ARTICLE

Genealogy, GEDCOM, and popularity implications

J. Tuomas Harviainen
University of Tampere
tuomas.harviainen@uta.fi
https://orcid.org/0000-0002-6085-5663

Bo-Christer Björk
Hanken School of Economics
bo-christer.bjork@hanken.fi
https://orcid.org/0000-0003-1545-9642

This paper examines the dominant genealogical file format GEDCOM and its implications. GEDCOM has come to influence the entire field of genealogy, including knowledge management and possibly even information seeking. The paper concludes with a note on the position of GEDCOM in genealogy, as well as with propositions on reasons for its popularity, and the ways in which a software format may in fact be a central contextual influence on some information practices.

Keywords: crowdsourcing; file formats; genealogy; information management

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

This short conceptual paper examines a file format of genealogical research, and the implications that format has for the practices of genealogy. Its central focus is on the way in which an essentially volunteer work such as genealogy can become standardised through the information systems and file formats used for it. Genealogy is an interesting case within the larger contexts of information seeking, categorisation, and cataloguing, in that it is a mostly volunteer work, creates results that are still useful hundreds of years later, and provides significant network benefits through information sharing. Central to this sharing process has been the adoption of one format, GEDCOM, as the de facto standard in the field (Gellatly, 2015). It functions as both the basis of the reporting and as the leading transfer standard used by information systems in

Persistent identifier: https://doi.org/10.23978/inf.76066

the field (e.g., "Ancestry," n.d.; "Genes reunited," n.d.). While it is not the only option, it dominates the area.

Systems that emerge to support information processes are technically embedded extensions of sociocultural practices (Land, 1992). It is therefore no surprise that genealogists have developed shared practices and collegial networks, as well as software for the task. What is a surprise is the extent to which they have done so (Yakel, 2004), and the way in which they tend to rely on each other more than on information professionals (Duff & Johnson, 2003; Yakel & Torres, 2007).

Information management is crucially important for genealogists (Friday, 2014; Yakel, 2004). It is therefore only logical that a basic system of coding has arisen, and GEDCOM had the advantage of early arrival and extended flexibility (in some of its facets). It is such a baseline now that new projects, even on a national level, are based on its use (e.g., the Finnish Suomi-tietokanta; "Suku forum thread on Suomi-tietokanta," 2014). Genealogy exemplifies the combination of layman crowdsourcing with structural system design. Genealogists were at first treated as "superficial" scientists by information professionals in e.g., the United States, but this trend has ceased (Yakel, 2004). This is a common phenomenon in connection to citizen science: its relationship to science done by scientists tends to fluctuate (Schrier, 2016). In the case of GEDCOM, however, the central support has come from a religious source, to which we turn next.

The unofficial standard: GEDCOM

This article answers the double question of *what exactly is GEDCOM, and what are its implications for genealogical research*. In the case of the former, we look at the format in the perspective of not just software review, but also as a kind of boundary object (as per Star & Griesemer, 1989) that enables and guides discourses around it. While allowing for direct 'translations' of the data and collaboration between practitioners from diverse background, because of its structure and limitations, the format also restricts interpretations. In Star and Griesemer's (1989) terms, it seeks to create coherence out of data that comes from many differing sources, by meeting the information requirements of each context in which it is used. The problem, however, is that it appears to meet only the *minimal* requirements of some of them, while leaving out many other factors of relevance to those contexts.

GEDCOM (.ged), short for Genealogical Data Communication, was first published in 1984. It was developed by The Church of Jesus Christ of the Latter-day Saints (LDS, for short), to work as a tool helping in their quest for

Listing 1: An individual-level GEDCOM file (by the authors)

```
0 @|1@ INDI
1 NAME John David /Phillips/
1 SEX M
1 BIRT 2 DATE 1 JAN 1800 2 PLAC London
1 DEAT 2 DATE 30 DEC 1855
2 PLAC London
1 FAMC @F1@ 1 FAMS @F2@
```

Listing 2: A family-level GEDCOM file (cited from Myllynen, 2007)

```
0 @F2@ FAM
1 HUSB @|1@
1 WIFE @|2@
1 CHIL @|3@
1 CHIL @|4@
1 MARR 2 DATE 1 JAN 1930
2 PLAC Jyväskylä
1 DIV 2 DATE 31 DEC 1960 2 PLAC Jyväskylä
```

genealogical data. Its key purpose is to enable both the coding of data and its transportation from one genealogical software to another without loss – even as loss may in fact take place because of e.g., conversion issues (Myllynen, 2007; see Sippu, 2000 for critique). The LDS's genealogical database, which since 1999 has been accessible online, unsurprisingly utilises GEDCOM as its basis (Mann, 1999; Mayfield & Brown, 1999).

