Open Data and the Academy: An Evaluation of CKAN for Research Data Management

Joss Winn, University of Lincoln, UK.[1]
jwinn@lincoln.ac.uk

1" http://staff.lincoln.ac.uk/jwinn

Table of Contents

- Introduction: RDM and openness in academia
- Orbital project
- A Brief History of CKAN
- Requirements for RDM
  - Curators and Librarians
  - Developers and IT support
  - Researchers and data re-users
  - Institutional requirements gathering
- Evaluation of CKAN
  - Development model
  - Features
    - First impressions (version 1.7)
    - Version 2.0
    - Security, permissions and role management
    - Activity data
    - Data previews and visualisation
    - CKAN APIs
    - Linked Data and RDF
    - Metadata
    - Datastore and Filestore
    - Usability
  - Gap Analysis
- Conclusions and recommendations

Introduction: RDM and openness in academia

> *"The management of research data is recognised as one of the most pressing challenges facing the higher education and research sectors. Research data generated by publicly-funded research is seen as a public good and should be available for verification and re-use. In recognition of this principle, all UK Research Councils require their grant holders to manage and retain their research data for re-use, unless there are specific and valid reasons not to do so. Research data can also be the subject of requests under Freedom of Information legislation or Environmental Information Regulations."*[2]

In the UK and elsewhere, research funders, researchers, information professionals and government ministers increasingly acknowledge that it is in the public interest to ensure that data which forms the basis of research findings should be made available for re-use and verification. 'Research Data Management' (RDM) refers to the development of new policy, research practices, technical infrastructure and professional support so as this relatively recent imperative within the research community is met.

Research Data Management can be seen as a continuation of the logic of the Open Access movement (OA), which as a distinct scholarly practice, pre-dates RDM by over a decade.[3] If openness within higher education is understood as a 'recursive public',[4] we can argue that the logic and limits of OA leads to a recursion in which open access to research outputs prompts questions around how those outputs can be verified and how participants in that 'public' can re-use and build upon existing research. i.e. Open Data.

Through the use of 'recursion' as an analytical tool, we can observe and predict that further recursions will be deemed necessary to support the logic of OA, e.g. 'Open Science', where the research process is conducted publicly;[5] Open Source, where the research tools and software algorithms are transparent and accessible;[6] and Open Peer Review, where the verification of research findings are themselves open to scrutiny for bias and inconsistency.[7] Although the concept of recursion suggests a series of steps or iterations, each recursive element can occur concurrently, deferring its limits to the next process while continuing to unfold. As Kelty has noted, "the 'depth' of the recursion is determined by the openness necessary for the project itself." (2008:30). The depth of recursion required by the logic of OA is still being worked out and

---

2" JISC (2011) Managing Research Data Programme 2011-2013. http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx

3" https://en.wikipedia.org/wiki/Open_access

4 The concept of a 'recursive public' has been developed in Kelty, Christopher M. (2008) Two Bits. Duke University Press. http://twobits.net

5" https://en.wikipedia.org/wiki/Open_science

6" https://en.wikipedia.org/wiki/Open_source

7" https://en.wikipedia.org/wiki/Open_peer_review

remains a contested public through which the nature and purpose of science is being questioned.[8]

This paper begins by briefly discussing a recent project to develop RDM policy, infrastructure and support at the University of Lincoln, UK. The author was the Principal Investigator and Project Manager of the 'Orbital project'. In the middle of the project, following several months of research and development, we adopted the open source CKAN software[9] as the main technical component for ingesting, describing and disseminating research data. As far as we are aware, this was the first time that CKAN had been adopted for RDM by a higher education institution and it has led to further interest in the potential of CKAN to meet the needs of the research community.

This paper offers a critical evaluation of CKAN for Research Data Management and is intended to be a constructive, useful discussion to inform the research community of CKAN's strengths and weakness for RDM and outline the further development of CKAN to meet the requirements of this community. In effect, this paper is a contribution to the recursive public of open access to science and the production of social knowledge.

Orbital project

The Orbital project was a pilot project funded by Jisc over 18 months to develop institutional policy, training and technical infrastructure for research data management (RDM) at the University of Lincoln.[10] The project team included staff from the Centre for Educational Research and Development (CERD), the Library, the Research and Enterprise Office, and worked with researchers in the School of Engineering and the School of Computing to understand their needs for RDM. The Orbital project built on the experience and technical outputs generated by previous Jisc-funded projects at Lincoln and employed two full-time web developers to examine the requirements, constraints and opportunities for the development of tools to enable and support RDM.

In retrospect, we can understand the progress of the project in six stages:

1. the first six months were spent learning the domain, when project staff developed their own expertise and reflected intensely on the requirements of researchers.
2. This led to the second stage, where we developed a 'Minimum Viable Product' based on our proposed Implementation Plan.
3. The third stage was marked by a significant shift in approach when we decided to adopt the CKAN software in place of our own product, leading to the development of the fourth stage of the project

---

8 For a "view from the trenches", see Suber, Peter (2012) Open Access. The MIT Press. https://mitpress.mit.edu/books/open-access

9 http://ckan.org

10 http://orbital.blogs.lincoln.ac.uk

4. during which we focused on systems integration and workflow for RDM.

