UK Data Service ReShare: CKAN and EPrints Gap Analysis

UK Data Archive, University of Essex

Title:	**UK Data Service ReShare: CKAN and EPrints Gap Analysis**
File Name:	ReShare_EPrintsCKAN_GapAnalysis_Public_01_00
Description:	Comparison of the EPrints and CKAN repository software packages, undertaken during the selection of a platform for the UK Data Service's ReShare repository. The gap analysis is based on requirements gathered during the CKAN for RDM workshop, run by the Jisc Managing Research Data Programme in February 2013 (see: http://www.dcc.ac.uk/blog/ckan-research-data-management).
Created By:	Tom Ensom, Alexis Wolton and John Shepherdson
Published:	2014-09-28
Last Modified:	2013-07-02

Requirement	User group	Need level	Gap analysis (green = possible, amber = probably possible, red = not possible	
			EPrints	CKAN
add/edit help pages	Administrator	Essential	pages added server side - user adds text as phrases	Pages added server side - ckan manages editing of page text in ui with sysadmin ('phrases' in eprints)
can capture metrics, e.g. no. of users downloading DIPS by project record, date, country, IP address. And reports?	Administrator	Essential	irstats plugin. irstats2 currently in beta - uses google charts, is extendable and exports to multiple formats	Metrics are stored (downloads , views per org and sitewide), and enabled via the stats extension.
can delete data - both individual files and entire project records	Administrator	Essential		Ability to delete datasets exists, but no record of deletion (i.e. audit trail)
can publish DIPs (project records) after review	Administrator	Essential		
can publish project records not yet submitted for review by a depositor	Administrator	Essential		
can review a submitted SIP before publishing, to check for disclosure risks in data files and verify the quality of documentation accompanying the data collection	Administrator	Essential		
can create snapshots of a project and its data files (versioning)	Administrator	Desirable	Eprints handles versioning of metadata very well. Get a full XML audit trail	
can upgrade ReShare data collections (AIPs) to UKDS data collections (AIPs), for long-term preservation, i.e. export metadata record and bundle of data/documentation files easily for UKDS Ingest.	Administrator	Essential	possible - use zip export plugin from edshare? Metadata will be mapped inline with UKDS profile	
receives automated messages at time of: - SIP submission for review - publication of DIP by UKDS - expiry of embargo period - request to remove project record	Administrator	Essential	Triggers would need to be added	
to have a 'dark archive' option for either data, metadata OR both	Administrator	Desirable	possible - push to dark archive first, then filter what should be publicly viewable for public archive	Already can manually control privacy and access settings. We can't currently have a public dataset (metadata) with *private* data - but would be possible to create this. On the CKAN roadmap to add publication/embargo dates. Opportunity to provide better, more fine-grained controls for both data & metadata - than currently available in publication repo software. NOt currently possible to have differen auth rules applied to metadata vs. data.
to have system by which adminstrators can review and approve submitted deposits	Administrator	Desirable		Doesn't currently exist, but will do by the end of March - editorial/approval step on ingest.

to link to data that researchers have in other places/repositories	Depositor	Desirable		Possible to harvest from external repositories by API (for regular updates) or whole-repo harvests. Need to map fields to CKAN logic. Would be possible to create and share plugins for common external data archives (e.g. subject-specific repositories of data). Possible requirement for standardisation interoperability between data repositories (CERIF4Data??)
to perform actions on behalf of depositors (i.e. editorial power)	Administrator	Essential		
to export all data (DIPs) and metadata records	Administrator / Developer	Essential	Many different export plugins available out of the box - easy to write extras	Yes! Can get as JSON, RDF. STFC have written an CERIF-CKAN export and import
Automated extraction of metadata from/about files e.g. mimetype	Depositor	Desirable	A core set of basic metadata fields are extracted by eprints	Currently does not recognise file format/extension. Requirement.
Batch operations for editing metadata and other features.	Depositor	Desirable	Core functionality to batch update metadata fields. Can also batch delete	Not currently avaialble.
can 'submit' a project for review after project record is final and all data/documentation files have been uploaded; this sends the SIP for review/approval by UKDS	Depositor	Essential		
can distinguish data files and documentation files, via a 'type' tag within the file-level metadata	Depositor	Essential	ReCollect functionality	
can easily assign Unique Identifier (not DOIs)	Depositor	Essential	Yes	Yes. Currently doesn't assign DOIs. Lincoln have done this and released the code
can edit the project record; except pre-populated metadata fields	Depositor	Essential		
can invite collaborators who are registered UKDS users, to contribute to a project; this invitation is sent by email, based on email address; a collaborator can upload data and documentation files and edit the editable fields of the project record; and submit a SIP for review	Depositor	Desirable		
can organize data collections in user-defined groups	Depositor	Essential	Collections plugin	Core
can point to a DOI / URL if data are already held in other repository	Depositor	Essential		
can pre-populate the project-level metadata record by harvesting metadata from ESRC Research Catalogue / ROS (or other source) via a research award number (except for CV metadata fields needed for Discover); the depositor does not need to be the PI or award holder	Depositor	Essential	Possible to harvest using OAI-PMH API. Can import transformed XML?	Possible to harvest using OAI-PMH API

