FRBR Applied to Scientific Data

Joseph A. Hourclé, Wyle Information Systems, <joseph.a.hourcle@nasa.gov>

The distinction between a creative work and the physical item that contains that work is clearly delineated in FRBR and other research by the Library Science community. A similar confusion exists in the scientific realm between the underlying scientific data and the digital objects that contain those data. We present a similarly scoped reference framework for sensor-based scientific data, drawing on the concepts in FRBR, and compare it with the application of FRBR for cataloging other non-book records.

Introduction

The science community has recognized a need for sharing of scientific data, both within a given scientific discipline and across typical scientific boundaries (NASA, 2007). With this sharing, we gain the opportunity for re-use and increased value from a given scientific experiment.

Scientific data may be processed into multiple derivative works, each having a different intended use or audience. The hierarchical relationship between the initial sensor data and the multiple ways in which the data may be translated, packaged and stored is similar to the FRBR (IFLA, 1998) concepts of *Work*, *Expression*, *Manifestation* and *Item* (Hourclé, 2007). For this paper, we assume that the catalog's primary audience is the scientists looking for data; we stress the "Functional" aspect of FRBR and the loose interpretations of FRBR may directly conflict with typical FRBR usage. We focus on the FRBR tasks of 'identify' and 'select', which are lacking in the Open Archive Information System [OAIS] Reference Model (CCDS, 2002). The necessary identification tasks include:

- Which sensor did the data come from?
- Does this data describe an event or object I am interested in?
- Is the data calibrated for the use I want?
- Has the data been reduced in any way?
- Which data package is best for my tools?
- Is a local copy available?
- Can I obtain the data from an alternate location?

An archive may be able to handle some of these questions, but the lack of coordination of identifiers between archives makes it difficult for designers of federated search systems to make decisions about how to present merged search results when there are two files available that seem similar:

- Should I show both of these records or a prototypical example?
- Can I cluster the records to keep from overwhelming the user?
- At what point is it necessary to disambiguate between two similar items?
- Which object is better to serve the user's needs?

The primary goal of this model is to allow for better differentiation and specificity in searching by decreasing the amount of duplication that is presented to the user. The secondary objective is to build catalogs that more accurately describe the relationships between instances of data, the articles they are used in, and the other supporting objects.

A Loose Interpretation of FRBR

For the casual user of a library catalog looking for an interesting fiction book to read, an information system based on FRBR allows a user to make decisions at higher levels without being overwhelmed by each individual *Item* in response to the initial search:

Work: Who wrote it? What is the subject?
Determine interest / Applicability
Expression: What language is it in?
Usability/Accessibility (of content)
Manifestation: What size is the font or book?
Usability/Accessibility (of content within carrier)
Item: Is the individual copy available to me?
Availability / Accessibility (of the carrier)

New Entities

Sensor

In order to model the data, we must consider the *sensor* that records the data. A *sensor* is created, maintained, operated, and owned by another Group 2 entity. The *sensor*'s sole purpose is to convert information about its environment into a digital signal. The nature of the data recorded by a *sensor* depends upon its design and operating mode.

For the sake of simplicity, this model does not consider the data storage and other system components necessary for a sensor to operate or other attributes of a sensor and will defer those to MMI (Greybeal 2008), SPASE (Thieman, Roberts & King 2004), SESDI (2007) and other similar efforts.

Observation

Sensor recordings are not a creative work, but might be considered an intellectual creation. This model uses a separate entity of *observation* to track the data created by the sensor, so that we can include the necessary relationships with differing interpretations of the same data.

Although this may be considered to be the subject in a typical FRBR implementation, there is the need to unambiguously track the specific *observation* on which the data is based, as we cannot rely on spatial and temporal metadata to uniquely and consistently identify a given observation (Hourclé *et al,* 2007).

