Maintaining object authenticity in very large digital libraries

Tobias Blanke, Stephen Grace, Mark Hedges, Gareth Knight, and Shrija Rajbhandari

Centre for e-Research, King's College London, UK
tobias.blanke@kcl.ac.uk
stephen.grace@kcl.ac.uk
mark.hedges@kcl.ac.uk
gareth.knight@kcl.ac.uk
shrija.rajbhandari@kcl.ac.uk

Abstract. Digital libraries are increasingly being used to manage research data, leading to a significant increase in the amount of material held in such systems. Much of this material is irreplaceable, and there is a pressing need to maintain long-term access to it; automation of curation is key if a scalable solution is to be found. The concept of significant properties is key to maintaining the integrity and authenticity of a digital object over time and technological change. We present a case study which addresses significant properties as they apply to a working scientific community active in experimental and computational research, and we describe an approach to automating this model by representing curation policies as rules that are implemented using the iRODS middleware.

1 Background

Research across disciplines is increasingly both a generator and user of very large quantities of data, the so-called "data deluge" [5]. This is not an issue only for sciences such as particle physics and astronomy, which have for many years been generating petabyte-scale data sets, but is increasingly the case in subjects such as the humanities, with the growth of digitisation projects producing high-resolution images, video and audio, as well as the existence of born-digital archives. The issue is also not just one of scale; in many disciplines, the information objects created by researchers may be highly complex, with many structural and semantic relationships both internal and contextual. For example, textual scholars may work with a variety of marked-up textual resources, databases and multi-media objects, as well as a number of secondary resources such as dictionaries and concordances; medical researchers may deal with large two-and three-dimensional image files that have detailed annotations and links to other resources. Thus we may speak of a complexity deluge as well as a data deluge. All this raises significant challenges for the curation of the data.

The creation of this digital material represents a considerable investment in intellectual effort, time and, in many cases, public funding. In recent years it has

been increasingly recognised that this investment in research must be protected through long-term management of its outputs. In contrast to the management of physical artefacts, it is considered infeasible to store digital artefacts in their original form and expect them to remain readable and usable when required at a later date. Instead, digital curation is built upon the premise that the environment in which research is accessed and used is likely to change over time and that specific activities, such as format conversion or emulation are required at distinct stages in a digital object's lifecycle to ensure that information remains accessible and usable.

Digital preservation is thus a major issue for research, and indeed for any domain that needs to ensure long-term access to digital material, for example archives, digital libraries and commercial sectors, and much work has gone into developing digital preservation standards such as the OAIS Reference Model [2] and the PREMIS Data Dictionary for preservation metadata [8]. In general, two distinct approaches to digital preservation are discussed; either digital resources are preserved for the future by transforming them into certain standard and normalised formats that one would expect to continue to be comprehensible as the environment changes, or the software environment in which the data is rendered and processed is maintained by enabling its emulation in future software environments [10] [6]. The latter poses significant challenges. Not only will we have to ensure that the code is future-proof and written in a language that will still be understood by future computer systems, we will also have to guarantee that the current code does not contain (fatal) logical errors that might break future systems and that it does not contain harmful code. Not only for these reasons, but also for reasons of scalability, we do not at the moment see this emulation approach as feasible for research data, and in the current work we have concentrated on improving the first option.

2 Significant properties

2.1 Overview

In the work described in this paper, we have followed a normalisation/migration approach; specifically, our approach involves converting a digital object to one of a range of preferred, standard formats at the time of ingest (format normalisation), combined with subsequent conversions of objects throughout their life-cycle as formats or rendering software tools become obsolete. Such technological migrations are not without risk, as it may result in information loss or some change to the way in which information is represented, so the experience of an object may not be identical after conversion. The level of data loss depends on the number and nature of preservation treatments applied to an object, the new data format(s) selected, and the level of human intervention and post-conversion analysis. In order to ensure the authenticity of a digital object throughout its life-cycle, detailed audit information must be captured and retained at each stage, in order to quantify the information loss and thus to provide a measure for the integrity of the data. The definition of the Significant Properties (SPs) of digital

material is key to ensuring that the integrity and authenticity of information is maintained, to enable subsequent access, use and understanding.

