EP2DC—an EPrints Module for Linking Publications and Data

TSP Austin [1], CM Scott [2], SJ Johnston [3], PA Reed [4], K Takeda [5]

[1] ***University of Southampton***
School of Engineering Sciences, University of Southampton, SO17 1BJ, Southampton, UK
EMail: taustin@soton.ac.uk

[2] ***University of Southampton***
School of Engineering Sciences, University of Southampton, SO17 1BJ, Southampton, UK
EMail: mark.scott@soton.ac.uk

[3] ***University of Southampton***
School of Engineering Sciences, University of Southampton, SO17 1BJ, Southampton, UK
EMail: s.j.johnston@soton.ac.uk

[4] ***University of Southampton***
School of Engineering Sciences, University of Southampton, SO17 1BJ, Southampton, UK
EMail: p.a.reed@soton.ac.uk

[5] ***University of Southampton***
School of Engineering Sciences, University of Southampton, SO17 1BJ, Southampton, UK
EMail: ktakeda@soton.ac.uk

ABSTRACT

Different research perspectives, contexts, and competencies almost certainly result in individual researchers interpreting data in different, sometimes radically different ways, and it is simply not credible to expect a single researcher or research group to gain the full potential from a given data set. Shared data thus represent an untapped resource of significant potential and, if the scientific community is to benefit to the full, it is essential that effective data conservation and sharing practices become embedded in mainstream research. To this end, EP2DC is a prototype EPrints plugin designed to support the submission of experimental data sets together with the manuscript to which they correspond. Its development recognises the worth and potential for reuse of high quality research data, and is consistent with trends in scientific publishing and funding agency policies that advocate a more responsible approach to conserving research data.

Keywords: ep2dc, data, reuse

INTRODUCTION

Recent years have seen a renewed interest in research data management, the motivations for which can be traced to emerging technologies, legal imperatives, and the scientific establishment. From the technology perspective, the emergence of a semantic web of data offers new opportunities for knowledge discovery, while established computational methods stand to benefit from the availability of an ever increasing body of quality data. More recently, this interest in data conservation and reuse has been further fuelled by increased scrutiny (as in the case of the UEA climate data controversy [1]) and regulatory compliance (as in the case of the tree rings FoI ruling [2] and more recently the UEA climate data itself). Together with the introduction of data conservation and sharing policies both by funding agencies [3] and publishing houses [4][5][6] the indications are that data management is destined to become an integral part of mainstream research.

Delivering Viable and Sustainable Data Management Solutions

At a time when there is renewed enthusiasm for conserving and sharing research data, it is crucial that effective and sustainable solutions are forthcoming. Otherwise there is the risk that a research community newly motivated to address long-standing research data management issues will become cynical of poorly implemented solutions and unrealistic expectations. Further, with the emergence of exciting new Semantic Web technologies, there is a very real temptation to deliver sophisticated solutions at the expense of the end-user community. As evidenced by a recent RIN report, the biological sciences appear to have suffered exactly this fate, with the research community overwhelmed by the extent and diversity of emerging data management technologies, leaving researchers dissociated from the technologies they are supposed to be using [7]. Unless there are successful and visible examples delivering tangible benefits, there is a very real danger of the research community becoming disillusioned with efforts to conserve and reuse research data.