5. This resulted in the final technical output of the project, the ‘Researcher Dashboard’ and a concurrent proposal for sustaining the work of the project through the formation of a new ‘Research Information Service’ at Lincoln.

6. At the same time as this intensive technical work, institutional policy and RDM documentation and training were being developed.

   The final outcomes of the Orbital project can be summarised as: Establishing expertise among academic support staff; the development of a technical design and product which meets some requirements for RDM and can be further developed to meet future needs; a body of documentation and a programme of training for post-graduate students and researchers; the development of institutional policy for RDM; and the requisite Business Case for a Research Information Service, which underpins the Policy and develops a sustainable roadmap for RDM at Lincoln. This has led to the creation of a new post of Research Services Developer to continue the technical work undertaken during the course of the project.

A Brief History of CKAN

> “CKAN is the world’s leading open-source data portal platform. CKAN makes it easy to publish, share and work with data. It's a data management system that provides a powerful platform for cataloging, storing and accessing datasets with a rich front-end, full API (for both data and catalog), visualization tools and more.”[11]

The development of the Comprehensive Knowledge Archive Network (CKAN) software began in March 2006[12] with the first public release in July 2007.[13] Today, there are over 50 documented CKAN ‘data hubs’ in place around the world.[14] CKAN is typically implemented by an organisation rather than by individuals and is increasingly popular among national and local governments worldwide. At the time of writing, there were eight official national government installations of CKAN as well as sixteen official regional instances. It is also used by the EU as a federated catalogue of open data from across the EU.[15] Since its use in 2010 by the UK government to publish data on http://data.gov.uk, interest and adoption of CKAN has increased significantly and this has led to a corresponding increase in development activity, too.[16]

---

11 Quote from the CKAN ‘README’ file https://github.com/okfn/ckan (22/05/2013)

12" http://web.archive.org/web/20061011175630/http://www.okfn.org/ckan

13" http://lists.okfn.org/pipermail/okfn-discuss/2007-July/005687.html http://blog.okfn.org/2007/07/04/the-comprehensive-knowledge-archive-network-ckan-launched-today/

14" http://ckan.org/instances/ http://ckan.org/case-studies/

15" http://publicdata.eu/

16" A summary of Google Trends data for CKAN can be found here: https://orbital.blogs.lincoln.ac.uk/2013/02/11/ckan-trending/ A graph of contributions is here: https://www.ohloh.net/p/ckan/commits/summary and a graph of lines of code here: https://www.ohloh.net/p/ckan/analyses/latest/languages_summary

The development of CKAN is led by the Open Knowledge Foundation (OKF)[17], a not-for-profit organisation created in 2004 "to promote the openness of all forms of knowledge."[18] It's original mission statement stated the objectives:

1. To promote freedom of access, creation and dissemination of knowledge.
2. We develop, support and promote projects, communities and tools that foster and facilitate the creation, access to and dissemination of knowledge.
3. We campaign against restrictions both legal and non-legal on the creation, access to and dissemination of knowledge.
4. We seek to act as an intermediary between funding institutions and projects that work in areas related to the creation and diffusion of knowledge, particularly those with a strong technological aspect.

CKAN was specifically created to address the second objective of OKF and was conceived early on as both a *tool* and hosted *service* to "provide a comprehensive database of open knowledge projects, tools and data sets."[19]

Around the time of its first public release, the hosted service at http://ckan.net contained over 40 'packages' (later called 'datasets') and was described as "the place to search for open knowledge resources as well as register your own. Those familiar with freshmeat or CPAN can think of CKAN as providing an analogous service for open knowledge. "[20] Within a year there were over 200 packages on ckan.net[21] and at the time of writing (May 2013) the number of datasets registered stands at 6400. To clarify its purpose and reflect its utility, the hosted service was rebranded as the Data Hub in August 2011.[22]

Development of CKAN has increased significantly during the last six years, having received over 10,700 commits to the source code repository from 73 recorded contributors.[23] More detail about the model and rate of development is given in a later section of this paper. Version 2.0 of CKAN was released in May 2013,[24] three years after the release of version 1.0. The changelog shows that releases have been made steadily since 2006 at an average of a new release every 2.5 months.[25] The latest version of CKAN can be tried by anyone at http://demo.ckan.org/

The utility of CKAN has evolved over this time from a simple registry and catalogue of datasets stored elsewhere to offer storage facilities for the upload of data 'blobs' or files[26] (since version 1.6) and most recently a 'datastore' (since version 1.7) for the ingest and visualisation of

---

17" http://okfn.org
18" http://web.archive.org/web/20040804174222/http://www.okfn.org/
19" http://web.archive.org/web/20061011175630/http://www.okfn.org/ckan
20" http://web.archive.org/web/20070626214036/http://www.ckan.net/

21" http://web.archive.org/web/20080828115341/http://www.ckan.net/
22" Initially http://datahub.org and from May 2012 at http://datahub.io
23" https://www.ohloh.net/p/ckan
24" http://blog.okfn.org/2013/05/10/announcing-ckan-2-0/
25" http://docs.ckan.org/en/latest/CHANGELOG.html
26" http://ckan.org/2011/05/16/storage-extension-for-ckan/

structured data.[27] Due to the design of its architecture, the file storage and datastore functionality were initially offered as optional extensions before being included as part of CKAN core software. Thus, the development of CKAN can be traced from a simple *catalogue* tool for the discovery of open data, to more recently become a *repository* and a *datastore* for the ingest and analysis of 'live' or 'active' data. Today, the design of CKAN remains extremely versatile: it can be either a catalogue, or a repository, or a datastore or a combination of all three. It's features are discussed in more detail in a later section of this paper.