The version currently most commonly in use is 5.5, developed and released in 1996 (Gellatly, 2015). After that time, it has not really been developed further. Standards like GEDCOM are created to help the processing of what Friday (2014) calls *genealogical facts*: the intersections of two or more pieces of historical information about names, dates, places and/or events. They assist in the coding, cross-referencing and aggregating of information. What makes GEDCOM remarkable is that it has never been defined as an official standard by any standardisation organisation, yet it is so prevalent that it is treated as a standard (Myllynen, 2007). It has received critique for both its age and format (Gellatly, 2015; Zandhuis, 2005), yet persists. It structures people into files of either individual or (nuclear) family types, with relations marked (Sippu, 2000, see listings 1–2 for examples). If one parent is unknown, multiple family files need to be made, one for each child.

Gellatly (2015, pp. 111–112) lists several advantages and critiques of the format:

> GEDCOM files are widespread as a means for storing and exchanging genealogical data. The reasons for this are clear:

- The format provides a systematic and standardised way of structuring information about individuals, their families and life events.
- Exchanging of family trees is a way that people may use to identify ancestral connections and expand their own family trees, so there is a demand for a common exchangeable file type.
- The files are in plain text format, which allows for easy development of software applications that can read or export to the files – hundreds of such applications have been developed.
- The format is flexible in terms of what can be added to a file, allowing users and software to easily make use of the format, even without having to conform to the correct standard specifications.

There are a number of criticisms of the GEDCOM format, in particular:

- There is a list of tags that can be used to describe events, but there is no allowance for additional tags, so unusual events typically have to be added in the notes. In some instances, people (or software) will use non-standard tags, which results in confusion or data loss when the file is read by a different person or software application. In these cases, the flexibility of the format is arguably problematic, even though the ability to add non-standard tags may have been useful at some point.
- There are no constraints over the data that may be entered under each tag. For example, it is possible to enter a date of birth from the future, or to enter e.g. '< 1900' or 'Born between 1900 and 1920' in a date field, which may not be understood by software. It is often the case that address identifiers are entered incorrectly, so that confusion may arise about geographical status, e.g. 'Washington' is both a city and a state in the US.
- here are technical constraints with the format, e.g. multiple people cannot be linked to the same record, so a single event, source citation or note may have to be replicated within each file. In technical terms, the data format is not normalised, leading to excessive replication of content and increased risk of data corruption.

Several other factors can both ease and complicate the process. For example, a single GEDCOM file can contain many people. This may for some purposes be very convenient, but prevents many forms of efficient data analysis and transfer. Since the format has no checks against typing errors or incomplete data (e.g., an author just inputting his own lineage nothing else), it is vulnerable to errors (Gellatly, 2015, p. 112). Two levels of the standard exist, the lower of which (the data format) is just a way of categorising information, while the higher (the Lineage-Linked GEDCOM Form) is specifically for genealogical information (Myllynen, 2007). This guides key use, even as it does not restrict other applications. Each field can be marked with a SOUR subfield, into which text on a citation source (e.g., a county registry) can be inserted, and a free text NOTE field is also available. Roughly put, GEDCOM may not be the best option available on the market, but its popularity makes it so dominant that it is difficult for other formats to enter the field. Here, we look at why this is so.

Method

Being mainly a metatheoretical work, this paper brings together strands of existing research (as per Galliers, 1992), which on this topic are still very rare. The sources that exist were usually aimed for use by information professionals, such as librarians and archivists (Yakel, 2004). They mostly focus on search strategies (Friday, 2012, 2014). Even as the combination of exiting works may bring forth important new data (Galliers, 1992), few articles on GEDCOM are available. In making this article, we have therefore sought not only to create new knowledge, but also to provide a reading list for future scholars on genealogical data formats, as well as access to key points that have been presented in publications that may be hard to reach due to language barriers (e.g., Finnish) or the publications being out of print (e.g., *Genealogical Computing Magazine*). Likewise, we have sought here to create propositions for further research (see e.g. Afuah, 2013) rather than empirically test hypotheses. Answering those propositions, however, is beyond the scope of this article.