Research Data Management in the Engineering Sciences

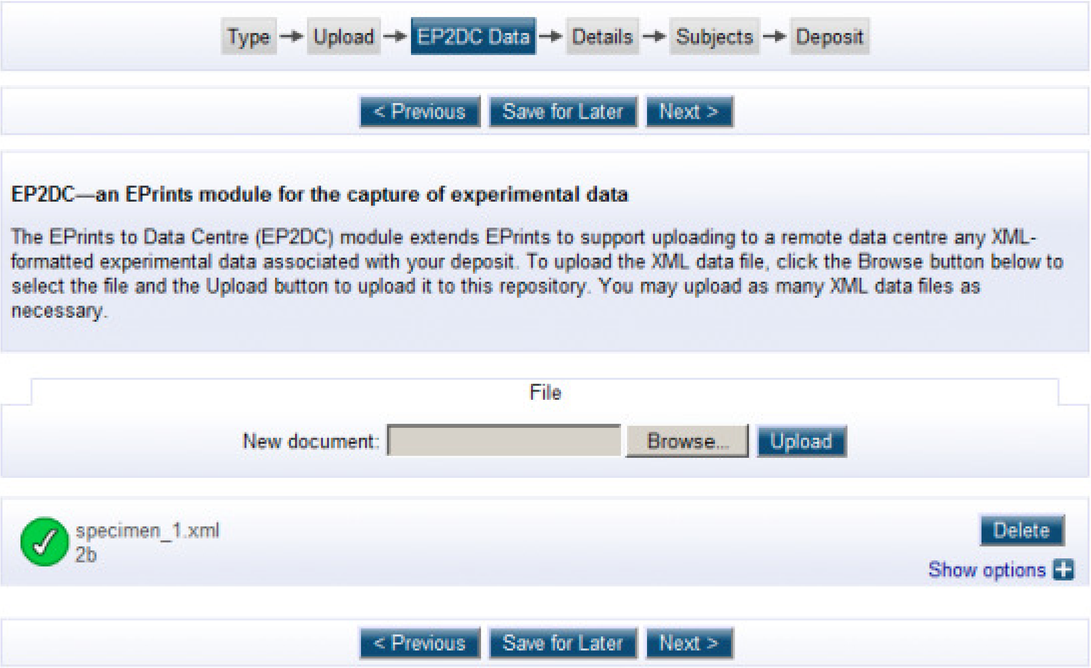
Compared to the natural sciences, research data management in the engineering sciences has received limited attention. Studies commissioned by the DCC indicate this circumstance can be attributed to a concern that the sharing of data has the potential to damage long-established industrial links [8][9]. In the UK, this problem is aggravated by the fact that of the eight research councils, the two that lack policy on research data management are the Engineering and Science Research Council and the Structural Funds Council. However, irrespective of these concerns, in the absence of services that allow researchers in the engineering sciences to capture, conserve, and share their data, any assertion about the state of research data management in the engineering sciences remains largely hypothetical and cannot be verified. There are indications though of a changing attitude, and a number of recent initiatives, including the ELSSI-EMD CEN Workshop and the JISC-funded MDC and EP2DC projects, have gone some way to contribute to a research data management infrastructure for the engineering sciences. CEN WS/ELSSI-EMD has delivered standards-compliant data formats that promote the capture and exchange of engineering materials data, and MDC provides a data centre for hosting and sharing a wide range of types of data. EP2DC leverages both CEN WS/ELSSI-EMD and MDC to deliver a service that promotes the collection and sharing of data. It is an extension to the EPrints repository that allows data sets hosted at specialized data centres to be referenced from the publications to which they correspond. Being easily integrated into any EPrints instance, EP2DC offers great potentially for promoting data capture and publication, not only in the engineering sciences but in any discipline, and provides a blueprint for extending existing publications repositories, whether they be open access or commercial.

EPrints

EPrints is used at more than 250 institutes world-wide. It belongs to a category of applications (which include Fedora and DSpace) designed primarily as repositories for various types of scholarly publication, including articles, book sections, patents, images, and audio files. Although there is support for conserving data sets, EPrints stores these in its own repository, and consequently is not able to take advantage of the services offered by specialist data centres. The EP2DC prototype is designed to overcome this limitation and allow researchers to submit data sets to remote data centres during the EPrints manuscript submission process. This is a new publishing model for data, providing seamless upload and retrieval of data using a federated architecture. This helps elevate data sets to first-class objects, with the potential to accelerate research and also increase citation rates—an important incentive for academics given the ever more stringent research audit processes.

