PREPARDE
D4.3 Roadmap for implementation of data publication at California Digital Library.

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences		
Project Hashtag	#preparde		
Start Date	1 July 2012	**End Date**	30 June 2013
Lead Institution	University of Leicester		
Project Director	Dr Jonathan Tedds		
Project Manager	Dr Sarah Callaghan		
Contact email	sarah.callaghan@stfc.ac.uk		
Partner Institutions	University of Leicester British Atmospheric Data Centre (BADC) US National Centre for Atmospheric Research (NCAR) California Digital Library (CDL) Digital Curation Centre (DCC) University of Reading Wiley-Blackwell Faculty of 1000 Ltd		
Project Webpage URL	http://proj.badc.rl.ac.uk/preparde/wiki		
Programme Name	*Managing Research Data*		
Programme Manager	Simon Hodson		

Document Information	
Author(s)	John Kunze
Project Role(s)	Project Participant

Date		**Filename**	
URL	*If this report is on your project web site*		
Access	□ Project and JISC internal		☒ General dissemination

Document History		
Version	**Date**	**Comments**
1	13 June 2013	First draft

CDL Data Publication Plan
June 2013-June 2014

In the next 12 months, the California Digital Library (CDL) plans to pursue four projects related to data publication with the goal of laying a foundation to support data publication for the University of California (UC). Our participation in the PREPARDE project has provided a wealth of information and experience to draw upon as we pursue this important new form of scholarly communication.

The broad aims are to increase research data sharing, archiving, and re-use by creating technical mechanisms and organizational support for moving researchers' data from their personal computers to robust, web-accessible servers. On top of this platform we would build services for obtaining persistent identifiers and generating metadata to support data citation, which we in turn expect to help establish incentives for researchers to take the trouble to publish their data.

The CDL already provides a number of services in this area, and we expect some co-evolution among them and CDL's data publication effort. The Merritt repository service offers domain- and format-agnostic curation storage and dissemination for versioned, hierarchical datasets. To support citation and reference, the EZID service helps you make and manage persistent identifiers (DOIs and ARKs), whether you deposit your data in Merritt or in another repository. CDL also is a member of the NSF DataONE federation of data repositories, which maintains APIs, data tools, and a central discovery index of data across all its repositories.

At a more granular level, meeting scientists where they work, the DataUp extension to MS Excel makes it easy to take spreadsheet data, add quality metadata, and publish it via the DataONE network; to support the latter, CDL built a special DataONE repository called ONEShare that anyone can publish to without requiring login. Wrapping it all up is the DMPTool, the Data Management Planning Tool, which CDL provides for capturing a description of the long view of your data publishing intentions that you would submit with a funding proposal and would be responsible for carrying out.

The four projects are:

1. Specify and implement a low-cost, low-effort "data paper" that is essentially an extended citation stored in the EZID citation service. This sort of paper would be generated automatically from an EZID record. Each record would identify a publishable dataset, complete with author, title, date, abstract, and links to stored data. These could be produced as a simple "view" of a record and made available to global search systems (eg, Google).

2. Develop a new data deposit interface for the Merritt repository system with an eye towards integrating deposit with CDL's eScholarship online publication platform. The focus of this effort would be to streamline publication of datasets supplemental to an online open-access publication (as opposed to stand-alone datasets).

3. Work with UC Berkeley's new data science initiative to investigate supporting publication of medium-large datasets (several hundred GBs). This will entail developing additional functionality, such as parallelized transfers and data reduction techniques that permit partial or sampled views of a dataset.

4. Begin work with the UC Berkeley Moorea Biocode project to pilot a data publication based on their extensive hierarchical data collections. A requirement to link between collecting events, specimens, and tissue samples is expected to generate large numbers identifiers.

The data projects will be coordinated by our new post-doctoral fellow in data curation, John Kratz. We will work closely with John to insure that the implementations will meet the emerging needs of the UC community and mandates for open access.