Requirements for RDM

As discussed above, during the Orbital project, we investigated the potential use of CKAN for RDM. We had been aware of CKAN since the start of the project and were watching its development quite closely. A summary of our initial evaluation has been written up elsewhere[28] and it was the inclusion of the Datastore in CKAN (i.e. catalogue+repository+datastore) that made us rethink the direction of our project. Based on our initial user requirements gathering, we had defined the 'Minimum Viable Product' for RDM as follows:[29]

- authentication
- data storage
- data publishing
- licensing
- persistent URIs
- analytics

These features alone allow an academic to reliably and permanently publish data to support their research findings and help measure its impact. CKAN meets these requirements. However, the requirements for a mature, sustainable research data management system extend far beyond these minimum requirements. We initially identified a number of further requirements, which CKAN is able to meet to some degree:

- Integration with the institutional research environment (e.g. hooks into Current Research Information System (CRIS), Institutional Repository, DMPOnline, networked storage)
- Capturing the research process/context/activity; notation, not just data

---

27" http://ckan.org/2012/03/27/ckan-datastore-and-data-api/

28" http://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/ Further justification for the adoption of CKAN can be found at http://orbital.blogs.lincoln.ac.uk/2012/08/17/hello-ckan/

29" https://orbital.blogs.lincoln.ac.uk/2012/05/23/a-minimum-viable-product-for-research-data-management/

- Controlled access to non-Lincoln staff e.g. research partners
- Good, comprehensive search tools
- Version control for data and metadata
- Customisable, extensible metadata
- Adherence to data standards e.g. RDF
- Multi-level access policies
- Secure, backed up, scalable file storage for anywhere access to files and file sharing
- Command-line tools and good web UI for deposit/update of data
- Permanent URIs for citation e.g. DOIs
- Import/export of common data formats
- Linking datasets (by project, type, research output, person, etc.)
- Rights/license management
- Commercial support/widely used, popular platform ('community')

Following a day-long meeting with three representatives from the Open Knowledge Foundation, the Orbital project was convinced that CKAN offered sufficiently compelling features to match many of our further requirements.

In February 2013, a workshop was held in London to discuss the use of CKAN for RDM in greater detail.[30] It was fully booked with over 40 delegates attending from various universities and other research organisations around the UK.[31] The workshop was based around a requirements gathering exercise which is summarised below.[32] In addition to this, the results of requirements gathering exercises held by other UK universities have also been synthesised below.[33]

30" http://orbital.blogs.lincoln.ac.uk/2013/02/27/ckan-for-rdm-workshop/

31" https://ckan4rdm.eventbrite.co.uk/

32 The full set of 'user stories' can be found here: https://docs.google.com/spreadsheet/ccc?key=0Arh4BnSV2XSIdGxibkhLXzhsUVZzVTl6MWdYOWVMcWc&usp=sharing

33" I am grateful to the following projects and institutions for sending me their RDM requirements documents: KAPTUR http://vads.ac.uk/kaptur/; ADMIRe http://admire.jiscinvolve.org/wp/; The University of St. Andrews http://research-computing.wp.st-andrews.ac.uk/category/rdm/ckan/; Research360 http://blogs.bath.ac.uk/research360/.

The process of gathering requirements at the 'CKAN4RDM' workshop was based around the collection of 'user stories'.[34] Delegates agreed to act as a 'proxy' for a set of RDM users on which we all agreed:

- Curator/Manager
- Researcher/user
- Researcher/re-user
- Developer
- IT support specialist

Delegates were divided fairly evenly between each type of user and joined groups where their own professional role matched the proxy user role most closely. As each proxy user group decided on a requirement, they were asked to write their 'story' in the following format:

> Who? What? Why?
>
> As a X, I want to Y, because Z.
>
> As a researcher, I want to upload my data, so that it can be cited by others.

Sufficient time was given to collect over 70 stories in this way from across the different user groups. Over lunch, the stories were written up into a spreadsheet[35] and then delegates returned to their proxy roles to to work through the spreadsheet and undertake a 'gap analysis', which is discussed later in this paper. For the gap analysis exercise, an experienced CKAN user joined each group to help determine whether CKAN already met each requirement.

The requirements gathered at the workshop can be summarised as follows:

Curators and Librarians

Curators are concerned with the provenance and good record keeping of data that is deposited. They need tools which are easy to use and provide a workflow of actions around the data that are recorded and can be audited and reported.

They want to be able to ensure that datasets and researchers have persistent IDs (e.g. Datacite[36], ORCID[37]) and that links between the two can be made.