FRBR Applied to Scientific Data

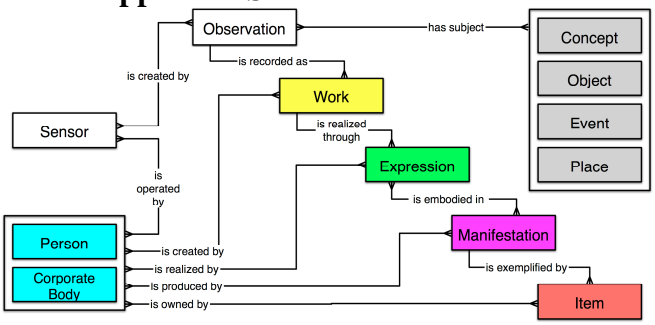

Figure 1: FRBR Applied to Scientific Data

Item vs. Manifestation

As most recent scientific data is digital, there may not be a specific physical manifestation. Although the issue of what is an FRBR item in terms of digital objects for the sake of library cataloging brings questions in how the concept is reconciled with physical objects (Floyd & Renear, 2007), this model considers a logical *item* to be a resolvable item, such as a URL might identify. This model does not consider the implications of moving the data across disks, the fundamental differences in file systems, differing byte order or RAID volumes.

We assume that any two *items* of the same *manifestation* would have the same effective bit sequences, considering the differences in byte ordering of the file system, and as such the two *items* could be considered identical copies. This definition may not suffice for "operational" data that is used to support decision-making and thus requires the tracking of provenance information.

Expression vs. Manifestation

An *expression* is the "realization" of the *work*, while the *manifestation* is the physical embodiment of the *expression* (IFLA, 1998). The medium and aspects of the container are considered attributes of the *manifestation* and not the *expression* that it contains.

As the scientific data is transmitted over digital networks, the physical container, be it on tape, hard drive or optical media is not significant to the user; the file format is much more significant, as the choice of distributing the file as a PDF or ASCII file affects the user's ability to extract the necessary content. We thus consider the *manifestation* to be a logical and not a physical embodiment, to include aspects of how each individual datum is organized within a file or other package for distribution.

In this model, a *manifestation* may contain multiple *expressions*, both of data and the complementary metadata for those data. We do not model the combination of multiple *expressions* in a given *manifestation* to be a "*unified work*" but as separate *works* that enhance each other (Velluci 2007, pp.138-41). Changing the metadata attached to the data packaging constitutes a change in the *expression* to the metadata and a change in the *manifestation* of the data package, but would not change the *expression* for the scientific data. This differs from Yee's (2007, p.120) convention when adding a commentary track to a DVD; although the supplemental information may change the interpretation of the data, it does not result in fundamentally new data.

Reprocessing the data, however, changes the values within the digital object, thus creating a new *expression*. This includes any form of translation of individual datum, to include change of units, down sampling or other forms of compression. The *expression* would not include aspects of the container, such as the file format or order of the individual datum within it. For example, images may be stored in row-major or column-major order, but they would be considered two *manifestations* of the same *expression* if the values in each cell remained the same.

Generally, changes in carrier but not content are simply new *manifestations* of a given *expression* or set of *expressions* while changes in content result in a new *expression* and resulting *manifestation*.

Work vs. Expression

This model assumes the calibrated state of a given observation to be a *work*, to include the CODMAC (NASA, 2005) "level 0" [raw sensor output] state. We define calibration for this model as the process of translating the sensor's output into units that remove the known issues with the sensor ["level 1a"] or to physical units ["level 2"]. Event catalogs, scientific metadata and other results from data analysis ["level 4"] would be a new *work*, but would have a subject that is the *expression* of the data collection, as opposed to being directly derived from the *observation*.

Data may be resampled in ways that do not maintain the original resolution of the original data, such as producing quarter-resolution images or hourly averages of time-series data ["level 1b"]. This change in resolution would be considered a new *expression* of the *observation*. Data may be further manipulated to alter the scaling of the values, such as to enhance features of interest, to alter the coordinate or other reference systems used, or fill in missing data ["level 3"]. These modifications would be considered new *expressions*, as opposed to Yee's (2007, p.119) interpretation of intentional loss of information through video editing to be a new *work* or Soergel's (2007) interpretation of *work* being the specific version of the text.

