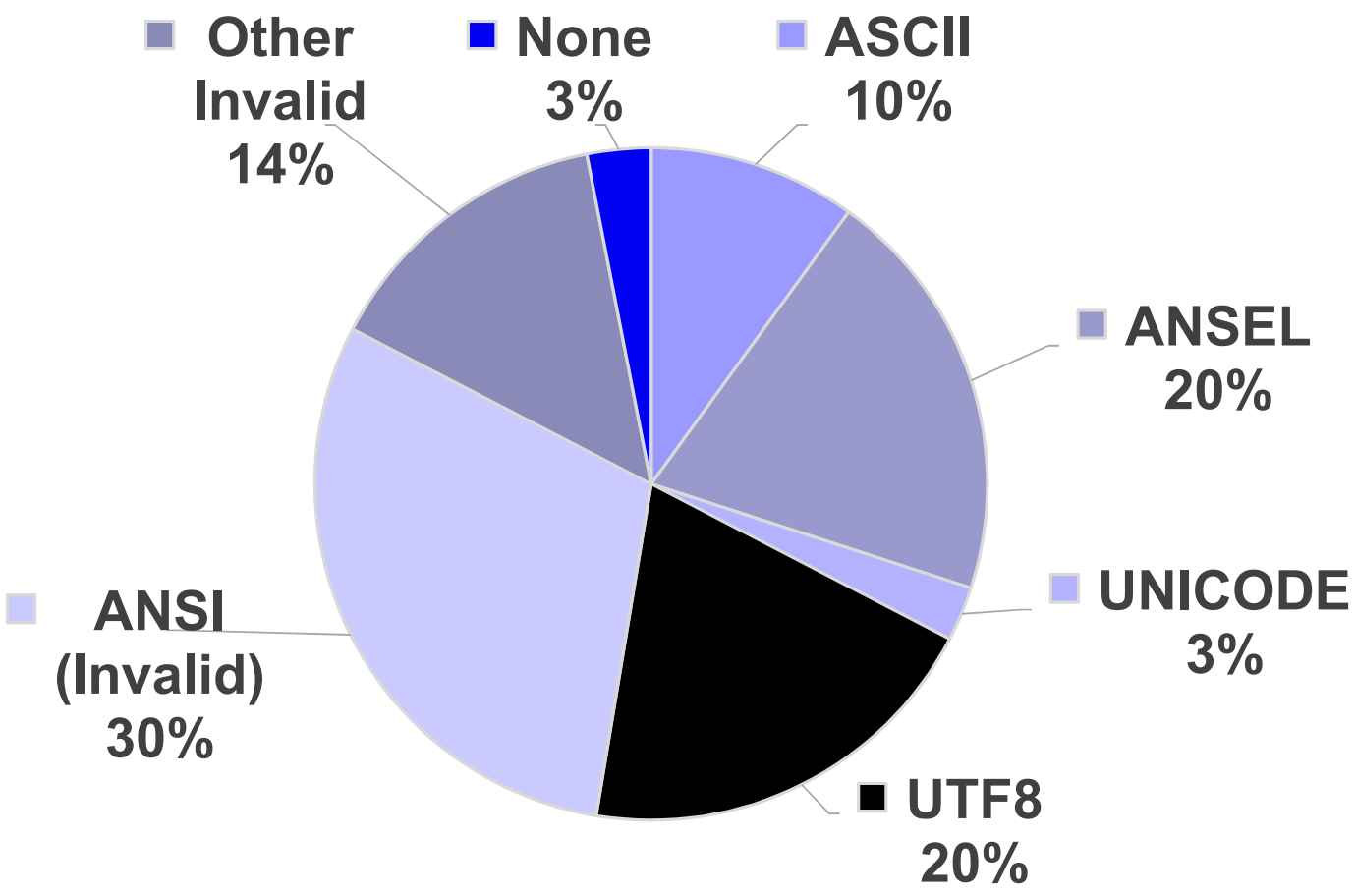
If you find UTF-8 with 5.5,
process it as the 5.5.1 file it really is

Invalid Character Sets

- None (~20)
- ANSI (30% - many different programs)
- IBM (1 – Reunion V3.0)
- IBM WINDOWS (~20 – Reunion V4.0, EasyTree)
- IBM_WINDOWS (2 – EasyTree)
- IBMPC (5% - Brothers Keeper, early FTW)
- CP1252 (1 – Lifelines)
- ISO8859 (1 – Genealogica Graphica)
- LATIN1 (2 - GenealogyJ)
- MACINTOSH (2- Reunion)

I use Encoding.GetString to interpret these

GEDCOM Character Sets



2. Structural Problems

GEDCOM validator: Reject all errors.

Choose your level

Behold's philosophy: Try to handle everything.