The deposit of data should be easy and there should be an 'approval' stage, where Curators can moderate the ingest of the data and check the integrity of the supplied metadata. Curators and researchers should be able to embargo the publication of datasets. That is, the data

---

34" https://en.wikipedia.org/wiki/User_story
35" https://docs.google.com/spreadsheet/ccc?key=0Arh4BnSV2XSIdGxibkhLXzhsUVZzVTl6MWdYOWVMcWc&usp=sharing
36" http://www.datacite.org/
37" http://orcid.org/

management system should support the requirement to restrict access to datasets for different types of users and for different time periods. The data management system should support versioning (‘snapshots’) for when updates are made to the dataset.

Curators are concerned with the long-term preservation of datasets and need to ensure that the data and its metadata can be exported and migrated. The system should support clear, open standards for this process.

The data management system should provide metrics on the use of the datasets e.g. number of page views, downloads. To support this, the datasets should be easily discoverable and metadata should be compatible with other institutional systems, such as the library catalogue.

In summary, Curators and Information Managers are concerned with the provenance and maintaining the integrity of their collections over time. They need simple but recognisable workflows to manage formal and ad hoc decision-making and all interactions with the data and metadata should be recorded and auditable.

Developers and IT support

Developers are concerned with flexibility and modularity in the design and extension of the data management system. It should be hardware and platform agnostic and interoperate with protocols and standards such as OAI-PMH, Bibtex, SWORD, CERIF and RDF, and it should have support for the creation of persistent identifers (Datacite DOIs). It should be possible to manage the resource requirements of the system by imposing user storage quotas, have multiple storage silos and run multiple systems from a single instance (‘multitenancy’).

Like Curators, they are concerned about the versioning of datasets, having experience with version-controlled source code repositories such as Git. They also recognise that the data management system should ensure the integrity and security of the data and metadata through checksums and encryption in transport and on disk, as well as facilitate easy publishing of datasets.

As well as a web-based deposit and discovery interface, they would like read/write APIs for the ingest, update and retrieval of data and metadata. They would also like APIs so as to integrate other methods of authorisation. They require a flexible and extensible permissions system and to be able to perform actions on behalf of other users.

Scripts for configuration management tools (e.g. Chef, Puppet), easy upgrades, and good documentation are important to these users as is the vitality of the developer community.

In summary, Developers prefer well-designed, modern systems that can be easily modified and extended. Interoperability is key. They appreciate an active and open developer community with a clear roadmap.

Researchers and data re-users

Researchers are both creators and owners of data, as well as re-users of other researchers' data. They have a variety of requirements relating to the deposit, management, discovery and re-use of research data. They are influenced by functionality found in other software, such as bookmarking, intuitive, advanced and faceted searching, across multiple catalogues. When they find data, they want to understand its provenance, how it was created and processed. Their requirements are also influenced by academic conventions such as the need for standardised citation formats and DOIs.

Researchers want help with using the data management system, through the provision of good documentation and training materials, an intuitively designed interface, and reassurance that they have done all they can to ensure their work can be easily discovered and cited. They also want help with choosing the correct license for their data.

When depositing data, they want the process to be simple and familiar though the integration of desktop file managers, cloud-based services such as Dropbox and a similar layout in the web interface. Uploads should continue as background tasks and should not be restricted by size. They want to deposit data directly from scientific instruments and be able to process it and query it on the server. Batch uploading and batch input/editing of metadata is an important requirement when working on datasets comprised of many resources.

Researchers want a personalised environment that pre-fills information about them and their work as much as possible and they want metadata to be automatically extracted from their data. When entering metadata, they want features such as controlled vocabularies.

Researchers want to be able to share the data to different degrees. Initially securely with colleagues and research partners and later to ensure that it is fully accessible to the public and optimised for ease of discover, download and re-use. They want metrics about how many times their data has been downloaded and cited as well as demographic information.

In summary, researchers want the data management system to support the whole lifecycle of the research process, offering tools to ingest, process, analyse and describe the data. Equally, they want good support and advice on how best to exploit the use of such a system.

Institutional requirements gathering

Requirements gathering exercises at several universities confirm the results of the above exercise. Multiple methods of access (web, API, networked storage), interoperability with existing academic systems, protocols and metadata standards (e.g. Current Research Information Systems (CRIS), SWORD2, OAI-PMH, Dublin Core), safeguards for the long-term preservation of and access to deposited data, the allocation of roles and access permissions to different users, versioning, and the ability to share data in different ways can be found across the reports from these exercises.

These documents also highlight the discipline specific requirements relating to the choice of metadata standards, the types and size of data being managed (e.g. video, tabular), and the level of security afforded to the data (e.g. medical, commercial and defence related).

Besides the functionality of the application, the sustainability of the data management system: the availability of commercial support, the openness of the company or community driving development, good documentation, the regularity of updates and fixes, and the opportunity to migrate away from any one system are all important considerations. No-one wants to feel locked into a poorly supported system.

