# Towards research data cataloguing at Southampton using Microsoft SharePoint and EPrints: a progress report

# Steve Hitchcock and Wendy White\*

JISC DataPool Project

Faculty of Physical and Applied Sciences, Electronics and Computer Science, Web and Internet Science, University of Southampton, UK

\* Hartley Library, University of Southampton, UK v1.0, 17 May 2013

#### **Abstract**

Researchers are increasingly required to manage research data effectively in terms of cataloguing and storage for visibility and access, and longer-term preservation. The IISC DataPool Project at the University of Southampton has been developing a data cataloguing facility based on Microsoft SharePoint, which provides a number of IT services at the university, and on EPrints repository software. We report progress along both repository routes, notably collaboration with the JISC Research Data @Essex project on ReCollect, a standards-based data deposit application for EPrints. The ReCollect plugin provides an EPrints repository with an expanded metadata profile for describing research data based on DataCite, INSPIRE and DDI standards. Repositories can customise the data deposit workflow, the ordering and paging of fields from the profile, without affecting standards compliance. An example of customising the workflow for the ePrints Soton institutional repository is compared with the ReCollect original. In the case of SharePoint, user interface forms for creating data records have been piloted and tested. The approach is distinctive in creating two linked forms, one to describe a project, the other to record a dataset, rather than a single workflow as in the case of EPrints. This development of SharePoint for description and storage of research data is part of a longer-term extension and integration of services provided on the platform at Southampton.

# 1 Introduction

A primary requirement for institutional research data management (RDM) services is a mechanism for data capture. This typically involves a means of supporting the upload of data to the designated storage, curation and presentation service, and describing the uploaded data by means of metadata. Some RDM projects have designated, or redesignated, institutional repositories

for this purpose, e.g. University of the West of England, Queen Mary University of London, or have built new services, e.g. DataFlow at Oxford. These approaches all support the key services of upload and description.

Metadata and EPrints customisation for the UWE data repository: objectives, requirements and standards for a data repository, Jan 2013, download from <a href="http://www1.uwe.ac.uk/library/usingthelibrary/researchers/manageresearchd">http://www1.uwe.ac.uk/library/usingthelibrary/researchers/manageresearchd</a> ata/managingresearchdata/projectoutputs/workpackage3.aspx
And the winning platform is..., Research Data Management at the Centre for Digital Music (C4DM), Queen Mary University of London, 05/01/2012
<a href="http://rdm.c4dm.eecs.qmul.ac.uk/platform\_choice">http://rdm.c4dm.eecs.qmul.ac.uk/platform\_choice</a>
DataFlow project, University of Oxford <a href="http://www.dataflow.ox.ac.uk/">http://www.dataflow.ox.ac.uk/</a>

The JISC DataPool Project at the University of Southampton has been developing both new and existing services for data capture, focussing respectively on Microsoft SharePoint and EPrints. This is principally because we have support groups at the university for these respective softwares. EPrints repository software was originally developed at Southampton in 2001 and continues to be developed and supported nationally and internationally by EPrints Services, a commercial organisation based here. SharePoint, though clearly not developed at Southampton, is extensively supported by the iSolutions institutional IT team as a platform for a wide range of services at the university, potentially also extending to RDM.

As existing repository software, EPrints provides native interfaces for data capture and description that can be adapted for RDM. SharePoint, while connecting to a more extensive underlying architecture of potentially useful services, does not come with such interfaces, which have had to be developed by iSolutions working with the project.

From the outset DataPool was concerned with the practicalities of RDM and is agnostic about the routes to achieve this, but has an expectation that any input routes would lead to the same, institutionally managed RDM services, notably for curation, storage and archiving. Underlying data deposit and access must be a robust and secure data management service, potentially including a mix of local and external storage services. Further investigation of cost options and capacity planning is ongoing.

It is anticipated that, as well as for local deposit, repository interfaces will be developed to support parallel deposit with external data repositories (e.g. Archaeology Data Service) using protocols such as SWORD, as well as to transfer data to and from such repositories. Special consideration will be given to handling of large datasets. Equally important, as content grows, are the search and access interfaces to find research data, well established for open access repository software such as EPrints. For data repositories there is a need to support greater control of access by creators and depositors than is assumed for open access repositories.