EP2DC DESIGN AND IMPLEMENTATION

The EP2DC system is implemented as an EPrints plugin that introduces an extra step in the EPrints workflow and supports a RESTful service that allows integration with remote data centres. As shown in Figure 1, the extra step in the workflow provides the user the opportunity to submit their data to a remote data centre when uploading a paper.



EP2DC—an EPrints module for the capture of experimental data

The EPrints to Data Centre (EP2DC) module extends EPrints to support uploading to a remote data centre any XML-formatted experimental data associated with your deposit. To upload the XML data file, click the Browse button below to select the file and the Upload button to upload it to this repository. You may upload as many XML data files as necessary.

✓ specimen_1.xml
2b

Figure 1. EP2DC workflow.

Key factors in the EP2DC design include ease-of-use, scalability, and data quality.

Scalability

With the objective of promoting its adoption across a range of disciplines, the EP2DC service layer is designed to be data centre agnostic, and as such provides methods that allow a data set to be accepted by any remote client that is compliant with its RESTful operations. The architecture is depicted in Figure 2.

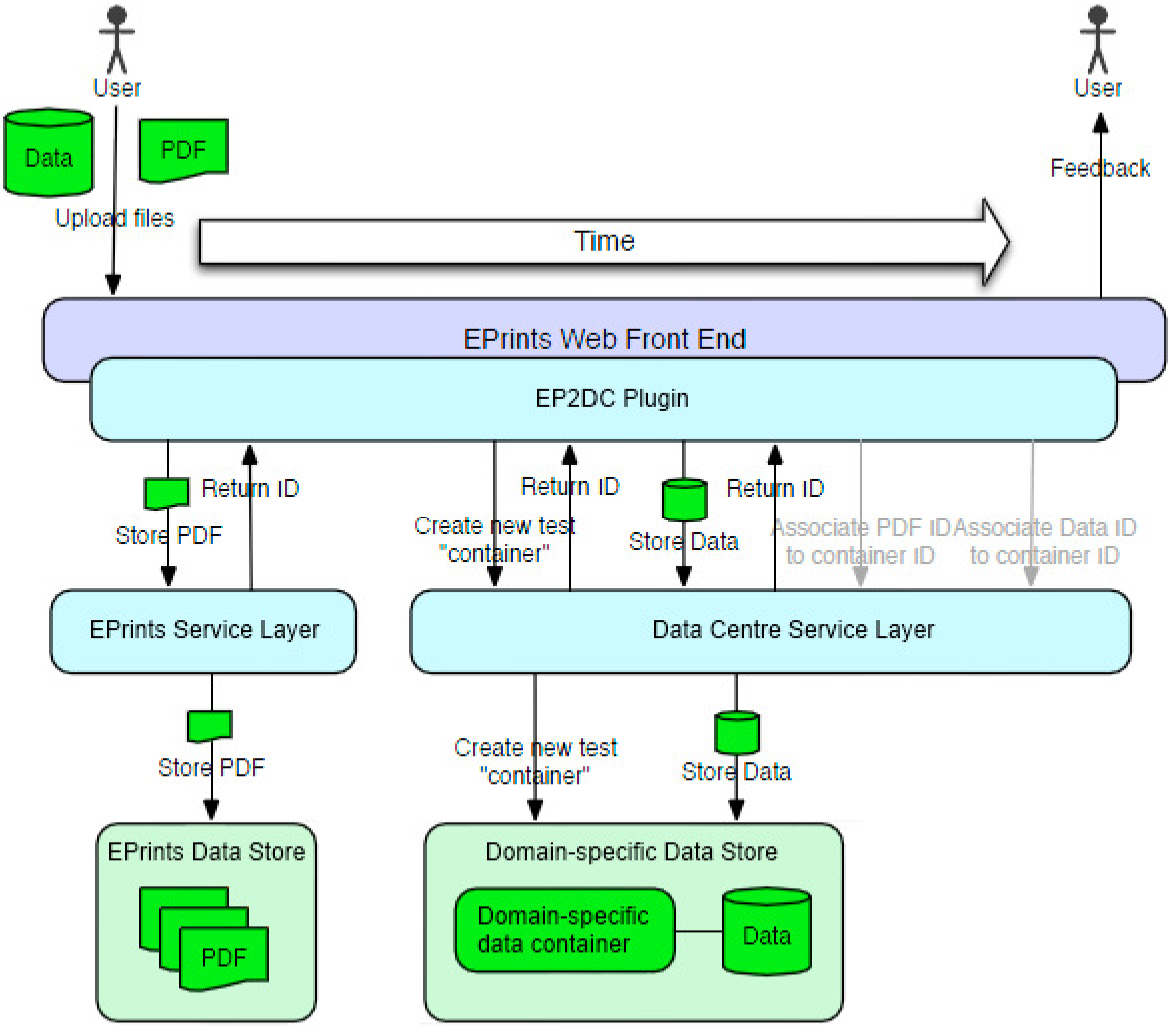

User
Data
PDF
Upload files
Time
User
Feedback
EPrints Web Front End
EP2DC Plugin
Store PDF
Return ID
Create new test "container"
Return ID
Store Data
Return ID
Associate PDF ID to container ID
Associate Data ID to container ID
EPrints Service Layer
Data Centre Service Layer
Store PDF
Create new test "container"
Store Data
EPrints Data Store
PDF
Domain-specific Data Store
Domain-specific data container
Data

Figure 2. EP2DC architecture.

EP2DC functionality is exposed through a RESTful service. This permits upload of data (with or without validation), associating items together, requesting information on existing links, and obtaining a list of data items that the data centre deems to be related using the metadata it has available. Technical information on the use of the RESTful service for data centre integration is available from the CodePlex EP2DC documentation page at http://ep2dc.codeplex.com/documentation.