Although the *work* vs. *expression* distinction may be subjective, it allows us to model the variations of sensor-specific data calibration separately from other forms of data processing that may not be sensor-specific. Generally, we model manipulations that retain the original resolution to be new *works*; those that result in loss of content as a new *expression* and those that change carrier but not the

individual datum are simply a new *manifestation* of the same *expression*.

Limitations

Scientific Discipline

This model has been designed and discussed in the realm of space physics data, with the author's experience being primarily solar-observing telescopes. This model has been designed to specifically address issues in the design of a "Virtual Observatory", a discipline-specific federated search system for space physics data (Hill, 2000). As each scientific discipline has a different mental model of their data, different attributes are necessary for each community of discourse to easily identify and select the data of interest to them (FGDC, 1998; SPASE, 2008; VSO; 2005). Defining those attributes is outside the scope of this paper and has been handled by discipline specific standards.

We believe that the general model would be applicable to other sensor-based "big science" data collection, to include earth observing *in situ* and remote sensing data. It may require modification for use with other types of scientific data.

Digital Objects

We assume that the data being cataloged are digital objects and do not have physical counterparts. For the space sciences, this is mostly accurate. We neither touch upon the generation of physical items based on the data, nor the generation of digital data from physical recordings, such as the digitalization of photometric plates.

Non-Human Workflow

There may be a need to model other non-human objects that generate or manipulate data to form *works, expressions, manifestations,* and *items*, to include software, data systems and other aspects of data workflow. This would allow users to identify data that have been similarly processed or two distinguish between two objects that were generated by different editions of the same software or completely different data pipelines. These entities would more accurately describe the relationships between Group 1 and Group 2 entities in the typical data workflow.

Data Collection vs. Data Granule

Scientific sensors record a number of successive data objects ('data granules') over time. The full set of aggregated data ('data collection') may be subsetted in some way other than time, creating a number of sub-collections to be served by an archive. This has similarities with tracking individual articles vs. the journal as whole (Delsey, 2003; Rosenberg & Hillman, 2004; Shaddle, 2007), but could also be modeled as *subexpressions* (Velluci, 2007, p.143). We look forward to the work of the FRBR Group on Aggregates (IFLA, 2007) for its recommendations on aggregates.

Individual Objects vs. Dynamic Packaging

Not all archives track individual objects for distribution, but instead store their data in a database and generate distribution objects the scientist requests. Each distribution may have data from different sensors, cover a different range of time, or be processed differently. It is outside of the scope of this model to determine how best to handle this situation or to decide if it is better to catalog at the data granule, the data collection, or some other level of aggregation. It is hoped that this issue can be revisited.

Data Archival Without Attached Metadata

Some of the higher cadence data systems have proposed to store their data separately from their metadata (Borne, 2007). This model tracks metadata as a supplemental work to prepare for that eventuality, but this fundamental change in data archiving may create additional complexities.

Acknowledgements

The author wishes to thank Ingbert Floyd and Karen Wickett for their assistance in better aligning my previous model with FRBR, although they debate there being a true manifestation or just two classes of expressions. Thanks to the scientists and programmers from VSO and SPASE for helping to characterize the issues as viewed by the scientists and data archives. Thanks also to Sunny Yoon for suggestions on early drafts of this paper and to Mary Ann Munn for help in keeping this under 8 pages.