An earlier presentation in November 2012 considered the development of SharePoint and EPrints as emerging research data repositories in the context of high-level 'architectural' and practical 'engineering' challenges. Here we report further progress along both repository routes, notably collaboration with the Research Data @Essex project to implement a standards-based data deposit application for EPrints, during the DataPool project to the end of March 2013.

To architect or engineer research data repositories, DataPool Project, Dec 17, 2012

http://datapool.soton.ac.uk/2012/12/17/to-architect-or-engineer-research-data-repositories/

# 2 Data repository vision at Southampton

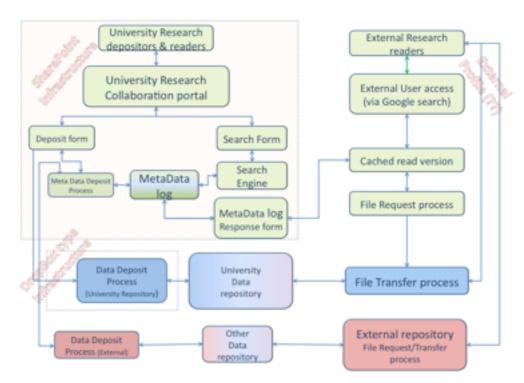


Figure 1. Architectural diagram, produced by Peter Hancock, director of the iSolutions IT services provider at the University of Southampton. Although it leans heavily towards referencing SharePoint, it can be viewed as a high-level reference model

Data repository development in DataPool, particularly of SharePoint, was guided from the outset by a high-level reference model (Figure 1). This showed how SharePoint, as a data description service, might interface to a Dropbox-like data deposit infrastructure, and also to external search and discovery services.

Data description and deposit workflow in this model was informed by a 3-layer metadata approach (Figure 2) elaborated in the Institutional Data Management Blueprint (IDMB) project, the JISC RDM project that preceded DataPool at the University of Southampton.

Takeda *et al.*, Data management for all: the Institutional Data Management Blueprint Project, *6th International Digital Curation Conference, Chicago*, December 2010 <a href="http://eprints.soton.ac.uk/169533/">http://eprints.soton.ac.uk/169533/</a>

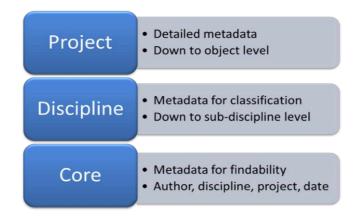


Figure 2. Three-layer metadata model, from the Institutional Data Management Blueprint (IDMB) Project

The 3-layer metadata model was adapted by Research Data @Essex in its work with EPrints, and can be seen quite clearly in the emerging user interface for data deposit built on SharePoint, as we shall see in the next section.

Re-using the IDMB toolkit (part 1), Research Data @Essex blog, May 28, 2012 <a href="http://web.archive.org/web/20130331063948/http://researchdataessex.posterous.com/report-on-the-idmb-toolkit">http://web.archive.org/web/20130331063948/http://researchdataessex.posterous.com/report-on-the-idmb-toolkit</a>

# 3 SharePoint at Southampton

SharePoint is an established Web application platform introduced by Microsoft in 2001. The platform provides a range of Web tools, including intranet portals, document and file management, collaboration, social networks, extranets, websites, enterprise search, and business intelligence.

Microsoft SharePoint, Wikipedia <a href="http://en.wikipedia.org/wiki/Microsoft\_SharePoint">http://en.wikipedia.org/wiki/Microsoft\_SharePoint</a>

In other words, it is powerful, flexible, and can be used to integrate tools and services that it supports, at a possible cost of complexity and long development times for new tools and services developed locally. An example of such a development would be the support for research data management built into SharePoint by the iSolutions team at the University of Southampton.