Ease-of-use

The EP2DC plugin is intended to be straightforward to install and easy to use. The plugin is compatible with EPrints version 3.1 and higher. The installation procedure is described at http://wiki.eprints.org/w/EP2DCOverview#EP2DC_Plugin_Installation. An instance of the EP2DC prototype is hosted at http://ep2dc-s1.soton.ac.uk.

The RESTful service makes communication with the data centre simple and independent of technology. The service itself is hosted on a Microsoft Windows Server, with IIS and Microsoft .NET Framework enabled.

While the EP2DC plugin supports integration with any data centre, the data centre chosen for this prototype, which the RESTful service uses to store received data, is the JISC-funded Materials Data Centre (MDC) hosted at https://mdc-s1.soton.ac.uk by the University of Southampton. The MDC was originally implemented using Microsoft SharePoint, but a later version simply used a file system in order to support larger data files. This is also documented at the CodePlex site.

To trial the service, sample files and data sets can be uploaded following the procedure described at http://wiki.eprints.org/w/EP2DCOverview#Trialing_the_EP2DC_Service.

Data Quality

The prototype system is designed to process structured data. For example, on receiving an XML file from EPrints, it will ensure the contents are valid by checking them against the corresponding schema definitions. A separate part of the data centre is available for storing XML schemas for validation or standard XML syntax is used to identify a schema stored on the Internet. On failure of the validation, the deposition fails and the EP2DC plugin is notified. The validation can also be turned off in circumstances where it is not required.