As a result of these limitations of existing research, we here first and foremost focus on GEDCOM as a case study on how a *de facto* standard is able to influence not only the software supporting it, but also the crowdsourcing discourses to which the standard and software options give form. Case studies such as this one can be used for three purposes: theory generation, theory testing and theory elaboration (Ketokivi & Choi, 2014). Here, we use GEDCOM to elaborate on the influence a particular file format's dominance has for an activity and the information systems connected to it. To supplement the core software

analysis, [First Author] also examined software-related threads on the Finnish genealogy sites *Suku Forum* and *Sukujutut*, and interviewed two experts who teach genealogy. Through this triangulation, we assessed the level to which the format influences information seeking processes in genealogy.

Analysis

The central challenge with GEDCOM is that it does not model the process of genealogy (described in e.g., Darby & Clough, 2013). It only contains set information. GEDCOM therefore forces the activity into particular forms. This correlates with Duff and Johnson's (2003) finding that genealogists first collect names, then details, and finally contextual information, a process we also observed on the national forum *Sukujutut*.

Practitioner discussion on the nature or change needs of the format (in contrast to how much its application is discussed) is practically non-existent. This ranges from the lack of dedicated forums to the fact that for example the Association of Professional Genealogists (2016) has only one single article on the topic in its quarterly journal's subject index. Sites like BetterGEDCOM have expired. Yet the format itself persists. The dominance of GEDCOM manifests in the ways in which even its suggested semantic replacements take it as a starting point, because of its present prevalence (e.g. Myllynen, 2007), even as its very essence goes against semantic web principles (see e.g., Hyvönen, 2018). Therefore, all new systems must at least for now take GEDCOM into account in their design. Tools exists to check for file quality, and they may not be exclusive to GEDCOM alone (e.g., "Genealogica grafica," n.d.), but they take it as the standard from which to expand and extrapolate into other, less popular options.

A good example of this is Geneanet's GeneWeb software. It offers two formats: the community's own .gw (Listing 3), and GEDCOM.

This is where the implications of GEDCOM truly manifest: one file type has reached a level of popularity so high that it effectively prevents the rise of market competitors. A genealogy tool that would not be GEDCOM-compatible will not find buyers, especially since the LDS freely distributes a tool of its own. So far, all other attempts have remained just minor disruptions.

Genealogical research is furthermore increasingly connected to genetic research. For example, the freemium-model system MyHeritage and its subsidiary Geni offer DNA tests as part of their services. Interestingly, Geni prohibits the uploading of GEDCOM files, as its database is already in the billions, and new data files from users would most likely just cause damage. Yet downloading files from it in GEDCOM is standard practice.

Listing 3: A .gw file (cited from .gw format). The gender indicators can be turned off, in case of non-heterosexual relationships, and variations exist (e.g., #nm, not married) for different types of spousal relations.

```
fam HusbandLastName FirstName[.Number] +[WeddingDate]
    [#sep | - DivorceDate] [#nm | #eng] [#noment]
    [#mp WeddingPlace [#ms WeddingSource]
  WifeLastName Firstname[.Number] # family arguments should be on a single line
  [wit [m|f|]: Witness (use Person format, see Person Information section)]
# possibly several witnesses, respect spaces
  [src Family source]
  [comm Family comments in free format]
  [fevt
    FamilyEvent (multiples)
   end fevt]
  beg
  - [h | f | ] Person # see detailed description at the next section
  end
```

At least in the United States, LDS provides courses on genealogy. Several respondents in one of the few earlier studies on information research and genealogy (Yakel, 2004) had attended such courses. LDS likewise provides a website toolkit for family history research (e.g., "FamilySearch," n.d.), and upholds the somewhat controversial, massive *International Genealogical Index* that grants access to data from also other religious organisations. They also offer partner discount to other providers' software tools – which work with GEDCOM (e.g., to "Geneanet," n.d.). These may be additional reasons for why LDS's GEDCOM has become the core template in genealogy.

Discussion

As a dominant standard, GEDCOM guides genealogical information seeking to certain directions, namely those desired by the LDS. It can thus be viewed as a religious instrument, a free easy access crowd science tool, or both. Its main advantages are convenience and portability, its major drawbacks inaccuracy, heteronomativity and lack of further development. It furthermore urges people to break data regulations, as some software utilising it wants also input from very recent times. On the other hand, developing a widely accepted standard through a regulatory office or committee work would have likely been impossible, so we can presume that without GEDCOM and the LDS, genealogy would not be nearly as advanced as it is now. The question however remains: how has it reached such a popularity despite its faults?