#### 3.1 Institutional case for SharePoint

Why would an institution embark on a large-scale exercise to build such critical infrastructure based on SharePoint? Simon Cox, professor of engineering at the university, gives a passionate case for SharePoint: for long-term project

management, collaboration and data management, for convergence of tools, concentration of intellectual property, integration of research and teaching experiences, and to provide controlled visibility for the university, instead of outsourcing content management to external organisations. In this context Simon explains the case for longer-term development of SharePoint as an institutional resource than can be completed in DataPool:

"The concept that formed part of SharePoint thinking from the very inception – there is one purpose ... that ability to use SP as a way to manage or at least collaborate as part of a 5-10 year programme of work.

"The other side is custodianship of data over a much longer period of time. This is about normalising and looking after your data, in the same way we have a publication mechanism for papers.

"We are feeling our way with the SP platform in a number of places - what do you get out of the box, what can you customise, where can you go with it? Otherwise the university might have bought 57 packages each of would have been the world's best at ONE part of the problem – supporting all of these becomes a real headache and potentially huge cost.

"SP provides certain things – not the answer to everything – but a common base to do quite a lot of things, as we decide how to take things forward in the university.

"The other side is what we're offering for students as an experience. I run individual and group design projects, and every single student says 'I just do it all on Dropbox'. The same is happening with our research. So I think we have at least to provide a level of service and a level of integration between our research experience and our teaching experience. Would these people go to the University of Southampton rather than University of Nowhereshire on the Web? These are deep questions for us but this technology can genuinely help further our research and enterprise-led teaching agenda."

# 3.2 SharePoint for research data management

Given these broad ideals intended to emerge over years, initial development of SharePoint for capturing research data is necessarily modest in comparison. Unlike EPrints, which already had an integrated workflow for research data deposit, to build similar workflow in SharePoint had to begin from scratch.

A two-stage approach was adopted, with linked forms designed to capture information about projects, and the collected datasets that emerge from those projects. Figure 3 a, b show these two forms, for project and datasets respectively, in their current state of development. These forms reveal the conceptual design and thinking behind this approach, which is in part determined by the structure and facility provided by SharePoint, but also represents a fresh perspective on the required dataset deposit workflow.

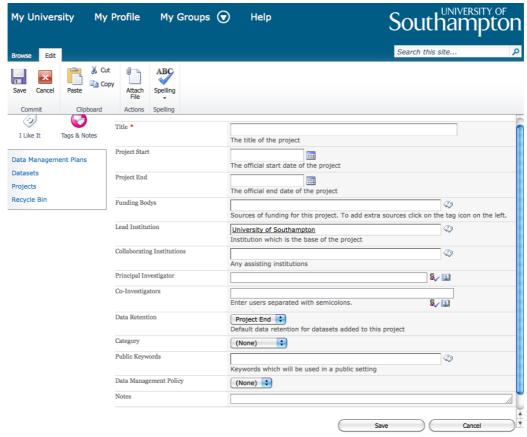


Figure 3a. Collecting metadata in SharePoint about a research project

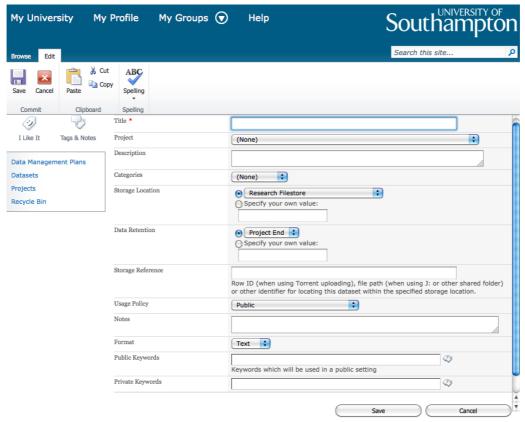


Figure 3b. Collecting metadata in SharePoint about a dataset produced by a research project

Details collected about a Project define its length, specify the research funder, indicate institutions and investigators involved, and provide simple, default descriptions for subsequent datasets that are linked to the Project.

One noticeable feature within both the Project and Dataset forms is the single mandatory field (indicated with a red asterisk) on each form. Mandatory fields have to be filled in for the form to Save successfully. In these cases you could feasibly submit a project or data description containing only a title.