Evaluation of CKAN

Development model

CKAN is open source software, licensed under the Affero GNU GPL 3.0 license.[38] This is a 'copyleft' license which requires other developers to make any code that they develop for a public CKAN instance available under the same license conditions. As such, it is not a permissive license like BSD, MIT or Apache, and requires 'reciprocity in perpetuity' among developers who work with it. In this way, the collective CKAN community benefits from the contribution of its users.[39]

The development of CKAN was started by Rufus Pollock, Founder of the Open Knowledge Foundation, who remained the main developer of the project until 2010. From 2007 to 2010, four developers are recorded as having contributed to the codebase.[40] Since 2010 (presumably relating to the adoption of CKAN for http://data.gov.uk), several developers have made especially large and sustained contributions. In almost six years, there have been 11,356 commits made by 75 contributors to create 243,878 lines of code. The project can be regarded as having "a well established, mature codebase maintained by a very large development team with increasing year-on-year commits"[41]

The opportunity to participate in the development of CKAN has been open since its first release in July 2007 and an invitation to join discussions about the project has been extended since 2005.[42] Today, there are two active email discussion lists for the project: CKAN-DEV[43] and CKAN-DISCUSS.[44] The developer mailing list is especially active. There is also a CKAN IRC channel for synchronous discussion[45] and weekly developer hangouts via Skype.[46]

---

38 https://www.gnu.org/licenses/agpl

39 For a discussion of 'reciprocity in perpetuity', see Peterson, Martin (2010) Property. Commoning and the Politics of Free Software. The Commoner, Vol. 14. https://commoning.wordpress.com/2011/01/04/misunderstanding-the-gnu-general-public-license-reciprocity-in-perpetuity/

40 https://github.com/okfn/ckan/contributors?from=2007-01-14&to=2009-01-14&type=c

41 http://www.ohloh.net/p/ckan

42 http://lists.okfn.org/pipermail/okfn-discuss/ (since 2005)

43 http://lists.okfn.org/mailman/listinfo/ckan-dev (since 2010)

44 http://lists.okfn.org/mailman/listinfo/ckan-discuss (since 2010)

45 #ckan on freenode

46 https://github.com/okfn/ckan/wiki/CKAN-Developer-Hangouts

Overall, CKAN has the attributes of a well-run, mature and active open source project that welcomes contributions.[47] The licensing is clear, the codebase is public, versioned and well documented;[48] standard tools are used for communicating with and within the developer community.[49]

This growing developer community reflects the increasing popularity of CKAN in the last three years. Open Data and the concept of a 'Data Management System' (DMS)[50] is relatively new. The OKF and its flagship project, CKAN, have both helped the growth of these respective initiatives as well as responded to them. The adoption of CKAN in 2010 by the UK government for http://data.gov.uk resulted in a significant increase in activity around open data and CKAN has become the *defacto* standard for publishing data by government bodies around the world. Most recently, the US government announced they are migrating http://data.gov to CKAN.[51]

Projects such as those led and funded by national governments and the EU determine how much development effort is put into CKAN and to some extent the direction of development, too. The CKAN website clearly states that new features can be 'sponsored'. Many aspects of data management are common across all domains, but from discussions with OKF it is clear that features are prioritised for development based on client's requirements. As such, the development of features in CKAN specific to the requirements of the academic community would have to be 'paid for' either in contributions of university developer time or by contracting OKF or a partner organisation to develop for the academic community.

Available support

One of the common requirements of institutions wishing to implement a system for RDM is that it is widely used and well supported, therefore reducing the risk of investing in an unsustainable infrastructure.

CKAN is open source software and therefore freely available to download and install. With a growing community of users and developers, opportunities for informal community support via mailing lists and IRC are likewise growing. OKF leads the development of CKAN and provides a number of commercial support options:[52]

- Three support options for *self-hosted* installations of CKAN, covering planning, installation, customisation and testing.
- Four support options for installations of CKAN *hosted* by OKF

47" https://github.com/okfn/ckan/blob/master/CONTRIBUTING.rst
48" http://docs.ckan.org; Basic Orientation to CKAN (for developers)
49" The CKAN project adheres closely to the guidance given by Fogel (2009) for running a successful open source project. See http://producingoss.com/
50" http://blog.okfn.org/2012/03/09/from-cms-to-dms-c-is-for-content-d-is-for-data/
51" http://www.data.gov/blog/ckan-horizon-datagov-20 http://blog.okfn.org/2013/05/23/u-s-governments-data-portal-relaunched-on-ckan/
52" http://ckan.org/datasuite/

- Ad hoc and pre-paid support on a daily basis. Support hours are normal working hours with a 24 hour response time.

In addition to support from OKF, there is also an international partner programme, with eight registered partners at the time of writing.[53]

Features

First impressions (version 1.7)

In our original evaluation of CKAN (v1.7.1) for the Orbital project, we identified a number of features which were of interest to us for a research data management system.[54] Principally, we were impressed with the design and flexibility of CKAN. It has an extensive API which its web UI is built on, providing the opportunity to integrate CKAN with other systems. Functions which are not yet designed into the web UI could still be programmatically possible through the API. Functionality can also be extended through developing 'extensions' to the core system, so that features can be added independent of the main direction of development.

As well as its extensibility, we were pleased with its versatility. CKAN can be used as a datastore (a relational database for the storage and analysis of data), a repository and a catalogue. Each of these features can be used together or independently when combined with other systems.