Byte Order Marks (BOM)

Valid with CHAR:

- UTF-8 with BOM (130)
- UTF-8 without BOM (23)
- little-endian Unicode (11) (PAF, GenealogyJ, GENprofi)
- big-endian Unicode (2) (MacFamilyTree)
- UNICODE without BOM (9)

Invalid with CHAR:

- ASCII / ANSEL / ANSI with UTF-8 BOM (6)

When CHAR mismatches BOM, use BOM

HEADing off the Wrong Way

- No header record
  - Non-GEDCOM files but with .ged extension

    Give an error for these

  - Test files by developers
  - Partial files – damaged by accident

    Try to process these,
    or reject if you choose to

Embedded GEDCOM files

- These don’t start with “0 HEAD”.
- Saving .ged webpages sometimes does this.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML><HEAD>
<META content="text/html; charset=windows-1252" http-equiv=Content-
Type></HEAD>
<BODY><PRE>0 HEAD
1 SOUR FAMILY_HISTORIAN
2 VERS 3.0
2 NAME Family Historian
…
0 TRLR
</PRE></BODY></HTML>
```

Try to extract the GEDCOM.
The “0 HEAD” might not start a line.

Empty Header Record

```
0 HEAD
0 @I0@ INDI
1 NAME Jacoba Adriana Johanna/Beijnen/
1 SEX F
1 BIRT
2 DATE 13-10-1876
2 PLAC Voorburg, Zuid-Holland, Netherlands
0 @I1@ INDI
1 NAME Cornelis Marius/Viruly/
1 SEX M
1 BIRT 2 DATE 11-11-1875
2 PLAC Vuren, Gelderland, Netherlands
1 DEAT
2 DATE 23-9-1938
2 PLAC Amsterdam, Noord-Holland, Netherlands
0 @F1@ FAM
1 WIFE @I0@
1 HUSB @I1@
1 MARR Y
0 TRLR
```

- You've got nothing to go on
- But there is data

Assume GEDCOM 5.5.1
But be flexible

See: *http://www.tamurajones.net/WieWasWieGEDCOM.xhtml*

Indenting / Blank Lines

"Some systems output indented GEDCOM data for better readability by putting space or tab characters between the terminator and the level number of the next line to visibly show the hierarchy. Also, some people have suggested allowing extra blank lines to visibly separate physical records. GEDCOM files produced with these features are ***not to be used*** *when transmitting GEDCOM to other systems"*
– GEDCOM 5.5, 5.5.1

```
0 HEAD
  1 SOUR FTW
    2 VERS 5.00
    2 NAME Family Tree Maker for Windows
    2 CORP Broderbund Software, Banner Blue Division
      3 ADDR 39500 Stevenson Pl.  #204
        4 CONT Fremont, CA 95439
      3 PHON (510) 794-6850
  1 DEST FTW
  1 DATE 18 FEB 2001
  1 CHAR ANSI
```

Process it anyway

28 of my sample files have blank lines in them.

Some people encourage indenting and provide methods to make it easier.

Booo!

A smart editor can show text Indented without adding spaces or tabs to the file.

Editing GEDCOM files with vim

March 5th, 2012 Gabriel Burca Go to comments Leave a comment

GEDCOM files sometimes have no indentation. That makes it difficult to read or edit them with a text editor. Using the following simple instructions, you can auto-indent the file so that it is more readable in the vim editor.

```
0 @PERSON1@ INDI
1 NAME Joseph Tag /Torture/
2 NPFX Prof.
2 SOUR @SOURCE1@
3 PAGE 42
2 NOTE These are notes about the first NAM
3 CONC embedded in the INDIVIDUAL record i
1 SEX M
1 BIRT
2 DATE 31 DEC 1965
2 PLAC Salt Lake City, UT, USA
2 ADDR St. Marks Hospital
3 CONT Salt Lake City, UT
3 CONT USA
2 AGNC none
2 OBJE
3 TITL Link to multimedia file
3 FORM tiff
3 FILE ImgFile.TIF
2 SOUR @SOURCE1@
3 PAGE 42
3 QUAY 2
3 NOTE Some notes about this birth source
4 CONC structure itself.
2 NOTE @N8@
2 FAMC @PARENTS@
```

```
0 @PERSON1@ INDI
    1 NAME Joseph Tag /Torture/
        2 NPFX Prof.
        2 SOUR @SOURCE1@
            3 PAGE 42
        2 NOTE These are notes about the f
            3 CONC embedded in the INDIVID
    1 SEX M
    1 BIRT
        2 DATE 31 DEC 1965
        2 PLAC Salt Lake City, UT, USA
        2 ADDR St. Marks Hospital
            3 CONT Salt Lake City, UT
            3 CONT USA
        2 AGNC none
        2 OBJE
            3 TITL Link to multimedia file
            3 FORM tiff
            3 FILE ImgFile.TIF
        2 SOUR @SOURCE1@
            3 PAGE 42
            3 QUAY 2
            3 NOTE Some notes about this b
                4 CONC structure itself.
        2 NOTE @N8@
        2 FAMC @PARENTS@
```

Compare the readability of the two formats

Save the next few lines into a file called `gedcom_indent`. Make the file executable and place it somewhere in your path.