The Dataset form does not provide a means of storing the actual data, just the metadata description of the data. Instead it captures pointers to the storage location and reference.

The dataset is described in two text boxes (Description and Notes), by simple drop-down categorization and by authored keywords. Selection of keywords is assisted by a linked collection. The length of time the data is to be retained, in terms of number of years or to project end, and who is allowed to see the data, are also specified by the depositor. The dataset is linked to the project via selection from a drop-down list informed by the prior Project form, and some incomplete fields will default to the values provided in that form.

# 3.3 Testing SharePoint RDM deposit interface

Five postgraduate students and a principal investigator from multiple disciplines performed initial formative testing of the SharePoint data deposit interfaces in November 2012, under the supervision of the SharePoint developers and DataPool team.

Participants were asked to bring examples of their research data so that they could complete first the Project and then the Dataset deposit forms using real examples, recording any comments or difficulties they encountered on a separate sheet provided. Most participants were unfamiliar with SharePoint and what it could do. The session ended with a demonstration of the collaborative space it provides and other key features.

The type of research being carried out by the participants had an impact on the potential usefulness of SharePoint for them as individuals, but gave the developers valuable insights into potential workflows and issues associated with the existing functionality. The results of this initial testing were documented and reported internally to the SharePoint team for further consideration.

# 4 EPrints research data apps

EPrints is software (eprints.org) used to build digital repositories, to describe, store, manage and access a wide variety of data types. It was developed at the University of Southampton and launched in 2001 as the first OAI-compliant software to support institutional repositories.

The first page of a standard EPrints deposit process asks the depositor to identify the type of content to be deposited, from a list including Article, Book or Conference item. In 2007 this list was expanded to include types such as Dataset and Experiment, both types that support the submission of research data. Selection of one type of data presents the depositor with a series of pages and fields that are designed to be appropriate for the description of the selected type. The order in which these pages and fields are presented defines the deposit 'workflow' for the data type, and is typically customised to specific repository implementations by institutions.

#### 4.1 EPrints Bazaar: store for one-click apps

The range of data types supported and associated forms for describing them are expanded through continued development and releases of the software. Additional repository functionality and interfaces were provided in the installed software until a major change in 2007, with the release of EPrints 3.0. This restructured EPrints to support a modular framework, allowing extra functionality to be developed separately of the main software and delivered as independent plugins or applications. With EPrints version 3.3 in 2012, applications that could be installed in an EPrints repository with one click could be advertised and distributed in the Bazaar, an EPrints 'app store'.

Bazaar Storerepository <a href="http://bazaar.eprints.org/">http://bazaar.eprints.org/</a>

# 4.2 Rise of research data apps

Among the apps in the Bazaar are some that will implement deposit workflow for research data deposit. Two original research data apps were: Data Core, which minimises the number of deposit fields and thus speeds up deposit but at the expense of the extent of description (Figure 4); and Datashare, resulting from the IDMB project, which added a couple of new fields to the default workflow for the dataset type but has now been superseded by the new Essex data app, ReCollect.



Changes the core metadata and workflow of EPrints to make it more focused for as a dataset repository. The workflow is trimmed for simplicity. The review buffer is removed to give users better control of their data.

Patrick McSweeney, 05 Jul 2012 <a href="http://bazaar.eprints.org/216/">http://bazaar.eprints.org/216/</a>

Figure 4. Icon and description for the Data Core research data app available in the EPrints Bazaar

If workflow for different data types is provided directly by EPrints, and can be customised to local repository needs, what is the need for an app that implements data deposit workflow? First, it can simplify and speed up customisation if the desired workflow can been implemented from an app. Second, and more importantly for research data workflow, it can lead to greater standards compliance, consistency and collaboration between repositories.

# 4.3 ReCollect: a standards-based app

Standards are the basis for a new research data app, ReCollect, with a metadata profile designed by Research Data @Essex, and packaged as an EPrints app in collaboration with Patrick McSweeney at the University of Southampton. Also collaborating were potential repositories interested in using the app, including the universities of Glasgow and Leeds, as well as Essex and Southampton.

