Information Systems 44 (2014) 34–53

Contents lists available at ScienceDirect

Information Systems

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

Managing heterogeneous datasets

Mark Scott*, Richard P. Boardman, Philippa A. Reed, Simon J. Cox

Faculty of Engineering and the Environment, University of Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

Article history:
Received 27 February 2014
Received in revised form
11 March 2014
Accepted 13 March 2014
Recommended by: D. Shasha
Available online 25 March 2014

Keywords:
Metadata
Heterogeneous databases
Interactive data exploration and discovery
Data sharing
Digital libraries
Cross-disciplinary applications

ABSTRACT

In disciplines that produce a wide variety of data – such as materials engineering – it can be difficult to provide an infrastructure for storing, managing, sharing and exploring datasets, particularly whilst that data is still in use. The Heterogeneous Data Centre (HDC) is an extension to a file server that provides scientists with tools for exploring their datasets, managing relationships between them and adding metadata. Many of the features evolved from close consultation with our users. In this paper, we evaluate the HDC's interface features for managing datasets using data provided by users from the materials engineering and human genetics domains. In particular, we show the simplicity of capturing data through a file share and the flexibility and extensibility of a system supporting hierarchical metadata, dataset relationships and plug-ins.

1. Introduction

Managing datasets within a research environment at a University whilst the data is still in use presents some challenges that are not addressed by a simple document repository. Many of the systems considered for a research group data management system for materials engineering data were found to provide excellent data management features suitable for the final publication of data. However, users can be daunted by the deposit processes particularly whilst still working with their data [1, pp. 55–56], and capture of data produced by equipment such as X-ray computed tomography scanners and scanning electron microscopes would also require manual interaction with the repository or API to supply any mandatory metadata.

Materials engineers produce very heterogeneous data when testing materials and investigating material failure, creating everything from text data to microfocus computed tomography data files in their day-to-day research, with a lot of these datasets having important relations to each other. The use of portable hard disks is common – for example, during visits to off-site synchrotron facilities or the use of on-site equipment – and many of these datasets are not being shared which is sometimes attributable to confidentiality concerns, but often is due to a lack of resources or motivation [1, pp. 59–63].

The Materials Data Centre (MDC) [2] was a UK Government (JISC) funded project to establish a repository promoting data capture and management in the engineering materials domain [1, p. 5]. We produced a system that would capture more of the materials engineering data by keeping the upload/ingest process as simple as possible, and then provided tools to work with the data to attract users [3]. Once data has been captured it becomes a great deal easier to preserve relevant works in an institutional data archive by transferring applicable data and accompanying metadata. The Heterogeneous Data Centre (HDC) extends this work, adding features and investigating data from the medical domain.

In this paper, we present a system that aims to discourage the use of personal storage or deletion of data,

* Corresponding author.
E-mail address: Mark.Scott@soton.ac.uk (M. Scott).

http://dx.doi.org/10.1016/j.is.2014.03.004
0306-4379/

increase data and metadata capture and to encourage sharing with features that concentrate on four areas:

1. Straightforward data deposition process.
2. Metadata and metadata tools.
3. Dataset discoverability.
4. Dataset tools adding value to users, to encourage uptake and induce users to provide data and metadata.

We show the system's use with materials engineering data and – due to the heterogeneous nature of the repository – we also test the system with medical data, specifically data from human genetic research.

The features are discussed further in Section 3.

2. Related work

There is no shortage of data management solutions, from large-scale data to discipline-specific repositories. The HDC aims to minimise the amount of metadata enforced, but provides tools to encourage users to supply metadata in preparation for later archival. This type of system is sometimes referred to as a local repository, collaboration repository or data staging repository [4,5].

Examples of data management on a huge scale have been shown by, among others, the LOFAR project [6], the Large Hadron Collider [7] and the Human Genome Project [8]. Many of these projects use Grid or cloud technologies to scale to the levels required.

The LOFAR project, a telescope array in the Netherlands, adapts the Astro-WISE system [9,10] for storing data in the order of petabytes. Metadata is stored in a database and files in a file system. The system uses Grid technologies to scale across multiple sites and tools in the Python programming language are supplied. The user interacts with the Astro-WISE system with the Python tools or a web interface to locate files (as it might not be known where files are physically stored), and can download them with HTTP. The Astro-WISE system has been used effectively to manage the masses of data produced in astronomy.

The Human Genome Project was a 15-year international project to identify all nucleotides in human chromosomes, with up to 100,000 genes each having up to 1 million nucleotides [8]. The Large Hadron Collider (LHC) is a superconducting hadron particle accelerator and collider at CERN. The four main detectors produced 13 petabytes of data in 2010 [7] and the ALICE experiment is required to process data at 1.25 GB/s [11]. The *ALICE Environment for the Grid* (AliEn) is a framework that is used by ALICE scientists to process data on Grid computers. The AliEnFS feature [12] of the AliEn Grid service provides the ability for tagging files with metadata. The user creates text files specifying the details of the metadata to be recorded and associates it with one or more folders. The system then automatically creates a table in a relational database with the columns specified and then the user is able to specify metadata values for any file under that folder [13]. This provides tremendous flexibility: folders can be associated with multiple metadata tables and metadata tables can be associated with multiple folders.

The HDC was originally aimed at materials engineers, although it has now been tested with other types of data; another materials repository, the Materials Atlas [14] project, sponsored by the Office of Naval Research (ONR) and the Defense Advanced Research Projects Agency (DARPA), uses collaborative software to collect materials data, experiments and simulation datasets, along with information on the software tools that users of the site may find relevant. The site is powered by Atlassian Confluence – commercial software that provides wiki-like pages, file sharing and other collaborative features. Atlassian Confluence supports files up to 2 GB [15] although it is unclear whether the Materials Atlas can support larger.

An example of other discipline-specific repositories include The Open Microscopy Environment project (OME) in biology [16] which identified the need to share multi-dimensional biological microscopy image data [16] and developed software and protocols allowing image data from any microscope to be stored, shared and transformed without loss of the image data or information about the experimental setting, the imaging system or the processing software [17]. It provides a data model implemented using a relational database where data and metadata are stored, and define an OME XML file format to permit data to be exchanged with other OME databases. The OME XML schema supports the storing of the image data, experiment metadata and results. The database stores the binary image data as well as the metadata about the image acquisition and any processing and analysis done.

2.1. Document repositories

Technologies such as EPrints [18] and DSpace [19] permit storage of publications and can be adapted for data as shown by the eBank project which established the eCrystals repository [20] that managed and disseminated metadata relating to crystal structures and investigated linking datasets from Grid/Cloud-enabled experiments to open data archives and through to peer-reviewed articles using aggregator services. This allowed crystal structure data to be provided with a paper for a reader to check validity.

Data Dryad [21] is a good example of a data archival repository that is built using DSpace. DSpace stores files on the file system or can integrate with Storage Resource Broker (SRB), a distributed file system. The API or the interface must be used to upload files into the file store as DSpace gives each file a unique name that it generates itself. Files cannot be loaded directly into the file store as the system will not recognise them. This is a similar approach to Jackrabbit [22] where names of files are generated using a hash of the content removing the possibility of duplicated files. This means that users must manage their files and metadata through the repository's interface.

2.2. Metadata formats

Archival repositories expect metadata to accompany the item being deposited. Dublin Core [23] is one of the more commonly used standards but, depending on the