CFPS 93

(Call for Papers Submission number 93)

Requirements for storing media objects

Submitted by: Smith, Richard

Created: 2013-10-30

URL: Most recent version: http://fhiso.org/files/cfp/cfps93.pdf
This version: http://fhiso.org/files/cfp/cfps93_v1-0.pdf

Description: This paper discusses the requirements for storing media objects, such as images, vidoes and audio recordings, and attaching them to items of genealogical significance.

Keywords: media objects, images, audio, container files, jar files, MIME types, metadata, Dublin Core

Abstract

Media objects, such as images, videos and audio recordings, are frequently generated during the research process. Examples include photographs of family members, scans of interesting source material, or recordings of interviews. This paper considers GEDCOM's handling of media objects and draws on that experience to propose functional requirements for the handling of media objects in a future FHISO specification.

1 Storage of media objects

In the existing GEDCOM specification [1], the `OBJE` tag can be used at top-level to embed the media object within the GEDCOM document (using an unorthodox form base-64 encoding to list the object's bytes in the `BLOB` sub-tag). Alternatively, the tag can be used as a link to an separate file on the local machine.

This paper suggests that generally, and where feasible, external storage of media objects is preferable to embedding them in a data file. When stored externally, they can be accessed directly by software to render or edit media objects. Removing large blobs of opaque binary data from the serialisation format improves readability and will often improve processing efficiency. If the serialisation format is to be XML, many parsers, especially DOM parsers which provide a simpler programming interface, can struggle to handle very large files. Furthermore, some means of encoding the object's data would be required, imposing a storage overhead. In the case of base-64, that overhead is 33%.

External files in GEDCOM are identified by a filename, given in the `FILE` tag. This paper proposes that a URL be used instead. For local files, the two are largely compatible as '`images/grandpa.jpg`' is a perfectly good relative URL. But by using URLs, it also becomes possible to reference media objects on a remote webserver, and in large quantities. There would be no particular problem if the files referenced ran to many gigabytes or even terabytes, which might be the case if scans of every source document were to be obtained. There is no technical reason why such images need be freely available, even if they have an `http` URL on a public webserver. A commercial data vendor could require users to log in before the images are displayed.

Where embedded media objects are necessary, the `data` URL scheme provides a way of encoding the object within the URL itself [2]. The use of this for large objects is not generally recommended; however the reasons for not recommending it are precisely the same as for recommending against embedded media objects in GEDCOM. Its implementation, by base-64 encoding the data, is very similar to GEDCOM's, though it is not identical. (The `data` URL scheme uses the standard form of base-64 encoding defined in RFC 4648 [3]; GEDCOM uses a different version found in appendix E of the GEDCOM 5.5 specification. The former has much better support in third-party libraries.) The `data` URL scheme also allows the image to URI-escaped, instead of base-64 encoded, which in some cases could be more concise.

The primary motivation for embedding media objects is to ensure that they remain with genealogical data during transit, and can be accessed by the recipient. The current draft GEDCOM X specification has an elegant solution to this problem [4]. It provides a means of bundling separate documents together into a *container file*, which also provides compression. The container file format used in GEDCOM X is compatible with the `zip` file format, but includes a *manifest file* for storing metadata about the file. The result is essentially a `jar` file, the file format used in the distribution of Java libraries. The FHISO may wish to consider adopting a `zip`-based container format similar to GEDCOM X's, though this paper does not make such a proposal. Individual files within a `jar` (or plain `zip`) file can be referenced in a URL by means of the `jar` URL scheme [5].

2 Media types

In GEDCOM, each media object must have a `TYPE` tag declaring what format the object is in [1]. In the version 5.5 specification, only seven formats are supported: `bmp`, `gif`, `jpeg`, `ole`, `pcx`, `tiff` and `wav`. Of these, `ole` and `pcx` are essentially obsolete, and two more (`bmp` and `gif`) are becoming less common. Conversely, more recent popular image and audio formats, such as `png` and `mp3`, are not present, and the list includes no video format. This illustrates the problem with specifying a short list of popular file formats.

This paper proposes that a media objects must instead declare the object's format using a standard MIME type [6], such as '`image/jpeg`'. These are extensible. A registration mechanism exists by which new file formats can be registered with IANA, and the '`x-`' and '`vnd.`' prefixes exists for unofficial file formats that have not been registered with IANA. The first component of the MIME type states whether the format is an image, audio file or video (together with a few other types less relevant to genealogists).

This paper advises against limiting the range of formats that are permitted. It is is difficult to foresee what formats will be popular in a few years' time, and different user communities will have different preferred formats. The FHISO may, however, wish to issue guidance notes discussing any interoperability concerns with various popular file formats and advising on how to select file formats to maximise the likelihood of them being readable in the future.

3 Metadata

In GEDCOM [1], media objects may be provided with a title (via the `TITL` tag) and arbitrary notes may be attatched (in a `NOTE` tag). A timestamp can be included in `CHAN` tag, and the `REFN` and `RIN` tags allow reference identifiers to be attached to object. The mechanism is not, however, extensible, and arbitrary metadata cannot be associated with the media object. It is not possible, for example, to give an author or copyright notice, other than as free-form text in a note.