References

Borne, K. (2007). LSST: Preparing for the Data Avalanche through Partitioning, Parallelization, and Provenance. *Science Archives in the 21st Century*, Adelphi, MD – April 25-16, 2007. http://nssdc.gsfc.nasa.gov/nost/conf/archive21st/presentations/posters/p02-borne.ppt

Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. http://public.ccsds.org/publications/archive/650x0b1.pdf

Delsey, T. (2003). *FRBR and Serials*. Retrieved 26 Feb 2008 from http://www.ifla.org/VII/s13/wgfrbr/papers/delsey.pdf

Federal Geographic Data Committee. (1998). *Content Standard for Digital Geospatial Metadata*. http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

Floyd, I.R. and Renear, A.H. (2007). What Exactly is an Item in the Digital World? *Proc. ASIS&T 2007*, Milwaukee WI, USA.

Greybeal, J. (2008), The Marine Metadata Interoperability Project. Retrieved 21 Feb 2008 from http://marinemetadata.org

Hill, F. (2000). The Virtual Solar Observatory: A White Paper. Retrieved 23 Feb 2008 from http://vso.nso.edu/vsowp.pdf.

Hourclé, J.A. (2007). FRBR in a Scientific Data Context, *Science Archives in the 21st Century*, Adelphi, MD – April 25-26, 2007. http://nssdc.gsfc.nasa.gov/nost/conf/archive21st/presentations/posters/p11-hourcle.pdf

Hourclé, J., Suárez-Solá, I., Davey, A., Tian, K., Yoshimura, K., Martens, P., Gurman, J., Hill F., and Bogart, R. (2007). Design Considerations for Data Catalogs, *Eos Trans. AGU, 88*(52), Fall Meet. Suppl., Abstract SH51A-0261

IFLA (2007). *Working Group on Aggregates*. Retrieved 22 Feb 2008 from http://www.ifla.org/VII/s13/wgfrbr/aggregates_wg.htm

IFLA Study Group on the Functional Requirements for Bibliographic Records, & International Federation of Library Associations and Institutions. (1998). Functional Requirements for Bibliographic Records: final report. München: K.G. Saur.

National Aeronautics and Space Administration. (2005). *Earth Science Data Terminology and Formats*, retrieved 25 Feb 2008 from http://science.hq.nasa.gov/research/earth_science_formats.html

National Aeronautics and Space Administration (2007). *NASA Heliophysics Science Data Management Policy*. http://lwsde.gsfc.nasa.gov/Heliophysics_Data_Policy_2007June25.pdf

Rosenberg, F. and Hillman, D. (2004). An Approach to Serials with FRBR in Mind: CONSER Task Force on Universal Holdings. Retrieved 21 Feb 2008 from http://www.lib.unc.edu/cat/mfh/serials_approach_frbr.pdf

SESDI. (2007) *Semantically-Enabled Science Data Integration: Introduction*. Retrieved on 23 Feb 2008 from http://sesdi.hao.ucar.edu/intro.php

Shadle, S.C. (2007). FRBR and Serials: One Serialist's Analysis. In A.G. Taylor (Ed.), *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*. Westport, CT: Libraries Unlimited.

Soergel, D. (2007). Digital Library Content Model. *First International Workshop on Digital Libraries Foundations*, In Conjunction with JCDL 2007, Vancouver, BC - June 23, 2007.

SPASE (2008). *SPASE Data Model Documents*. Retrieved 21 Feb 2008 from http://www.spase-group.org/data/doc/

Thieman, J., Roberts, A. and King, T. (2004). The Space Physics Archive Search and Extract (SPASE) System for Space Physics Data, *Eos Trans. AGU, 85*(47), Fall Meet. Supl., Abstract SA54A-04

Vellucci, S.L. (2007). FRBR and Music. In A.G. Taylor (Ed.), *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*. Westport, CT: Libraries Unlimited

Virtual Solar Observatory (2005). "VSO Data Model: Version 1.8," Retrieved on 21 Feb 2008 from http://vso.stanford.edu/datamodel.html

Yee, M.M. (2007). FRBR and Moving Image Materials: Content (Work and Expression) versus Carrier (Manifestation). In A.G. Taylor (Ed.), *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*. Westport, CT: Libraries Unlimited.