Traditional approaches to this audit information (or preservation metadata) have been data-centric, concerned with recording those aspects of the digital object necessary to maintain access to it. However, recent work, by projects such as InSPECT, PLANETS and CASPAR[1], has advocated a sociological methodology that examines the epistemological interpretation of the digital material's creator and its 'designated community', together with the purposes for which it was created and is being used [7]. The challenge for a data curator, therefore is to analyse the requirements of this designated community and identify the characteristics of the digital artefact that enable these requirements to be satisfied. An example of such an approach in given in [3], which presents a preservation system for cultural heritage material that is based on migrating files to preservation formats. After migration, the information loss incurred is determined by applying comparison functions based on significant properties.

These projects have applied an SP approach to a range of generic categories of digital object, from the point of view of data curators. The work described in this paper differs in that we focus on a specific discipline, specific communities of practice within that discipline with their own workflows and objectives, and specific categories of digital object used by these communities, with the aim of investigating how the definition of SPs can contribute to the preservation of these communities' research data.

2.2 Case study

In our case study, we are addressing SPs as they apply to a working scientific community that is active in experimental and computational research and that uses complex digital material on a day-to-day basis. We are not only analysing the data, activities and objectives of this community at a theoretical level, we are also developing a practical demonstration of the application, impact and importance of the significant properties approach to digital curation within this targeted environment.

The discipline is cell and molecular biophysics, an area with a strongly interdisciplinary focus, operating at the interface between the health, biomedical and physical sciences, and which has links to numerous other areas of research within biomedical and health sciences. A variety of digital objects is produced in this research - in the current work we are focusing on information objects that may be broadly described as images. In many cases the raw images are obtained by advanced microscopy or nanoimaging techniques, but these may also include 3D images produced from 'stacks' of 2D images, and time sequences of images capturing temporal development or processes. In other cases the objects are less conventional images generated by detectors specific to particular physical processes, for example diffraction images generated by crystallographic experiments

[1] http://www.significantproperties.org.uk/, http://www.planets-project.eu/, http://www.casparpreserves.eu/

on large molecules such as proteins. Raw objects are frequently in proprietary formats that are dependent on the equipment used to capture them, although the formats are generally open, facilitating the development of software for rendering or processing them. In all cases, these objects are complex in nature and possess properties whose significance is often specific to the domain.

The designated communities are (i) the researchers, who generate and transform digital material in the course of their research activities, and (ii) staff concerned with the management, curation, archiving and preservation of the digital objects generated by the research. The objectives of these two communities are not identical, nor are the workflows in which they involve the digital objects. Researchers are concerned with the rendering of images, and with a variety of transformations of images, in order to facilitate particular research and teaching activities; curation staff are concerned with maintaining the autheniticity and integrity of an object over time. Nevertheless, their objectives are of course strongly connected, as one of the aims of archiving the digital material is to enable subsequent verification of existing research, or future research based on existing experimental data.

As a practical demonstration of our approach, we built services that can be used to (i) extract SPs from an object, (ii) convert an object from one format to another, and (iii) validate the SPs of a converted object against its original profile. These services have been combined to develop a software demonstrator, based on the iRODS system (see below), which implements an SP approach within our targeted area.

As remarked above, any format transformation involves some degree of information loss. Our aim was to identify the properties of the information objects (in our case raster images) that were deemed significant with respect to the future interpretation of those objects by the designated user community, and to express them quantitatively or formally, so that their persistence through images transformations can be verified. Although at a quantitative level there is information loss, this information is not relevant from the point of view of the searchers' visual or processing experience.

There are a number of ways of quantifying information loss. A simple example is to calculate the *root mean square* error between two images. However, as such measures are very generic and give equal weight to all parts of the image, they may not be very helpful in ensuring the persistence of the image characteristics that are of significance in particular contexts. Their limitations become quite clear in the context of image features such as the existence of fine lines, where the rms error measure may manifestly fail to quantify information loss that is immediately evident to visual inspection[2].

Instead of attempting to apply more complex, but still generic, measures of quantifying information, we took the approach of working closely with the researchers that create and use the images, and identifying the characteristics of various categories of image that were of significance to them. The example in the previous paragraph is apposite, as in certain of the images considered

[2] See, for example, [4], p. 536

the existence and location of quite fine lines was of great significance to the researcher, whereas some parts of the image were of little or no relevance.

