Future Generation Computer Systems 25 (2009) 446–452

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Rule-based curation and preservation of data: A data grid approach using iRODS

Mark Hedges [a,*], Tobias Blanke [b], Adil Hasan [c]

[a] Centre for e-Research, King's College London, UK
[b] Arts and Humanities e-Science Support Centre, King's College London, UK
[c] Department of English, University of Liverpool, UK

ARTICLE INFO

Article history:
Received 16 June 2008
Received in revised form
23 September 2008
Accepted 1 October 2008
Available online 17 October 2008

Keywords:
Data management
Digital preservation
Digital curation
Data grid
Rule-based programming
iRODS

ABSTRACT

Research is generating large quantities of digital material, much of it irreplaceable, and there is a pressing need to maintain long-term access to it. Not only is the quantity of data growing in size, it is becoming much more diverse and complex, significantly complicating the issues around its curation. Automation of curation is key if a scalable solution is to be found. We describe an approach to automation in which digital curation policies and strategies are represented as rules, which are implemented in data grids based on the iRODS middleware.

1. Introduction

An increasingly large quantity of digital data is being produced by research projects, and in addition the data is growing in complexity and diversity. This data represents a significant investment, and is in many cases irreplaceable; consequently there is a pressing need for the digital curation and preservation communities to implement strategies for ensuring long-term access.

This growth in size, complexity and diversity, together with the fact that curation often requires discipline-specific knowledge and experience, implies that, if such strategies are to be scalable, then the systems that implement them must be as automated as possible. One approach to this is to define these policies formally as rules, which specify the sequences of actions that are taken in particular circumstances, together with their pre- and post-conditions, the latter allowing verification of any actions that have taken place.

Here we outline an approach to implementing such automated, rule-based preservation strategies in data grids based on the iRODS (Rule-Oriented Data management System) middleware. The iRODS system incorporates a Rule Engine that allows pre-defined sequences of actions to be executed in particular circumstances, or when particular events occur. The Rule Engine allows iRODS data grids to go beyond the limitations of some other data grid systems, providing a simple, flexible mechanism for implementing application-specific processing, and in particular supporting integration with external systems that can provide the specialised processing or metadata management required for preserving complex digital content. The approach is illustrated by means of a number of simple examples that implement particular preservation workflows and illustrate the various features of the Rule Engine that can be utilised.

The programme of work outlined here represents work being carried out by the Centre for e-Research (CeRch)[1] at King's College London (KCL), and continuing investigations begun by the former Arts and Humanities Data Service (AHDS),[2] which has been partially incorporated into CeRch. The Centre is responsible for the curation and preservation of a quantity of highly diverse and complex digital resources, mainly produced as the outputs of research projects in arts and humanities disciplines, and formerly managed by the AHDS. The AHDS built up a considerable body of knowledge and experience in the curation of this material, which is encapsulated in a well-documented body of policies and procedures, albeit in a form that addresses a largely manual approach. The programme of work on which we are currently

* Corresponding author. Tel.: +44 20 7848 1970.
E-mail address: mark.hedges@kcl.ac.uk (M. Hedges).

[1] www.kcl.ac.uk/iss/cerch/.
[2] www.ahds.ac.uk.

0167-739X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2008.10.003

engaged involves the extraction and formalisation of these policies and procedures as abstract rules, and the development of a concrete implementation of these rules by mapping them onto rules within the iRODS system.

2. Motivations

2.1. Managing the data deluge

Research across the academic disciplines is increasingly both driven by and a generator of data on a large scale, the so-called "data deluge" [6]. The curation of this research data raises significant challenges. In the "big" sciences such as particle physics and astronomy, the primary emphasis has been the management of very large, petabyte-scale data sets, and the support of distributed access to such data, in particular by some form of data grid middleware.

But the issue is not just one of scale. In some disciplines, the emphasis is more on the complexity and richness of context of the data; a digital resource may comprise a number of files of different formats with relationships (possibly semantically tagged) between them. For example, research in the arts and humanities may use combinations of textual resources (enriched with various degrees of mark-up), databases and multi-media objects [4]; medical researchers may use large images with complex, detailed annotations and links to other resources. The complexity of the data is reflected in the complexity of the metadata that must be created and managed to support discovery and re-use. Thus we may speak of a complexity deluge as well as a data deluge.

One approach to managing this complexity has been to use some form of digital repository software [2]. Earlier implementations of digital repositories were used for the most part to manage relatively simple document-based digital objects, such as pre-prints, post-prints and e-theses, where the forms taken and formats used are comparatively standard across academic disciplines. However, they are increasingly being used to manage the complex and diverse data generated by and used in research [10], and in particular scientific data sets [11,12], as well as the output from digitisation programmes. Nevertheless, while such systems facilitate management of complex digital resources and their metadata, they are less focused on addressing issues of scale and persistence of data across time.

2.2. Digital curation

The increasing quantities of digital research data and publications being produced represent a major investment, both of public and private funds, and of the time of the expert researchers and scholars that created them. It would be highly infeasible, and in some cases, such as archaeological excavations and climate records, impossible, to reproduce these digital resources. 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 national archives, digital libraries, and commercial sectors such as banking.[3]

Preservation of digital content implies much more than preserving the constituent bit streams. Binary data remains useful only as long as it can be correctly rendered (displayed, played-back, interacted with) into meaningful content such as text, images and video clips. The process of rendering is performed by a potentially complex combination of hardware and software, which is subject to rapid obsolescence as technology advances, possibly resulting in the objects becoming unusable. Preservation implies a set of coordinated activities that act upon the objects in a repository, over and above simply maintaining the bit streams. These activities may aim to preserve the information content (words, images, sounds etc.) represented in a bit stream, or to maintain the experience (speed, layout, display device, input device characteristics etc.) of interacting with the information content [8].

There are two broad approaches to this: either data must be modified to operate in a new technical environment (format normalisation and migration), or the new environment must be modified so that it can render the old data and replicate the behaviour of obsolete software (emulation) [14]. 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 migrations of objects throughout their life-cycle as formats or rendering software tools become obsolete.[4] Some attributes of a digital object may be lost during conversion, so the experience may not be equivalent after migration. The level of data loss through migration 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-migration analysis. In order to ensure the authenticity of a digital object throughout its life-cycle, detailed audit information must be maintained, and in the case of complex objects with multiple internal and external relationships, all this relational information must also be preserved.

To the creators and users of data, preservation activities should be transparent. Researchers want to deposit the results of their work with the assurance that these results will persist after the research project is complete; users want to be able to discover the data and access it in a form that remains usable throughout technological and cultural change. Responsibility for ensuring this continued accessibility is taken on by the preservation archive. However, digital preservation can involve a great deal of complex activity on the part of archive staff, and a largely manual approach to preservation is not sustainable as deposits continue to increase in number, size and complexity. In particular, as data becomes more diverse and complex, specialised knowledge of the relevant subject domains is needed to curate it. Even if it were in principle possible to deal with this manually, it is likely that staff cost considerations would be prohibitive. Consequently, approaches must be developed that as far as possible automate the processing and other actions involved in digital preservation, and involve archive staff only when a human decision is necessary.

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.

The policies and procedures that form the basis of this representation are derived from the Preservation Handbooks and Ingest Manuals created by the AHDS, which between 1996 and 2008 built up a considerable body of knowledge and experience in the curation of the complex and diverse digital material

[3] www.nationalarchives.gov.uk/preservation/digital.htm, www.digitalpreservationeurope.eu.

[4] Typically, pre-migration versions are not deleted from the archive, so legacy software can continue to make use of them.