

Formats, Format Registries and Digital Archiving/Preservation: David Gewirtz November 2006

File Formats and the Need for Format Registries:

Everybody has noticed that when they create computers files the name of the file almost always has two parts. One part is the name given by the creator that is followed by a dot (“.”) and than some other three character extension like “.doc” or “.jpg” These extensions are abbreviations of names of file formats¹. The name of the format refers to the special encoding or arrangement of electronic information on media (like a hard drive) called bits that can be understood by a software program to access and render a digital file. For example a text file like MS Word has the extension “.doc” that is understood by Microsoft Word to display text on your computer screen. Similarly a software product like MS Photo Editor understands how to decode a “.jpg” file so that an image is rendered on your computer screen. Unfortunately, file formats like many of our everyday technologies become obsolete over time. The consequence of this is that digital objects stored in a version of a file format can become unreadable and inaccessible to rendering applications over time. This is a special problem to preservationist and archivist that have the responsibility to preserve digital information for the long term. Librarians and technologist believe that if technical information about a formats encoding, sometimes referred to as *representation information*², can be permanently stored in a Registry than preservationist or archivist will always have a means to render a digital file using future hardware and software.

Preserving a format in a registry is then a preservation strategy that ensures that as file formats evolve older version of formats can remain readable by current software application and therefore useful to an end-user. In addition file formats are often drivers for digital repository policies that define the major process of a repository such as ingest data managements, administration, preservation, access and interoperability. Across academic domains there are so many different file formats and versions that it would be impossible and economically impractical for any one organization to manage a complete registry of file formats. Therefore the Digital Library and Archival communities have expressed their need for a sustainable, global registry of digital formats for the purpose of protecting and preserving digital objects for the long term. This Global Registry would be collaborative effort sponsored and supported by the Cultural Heritage Community and based upon network technologies. For a more in-depth and technical discussion of the need for file formats and global registries please see the reference below³.

¹ Technically a file format is defined expansively as a fixed, byte-serialized encoding of an information model.

² Representation information is an OAIS concept and refers in this case to the mapping of typed formats into more meaningful concepts by capturing the significant syntactic and semantic properties of those formats. Significant properties are defined as those aspects of a format that are the primary carriers of the format’s intellectual value.”

³ http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Current Research in the Preservation of File Formats:

The principal objective of research in file formats is to create an authority control of identifiers for file formats used in the larger cultural heritage community. To meet the preservation requirements for file formats and to promote interoperability between repositories authority control research focuses upon the **representation** of digital formats so that obsolete digital images can be **rendered** in the future. Since the beginning of network computing MIME⁴ types were used to identify file formats. This global typing system is deficient for preservation purposes because it is not granular enough to differentiate important variations in file formats. Two good examples of this are tiff and pdf files which have many versions but are all categorized by MIME as file formats of type “.pdf” or “.tif”. This gives the end user the false impression that any pdf or tiff reader can render a “.pdf” or “.tif” file.

Like many other business processes in the Library the creation of an authority control function for file format identifiers is based upon specialized metadata for file formats that is needed to create future rendering applications. Research investigations in file formats focus upon the development of a data model that defines elements and attributes of a file format that are needed to unambiguously differentiate versions of a typed format and that are needed to preserve the encoding of the format over the long term. For example the CAMiLEON⁵ research project sponsored by JISC⁶ “clearly demonstrated the need for advanced rendering technologies that offer accurate and economical preservation of digital materials”. PRONOM⁷ another research project by JISC and the Netherlands developed a network service that enables a user to obtain “impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.” In the United States, and at a peer institution, the University of Pennsylvania, Ockerbloom⁸ is developing systems to manage diverse data formats. In a joint effort between Harvard University Library and OCLC with funding from the Andrew W. Mellon Foundation a research project has been established to create a Global Digital Format Registry (GDFR). The GDFR⁹ will provide services for (1) The centrally-organized collection of format representation information and (2) The distributed storage, discovery and delivery of that information.

⁴ Multipurpose Internet Mail Extensions is a global but deficient mechanism for preservation and identification of file format types.