We propose that there are three reasons for this. The first of these is the fact that the format fits extremely well with Davis' (Davis, 1989, p. 320) two

classic criteria for the adoption of information systems: Perceived Usefulness, in which *"people tend to use or not use an application to the extent they believe it will help them perform their job better",* and Perceived Ease of Use, *"the degree to which a person believes that using a particular system would be free of effort."* The second reason is that for hedonic purposes – serious hobbies such as genealogy included – people tend to choose systems based on enjoyment and ease of use, rather than on maximal usefulness (van der Heijden, 2004). In essence, they are *satisficing* rather than accommodating all of their current and probable future needs (as per Simon, 1956). The third is that social influence affects the adoption of any particular information systems technology (Venkatesh & Davis, 2000). Simply put, popularity increases popularity. GEDCOM is simple, it saves the data, it is free, it is portable – and many others are using it already. It is therefore rather logical that it has become the proverbial, popular "VHS" that wins over any technologically superior "Betamax".

At the same time, this ease-of-use popularity is the very cause of its risks. By dominating the field, GEDCOM implies through its simplicity that only certain parts of genealogical information are important. Everything else is just "NOTEs", extra topping. This is in line with Buckland's (1991, p. 159) finding that information systems consist of three facets: the *cognitive* part that leads to users becoming informed, the *economical* part that makes the system's use seem worthwhile in cost-effect terms, and the *managerial* part that defines access to tools and perceived benefits. GEDCOM is excellent in reducing economical and managerial strain at the cost of results on the cognitive part, paring those down to the bare minimum required by its producers, for their purposes. Information seeking becomes thereby limited, if one uses computer-assisted systems for genealogical research. Yet GEDCOM based seeking prevails, at the cost of more widely formed queries.

The LDS' archival suggestions dominate over potentially more fruitful wide-range seeking. Thereby, crowdsourcing is implicitly restricted to the bare minimum, not by rules but by conventions. While it should optimally create, in Star and Griesemer's (1989) terms, modular *repositories* of *standardized form* boundary objects (here: useful, informative files), it appears that the LDS sees it also as a boundary object of the *ideal type* form, stating what is and is not necessary for genealogical data collection. Boundary objects do not just help, they can also hamper and impede collaboration (Oswick & Robertson, 2009). Instead of trying to fulfil the needs of many potential stakeholders, as many information systems boundary object do (see Mark, Lyytinen, & Bergman, 2007), GEDCOM collects what the LDS needs, and creates its value-in-use for genealogists by being free, not by being the best. Its implied purpose is to

provide minimal data to the creators of the format, not to develop wider toolkits for information seekers. And in that purpose, it excels.

Conclusions

Currently, the only real challenger to GEDCOM is the fact that many genealogists are senior citizens, who sometimes still prefer a notebook to a computer. If, however, its development continues to be completely halted, in order to serve LDS' backward compatibility issues, this may well soon change. No matter how popular and how easily perceived as useful it is, at some point a competitor will take its place. The question, therefore, is whether its advantages will provide sustainable benefits, and if its severe limitations, such as incompatibility with the principles of e.g., semantic web principles, or implicit heteronormativity, will cause lasting damage to genealogical research habits.

As a boundary object, GEDCOM exemplifies the ways in which convenience, easy access, and perceived ease of use may triumph over systemic quality. Of particular interest is the way in which the LDS church has apparently decided on a strategy of backwards compatibility over continual – or even sporadic – development. It speaks volumes about their probable interest on their own religious goals rather than on the needs of the genealogists who use the software format, or those of the tools that have been designed based upon it.

In addition, it is important to note that people react very differently to boundary objects. Some let them guide their actions, other keep the objects (in this case, the software format) in the background, without allowing those to influence themselves much (Nicolini, Mengis, & Swan, 2012). It therefore requires further research to see how far the impact of GEDCOM's limitations actually reach. We hope that we have here laid the groundwork for such research to continue.

References

Afuah, A. (2013). Are network effects really all about size? The role of structure and conduct: Are network effects really all about size? *Strategic Management Journal*, *34*(3), 257–273. https://doi.org/10.1002/smj.2013

Ancestry. (n.d.). https://www.ancestry.com/

Association of Professional Genealogists. (2016). APGQ subject index 1979-2016 (September). https://www.apgen.org/publications/quarterly/archives/1979-2016SeptAPGQindex.pdf

Buckland, M. K. (1991). *Information and information systems*. Westport (Conn.): Praeger.