Due to interest in this project and at Southampton in developing SharePoint for RDM, the JISC Research Data @Essex project at the University of Essex took the lead in extending EPrints for RDM, beginning with an EPrints for data webinar in March 2012. Early thoughts on adapting standard metadata schema for research data were first mapped in some detail, drawing on elements from DataCite, INSPIRE, DDI and DataShare schema, a couple of months later, culminating in the release of a metadata profile, with the metadata elements "further fleshed out with field-types, help labels and other information" on a beta instance of an EPrints repository at Essex. These developments were each reported on the Research Data @Essex blog:

EPrints for data: webinar report, Apr 5, 2012 <a href="http://web.archive.org/web/20130331064037/http://researchdataessex.posterous.com/eprints-for-data-webinar-report">http://web.archive.org/web/20130331064037/http://researchdataessex.posterous.com/eprints-for-data-webinar-report</a>

Adapting EPrints for research data: metadata and design, May 28, 2012 <a href="http://web.archive.org/web/20130331063936/http://researchdataessex.posterous.com/metadata">http://web.archive.org/web/20130331063936/http://researchdataessex.posterous.com/metadata</a>

Repository beta & metadata profile released, Oct 19, 2012 <a href="http://web.archive.org/web/20130331063859/http://researchdataessex.posterous.com/repository-beta-metadata-profile-released">http://web.archive.org/web/20130331063859/http://researchdataessex.posterous.com/repository-beta-metadata-profile-released</a>

Work began on packaging the app at Southampton soon after the release by Essex of its profile and demonstrator repository, the resulting app first appearing in the EPrints Bazaar on 27 March 2013 (Figure 5).



The ReCollect plugin transforms an EPrints install into a research data repository with expanded metadata profile for describing research data (based on DataCite, INSPIRE and DDI standards) and a redesigned data catalogue for presenting complex collections. Developed by the UK Data Archive and the University of Essex, as part of the JISC MRD Research Data @Essex project

Alexis Wolton <a href="http://bazaar.eprints.org/280/">http://bazaar.eprints.org/280/</a>

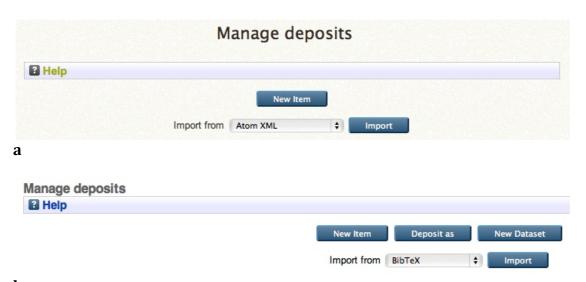
Figure 5. Icon and description for the ReCollect research data app available in the EPrints Bazaar

A metadata profile is the specification of all the data needed to describe a digital object, in this case a research dataset, according to that profile, and which can be represented by defined fields in the repository deposit workflow. The profile includes the specification of mandatory and optional fields, as we saw in the

SharePoint RDM example above, although typically with more mandatory fields than was the case there.

The implementation of the ReCollect app is in strict conformance with the metadata profile specified by Essex, that is, every installation defaults to the same profile. Repository administrators, however, are encouraged to customise the resulting workflow, the ordering and paging of fields from the profile. An example of customising the workflow for the ePrints Soton institutional repository is compared with the ReCollect original in the Appendix.

The first change to the process of standard EPrints deposit (Figure 6a) due to ReCollect is immediately apparent with the appearance of a 'New Dataset' button to initiate a dataset deposit (Figure 6b). Clicking on New Item takes the depositor to a screen offering a range of data types. The New Dataset button circumvents this for repositories expecting high numbers of datasets, instead taking the user straight to the dataset deposit screen.



**b**Figure 6. Start of EPrints deposit process: **a**, standard install; **b**, with Recollect app a 'New Dataset' button appears

Moving into the workflow, repository deposit typically has two stages: a file-level deposit, and creation of a record. This recognises that a record may describe a digital object that consists of multiple files, for example, a dataset may have many data files. In this way each file can be described separately, while still being part of the same record. Figure 7 shows the file-level deposit form provided by ReCollect.



