Information Systems 42 (2014) 36–58

Contents lists available at ScienceDirect

Information Systems

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

A framework for user driven data management

Mark Scott [a,*], Richard P. Boardman [a], Philippa A. Reed [a], Tim Austin [b], Steven J. Johnston [a], Kenji Takeda [a], Simon J. Cox [a]

[a] Faculty of Engineering and the Environment, University of Southampton, SO17 1BJ, United Kingdom
[b] European Commission, Joint Research Centre, Institute for Energy and Transport, PO Box 2 1755 ZG Petten, The Netherlands

ARTICLE INFO

Article history:
Received 15 October 2013
Received in revised form
25 November 2013
Accepted 29 November 2013
Recommended by: D. Shasha
Available online 10 December 2013

Keywords:
Digital libraries
Document management
Heterogeneous databases
Scientific databases
Dissemination
Data models

ABSTRACT

Scientists within the materials engineering community produce a wide variety of data, with datasets differing in size and complexity. Examples include large 3D volume densitometry files (voxel) generated by microfocus computer tomography (μCT) and simple text files containing results from a tensile test. Increasingly, there is a need to share this data as part of international collaborations. The design of a suitable database schema and the architecture of a system that can cope with the varying information is a continuing problem in the management of heterogeneous data. We present a model flexible enough to meet users' diverse requirements. Metadata is held using a database and its design allows users to control their own data structures. Data is held in a file store which, in combination with the metadata, gives huge flexibility. Using examples from materials engineering we illustrate how the model can be applied.

1. Introduction

With the volume of scientific research data increasing rapidly [1,2], and institutions and funding councils spending large amounts of money on the resources required to produce it, proper curation and dissemination could have huge advantages for future work.

The Materials Data Centre (MDC) [3] is a UK Government (JISC) funded project to establish a repository promoting data capture and management in the engineering materials domain [4, p. 5].

Responses to a questionnaire distributed at the start of the project identified the use of many types of data in materials engineering [4]. Some engineers need to store very structured data such as MatDB [5], a standard schema able to represent data from tensile, creep and fatigue tests as XML. Others just want a large data store where they can save raw, unprocessed image data until they are ready to use it. A few require the ability to create information sheets about the materials they use and link them to the results of experiments, and some just want to store raw data CSV files containing their results in order to support findings in a paper.

The variety of dataset types, sizes, metadata and relationships between datasets create a complex problem in terms of storing, managing and retrieving. The model we present in this paper, created for the Materials Data Centre project and aimed at materials engineers, addresses these issues by permitting storage of heterogeneous data, accompanying metadata and relationships.

* Corresponding author. Tel.: +44 2380598374.
E-mail address: Mark.Scott@soton.ac.uk (M. Scott).

0306-4379/$ - see front matter
http://dx.doi.org/10.1016/j.is.2013.11.004

1.1. Guiding materials datasets

The data from the following materials engineering tests were identified for storage in a prototype system. The references provided are to materials engineering research produced by engineers (the prospective users of the Materials Data Centre) giving concrete examples of the data they produce and how it is used in their work.

- Tensile test [6,7].
- S-N fatigue test [8,9].
- Fatigue crack growth [10,11].
- Impact/toughness test [12,13].
- Fractography [14,15].

The datasets generated by these tests are diverse in size and complexity. Results produced by tensile and fatigue tests can be as simple as comma separated numeric data while fractography uses computed tomography and produces files containing 3D voxel densitometric data which can be 32 GB each, and this is likely to increase as technology improves.

The relationships between these tests are also important as can be seen in Fig. 1. For example, the results of a tensile test can be used to calculate how a crack will grow in a fatigue crack growth test, and a material's composition and other particulars are useful when performing or reviewing the results of most tests.

1.2. Architecture

To fully support the wide ranging data types and relationships, we combined a file system with a database for holding metadata. A synchronisation service keeps the metadata held in the database consistent with the file system and the database allows users to create their own data structures without affecting others or requiring an administrator.

Choosing a file store as the base of the design ensures flexibility and reliability and allows the system to seamlessly take advantage of developments in file system technologies and benefit from more advanced file systems such as a distributed file system.

The final system we present provides the following features:

- Storage and sharing of any data files, limited in size only by the file system.
- The ability for users to define their own data structures, with nesting, for situations where the data is not defined in a data file.
- Addition of metadata to a dataset.
- Support of the relationships between datasets.
- Allow for predefined metadata to be created as a template and for metadata to be copied between datasets.
- Encourage use of the system by providing dataset recommendations, blogs, wikis and message boards.
- Allow plugins, in order for the system to provide customised reports and tools depending on the data.

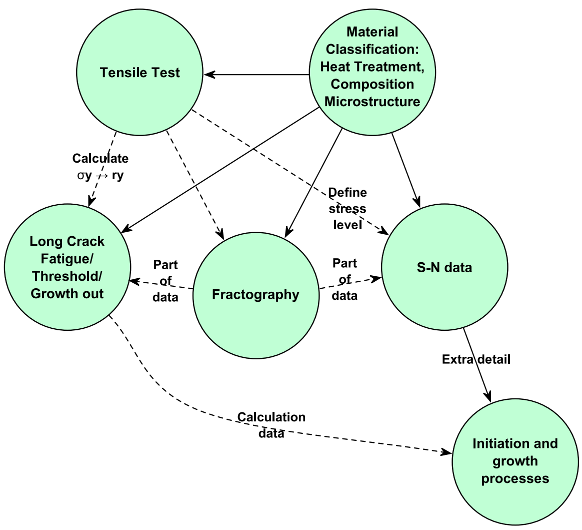

Fig. 1. Relationships between the different types of test data, as represented in the Materials Data Centre.