Darby, P., & Clough, P. (2013). Investigating the information-seeking behaviour of genealogists and family historians. *Journal of Information Science*, *39*(1), 73–84. https://doi.org/10.1177/0165551512469765

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Duff, W., & Johnson, C. (2003). Where is the list with all the names? Information-seeking behavior of genealogists. *The American Archivist*, *66*(1), 79–95. https://doi.org/10.17723/aarc.66.1.l375uj047224737n

FamilySearch. (n.d.). https://familysearch.org/

Friday, K. (2012). Learning from e-family history: Online research behaviour and strategies of family historians and implications for local studies collections. https://openair.rgu.ac.uk/handle/10059/734

Friday, K. (2014). Learning from e-family history: A model of online family historian research behaviour. *Information Research*, *19*(4). http://www.informationr.net/ir/19-4/paper641.html

Galliers, R. (ed.). (1992). Choosing information systems research approaches. In *Information systems research: Issues, methods and practical guidelines* (pp. 144–162). Oxford: Blackwell.

Gellatly, C. (2015). Reconstructing historical populations from genealogical data files. In G. Bloothooft, P. Christen, K. Mandemakers, & M. Schraagen (eds.), *Population reconstruction* (pp. 111–128). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-19884-2_6

Genealogica grafica. (n.d.). http://www.genealogicagrafica.nl

Geneanet. (n.d.). http://www.geneanet.org

Genes reunited. (n.d.). http://www.genesreunited.co.uk/

Hyvönen, E. (2018). *Semanttinen web : Linkitetyn avoimen datan käsikirja*. Helsinki: Gaudeamus.

Ketokivi, M., & Choi, T. (2014). Renaissance of case research as a scientific method. *Journal of Operations Management*, *32*(5), 232–240. https://doi.org/10.1016/j.jom.2014.03.004

Land, F. (1992). The information systems domain. In R. Galliers (ed.), *Information systems research: Issues, methods and practical guidelines* (pp. 6–13). Oxford: Blackwell.

Mann, A. E. (1999). FamilySearch: What it is and how to use it. *Genealogical Computing*, *19*(1), 8–11.

Mark, G., Lyytinen, K., & Bergman, M. (2007). Boundary objects in design: An ecological view of design artifacts. *Journal of the Association for Information Systems*, *8*(11). https://aisel.aisnet.org/jais/vol8/iss11/34

Mayfield, D. M., & Brown, A. G. (1999). FamilySearch. *Genealogical Computing*, *10*(1), 1, 8–12.

Myllynen, J. (2007). *Semanttinen web ja sukututkimus* (Master's Thesis). University of Jyväskylä, Jyväskylä. http://urn.fi/URN:NBN:fi:jyu-2007243

Nicolini, D., Mengis, J., & Swan, J. (2012). Understanding the role of objects in cross-disciplinary collaboration. *Organization Science*, *23*(3), 612–629. https://doi.org/10.1287/orsc.1110.0664

Oswick, C., & Robertson, M. (2009). Boundary objects reconsidered: From bridges and anchors to barricades and mazes. *Journal of Change Management*, *9*(2), 179–193. https://doi.org/10.1080/14697010902879137

Schrier, K. (2016). *Knowledge games : How playing games can solve problems, create insight, and make change*. Johns Hopkins University Press.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Sippu, S. (2000). Tietokoneavusteinen sukututkimus. *Genos*, *71*(3), 127–141. https://www.genealogia.fi/genos-old/71/71_127.htm

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, "translations" and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Suku-forum thread on Suomi-tietokanta. (2014). http://suku.genealogia.fi/showthread.php?t=31748

van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly*, *28*(4), 695–704. https://doi.org/10.2307/25148660

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Yakel, E. (2004). Seeking information, seeking connections, seeking meaning: Genealogists and family historians. *Information Research*, *10*(1). http://www.informationr.net/ir/10-1/paper205.html

Yakel, E., & Torres, D. (2007). Genealogists as a "community of records". *The American Archivist*, *70*(1), 93–113. https://doi.org/10.17723/aarc.70.1.l15414u736440636

Zandhuis, I. (2005). Towards a genealogical ontology for the semantic web. In *Humanities, computers and cultural heritage. Proceedings of the XVI international conference of the Association for History and Computing 14-17 September 2005* (pp. 296–300). Amsterdam: Royal Netherlands Academy of Arts and Sciences. https://www.zandhuis.nl/sw/genealogy/