⁵ http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/project_fileformat.aspx

⁶ In the UK The mission of the Joint Information Systems Committee (JISC) is to provide world-class leadership in the innovative use of Information and Communications Technology to support education and research.

⁷ More information on PRONOM is available at <http://www.nationalarchives.gov.uk/pronom/>

⁸ More information on the Typed Object Model is available at <http://tom.library.upenn.edu/>

⁹ More information on the GDFR can be found at <http://hul.harvard.edu/gdfr/>

Towards Best Practices to Protect File Formats

“Technical documentation about digital formats will necessarily be a core part of any preservation program. In the absence of a generally accessible, reliable, and persistent registry of such data, each individual preservation program will need to collect and maintain its own documentation. Not only is this wasteful in terms of large-scale duplication of effort, but it would also require each program to have access to highly sophisticated staff with the skills to document each format the program ingests. Such expertise is scarce and expensive and many programs will likely be unable to support the activity at an appropriate level”¹⁰. In concept than the GDFR represents a best practice for large research libraries like Yale that shows promise for preserving representation information about file formats. The benefits of this practice to large research libraries like Yale are:

- A common mechanism to pool and share scarce technical expertise on a global basis, reducing the necessity for duplicative local effort
- A channel for the widest possible distribution of the fruits of that expertise to all actors engaged in preservation activities
- A process for generating community-wide agreement as to the normative definitions of format syntax and semantics, promoting best practices and effective interchange of digital assets between preservation institutions, programs, and systems
- A foundation for additional value-added services requiring detailed knowledge of digital formats

The Architecture for a GDFR and a Simple Explanation:

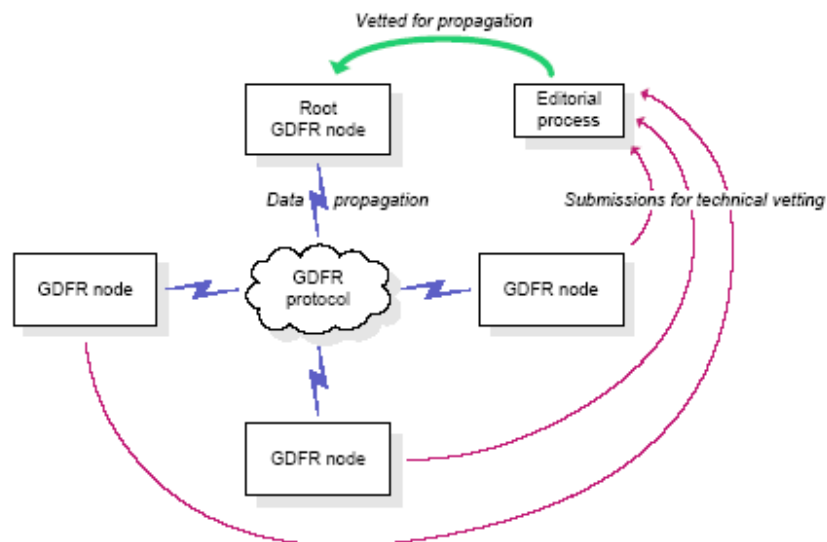


Figure 1. GDFR architecture

¹⁰ http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Each cooperating format registry is a node in the GDFR network. Each node must conform to a standard to store data and to provide services to end-users and machine processes. Data stored by a participating repository (GDFR node) identifies a registered format and provides descriptive, administrative, and technical information or metadata about the format. This metadata might include the official format name, the common format file extension (for example “.jpg” for images and “.doc” for a MS word file) and a document that authoritatively describes the formats encoding or specification.

Data that is stored must also be managed by a node. Data management tasks are designed to preserve, for the long run, information about the stored format. Beyond storage other management tasks include the function to participate in an inter-nodal process to approve the registry of a format. A synchronization service ensures that all participating registries can have the same and most current definition of a format. Also data replicated across geographically dispersed nodes promotes sustainability of the data and preservation services that rely upon format information.

These management services form the basis of access services that enable the discovery and the dissemination of format information across the global format registry network. The GDFR protocol provides the communication necessary for a machine process or end-user to request an access service. Requestors might typically ask for a format description, an alert message to notify the registry of a new format in the network and a bulk export of formats to populate a new registry in the network.