```
#!/usr/bin/env python from __future__ import print_function import sys
 for line in sys.stdin:
    line = line.lstrip()
    try:
        level = int(line.split(' ', 1)[0])
        print('\t' * level, end='')
    except:
        pass
    print(line, end='', sep='')
```

Now add the following 2 lines to your ~/.vimrc file:

```
autocmd BufReadPost,FileReadPost *.ged %!gedcom_indent
autocmd FileType gedcom set foldmethod=indent nolist ts=4
```

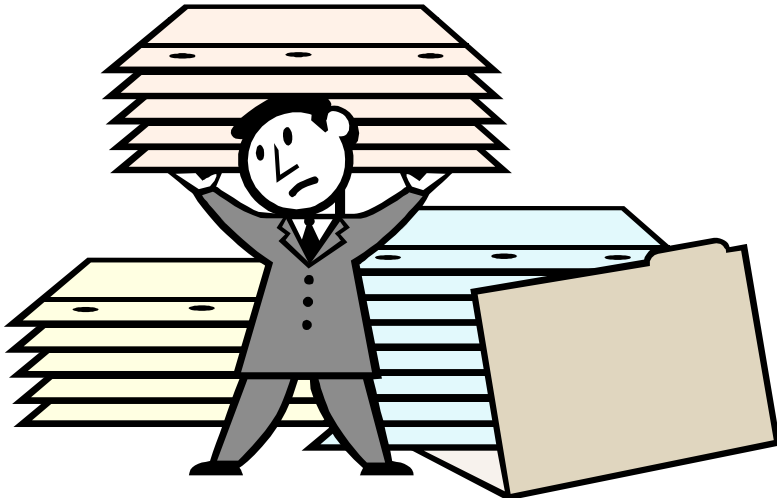
This tells vim to filter any file with a "ged" extension through the small gedcom_indent filter (which will add leading tabs to the file). The second line tells it to make each TAB count for 4 spaces, and to fold based on indentation.

Length Limits

- 38 of my sample files have lines longer than 255 characters in them.
- Lots of line items exceed their maximum length specified in GEDCOM.

Do not restrict lengths when reading or you risk losing data.

3. Level 0 Records



Top Level 0 Records (664)

0 HEAD (655)

0 TRLR (612)

0 @I1@ INDI (652)

0 @F1@ FAM (613)

0 @S1@ SOUR (291)

0 @R1@ REPO (154)

0 @O1@ OBJE (25)

0 @SUB1@ SUBM (submitter – reqd - 439)

0 @C1@ SUBN (submission – opt - 48)

0 @N1@ NOTE note text (155)

Level 0 Event Record

(never encountered)

- In GEDCOM 5.3. Eliminated in 5.4

```
0 @EV13@ EVEN
   1 TYPE CHR
      2 DATE 17 NOV 1830
      2 PLAC Littlehampton, West Sussex, England
         3 ADDR 9 Chiltern Close
            4 CONT East Preston
      2 @EV13!1@ CHIL
         3 NAME Jason \Wilde\
         3 AGE 4 yrs
      2 @EV13!2@ MOTH
         3 NAME Wilma \Wilson\
         3 BIRT
            4 DATE 15 MAY 1810
            4 PLAC Nottingham, England
```

"This context was intended to support the evidence record concept … which ended up being more complicated than first supposed … requires further study."
- GEDCOM 5.4, 5.5, 5.5.1

Yes, the sample code shown in GEDCOM 5.3 is indented, in violation of itself

Level 0 Place Records

- GEDCOM includes PLAC as a tag, but not as a record.
- Some programs have added place records
- _PLAC used by RootsMagic
- _PLAC_DEFN used by Legacy
- _LOC used by GEDCOM EL (extended locations)

See: http://www.beholdgenealogy.com/blog/?p=899 – The Place Record in GEDCOM

RootsMagic

```
0 _PLAC Ballyduff, Kerry, Ireland
1 MAP
2 LATI N52.3044444
2 LONG W9.4044444
```

```
1 BIRT
2 PLAC Ballyduff, Kerry, Ireland
```

Legacy

```
0 _PLAC_DEFN
1 PLAC Manila, , , Philippines
2 ABBR Manila, , , Philippines
2 MAP
3 LATI N14.5862344444444
3 LONG E120.992484444444
```

```
1 BIRT
2 PLAC Manila, , , Philippines
```

All of this for basically just for a latitude and longitude.

These structures require custom programming

GEDCOM EL (extended locations)

```
1 GEDC
2 VERS 5.5
3 _EXTENDED_LOCATIONS
2 FORM LINEAGE-LINKED
```

PC Ahnen puts in the HEAD the _EXTENDED_LOCATIONS line.

But for others, there's no simple way to identify GEDCOM EL.

To handle GEDCOM EL
You have to handle its constructs in any file

High-Level-Entitiy "Location"

The High Level Entity "Location" is introduced of following structure:

```
0 @P65@ _LOC
   1 NAME <PLACE_NAME> 1:M
      2 DATE <DATE_VALUE> 0:1
      2 SOUR <SOURCE_CITATION>
      2 NAMC <PLACE_NAME_ADDITION> 0:1
      2 _LANG <LANGUAGE> 0:1
   1 _FPOST <FOKO_POSTCODE> 0:M
      2 DATE <DATE_VALUE> 0:1
   1 POST <POSTCODE> 0:M
      2 DATE <DATE_VALUE> 0:1
      2 SOUR <SOURCE_CITATION> 0:M
   1 _FSTAE <FOKO_TERRITORY_IDENTIFIER> 0:1
   1 _FCTRY <FOKO_STATE_IDENTIFIER> 0:1
   1 _FOKOID <FOKO_IDENTIFIER> 0:1
   1 MAP <GEOGRAPHICAL_COORDINATES> 0:M
      2 TYPE <GEOGRAPHICAL_COORDINATE_TYPE> 1:1
   1 EVEN <EVENT_DETAIL> 0:M
   1 _LOC @<XREF:_LOC>@ 0:M
      2 TYPE <HIERARCHICAL_RELATIONSHIP> 1:1
      2 DATE <DATE_VALUE> 0:1
      2 SOUR <SOURCE_CITATION> 0:M
   1 _DMGD <DEMOGRAPHICAL_DATA> 0:M
      2 DATE <DATE_VALUE> 0:1
      2 SOUR <SOURCE_CITATION> 0:M
      2 TYPE <TYPE_OF_DEMOGRAPICAL_DATA> 1:1
   1 _AIDN <ADMINISTRATIVE_IDENTIFIER> 0:M
      2 DATE <DATE_VALUE> 0:1
      2 <SOURCE_CITATION> 0:M
      2 TYPE <TYPE_OF_ADMINISTRATIVE_IDENTIFIER> 1:1
   1 <<MULTIMEDIA_LINK>> 0:M
```

See: http://www.tamurajones.net/DetectingGEDCOM5.5EL.xhtml
See: http://wiki-en.genealogy.net/Gedcom_5.5EL - GenWiki

To improve the readability lines are inserted. This should (and will) not be performed at real Gedcom-output.

PCAHNEN

```
0 @P13@ _LOC
1 NAME Horst bei Elmshorn
1 _FCTRY D
1 POST 25358
1 _FSTAE SH
1 _FOKOID HORRSTJO43TT
1 MAP JO43TT
2 TYPE MAIDENHEAD
1 NOTE letzte Änderung: 17.05.2007
2 CONT Gemeinde im Amt Horst, Kreis Steinburg, Schleswig-Holstein, Bundesrepublik Deutschland
2 CONT Postleitzahl: 25358
2 CONT GOV-Kennung: HOREINJO43TT
```

```
1 MARR
2 TYPE RELI
2 DATE 03 MAR 1903
2 PLAC Horst bei Elmshorn
3 _LOC @P13@
```

Other Level 0 Records

- RootsMagic: _EVDEF
- Legacy: _EVENT_DEFN, _TODO
- GenoPro: BOTTOM, CONTACT, DATE, EDUCATION, GENOMAP, GLOBAL, LABEL, MARRIAGE, OCCUPATION, PEDIGREELINK, PICTURE, PLAC, SHAPE, SIZE, SOCIALENTITY, TITLE, TWIN, _INDI

It's best to build a generalized
Level 0 record,
so you can handle anything

4. The CONC Tag

Th e Horr ible CONC t ag

- Confusion due to Specification Flip Flop and incorrect examples

CONC: “An indicator that the additional value information follows and is to be connected to the value of the superior preceding line without a new line.” – GEDCOM 5.3

That’s all the GEDCOM 5.3 said. No example was given in GEDCOM 5.3.

Update in GEDCOM 5.4

The following example was given in GEDCOM 5.4:

```
2 SOUR Waters, Henry F., Genealogical Gleanings in England: Abstracts of
3 CONC Wills Relating to Early American Families. 2 vols., reprint 1901, 1907.
3 CONC Baltimore: Genealogical Publishing Co., 1981.
3 CONC Stored in Family History Library book 942 D2wh; films 481,057-58
3 CONC Vol 2, page 388.
```

This implies you would add a space when concatenating, and display it as:

Source: Waters, Henry F., *Genealogical Gleanings in England: Abstracts of Wills Relating to Early American Families. 2 vols., reprint 1901, 1907. Baltimore: Genealogical Publishing Co., 1981. Stored in Family History Library book 942 D2wh; films 481,057-58 Vol 2, page 388.*

However, in GEDCOM 5.5.1

The example was changed to the following:

```
2 SOUR Waters, Henry F., Genealogical Gleanings in England: Abstracts of W
3 CONC ills Relating to Early American Families. 2 vols., reprint 1901, 190
3 CONC 7. Baltimore: Genealogical Publishing Co., 1981.
3 CONT Stored in Family History Library book 942 D2wh; films 481,057-58 Vol 2, pa
3 CONC ge 388
```

This implies you would NOT add a space when concatenating, and display it as:

Source: Waters, Henry F., *Genealogical Gleanings in England: Abstracts of Wills Relating to Early American Families. 2 vols., reprint 1901, 1907. Baltimore: Genealogical Publishing Co., 1981.*
Stored in Family History Library book 942 D2wh; films 481,057-58 Vol 2, page 388

GEDCOM 5.4 tried to clarify…

"*The information from the CONC value is to be connected to the value of the superior preceding line without a carriage return and/or new line character. If a space is to be inserted between the end of the previous value and the CONC value then the space must be the first character of the CONC value because many GEDCOM values are trimmed of trailing spaces. – GEDCOM 5.4*

The GEDCOM 5.4 example is wrong according to this definition.

In the example, they split the line at a space.

But this has an additional problem. Line_values cannot begin with a space.

And clarify… (GEDCOM 5.5)

"The information from the CONC value is to be connected to the value of the superior preceding line without a space and without a carriage return and/or new line character. Values that are split for a CONC tag must always be split at a non-space. If the value is split on a space the space will be lost when concatenation takes place. This is because of the treatment that spaces get as a GEDCOM delimiter, many GEDCOM values are trimmed of trailing spaces and some systems look for the first non-space starting after the tag to determine the beginning of the value." – GEDCOM 5.5

But they included the example unchanged from GEDCOM 5.4

– splitting the line at a space. Whoops!

In GEDCOM 5.5.1

They eliminated the detailed CONC description they had in 5.5.

They included only the following:

“The CONC tag assumes that the accompanying subordinate value is concatenated to the previous line value without saving the carriage return prior to the line terminator. If a concatenated line is broken at a space, then the space must be carried over to the next line.”

So 5.5.1 now, like GEDCOM 4.0, allows it either way, at a space or non-space.

And they changed the example to the one that splits the line at a non-space.

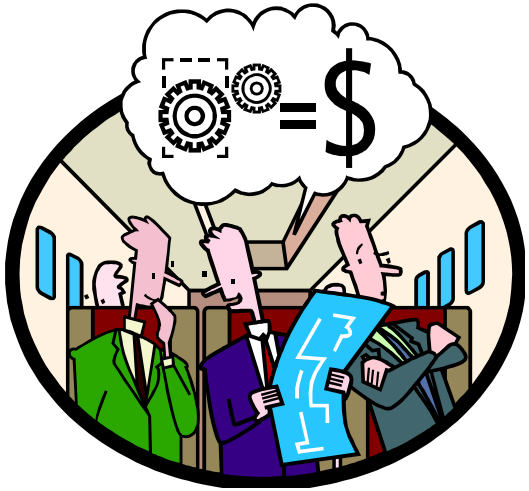
Still: a line_value in a gedcom_line cannot begin with a space, can it?

Specification versus Example

Vers	Specification	Example	
5.3	Connect line to next line.	No example	
5.4	Connect line to next line. If space wanted, it starts CONC value.	The lines break at a space	✗
5.5	Connect line to next line. Always split at non-space.	The lines break at a space	✗
5.5.1	Connect line to next line. If space wanted, it starts CONC value.	The lines break wi thin a word	OK

CONCfusion

- Most programs decided to break at a non-space.
- A few, like TMG, has an option to output CONC either way.
- All in all,
  mass confusion reigns.



If GEDCOM Splits at a Space

```
2 SOUR Waters, Henry F., Genealogical Gleanings in England: Abstracts of
3 CONC Wills Relating to Early American Families. 2 vols., reprint 1901, 1907.
3 CONC Baltimore: Genealogical Publishing Co., 1981.
3 CONC Stored in Family History Library book 942 D2wh; films 481,057-58
3 CONC Vol 2, page 388.
```

If assuming split at non-space, i.e. no space added when concatenating:

Source: Waters, Henry F., *Genealogical Gleanings in England: Abstracts ofWills Relating to Early American Families. 2 vols., reprint 1901, 1907.Baltimore: Genealogical Publishing Co., 1981.Stored in Family History Library book 942 D2wh; films 481,057-58 Vol2, page 388.*

If assuming split at space, i.e. space added when concatenating:

Source: Waters, Henry F., *Genealogical Gleanings in England: Abstracts of Wills Relating to Early American Families. 2 vols., reprint 1901, 1907. Baltimore: Genealogical Publishing Co., 1981. Stored in Family History Library book 942 D2wh; films 481,057-58 Vol 2, page 388.*

If GEDCOM Splits at Non-space

