Using CKAN: storing data for re-use

the Open Source Data Portal Software

Mark Wainwright, Open Knowledge Foundation

Introduction

CKAN is a free, open-source data hub software package, written in Python, developed by the non-profit Open Knowledge Foundation. It is used to power local, national and supranational 'open government data' portals around the world, as well as community data hubs in various countries. Examples are the UK's data.gov.uk and the European Union's publicdata.eu, the Brazilian dados.gov.br, and city and municipal sites in the US, Argentina, Finland and elsewhere. Community instances such as the DataHub (thedatahub.org) allow anyone to publish data for free.

CKAN is not a repository

A repository is sometimes seen as a place to deposit your research and then forget about it. In this sense CKAN is not a repository. It can certainly do what is needed from a repository, but it is also a place where data will carry on working for the research community.

It is also not an *institutional* repository in that it has not yet been widely used as one. Setting it up as one would take some work, but we'd be happy to talk to an institution that was interested in trying it. Another possibility would be to use it as a datastore alongside an existing repository. It can be and is used now to publish research outputs on the DataHub.

CKAN is a repository

CKAN has the essential features for an academic repository: rich configurable metadata, datasets to which resources can be added, a datastore with preview, fine-grained options for authorisations, curated groups of datasets (e.g. for different departments), versioned history, faceted search, and an easy and intuitive web interface. It also has other features that could add value in various ways, some of which are mentioned in the sections below.

Web, command line and API interfaces

CKAN has an intuitive and user-friendly web interface for uploading, editing and searching: a user can create a dataset in a couple of minutes. The search is heavily road-tested on portals like data.gov.uk and allows free text search or faceting by group (department), document type, etc.

Heavy users can also make use of the Open Knowledge Foundation's open-source command-line data package manager, dpm. (dpm could also be adapted for use with other data repositories.)

Collect data resources together

CKAN has 'datasets' each containing any number of 'resources'. A paper could be catalogued as one dataset with resource such as: different versions of the printed paper (e.g. TeX and a PDF); a link to the paper's page on a journal website; spreadsheets of experimental results; the source code to process the results; and others, such as separate image files of graphs and diagrams.

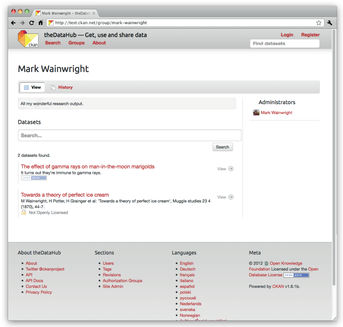
Resources

1. Journal page for article — html
2. Source code repository — html
3. CCC Temperature anomaly data — text/csv
4. PDF preprint ⓿ 5 — application/pdf
5. Comparison graph: GISTEMP/ccc — image/png

1. Published article (external link)
2. Source code at Google Code (external link)
3. Data file - previewable and queryable
4. Authors' PDF preprint
5. Image files can be separately stored

Community repository: Personal publication lists

A repository need not be run by an institution to be useful. Got a piece of data or research you want to share via CKAN at a permanent address? You can do it right now at thedatahub.org, a CKAN repository where anyone can register and upload datasets. Start a group for all your own papers, giving your output a permanent address even when you move department. Or start a group for your department's papers. Permissions are configurable for each dataset, for example allowing all co-authors of a paper to update it.



A permanent address for your research output

Rich metadata

Each resource, including external links, has its own associated metadata, as does the whole dataset. The default configuration includes standard fields such as author, title, description, licence, etc. Arbitrary extra fields can be added for each dataset. A CKAN site specialised for research could include such fields as DOI, Journal, etc, by default, and these can vary according to dataset type. (For example, a thesis could have required fields such as 'Supervisor'.)

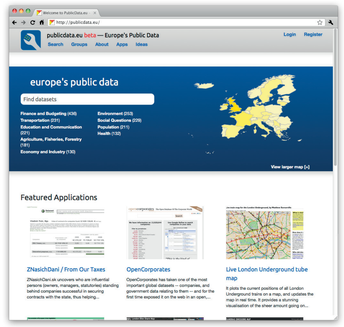
Additional Information

Field	Value
Author	Nick Barnes and David Jones
Maintainer	Maintainer not given
DOI	10:1109/MS.2011.113
Issue no	6
Journal	IEEE Software
Journal homepage	http://www.computer.org/portal/web/computingnow/software
Publication date	Nov-Dec 2011
Volume	28

Metadata can include arbitrary fields, with configurable defaults

Federation and linking

CKAN's 'harvesting' feature can federate datasets between different servers. For example, a research council could run its own repository, and harvest metadata from institutions about research it has funded.



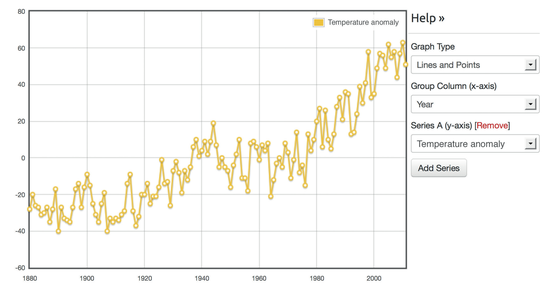
publidata.eu harvests data from 18 European data catalogues

CKAN's metadata can also be exported in standard formats including the W3C data catalogue standard DCAT, and RDF (Linked Data) output is built in. Because CKAN has not been widely used for academic repositories, there is no support at the moment for OAI-PMH. This would be an excellent area for a CKAN extension (see Future work).

Maximising re-use: raw data now!

CKAN's datastore can store structured data and provide access to it via an API. This means a data file can be linked to or uploaded as a CSV or spreadsheet, and users - as well as downloading it - can query it directly on the server. This could make life easier for researchers checking and re-using data from earlier research - their own as well as others' - as large datasets can be explored without the need to download and build interfaces for them.

CKAN creates interactive data visualisations, using the built-in Recline data viewer, which can be embedded elsewhere on the Web - for example, in a blog post about the research that produced the data. Visualisations also include map plots of geo-coded data. Image files are displayed on their resource pages.



Write a blog post and include an interactive view on your data

Future work

CKAN is highly extensible, with a standard interface for writing extensions, which can also do background processing. While the DataHub can be used to store research right now, it would be interesting to see how a widely-used CKAN instance specialised for research data would develop. To mention just one additional aspect, CKAN dataset metadata can include links to other datasets. This could be used to implement a system of references as outward links, with inward links displayed automatically as citations.

Find out more

For further information, contact **info@ckan.org** or visit **www.ckan.org**