DISCUSSION

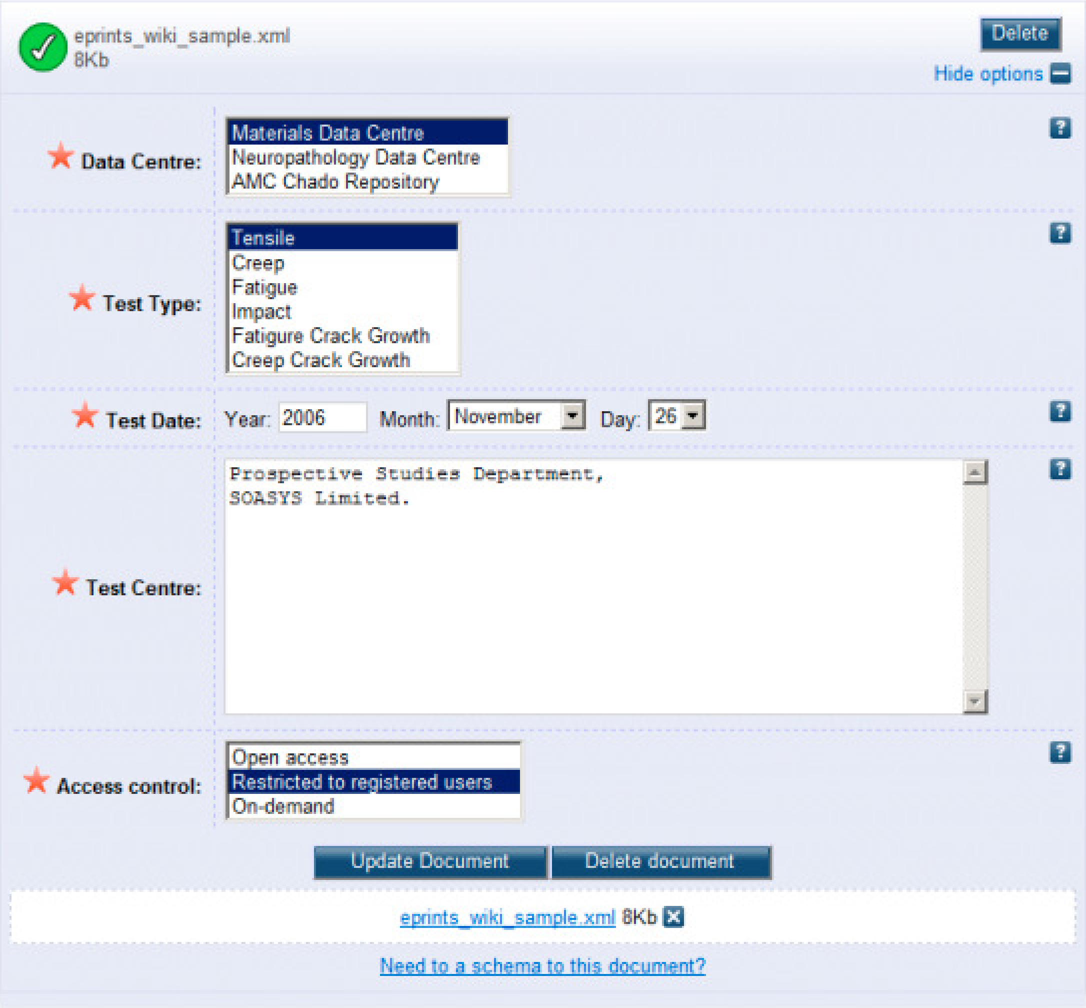
Requirements gathering efforts and reviews indicate that researchers are concerned with security and ease of use. These requirements led to the decision to extend EPrints and leverage its inherent security model and usability features.

EP2DC Utilisation

The RESTful service ensures that the EP2DC plugin is data centre agnostic, thereby allowing any data centre to integrate with the EP2DC plugin irrespective of content type, thereby delivering a scalable, high availability service. In the engineering materials sector, such data centres include the Online Data and Information Network hosted at https://odin.jrc.ec.europa.eu by the European Commission Institute for Energy and Transport and the JISC-funded MDC hosted at https://mdc-s1.soton.ac.uk by the University of Southampton. There are, however, many other data centres serving the data management needs of other disciplines. With relatively little effort, such data centres could be enabled to support integration with the EP2DC plugin, and hence enable institutes that host an EPrints instance to use the EP2DC plugin to promote the conservation of data sets together with the publications to which they correspond.