```
2 SOUR Waters, Henry F., Genealogical Gleanings in England: Abstracts of W
3 CONC ills Relating to Early American Families. 2 vols., reprint 1901, 190
3 CONC 7. Baltimore: Genealogical Publishing Co., 1981.
3 CONT Stored in Family History Library book 942 D2wh; films 481,057-58 Vol 2, pa
3 CONC ge 388
```

If assuming split at non-space, i.e. no space added when concatenating:

Source: Waters, Henry F., *Genealogical Gleanings in England: Abstracts of Wills Relating to Early American Families. 2 vols., reprint 1901, 1907. Baltimore: Genealogical Publishing Co., 1981.*
Stored in Family History Library book 942 D2wh; films 481,057-58 Vol2, page 388.

If assuming split at space, i.e. space added when concatenating:

Source: Waters, Henry F., *Genealogical Gleanings in England: Abstracts of W ills Relating to Early American Families. 2 vols., reprint 1901, 190 7. Baltimore: Genealogical Publishing Co., 1981.*
Stored in Family History Library book 942 D2wh; films 481,057-58 Vol 2, pa ge 388

So What the CONC Do We Do?

- For these programs, assume split at space:

  HEAD.SOUR = AncestQuest, CFTree, FamilyOrigins, FamTiesDlx, FamTreesQE, FTM

- For others, assume split at non-space

- You could inspect line breaks and guess. (Nah!)

- Allow user to override if it is done wrong.

See: http://www.beholdgenealogy.com/blog/?p=739 – CONC Me On The Head

5. User-Defined Tags

User-defined Tags

"We do not encourage the use of user-defined tags. Applications requiring the use of non-standard tags should define them with a leading underscore so that they will not conflict with future GEDCOM tags."

"Systems that read user-defined tags must consider that they have meaning only with respect to a system contained in the HEAD.SOUR context."

???

Try using HEAD.DEST, else HEAD.SOUR to interpret tags

The _UID Tag

- Never was in GEDCOM
- But PAF has it, so many others followed
- UID without _ (Ancestry.com Family Tree)
- In 183 of my test files

```
1 _UID FD43E4D58EBE47298D58627884A58F8CB82C
```

You should handle this.

See: http://www.tamurajones.net/The_UIDTag.xhtml

Level 1 Schema

(encountered in files from FTW and GENprofi)

- In GEDCOM 5.3. Eliminated in 5.4

```
1 SCHEMA
2 INDI
3 _FA1
4 LABL Fact 1
...
3 _MREL
4 LABL Relationship to Mother
…
2 FAM
3 _FA1
4 LABL Marriage fact
…
3 _MSTAT
4 LABL Marriage Beginning Status
```

"Although the schema concept is valid and essential to the growth of GEDCOM, it is too complex and premature to be implemented successfully into current projects"
- GEDCOM 5.4, 5.5, 5.5.1

Schema-defined tags in use

```
0 @F03@ FAM
1 HUSB @I187@
1 WIFE @I201@
1 CHIL @I202@
2 _FREL Natural
2 _MREL Natural
1 MARR
2 DATE 3 Feb 1979
2 PLAC Gladwyne, Montgomery Co, PA
1 _FA1
2 DATE 1991
1 _MEND Divorce
```

Don’t handle the schema.
If you want, handle the common schema-defined tags.

Why Oh Why?

Reunion

```
1 BIRT
2 DATE 25 MAY 1829
2 PLAC Blackerstone, Longformacus, Berwickshire, Scotland
2 SOUR @S61@
2 SOUR @S1690@
1 _BTH
2 DATE 25 MAY 1830
2 PLAC Blackerstone, Longformacus, Berwickshire, Scotland
2 NOTE Year is in error - should be 1829
```

Reunion 9.0 (2007): *http://roger.lisaandroger.com/WilliamMoffat.ged*

Reunion's User-Defined Tags

DATV, FRAM, ELEC, HEAL, LOCA, REPT, URL, …
_ALT, _AWD, _BTH, _DTH, _EMI, _GRO, _HAM, _HAN, _JOI, _MED, _MVR, _OBT, _PAG, _PRIM, _SEC, _SIZE, _TYP, _TYPE, _UID, _WIL, …

```
0 @S329@ SOUR
1 TYPE Newspaper
1 TITL Obituary of Ellen Houliston
1 AUTH author unknown
1 PERI from The Clutha Leader
1 PLAC Balclutha, Otago, New Zealand
1 DATE after 15 May 1919
1 _MED photocopy of transcription from newspaper
1 LOCA filed under Source 329
```

```
0 @I16@ INDI
1 NAME William /Moffat/
…
1 ORIG 3
1 _HAN @N357@
1 _EMI @N18@
1 _WIL @N415@
1 _OBT @N418@
```

Reunion is one of a few programs that lets users define their own custom tags, so anything goes.

Getting Carried Away with User-Defined Tags

GenoPro:

_XREF, _INDI (in addition to an INDI) with a host of non-standard subtags under them:

ACTION, BOTTOM, BOUNDARYRECT, COLORS, DISPLAY, FAMC, FAMS, GENDER, GENOMAP, HYPERLINK, INDIVIDUALINTERNALHYPERLINK, LABEL, NAME, POSITION, SYMBOL, Z

User-Defined Tags - GenoPro

