RDA Interest Group
Scientific File Formats

June 30, 2014

Whether data should be archived, shared between scientists or fed into an analysis program, at some point in time, data has to be stored on a physical medium using a particular file format. Although it sounds like a trivial job to do, choosing the right file format can be a difficult task and a wrong choice can cause serious problems for a research project on the long run. While the number of available file formats is compelling, they can differ largely on the level of support for different platforms, the availability of libraries for different programming languages, and their status of maintenance. In addition to this rather technical problems, not all formats are equally well suited for a particular scientific application. Even if several formats would do well for an application scenario some of them may not be compatible with other applications or different platforms and thus cause a problem for interoperability. A compelling example is the interoperability between CIF and NeXus/HDF5, which could be achieved in co-operation between the standards defining bodies (the International Union of Crystallography (IUCR) and the NeXus Internation Advisory Comittee (NIAC)).

Due to the complexity of the problem many scientists refuse to choose an existing format and develop their own standard. Such approaches have serious implications for the long term availability of data, since custom formats are typically badly maintained and documented due to the few of resources a scientific project can afford to spend on such a problem.

The aim of this interest group (IG) is to establish some structure in the Babylonian abundance of file formats. It should help scientists to quickly find the file format which suites their purpose the best and thus deter them from developing custom formats. In addition the IG should identify requirements a file format has to satisfy in order to become an advisable format for storing scientific data. These requirements not only help users to evaluate existing standards but also defines project goals for format developers which want to have their formats advised by the IG.

It is obvious that the IG cannot achieve all its goals by itself. It should thus be considered as the long-living organizational guidance initiating and co-ordinating working groups (WGs) with a life-time of typically 12-18 month focussing on well defined issues. A possible situation is depicted in Fig. 1 where the IG coordinates the work of three WGs including one which defines the technical requirements the IG demands of file formats.

The IG will presumably favor to focus on widely adopted, well defined open standard data formats, but is entirely open. The IG aims to promote the participation of scientific user and developer communities, standards defining organization and vendors or developers for example of scientific instruments or applications. An additional aim of the IG should be to

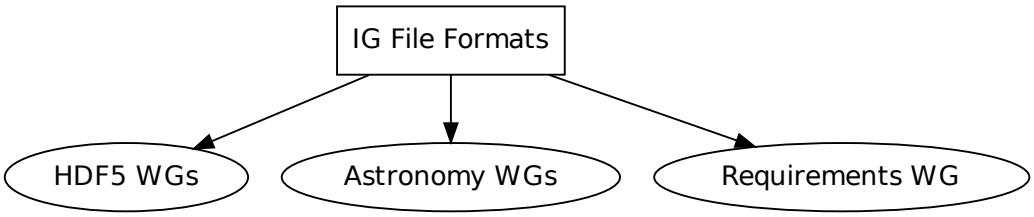

Figure 1: The IG file formats acts as the organizational framework for working groups concerned with a particular format or with the formats of a particular field of application.

help in the development of interoperable file formats. An important issue here is to specify what the term *interoperability* means in the world of file formats and how the design of a file format can help in the development of interoperable applications.

It must however be noted, that the activities of the IG should be currently constrained to formats used to store data on a physical medium rather than those used to transfer data (for instance over a network).

1 Emphasis on patent free and open source technologies

The IG should focus on open source and patent free formats. This restriction is less based on ideological rather than on practical considerations. For long term storage, proprietary technologies can become a problem when the company behind it ceases. In such a situation it is usually impossible to get a format ported to new platforms or environments. Continuation of the development by third parties might be prohibited due to missing documentation or by claims of patent holders. Patents by themselves are a source of all kinds of problems. Though patents usually should describe the technology they are protecting this documentation is in many cases far to general for a third party to continue development once the patent has run out. Additionally, as patent laws are national laws and thus not necessarily compatible among all countries involved in a scientific project. However, there is a chance that a patented technology becomes advisable if the patent holder allows third party implementations of its technology as open-source and for non-profit purposes.