Data Sharing Practices

While the merits of sharing data are generally acknowledged, it is naïve to expect researchers to share data sets without having control over the terms of their use. EP2DC leverages the inherent access management features of EPrints to ensure that the owner of the data has the opportunity to benefit from making the data available to the broader scientific community. Thus, as shown in Figure 3, the EP2DC stage for uploading experimental data includes a mandatory field to define the access control.



eprints_wiki_sample.xml
8Kb

★ Data Centre: Materials Data Centre
Neuropathology Data Centre
AMC Chado Repository

★ Test Type: Tensile
Creep
Fatigue
Impact
Fatigure Crack Growth
Creep Crack Growth

★ Test Date: Year: 2006 Month: November Day: 26

★ Test Centre:

```
Prospective Studies Department,
SOASYS Limited.
```

★ Access control: Open access
Restricted to registered users
On-demand

eprints_wiki_sample.xml 8Kb

Figure 3. EP2DC access control.

In Figure 3, the access control field affects the data retrieval process, as follows:

- Open—allows data retrieved from the EPrints repository to be downloaded by anyone.
- Restricted to registered users—allows data retrieved from the EPrints repository to be downloaded by registered users.
- On demand—data is supplied on request direct from the owner.

Equally, for the end-users of shared data, some guarantee of quality is needed to ensure that the data add value to their research. For this reason, EP2DC is designed to process structured, validated data.

Adding Value to the Research Process

Motivations for researchers to share their data include establishing a means by which their contribution to the research process can be acknowledged in a tangible way. The EP2DC prototype allows publications stored at an EPrints repository to be linked to corresponding data sets at remote data stores. These data sets can then be made available under conditions that suit the owner, such as requiring the data set to be cited in any derivative works.

Knowledge Discovery

With the emergence of a semantic web of data and its associated technologies, the availability of an ever-increasing body of high quality, linked data offers the prospect of exciting new opportunities for knowledge discovery. At present, this potential is manifest in EP2DC simply as support for disproportionate feedback (the mechanism by which information about data of potential interest is returned in the response to the original request) but there is clearly potential for innovation.

CONCLUSION

EP2DC is an EPrints plugin that links data to knowledge reported in scientific publications and provides the opportunity for researchers to benefit from sharing their data. Its design and implementation make it amenable to integration with any data centre, in which case institutes that host an EPrints instance can use the EP2DC plugin to promote the conservation of data sets together with the publications to which they correspond.

REFERENCES

[1] M. Russell, J. Boulton, P. Clarke, D. Eyton, and J. Norton, The Independent Climate Change E-mails Review, 2010.

[2] Information Commissioner’s Office, Freedom of Information Act 2000 (Section 50) Environmental Information Regulations 2004, Decision Notice, Reference FS50163282, 2010.

[3] Digital Curation Centre Overview of Funders' Data Policies. Retrieved 12th September, 2011 from http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies.

[4] Data's shameful neglect. Retrieved 12th September, 2011 from http://www.nature.com/nature/journal/v461/n7261/full/461145a.html.

[5] Nature Publishing Group – Authors and Referees - Availability of data & materials. Retrieved 12th September, 2011 from http://www.nature.com/authors/editorial_policies/availability.html.

[6] Open Access Directory. Retrieved July 7, 2011 from http://oad.simmons.edu/oadwiki/Main_Page.

[7] Patterns of information use and exchange: case studies of researchers in the life sciences. A report by the Research Information Network and the British Library, November 2009. Retrieved 12th September, 2011 from http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/disciplinary-case-studies-life-sciences.

[8] Curation of research data in the disciplines of Engineering (2010). Retrieved 12th Septmeber, 2011 from http://www.dcc.ac.uk/sites/default/files/documents/publications/case-studies/SCARP_B4812_EngCase_v1_2.pdf.

[9] Howard, T., Darlington, M., Ball, A., Culley, S., McMahon, C., 2010. Understanding and Characterizing Engineering Research Data for its Better Management. Project Report. Bath, UK: University of Bath, (ERIM Project Document erim2rep100420mjd10).

The EP2DC plugin is the realization of the JISC-funded EP2DC Rapid Innovation Project, and the support of JISC, both financial and managerial, is gratefully acknowledged.