```
0 @ind00005@ INDI
1 BIRT
2 CEREMONYTYPE Baptism
2 DATE 18 JAN 1905
2 PLAC Solbjerg sogn, Løve herred, Holbæk amt (Country: Danmark)
3 _XREF @place00020@
```

```
0 @place00020@ PLAC
1 NAME Solbjerg sogn, Løve herred, Holbæk amt
1 COUNTRY Danmark
```

```
0 @ind00239@ _INDI
1 INDIVIDUALINTERNALHYPERLINK @ind00005@
1 NAME
2 DISPLAY Karen Marie  Bendixen
1 POSITION -980,-70
2 Z 120
2 GENOMAP Larsen
2 BOUNDARYRECT -1017,-37,-943,-131
```

RootsMagic's _TMPLT tag

```
0 @I100000@ INDI
1 NAME William /Ewing/
1 SOUR @S908@
2 _TMPLT
3 FIELD
4 NAME Page
1 BIRT
2 DATE ABT 1664
2 _SDATE 1 JUL 1664
2 PLAC Stirling,
Scotland
2 SOUR @S908@
3 _TMPLT
4 FIELD
5 NAME Page
```

```
0 @S908@ SOUR
1 TITL Burt, Dorothy Cook, Rootsweb GEDCOM
1 _SUBQ Burt, Dorothy Cook, Rootsweb GEDCOM
1 _BIBL Burt, Dorothy Cook. Rootsweb GEDCOM.
1 _TMPLT
2 TID 0
2 FIELD
3 NAME Footnote
3 VALUE Burt, Dorothy Cook, Rootsweb GEDCOM
2 FIELD
3 NAME ShortFootnote
3 VALUE Burt, Dorothy Cook, Rootsweb GEDCOM
2 FIELD
3 NAME Bibliography
3 VALUE Burt, Dorothy Cook. Rootsweb GEDCOM.
```

User-Defined Tags - Family Historian

```
0 @I6@ INDI
1 NAME Emma Kathleen /Wright/
2 _USED Kate
```

```
1 EMIG
2 DATE 19 OCT 1956
2 PLAC Liverpool, Lancashire, England
2 AGE 32y
2 NOTE Sailed on the Empress of Scotland
2 _PLAC Montreal, Quebec, Canada
```

```
1 FAMC @F728@
2 _PEDI Step
```

```
1 NAME Elizabeth /Crocker/
2 _AKA Liz Crocker
2 _MARNM Elizabeth Price
```

Handling User-Defined Tags

- Read any tag
- Warn if non-standard but not user-defined
- Dumb but easy method:

  Allow user to specify text to display
- Smart but hard method:

  Custom program what you can

See: http://www.beholdgenealogy.com/blog/?p=876 – A Plethora of Extra GEDCOM Tags

6. Odds and Ends

Witness Tag (WITN)

- Eliminated in GEDCOM 5.4
- In 8 of my example files.
- They replaced it with _WITN

  (GENBOX, PCAHNEN, GESW, PRO-GEN)

```
1 BIRT
2 DATE 3 AUG 1780
2 _WITN @I4@
3 _ROLE Saw birth
2 _WITN @I3@
3 _ROLE Godmother
```

You should handle this.

The Association_Structure

- Added in 5.5. Does what WITN did & more
- In 29 of my sample files
- Not “wrong” GEDCOM, but “new” GEDCOM

```
1 BIRT
2 DATE 3 AUG 1780
1 ASSO @I4@
2 RELA Saw birth
1 ASSO @I3@
2 RELA Godmother
```

You should handle this.

Embedded Characters in Line Values

Hex 0B (a line feed), Hex 09 (a tab), …

(in 23 files, Legacy, RootsMagic and PAF)

These need to be detected and handled so they don't wreck the display of your reports.

Hex 00 (end of string) - RootsMagic

- Problem if reading text file in as string
- Need to get the file size and load a buffer

See: http://www.beholdgenealogy.com/blog/?p=1070 – Literally Nothing from RootsMagic

The horrors of Markup

RootsMagic – HTML in notes

```
2 NOTE In a publication of the Mahoning County Chapter of the OGS ("Springfiel
3 CONC d Township Veterans before the World Wars: Petersburg Cemetery) - found a
3 CONC t www.<b>mahoningcountychapterogs</b>.org/<b>Veterans</b>%20Page/...  t
3 CONC he following information is listed:
3 CONT <b>Welk</b>, Henry A. <u><i>Civil War</i></u> Infantry Private Company D
3 CONC ?, 196th OVI 759Fairgreen Ave, Youngstown, Ohio, born 14 February 1835 L
3 CONC ittlestown,Pennsylvania - died 10 May1923 in Youngstown, Next of kin: H
3 CONC elen<b> Welk</b>
```

EasyTree (Sierra On-Line)

```
0 @N95@ NOTE
1 CONC <a href="browse.cfm?dbname=royals&ID=337">Link to Royals
1 CONC Database</a>
```

See: http://www.beholdgenealogy.com/blog/?p=808 – Markup in GEDCOM

The horrors of Markup

Reunion – Preformatted information