This paper proposes as a requirement for a future FHISO specification that it should be possible to attach arbitrary structured metadata to a media object. This paper further proposes that the FHISO defines a core vocabulary of common metadata terms (such as author, copyright, creation date, etc.), which could be extended by third parties. In the terms of CFPS 20, metadata terms would form a partially conrolled vocabulary [7]. It may be advantageous to align the core metadata terms with an existing standard such as Dublin Core (standardised as ISO 15836:2009) [8]. The current draft GEDCOM X specification allows arbitrary metadata to be included in the manifest file [4], and has aligned a number of its metadata terms with Dublin Core.

Care will be required in interpretting certain metadata terms to determine what, precisely, the term is being applied to. Consider, for example, the Dublin Core 'creator' term (included in GEDCOM X using the `X-DC-creator` manifest header [9]) applied to an image of a will. The wording of will was perhaps written at the direction of the testator by a lawyer, and a registered copy made during probate by an anonymous clerk. Maybe the archive holding the registered copy paid a company to make a microfilm from which a copy was printed by, say, the researcher's cousin, which finally gets digitised and imported by the researcher.

To which of those parties should 'creator' refer? The testator? His lawyer? The probate court? The probate clerk? The archive? The scanning company? The cousin who printed it? Or the researcher who digitised it? This is a question the FHISO must address (and indeed that the GEDCOM X specification should, thought currently fails to). This paper does not attempt a complete answer, other than to distinguish between a source and a media object that reproduces it.

A source is often (though not invariably) a physical item, created many years ago, and located in an record office. A media object is a sequence of bytes that attempts, with some degree of faithfulness, to reproduce an object such as as source (or a person, in the case of a digital photo of a person). The testator, lawyer, probate court and clerk all had a part in the creation of the source; and the archive, scanning company, cousin and researcher all had a part in the creation of the media object from the source. This paper therefore suggests that only the latter four parties should be considered as the possible 'creator' of the media object.

4 Linking to other entities

In GEDCOM, media objects can be referenced from almost any record [1], namely:

- individuals;
- events (whether family or individual) and individual attributes;
- family groups (the structure used to associate two spouses, and/or a child with a parent);
- source records and source citations; and
- submitter records (giving information about the researcher).

This paper supports GEDCOM's liberal scope for attaching media objects. However the GEDCOM standard is silent on what it means to link to a media object in each of these ways. This paper proposes to rectify this by making the link between a media object and the entity to which it is attached explicit, and giving the link a name. This fits well into the framework proposed in CFPS 4 [10]. A media object would be a *thing node*, as are the entities to which media objects may be attached. The directed link between them is a *connection* and in CFPS 4 requires a *type label* to ascribe meaning to the particular connection.

One possibility is to give the connection a nondescript type label like 'has attachment'. This paper does not recommend such an approach, and intead proposes a more specific 'is depicted by' type label (where 'depict' is should not be understood to preclude audio or audio-visual media). Or if a link from, rather than to, the media object is preferred, 'depicts' is proposed as its type label. This direction of link would make sense if the conceptual model is to be one of tagging people and events in photos and other media, rather than attaching media to people and events.

In current GEDCOM, if an image attached to an individual, it could be a photo or painting depicting that individual, or it could be a scan of a document talking about the individual, or it could be the building in which they lived or worked, or even a work of art created by that individual. Under this paper's proposal, only the first of these possibilities would be permitted. In the other cases, the scan of a document depicts a source not an individual, and should be attached to the source (which may itself will perhaps be referenced by the individual); the picture of a building should be attached to the attribute containing his place of residence or work; and the work of art created by the individual might be attached to his occupation, if he were an artist.

In order to allow for the automated translation of present-day GEDCOM files, it may prove beneficial to provide a connection type more general than 'is depicted by', perhaps called 'has relevant media'. If such a connection is provided, its use should be deprecated.

References

[1] Church of Jesus Christ of Latter-day Saints, 1996, *The GEDCOM Standard (Release 5.5)*, `https://devnet.familysearch.org/docs/gedcom/gedcom55.pdf`

[2] L. Masinter, 1998, *The "data" URL scheme* (RFC 2397), `http://www.ietf.org/rfc/rfc2397.txt`

[3] S. Josefsson, 2006, *The Base16, Base32, and Base64 Data Encodings*, `http://www.ietf.org/rfc/rfc4648.txt`

[4] Intellectual Reserve Inc., 2013, *The GEDCOM X File Format*, `http://www.gedcomx.org/Specifications.html`

[5] Oracle, 2011, *Class JarURLConnection* in *Java™ Platform, Standard Edition 6 API Specification*, `http://docs.oracle.com/javase/6/docs/api`

[6] N. Freed & N. Borenstein, 1996, *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types* (RFC 2046) `http://www.ietf.org/rfc/rfc2046.txt`

[7] Tony Proctor, 2013, *Proposal for Handling Partially Controlled Vocabularies* (CFPS 20), `http://fhiso.org/files/cfp/cfps20.pdf`

[8] Dublin Core Metadata Initiative, 2012, *Dublin Core Metadata Element Set, Version 1.1*, `http://dublincore.org/documents/dces/`

[9] Intellectual Reserve Inc., 2013, *The GEDCOM X Standard Header Set*, `http://www.gedcomx.org/Specifications.html`

[10] Luther Tychonievich, 2013, *Modeling Research, not Conclusions* (CFPS 4), `http://fhiso.org/files/cfp/cfps4.pdf`