Finally, we were impressed with the way CKAN retains a history of changes to dataset metadata as well as keeping a record of user activity. It also includes a useful browser tool for visualising previews of datasets.

We also found it lacking in some areas, too. As a data management system principally designed to publish public open data, the security model was too simple for all use cases in higher education. Related to this was the way in which datasets and resources are organised, which did not immediately lend itself to the concept of 'research group' and 'project'. In version 1.7.1, the workflow was quite minimal, clearly intended to facilitate the *publishing* of open data rather than *manage* research data over the lifecycle of a project.

Version 2.0

Development of version 2 of CKAN was already underway when we undertook our first evaluation and we were reassured that forthcoming features would begin to address some of the issues identified above. Version 2 was released on May 10th 2013[55] and what follows is a re-evaluation of CKAN based on v2.0. Readers should bear in mind that new versions of CKAN are released quite regularly and new features are being added each time. As a result, parts of this document will be out-of-date within weeks of its completion, and you are encouraged to review the Changelog for CKAN to monitor its development.[56]

53" http://ckan.org/datasuite/partnerships/
54" http://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/
55" http://blog.okfn.org/2013/05/10/announcing-ckan-2-0/
56" http://docs.ckan.org/en/latest/CHANGELOG.html

The main features of CKAN are clearly presented on the project website http://ckan.org/features/ and release announcements are made on the CKAN blog, highlighting the main changes.[57] The evaluation below focuses on a *selection* of features which will be of interested to the research community. It was undertaken on May 20th 2013, using the public test version of CKAN at http://beta.ckan.org

Installation

The recommended Operating System for CKAN v2.0 is Ubuntu 12.04 64-bit. A pre-built package is available to download.[58] It is also possible to install CKAN from source on other versions of GNU/Linux. CKAN is primarily written in Python with dependences on Solr for search and PostgreSQL for its relational database.

Security, permissions and role management

The authorisation model for datasets changed with the release of version 2.0.[59] The concept of an 'organisation' was introduced with a dataset belonging to this new entity rather than an individual user. Roles and associated permissions over datasets are assigned by the organisation, making it easier to distribute responsibilities among a number of people. Individuals can be administrators, editors or simply members of the organisation.

For research data management, an organisation is likely to correspond to a research group, whereas in prior versions of CKAN, the only way to identify the ownership of a collection of datasets was through the use of CKAN 'groups', which are effectively labels that can be applied to multiple datasets. There is no sense of ownership of a dataset by a group since all datasets must belong to an organisation. Groups continue to exist in CKAN and are aimed at the ad hoc creation of collections of datasets. This might correspond to projects within research groups.

Activity data

A significant concern for research data management is the recording of actions made to the data. This relates to the requirement for understanding the provenance of a dataset. Without an understanding of the context in which the data has been derived and manipulated, it may be impossible to verify the research results. In CKAN v2.0, activity relating to a user, dataset, group and organisation is displayed in an 'Activity Stream'. Changes to datasets are reflected in the organisation's activity stream as well as any groups they are part of. Furthermore, users can 'follow' activity around another user, a dataset, a group or an organisation and are notified when changes are made. Activity that is 'followed' is aggregated into a personal 'News Feed' for each user.

Data previews and visualisation

Certain types of data held in CKAN can be automatically previewed and to some extent interactively analysed in the browser prior to download. This functionality has been included since CKAN v1.6 and steadily improved in subsequent versions. CKAN uses the Recline.js 'data

57" http://ckan.org/category/releases/
58" http://docs.ckan.org/en/ckan-2.0/install-from-package.html
59" http://docs.ckan.org/en/latest/authorization.html

explorer' library, developed by the OKF.[60] Depending on the nature of the data, it can be displayed in tabular form, graphed, displayed in a timeline, or plotted on a map. Since v2.0, custom previews for other filetypes can be created via CKAN extensions, making this area of CKAN functionality extensible.[61]

CKAN APIs

One of the key features of CKAN is its RESTful API.[62] Version 2.0 of CKAN provides a very rich interface to programmatically interact with CKAN with greater functionality available via the JSON-based API than through the web user interface itself. Furthermore, the CKAN Datastore[63] and Filestore[64] both have APIs, too, providing specific interfaces to create, read, update and delete tabular data held in the Datastore and upload and modify files held in the Filestore. All CKAN APIs can use authenticated requests when required. The extensive provision of APIs benefits all users of CKAN, allowing for integration with other institutional systems, and methods for researchers and data re-users to deposit, query and retrieve data. A basic command line client is available.[65]

Linked Data and RDF

Support for Linked Data and RDF has been included in CKAN since v1.7 and prior to this as an extension.[66] No special configuration is required to retrieve a dataset in RDF. Linked Data can be retrieved by requesting 'application/rdf+xml' or 'text/n3' from the API or more simply by adding the mimetype to the end of the dataset URI e.g. http://thedatahub.org/dataset/gold-prices.rdf Dublin Core and DCAT schemas are used by default, but other templates can be created.