Can register for a UKDS account	Depositor	Essential		
can remove files whilst deposit is not yet submitted	Depositor	Essential		
can request UKDS to delete a project record, but not able to delete a project record him/herself	Depositor	Essential		
can return at a later stage to a saved project record to make changes to editable fields (those not pre-populated by harvest)	Depositor	Essential		
can see (or request) metrics about DIP view/download/citation	Depositor	Essential		
can see demographics about who has viewed/downloaded dataset	Depositor	Desirable	irstats graphs can be displayed on individual records, or as widgets on user profiles	Not available; can be run from server logs or from third party analytic services, but these can be plugged in and there has to be a method for displaying the information.
can select/choose a licence agreement for each file (or entire SIP ?) via tick box	Depositor	Essential	ReCollect adds per file	Could be added to file metadata?
can set access level of each file (or entire SIP ?) as: Open access; registered user; embargo	Depositor	Essential		
can share stable links to data collection	Depositor	Essential		Also to individual files (relies on stable naming)
can upload data and documentation files, and tag them via a 'type' metadata field as data file or documentation file	Depositor	Essential	ReCollect functionality	
can use status tracker to see which part of the deposit process has been completed or not, with flexible 'resume workflow' points	Depositor	Essential	Core functionality	
cannot upload .exe files	Depositor	Essential	This is difficult to achieve when files are bundled (e.g. zipped), and is currently not supported	Same logic as eprints
has link from project record to the published output	Depositor	Essential		
Multiple file and/or batch upload	Depositor	Essential	Indirectly possible through uploading and unarchiving zips	Not currently available through browser.
Receive automated system notifications and alerts	Depositor	Essential		Audit trail exists, report/display does not exist
wants data to be discoverable - page rank, etc.	Depositor	Essential		State of CKAN SEO?

can set embargo to schedule the accessibility of data files in the DIP; this is set at the project level and makes the project record visible in the catalogue and the documentation files available for download, but the data files unavailable for download; embargoes expire automatically	Depositor / Developer	Essential		
Include in project-level metadata record a link to research outputs held in e.g. ESRC research catalogue or ROS	Developer	Essential		
virus checking of all uploaded files	Developer	Essential	?	?
any metadata field that relates to a Discover facet should be entered from the CV list, rather than by harvesting (anlyUnit, dataKind, geogUnit, topcClas, nation, datatype)	Developer	Essential		
data and documentation files have minimal metadata	Developer	Essential	ReCollect functionality	
Depositor ID metadata can be harvested from user account (name, institution, email,….)	Developer	Desirable		
software system has longterm community support	Developer	Essential		
Metadata are pushed to Discover after the DIP has been published	Developer	Essential		
publish metadata as OAI streams	Developer	Essential	OAI-PMH metadata harvesting	
system to be hardware and platform agnostic	Developer	Essential	Preferred is Debian, Ubuntu, RedHat or Fedora Linux OS. Eprints can run on Windows, but is not advised in a production environment	Ubuntu and Red Hat Linux, Postgres, Python. Not Officially supported on Windows, but can be run as a VM?
to either handle large files or have a limit of file size upload	Developer	Essential	Apache sets the max size for file uploads. File upload size limited when using UI by browser capability. Eprints has a bitorrent plugin that attempts to circumvent this problem when using Sword2.	Apache sets the max size for file uploads. File upload size limited when using UI by browser capability.
to have proper documentation	Developer	Essential	API that is exposed is documented	Started but unfinished, needs a mass effort to bring up to standard.
to have versioning for data	Developer	Desirable		Implementation - requires new storage layer to handle deltas.
to plugin to external authorisation API (such as Active Directory)	Developer	Essential	LDAP and shibboleth authorisation straightforward	Shibboleth plugin available, LDAP plugin needs implementation
to serve open data, registered user access data, embargoed data and secure data	Developer	Essential		Requires development and possible certification of integrity. Requires new security options in terms of encryption, stricter access controls including audit eg for medical data.