Figure 7. File level deposit with ReCollect data app (25 March 2013)

On completion of the file deposit the Next screen provides the form to complete the record, presented, in effect, in a single page (Details – see workflow represented in Figure 8). A detailed list of the fields provided by ReCollect on this page - a screenshot would be too large to capture and include here - can be found in the Appendix. Completing the workflow is a separate classification page (Subjects), followed by confirmation by the depositor of the repository licence and deposit (Deposit).



Figure 8. Simple graphical representation of workflow from ReCollect data app

#### 4.4 ePrints Soton implementation of ReCollect data app

The principle of the ReCollect app is it defines a metadata profile for a dataset to be deposited in a repository. How that profile is implemented in practice, that is, the order of the workflow and the inclusion of specific fields, can be adapted by the individual repository. In other words, the app is a starting point for an implementation.

Here we review the implementation for ePrints Soton, the institutional repository of the University of Southampton, based on the test version of the repository, on 26 March 2013, prior to release on the live repository server.

For comparison with the unamended ReCollect app, Figure 9 shows the file-level deposit form provided by ePrints Soton. This has a mandatory Title field and requests a reason if an embargo period, restricting visibility of the data, is specified, and the License field is optional; elsewhere this file deposit form follows the ReCollect original.

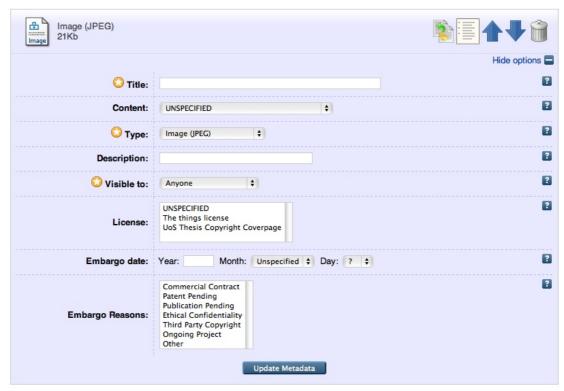


Figure 9. File level deposit with ePrints Soton implementation of ReCollect (viewed 26 March 2013)

Following file-level metadata, we look at the workflow for creating a record. As we have described, the workflow for creating a record originally implemented by ReCollect is essentially presented in a single page (Details, Figure 8). Selected mandatory fields occur throughout the workflow on this page, including the first and last fields. ePrints Soton essentially turns this into a two page workflow (Dataset information and Additional details, shown in Figure 10), with all the mandatory fields collected in the first of these pages. This split between core and optional metadata description supports rapid creation of minimal metadata for, in some cases, draft records. As with Recollect, there are also classification (Shelves) and Deposit confirmation pages to complete. In this way a user can complete the deposit in a single page, with other pages leading to Deposit entirely optional. It is believed this will improve efficiency of deposit.

Again, details of the fields provided in this workflow can be found in the Appendix.



Figure 10. Simple graphical representation of workflow from ePrints Soton implementation of ReCollect

# **5** Conclusion

Policy specified at the University of Southampton requires its researchers to manage research data effectively in terms of cataloguing and storage for visibility and access, and longer term preservation. Work has been ongoing to provide a cataloguing facility at Southampton based on Microsoft SharePoint, which provides a number of IT services at Southampton, and on EPrints repository software, originally developed at Southampton and which can be used to create records and provide access to many types of digital objects, including research data.

University of Southampton, University Calendar 2012/13, Section IV: Research Data Management Policy http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html

This report has described progress with both platforms. In the case of SharePoint, user interface forms for creating data records have been piloted and tested. The approach is distinctive in creating two linked forms, one to describe a project, the other to record a dataset, rather than a single workflow as in the case of EPrints. This development of SharePoint at Southampton remains incomplete and is now seen as part of a longer-term extension and integration of services provided on the platform.

In the first instance research data cataloguing at Southampton uses EPrints, extending the existing institutional repository by installing the Essex ReCollect data app. The service went live on ePrints Soton in April 2013.