Metadata

CKAN's default metadata schema is very simple but can be extended. By default, depositor's are asked for the following information:

Dataset:

- Title
- ID
- Description
- Group
- Organisation
- Tags
- License
- Author
- Author Email

60" http://reclinejs.com/; http://docs.ckan.org/en/latest/data-viewer.html
61" http://ckan.org/2013/03/13/custom-previews/
62" http://docs.ckan.org/en/latest/api.html
63" http://docs.ckan.org/en/latest/datastore-api.html
64" http://docs.ckan.org/en/latest/filestore-api.html
65" https://github.com/okfn/ckanclient
66" http://docs.ckan.org/en/latest/linked-data-and-rdf.html

Maintainer
Maintainer Email
Resource:
Title
Description
Format
File ID

Other system information such as date created/modified are also generated. Any number of ad hoc key-value pairs can also be created at time of deposit to accommodate subject-specific and unique metadata requirements.

Institutions will likely wish to extend the default schema and web input form to map to subject-specific standards or CERIF.[67] This is possible through the development of a CKAN form extension and creating a custom schema. Examples are given in the CKAN documentation.[68]

Datastore and Filestore

As previously discussed, the development of CKAN has moved from its focus on a data catalogue, to include file storage and most recently a datastore for tabular data.[69]

Unlike the upload and storage of a spreadsheet file (e.g. .xls), the Datastore (together with an extension[70]) allows such data to be ingested directly into a PostgreSQL relational database or optionally converted from an appropriate file and imported into the Datastore on deposit. Prior to version 1.8, CKAN used Elasticsearch for its Datastore, but now uses PostgreSQL and therefore provides a full SQL query interface over its own APIs.[71] Since the Datastore uses PostgreSQL, it can be managed and resourced like any other relational database.

Similarly, the CKAN Filestore[72] allows CKAN to use any local or remote storage medium (e.g. Amazon S3) and has an API for uploading data and modifying metadata. Currently, only one storage location can be used, rather than multiple locations.

Usability

Version 2.0 of CKAN introduced a completely new design for CKAN's web user interface. It introduces a simple workflow for adding a new dataset and resources. On the whole, it is an improvement over earlier versions but at this stage, lacks some of the functionality that was previously available (e.g. browsing metadata history).[73] Furthermore, the workflow remains too

---

67" http://cerif4datasets.wordpress.com/
68" http://docs.ckan.org/en/latest/forms.html
69" http://docs.ckan.org/en/latest/datastore.html
70" https://github.com/okfn/ckanext-datastorer
71" http://ckan.org/2012/10/22/ckan-1-8-released/
72" http://docs.ckan.org/en/latest/filestore.html
73 In the re-design, some user-facing functionality has been 'removed', but is still accessible via hidden URLs e.g. dataset history http://demo.ckan.org/dataset/history/climatologia-espana This underlines my point that CKAN offers more functionality than is apparent from its web UI.

simple to support curatorial processes, where research data is being deposited with an institution for long-term preservation. CKAN's workflow implicitly assumes that data owners themselves wish to publish their data to the web, rather than deposit their data with an institution to curate and manage.

Gap Analysis

The functional emphasis in CKAN v2.0 on data publishing over data management and curation should not be seen as problem with its design, but rather a reflection of its existing use and development over the last six years. Led by the OKF, whose mission is to "open up knowledge around the world and see it used and useful", the requirements for CKAN, starting as a catalogue of data already published elsewhere, have always been focused on data owners publishing data. Development has been sponsored largely by public authorities and governments with an agenda to improve the transparency of their public service. As such, CKAN has not been developed as a primary archival system for the permanent deposit and preservation of an organisation's data. The assumption, so far, seems to be that this work continues to be undertaken by record managers and archivists using existing processes and systems.

With this in mind, the research community's requirements for RDM extend the use of CKAN, while making full use of its current state of development. In the CKAN4RDM workshop, we undertook a 'gap analysis' exercise by reviewing the user stories and considering the work required to develop CKAN so as to meet those requirements.[74] The most significant gaps appear to be related to workflow and to bulk operations via the web user interface.

Based on our experience in implementing and using institutional repositories, institutions are cautious about academic staff publishing directly to the web without some form of moderation and there is good reason for this. Institutions are being mandated to take responsibility for the long-term accessibility of their researchers' data. UK funding councils expect research data to remain accessible for many years, even decades after the end of the research project.[75] Institutions must develop RDM solutions based on the assumption that they may be responsible for the data in perpetuity and are acutely aware of the associated costs involved in doing so. Therefore the deposit of research datasets, which can often amount to several gigabytes or even terabytes at a time, necessitates greater controls over the ingest, management and dissemination of the data than is currently designed into the existing CKAN workflow. Furthermore, these controls should be supported by improved auditing and reporting functionality so that data managers have adequate oversight of their dataset collection and are able to prove the provenance of the data they safeguard, the impact it is having, and to aid decisions about its disposal.