```
2 PLAC Renton Cottage, Coldingham, Berwickshire, Scotland
2 NOTE <pre>Parish of Houndwood Quoad Sacra (formerly Coldingham) Page 5-6<br />Ren…
3 CONC 5   Ag Lab    Born in Berwickshire<br />Isabel Moffat         45             Born in Berwic…
3 CONC t      20   Ag Lab    Born in Berwickshire<br />Margaret Moffat       15   Ag Lab    Born…
3 CONC ffat          15   Ag Lab    Born in Berwickshire<br />William Moffat        10            Bor…
3 CONC eorge Moffat          8             Born in Berwickshire<br />Catherine Moffat       5
…
3 CONC </pre> .
```

Preformatted Text:

- It wants us to line up the information its way.
- That’s just not possible most of the time.
- Requires a fixed-width font (ugly, takes extra space, may not wrap well).
- May not fit properly in the space available for display.
- Lots of headaches.

The horrors of Markup

RootsMagic – And the special characters can even be encoded

```
2 NOTE Source Information:
3 CONT   &lt;b&gt;Census Place&lt;/b&gt;      District 3, Edmonson, Kentucky
3 CONT   &lt;b&gt;Family History Library Film&lt;/b&gt;
         &lt;../../library/fhlcatal
3 CONC og/supermainframeset.asp?display=filmhitlist&amp;columns=*%2C180%2C0&am
3 CONC p;filmno3 =1254411&gt;&lt;/u&gt;
```

User-Defined tags might be used:
(Legacy, PAF or AncestralQuest)

```
0 @S103@ SOUR
1 TITL 1841 Scotland Census
1 _ITALIC Y
1 _PAREN Y
```

Handling markup is hard.
Harder than you think.
Try it …if you're a masochist.

See: http://www.beholdgenealogy.com/blog/?p=808 – Markup in GEDCOM

Bad Dates

FTM

```
2 DATE BEF. 5 SEP 1647
2 DATE BET. 1547 - 1598
```

Easytree

```
2 DATE 1858/1878
```

PAF 4.0

Reunion

```
2 DATE STILLBORN
2 DATE abt 1950 (?)
2 DATE 22 JAN
2 DATE while a student at college
2 DATE NOT MARRIED
```

RootsMagic

```
2 DATE INT 1861 ()
```

See: Behold Version 1.0.4.3:
http://www.beholdgenealogy.com/blog/?p=1304

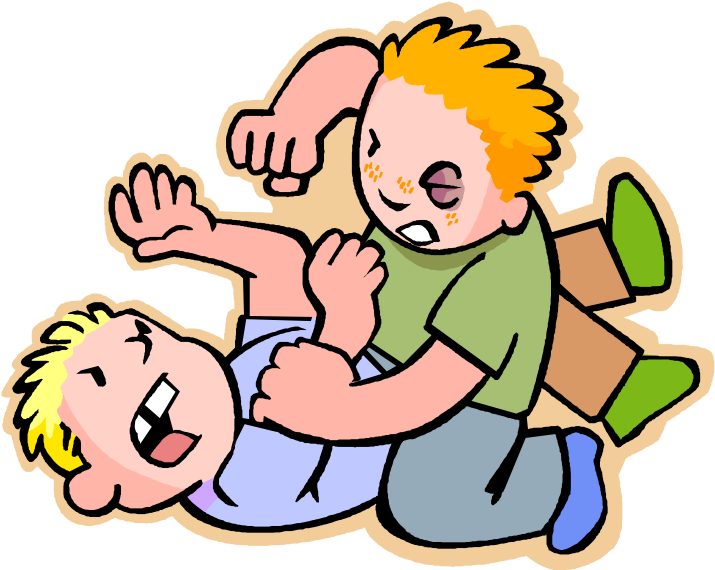
This overlaps with consistency checking.
Tackle it once you're there

A Sampling of Invalid Dates

- *29 FEB 1897 – 1 result, MyHeritage*
- *29 FEB 1903 – 2 results, Family Origins and PAF*
- *29 FEB 1906 – 1 result, PAF*
- *29 FEB 1909 – 1 result, Ancestry.com Family Trees*
- *29 FEB 1910 – 1 result, PAF*
- *29 FEB 1911 – 2 results, Family Origins and PAF*

DATE 30 FEB – 24 results, Brother's Keeper (3), PAF (11), Family Origins (3), RootsMagic, Legacy (4), BasGen and one stated to be from: AAAAAA (Eh what?)

- *DATE 31 FEB – 14 results, plus Family Treasures, Heredis and Pro-Gen 2*
- *DATE 31 APR – 40 results, plus EFTree and Holger*
- *DATE 31 SEP – 32 results, plus GenoPro*
- *DATE 31 NOV – 37 results, plus Ancestral File and CFTree*
- *DATE 32 – 1 result, EasyTree that gave 32 Dec 1841*

See: http://www.beholdgenealogy.com/blog/?p=896 – Out on a Bad Date

Discussion