3 Rule-based automation of preservation policies

To the researchers who creates and use this data, curation activities should be transparent. Data creators want the assurance that their results will persist after the research project is complete, and data users want continued access to the data in a form that remains usable through technological and cultural change. It is the responsibility of archival staff to ensure this persistence. However, as datasets increase in size and complexity, an approach to curation that involves significant manual activity is not sustainable, in particular since specialised knowledge is frequently required to curate data in particular disciplines. Consequently, there is a need to develop approaches that maximise the automation of preservation activities and involve archive staff only when required.

The methodology followed in the work described here is to represent the preservation policies and procedures formally as rules, which specify the sequences of actions that are taken in particular circumstances, or when certain pre-conditions are satisfied. These pre-conditions may include the occurrence of a triggering event, and assertions about the current state of the preservation system or of objects within it. In addition, the rules can incorporate post-conditions, which support verification of any actions that have taken place, for example that the preservation environment is in a consistent state or that authenticity of the preserved objects has been maintained.

These rules are implemented using the iRODS[3] data grid middleware developed by DICE[4] [9]. A particular feature of iRODS is its Rule Engine, which allows data management policies to be represented in terms of rules comprised of pre-defined sequences of actions that are executed in particular circumstances. Rule execution results in the creation of persistent state information, which can be accessed from within rules to track and control subsequent rule execution. These rules can be executed automatically by iRODS as part of its normal execution, in response to certain conditions or triggers. These rules have great potential for implementing data management strategies that are to take place "under the hood", where the data owners need to be confident that certain processing is occurring, but do not want to concern themselves with it.

3.1 Example

Let us now give an example of the use of these rules for implementing digital curation strategies that integrate the idea of maintaining the significant properties of information objects, as described above. Format conversion takes place typically when an information object is ingested into an archive, and on subsequent

[3] http://irods.sdsc.edu
[4] http://dice.unc.edu/

occasions when it is judged that file formats or associated software packages are in danger of obsolescence, putting information content at risk. In this example, we consider the former scenario.

An iRODS rule is defined as follows[5]:

actionDef|*condition*|*workflow-chain*|*recovery-chain*

where *actionDef* is the identifier of the rule, *condition* defines the circumstances under which the rule will be invoked, *workflow-chain* is the sequence of actions that the rule will execute (separated by ##), and *recovery-chain* is the sequence of actions to be executed in case a failure occurs within *workflow-chain* (that is, it defines how a partially executed rule will be rolled back). The actions in a *workflow-chain* can be either atomic actions, know as "micro-services", or rules, thus a rule can be built up cumulatively from other rules.

It may be represented as an iRODS rule as follows:

acPostProcForPut||
acCheckObjectIntegrity##acAnalyseObject##
acNormaliseObject##msiSysReplDataObj(PresRescGrp,all)|
nop##nop##nop##msiCleanUpReplicas

where *acPostProcForPut* is a system action that is executed automatically when an object is put into an iRODS system, and *nop* indicates that no *recovery-chain* component is executed for the corresponding *workflow-chain* component. The components of the *workflow-chain* indicate the various activities that are carried out when the object is ingested, and may be further broken down until the base components are atomic actions or "micro-services", for example:

acNormaliseObject||
acCharacteriseObject##acConvertObject##
acCharacteriseConvertedObject##acValidateConversion|
nop##nop##nop

Here *acCharacteriseObject* and *acCharacteriseConvertedObject*, each of which may comprise several actions, extract the SPs of the original and converted objects, and *acValidateConversion* verifies that the degree of information loss lies within acceptable limits. iRODS also allows conditional execution of rules (see the rule format described above), allowing different rules and conversion services to be configured for different categories of object.