Furthermore, both researchers and data managers need to undertake bulk operations across most aspects of the workflow via the web browser. The ability to upload multiple resources at a

74" https://docs.google.com/spreadsheet/ccc?key=0Arh4BnSV2XSIdGxibkhLXzhsUVZzVTl6MWdYOWVMcWc&usp=sharing (Column five)
75" http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

time and batch edit the metadata is a basic requirement of any archival system which CKAN does not currently meet. It is our understanding that work is currently underway to address both improvements to workflow and batch operations.[76]

A number of other gaps were identified of varying importance and difficulty. Examples such as ensuring CKAN meets specific metadata requirements and is compatible with protocols commonly used by the research community, are less of a concern (in terms of resourcing the development) than significant changes to workflow, for example.

Conclusions and recommendations

The specific 'gaps' between the current state of CKAN and institutional requirements for RDM can be explained by the use-case requirements for open data publishing in recent years and the level of sponsorship that OKF have received from public bodies to meet their specific requirements. While a number of gaps have been identified in this paper and the related workshop, there is much cause for optimism about how CKAN can be currently used as part of a RDM infrastructure and how it may be developed to meet the research community's requirements.

At the University of Lincoln, we have begun to integrate CKAN with other existing research systems, such as an 'Awards Management System' containing project and funding details, staff directory, and ePrints repository.[77] Currently, CKAN provides *part* of a pilot RDM solution and offers functionality that no other system can provide, such as the Datastore, visualisation preview tools and activity data. Furthermore, CKAN's rich APIs and modular design through the use of extensions ensures that it can be more tightly coupled and integrated with other systems without requiring changes to the core system. We need to recognise where other systems are better suited to aspects of RDM and work on their integration with CKAN.

CKAN should not simply be judged on what user functionality it offers today but rather the potential in its technical design and user community to support the functionality of tomorrow. Research Data Management is a relatively new practice within higher education. The requirements for RDM are only just being made clear to information professionals, developers, researchers and funders. Being central to wider efforts to curate and publish collections of open data, CKAN can be seen as a good reflection of where the open data movement stands and CKAN's current functional limitations point to the rapid changes taking place as different domains join the open data movement with new requirements.

The design of CKAN should be seen as its greatest strength. It is free and open source software with a robust community of developers and high profile users across the world. All functionality is accessible over mature APIs and an 'extension' framework ensures that bespoke development to meet local requirements can be accommodated. In this respect, as a *technology platform* for managing data, it is versatile, flexible and extensible. Because of this, it

76" http://ckan.org/2013/02/27/ckan-2-0-beta-has-arrived
77" https://orbital.blogs.lincoln.ac.uk/2013/05/03/the-researcher-dashboard/

is also a *low risk* option for institutions to adopt, and furthermore, the services provided by OKF and its partner programme ensure that support is contractually available on an on-going basis.

On the basis of this initial evaluation of CKAN for Research Data Management, the following recommendations to higher education institutions, funding bodies and the research community, can be made:

- The requirements for RDM differ across different research disciplines, but many common requirements can also be identified. These common RDM requirements are derived from work and experience within the higher education sector over the last two decades to define standards, protocols and best practice for managing research information and research outputs in general. **The development of a research data management system, such as CKAN, should build upon this experience and extend that work. We are not starting from scratch.**

- Relatedly, the 'RDM community' is in its youth, having grown out of major national funding programmes.[78] However, it also has roots in an older community of information professionals working in higher education. Communication among people working in this area is well established and **those of us interested in the use of CKAN for RDM should use our existing networks to drive the development of CKAN for our community's needs.** This work has already begun (i.e. the London workshop, this paper, the adoption of CKAN at Lincoln and Bristol[79], etc.) but we should be aware that as leaders of the CKAN project, the OKF cannot steer the development of CKAN for RDM without leadership, contributions and resourcing from the HE sector. How this work might be undertaken and evolve can be discussed on a dedicated mailing list.[80]

- CKAN is the *defacto* standard software for publishing open data. In a very short time, it has been widely adopted by high profile public projects, such as http://data.gov.uk (UK), http://data.gov (USA) and http://data.gov.au (Australia). It is a mature, open project with a sound model of development and a growing community of developers. As such, the adoption of CKAN for RDM allows a research institution to benefit from the experience gained and work already undertaken by other clients and projects. It is **a low risk option with immediate benefits.**

- The licensing and design of CKAN offers a very flexible technology solution for meeting common and local RDM requirements. Research institutions and funding bodies should identify key requirements that are common to the sector and **collectively resource the development of CKAN to improve its utility for RDM**. This may be through the active and recognised participation of higher education developers in the CKAN community and/or the funding of specific projects which address agreed objectives.

- **A closer relationship between the higher education sector and the Open Knowledge Foundation should be developed**. The principles, objectives and overall public mission of the

78" Mailing lists include RESEARCH-DATAMAN https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=RESEARCH-DATAMAN and JISCMRD https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=JISCMRD
79" http://data.bris.ac.uk/2012/12/18/ckan-and-data-bris/
80" http://lists.okfn.org/mailman/listinfo/ckan4rdm

OKF[81] are shared by the HE sector and therefore greater effort should be made to formally support and complement the work of the Foundation. There are a growing number of academics actively engaged with the work of OKF and reciprocally receive the Foundation's support. This should be recognised and reflected through support from HE institutions and funding bodies, so as to leverage the enthusiasm of the open knowledge community.

81" e.g. http://pantonprinciples.org/