A community of potential users, including the EPrints repositories at Glasgow University and Leeds University, has been established through webinar-based conference calls. Formal user testing of the original profile and workflow on a demonstrator repository was performed by the Essex project. Informal user feedback has been collected at Southampton based on a test implementation of a locally adapted version of the ReCollect data app. On this basis further modifications to the version described here are likely, including the facility to automate minting and embedding of British Library DataCite DOIs, designed for data citation, for each data record.

Motivated by the JISC RDM Programme many UK research institutions are implementing research data repositories. A variety of repository platforms, from original solutions such as DataFlow at Oxford, as well as established digital repositories such as DSpace and EPrints, has been adopted. Southampton is unusual in pursuing two repository options.

What is needed now, however, is practice and experience with real data collections. In this respect many questions about the use of data repositories remain open. These early implementations are likely to change significantly as that process evolves.

# **Acknowledgements**

**SharePoint** developer team: Peter Gibbs, Rich Gooding and Chris Yorke, all from from iSolutions, University of Southampton Academic champion for Sharepoint: Simon Cox

**ePrints Soton**: development and implementation of ReCollect-based workflow, Patrick McSweeney and Matt Smith

**ReCollect EPrints app**: profile developed by Alexis Wolton, Tom Ensom and Louise Corti at Research Data @Essex

# Appendix. Table of fields for two implementations of ReCollect data app: ReCollect original, and ePrints Soton with ReCollect

This table is presented to show the full data deposit workflow provided by the ReCollect EPrints app, and to demonstrate how this workflow can be customised with reference to the implementation by ePrints Soton repository.

It should be noted that for full compliance with the DataCite standard the dates and times a file is deposited and made public are recorded automatically within ePrints Soton and are not shown in the fields below.

Fields, ReCollect original (25 March)	Fields, ePrints Soton ReCollect (viewed on test server, 26 March 2013)
Details page  *Title (toyt field)	Dataset Information page
*Title (text field)  *Collection description (text field)  *Keywords (text field)  *Divisions (multi-selection list)  +Alternative title (text field)  *Creators (family name, given name, email – all text fields)  +Corporate creators (text field series)  *Type of data  Contributors (contribution - single-selection drop-down list; family name, given name, email – all text fields)  *Research Funders (text field)  Grant number (text field)  +Research project title (text field)  *Time period  *Collection period (From	*Title (text field)  *Type of Data (text field)  Description/Abstract (text field)  Resource Language (text field)  *People and Groups  Contributors (contribution - single-selection drop-down list; family name, given name -text fields, Cited? - single-selection drop-down list)  *Divisions (multi-selection list)  Research Funders Grant Number (text field) Research Funder (text field)  Additional details page
Y/M/D, To Y/M/D) *Temporal coverage (From	Additional dataset information

Y/M/D, To Y/M/D)	Alternative title (text field)
+Geographic coverage (text field)	Keywords (text field)
+Geographic location (text field)	Research project title (text
Bounding Area (North Latitude,	field)
East Longitude, South Latitude,	Related resources (URL – text
West Longitude – all text fields)	field, Type - drop-down list)
+Data collection method (text field)	Copyright Holders (text field)
+Legal and ethical issues (text field)	Original Publication URL (text
+Data processing and preparation	field)
activities (text field)	Dataset collection details
*Resource language (text field)	Collection period (From Y/M/D,
+Additional information (text field)	To Y/M/D)
Related resources (URL – text field,	Temporal coverage (From
Type – drop-down list)	Y/M/D, To Y/M/D)
*Original Publication Details	Geographic coverage (text field)
Publisher (text field)	Bounding Area (North Latitude,
*Status (radio button selection)	East Longitude, South Latitude,
Original publication URL (text	West Longitude – all text fields)
field)	Data Collection Method (text
Date (YMD)	field)
Date type (radio button	Legal and Ethical Issues (text
selection)	field)
*Copyright holders (text field)	Data Processing/Preparation
*Contact email address (text field)	Activities (text field)
+Administrative note (text field)	Additional Information (text field)