A second feature of iRODS rules that we can exploit is the ability to specify a number of rule definitions corresponding to the same goal, which can be executed in turn (in a preferred order) until one is successful. Unsuccessful rule executions are rolled back by executing the associated *recovery-chain* workflows. To see why this may be useful, consider again the scenario of format conversion. Conversion tools will not perform perfectly in all cases, so in practice it may be necessary to try several before an acceptable result is achieved. Using iRODS rules we can automate this as in the following example, where *msiConvertImage*<*n*> are two micro-services invoking distinct conversion services, and <*imageCategoryA*> is the category of images being addressed:

[5] See http://www.irods.org/index.php/Rules

acNormaliseObject|
$format == <*imageCategoryA*>|
acCharacteriseObject##msiConvertImage1##
acCharacteriseConvertedObject##acValidateConversion|
nop##nop##nop##msiCleanupNormalisation
acNormaliseObject|
$format == <*imageCategoryA*>|
acCharacteriseObject##msiConvertImage2##
acCharacteriseConvertedObject##acValidateConversion|
nop##nop##nop##msiCleanupNormalisation

If an object falls into the appropriate category, the iRODS Rule Engine will invoke the first matching rule. If one of the micro-services within *acValidateConversion* determines that the conversion was unacceptable (e.g. in terms of information loss), it can return an error, in which case the Rule Engine will roll back the rule execution by calling *msiCleanupNormalisation*, and will execute instead the next matching rule in the list[6]. This procedure can be repeated, executing a number of different conversion implementations in turn (in a preferred order) until one is successful. If no rule is successful, curation staff can be notified that manual action is required.

4 Current status and future work

In this paper we have outlined how to implement an iRODS-based data grid system to support digital curation functionality in large archives of potentially complex scientific data. The iRODS Rule Engine allows complex preservation strategies to be written as rules, which can be triggered automatically when certain events occur, for example the ingest of an object into the repository. As these rules can be implemented conditionally, iRODS implements the event-condition-action model known from active database management systems, providing a great degree of flexibility for implementing automated curation applications. In particular, we can use this method to develop concrete implementations of significant property approaches to digital preservation. The initial prototyping demonstrated the feasibility of this approach, and we were encouraged to begin a deeper analysis, developing more extensive sets of rules and applying our approach of identifying significant properties to more categories of research data

Another issue is the difficulty of expressing the significant properties of an information object in a systematic way across different file formats, which may incorporate quite different data structure. While the more generic characteristics of an object can be expressed using standard schemas such as PREMIS [8], this is more difficult with the varied domain- and community-specific characteristics that we are addressing in our current work. However, this is necessary if we are to be able to extract and compare such characteristics automatically. At present,

[6] Note that in a real-life situation each component of *workflow-chain* will have a corresponding *recovery-chain* component; here for clarity all but the last one are omitted.

this is done in a somewhat *ad hoc* way, indeed some knowledge is embedded in the micro-services themselves (i.e. in the software), which is not desirable. In future work, we plan to look at more general ways of describing these, and in particular the XCL (eXtensible Characterisation Language) ontology developed as part of the PLANETS project [1].

References

1. C. Becker, A. Rauber, V. Heydegger, J. Schnasse and M. Thaller. A Generic XML Language for Characterising Objects to Support Digital Preservation. In *SAC08*, Fortaleza, Ceara, Brazil, 2008.
2. Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS, Washington, DC, 2002.
3. M. Ferreira, A.A. Baptista and J.C. Ramalho. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*, 6(4):295–304, 2007.
4. R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Third Edition, Prentice Hall, 2007.
5. T. Hey and A. Trefethen. The data deluge: an e-Science perspective. In *F. Berman, A. Hey, G. Fox (eds.), Grid Computing: Making the Global Infrastructure a Reality*, John Wiley and Sons, Hoboken, NJ, 2003.
6. J. van der Hoeven, B. Lohman and R. Verdegem. Emulation for digital preservation in practice: the results. *Int. J. of Digital Curation*, 2(2), 2007.
7. G. Knight. Framework for the definition of significant properties. Project report, Arts and Humanities Data Service, London, UK, 2008.
8. PREMIS Data Dictionary for Preservation Metadata. Version 2.0, PREMIS Editorial Committee, Washington DC, 2008.
9. A. Rajasekar, M. Wan, R. Moore, W. Schroeder. A Prototype Rule-based Distributed Data Management System. In *HPDC workshop on Next Generation Distributed Data Management*, Paris, France, 2006.
10. K. Thibodeau. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In *Proceedings of The State of Digital Preservation: An International Perspective*, Washington DC, USA, 2002.