to use Oracle/MySQL/our supported RDBMS on the back end	Developer	Essential	MYSQL	Postgres
to work with other standards (OAI-PMH, SWORD, CERIF, INSPIRE)	Developer	Desirable		Not implemented (bar RDF?), but feasible
XML-based metadata are pushed to the UKDS administrative database (currently Mirage) for administrative purposes, after a project records has been created	Developer	Essential		
Is skinnable, so can be branded in-house	Developer	Essential		
Has plug-in architecture so can easily be extended using components developed by 3rd-parties or in-house	Developer	Essential		
Has product roadmap and support/update options	Developer	Desirable	EPrints roadmap: https://github.com/eprints/eprints/issues/milestones Free support via techlist - http://www.eprints.org/tech.php/ - paid support via Eprint Services - http://www.eprints.org/services/	Roadmap available - http://trac.ckan.org/roadmap - http://lists.okfn.org/pipermail/ckan-dev/ - ckan services - http://okfn.org/services/
Has active online Developer community, with frequent bug fix releases	Developer	Essential	Eprints services at Southampton develop core releases. 3rd party plugins made available via bazzare	mailing list - http://lists.okfn.org/mailman/listinfo/ckan-dev
Has web services API	Developer	Desirable	CRUD/REST API offers access to Eprints objects. Sword 2.0 supported	REST API provided with wrappers in a number of languages.
to provide external access to our metadata	External Data Portal	Essential		API key – allows access every metadata field of the dataset and ability to change the data if you have the relevant permissions via API. - See more at: http://ckan.org/features/metadata/#sthash.3ty4XKLR.dpuf
to check that data has been deposited	Funder	Desirable		
can agree to Use terms & conditions (tick box) before download of a DIP	User	Essential		
can receive standard citation for each DIP in DataCite format, incl. unique and stable ID,either via an ‘export citation’ button or as note file within the DIP.	User	Essential	Will need to write an export plugin	Add 'export citation' button that combines data from the relevant metadata fields. (Maybe with options for different citation formats.)
can search/browse within data and documentation files within ReShare system	User	Essential		Does in file searching?
can send queries to depositor via ‘data contact’ ; email address is needed for each project in the project record	User	Essential		
clear presentation of projects (e.g. file browser)	User	Essential		

Want to query/process actual data and doc contents on the server and get results	User	Essential	Xapian search indexes all text readable files inc pdf	CKAN's Filestore indexes files. In contrast to the the FileStore which provides 'blob' storage of whole files with no way to access or query parts of that file, the DataStore is like a database in which individual data elements are accessible and queryable. (http://docs.ckan.org/en/latest/datastore.html)
Want to see the existence of, and request access to restricted data	User	Essential	Done via stub record and email request	
Want to see where this dataset has been re-used, which papers have cited it, etc.	User	Desirable	Related Items' allows this, but not automatically.	Related Items' allows this, but not automatically. How about implementing a pingback for DOIs to CKAN?
Has active online User community and suppor/discussion fora	User	Desirable	Eprints techlist - http://www.eprints.org/tech.php/	Mailing Lists - http://lists.okfn.org/mailman/listinfo/ckan-discuss
Tech Stack			Linux Apache MySQL Perl Xpage XML	Linux Apache Postgres Python Node.js Jinja2 LESS CKAN 2.0 Deb package requires Ubuntu 12.04 64Bit
Active instances			ROAR repository registry lists 505 instances	Estimated at 50 official data hubs, which does include national services such as data.gov.uk