NERA

**Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation**

# Report

# D9.2 Architecture Specification for data workbench

| | |
|---|---|
| Activity: | *European-Mediterranean Earthquake Portal and Services* |
| Activity number: | *NA9, Task 9.2* |
| Deliverable: | *Architecture specification for data workbench* |
| Deliverable number: | *9.2* |
| Responsible activity leader: | *EMSC* |
| Responsible participant: | *EMSC* |
| Author: | *Laurent Frobert, ed.* |

**Seventh Framework Programme**
**EC project number: 262330**

## Summary

The Seismic Portal (www.seimicportal.eu) provides several small web applications to search and present data about earthquakes (like parametric events, broadband waveforms …). Users can submit queries via several applications providing access to different sort of information. The obtained results are then saved into dedicated "data-carts", which are not offering an homogeneous management of the data and the metadata, showing the limitations of the current approach either in terms of interactivity and sustainability.

A solution proposed for the NERA project is to merge all carts into a unique place where the users can see all their saved datasets and queries. The users will be able to organize (create groups, add tags, add metadata, remove data), share (with others users) and launch processes workflow in the cooperation with developments of the EC project VERCE on his dataset.

Unlike the current seismic portal's cart, the user workbench will be accessible not only by a web user interface but also by third party services. Moreover, it will allow future integration of new data management tools, based on the users' requirements.

The workbench will be useful for the submission of user-defined data queries like "all broadband waveforms from the last event". Notifications will be sent to subscribers by using different kind of systems like email or web-callback.

In this document we will present in more details the features that the workbench will offer to the user but also to the data and services providers and we will describe the technology constraints to solve to meet the workbench's requirement.

This workbench idea is a key part of the seismic portal and we can find a similar approach in other European project like EUDAT[1].

[1] http://www.eudat.eu/simple-store

# Introduction

The NERA project aims to enhance the work done during the NERIES past European project by using standardisation for web service to deliver data from the different data provider in the seismologic field. It aims also to spread data in a more automatic way when they are available and to search them efficiently. Moreover, a dataset created by a user should be easily be shared between researcher.

During the production phase of the seismic portal, more than 1000 users registered from 100 countries. And after a period of activity a survey from EMSC has been submitted to their members to have feedback from their portal use.

As we can see the cart paradigm is not often used, because lack of features. Currently a user can see his saved dataset but without organise or search them. Moreover, there is one cart per application, each cart with a different user interface.

Event cart

Accelerometric waveform cart

Broadband waveform cart

In the NERA project, an accent has been made to let the user share their work for other people in an easy way, the current ‘cart paradigm’ in the seismic portal is not well adapted to this task so a new idea has been proposed for the NERA project : the Workbench.

# The Workbench

## *The concept*

A virtual data workbench, an extension of the current data cart concept, will be one of the key elements in this integration part. It will provide a user friendly interface to work with selected data and an optimal interface to additional web services, processing services and portlet applications, i.e. provide interoperability (GEO[2], GEM[3] ..) and promote multidisciplinary data-sharing.

The workbench will be provided as a hosted web service. The actual data search result remains stored outside of the workbench whilst the workbench will provide the information to the user on the status and on the location of the heterogeneous data products requested, that will be collected and packed either at the data or processing center, or at a participating storage provider. The workbench will support a query interface, while individual result set items are accessed via a URL. Data processing services locate their input data sets through these URLs. The virtual data workbench handles the metadata management both related to the original query, as well as meta information on the data processing steps, as an integral part of the overall data provenance and workflow strategy followed within the portal and services design. This design will enable:

1) an integrated portal encompassing access to a comprehensive set of data and data products, as well as computational resources;

2) targeted portals for specific well-defined user groups to be built on top of the core set of data and computation web services ;

3) the next generation researchers to develop custom analysis applications that access the available catalog of services.

Users will have more and more dataset processed available (from the Verce[4] European project and Seismic portal Waveform explorer for example), they can't store these dataset on their desktop storage because of high data volume and asynchronous result of Process

The idea of an online workbench is to let the dataset to be stored on their own repository where they are created, like datacenter or HPC, and let the users see a short description (described by metadata) into this workbench.

The workbench will present a list of all available and ongoing dataset or processes start by the user.
The workbench will act as a virtual desktop available all over the world to the user.
The user will be able to filter the view by using different criteria (date,size, metadata, ...) and to manage dataset (download,remove, start workflow process ...)

[2] http://www.earthobservations.net
[3] http://www.globalquakemodel.org
[4] http://www.verce.eu/

Additional feature can be to send notification to the user when a dataset is ready, this feature need a notification of dataset updated by the service provider and the definition of a common protocol to communicate between the service providers and the workbench service. A common user authentification service should be used to identify the user.

The workbench should be a central and unique point to manage dataset and processes from different Service and Data Providers.

Benefits for the service/data providers:

- dataset are stored on their own repository (where they are created, so avoiding downloading time to the user)
- service/data providers keep control of their data
- service/data providers don't have to provide graphic user interfaces to let the user manage the dataset

Benefits for the users :

- users will access their dataset everywhere, every time
- users will have an unique central access to all of their dataset
- users will be able to share their work
- dataset can be shared with other catalog service from different domains

Some basic requirement must be present to fulfil our workbench concept :

### *Features needed*

A user should be able:

- To add, remove and manage data (group data, tag data).
- To provide data (photo, spreadsheet…) or a reference of a data (URL) itself, not only by the data provider's
- To keep private some data and share others
- To pre-visualize the dataset or a subset of itself
- To launch workflow process
- To be notified when a "registered" dataset is up to date
- To be notified when a new dataset matching certain criteria is created

Data are provided either by the user but more by the data provider of the seismic portal or other service application from a user interaction so the user's workbench should be fed directly by the data provider's, an Application Programming Interface (API) must be provided. Basically, the API should provide a way to add, update and remove dataset into the user workbench.

To be able to add data to a user's workbench, the application must have user identification

### User authentification and authorisation

A user can search data without been authenticate because some data are public but for personal data a user must logged on the system. More over, The authentification mechanism should be compatible with the one used on the seismic portal: a user already authentificate on the seismic portal should not re-log on the workbench.

User’s data should be protected and visible only by the owner unless he has published his data to be shared

At least simple role management should be provided perhaps group management. For a first implementation we think about using the OpenID[5] authentification mechanism.

## Technical solutions

### Which kind of software ?

Instead of reinventing the wheel we have investigated on possible open source software fulfilling our needs. Browsing the internet we have found a category of software responding to our requirements: Document Management System.

From Wikipedia, the free encyclopedia:

“A **document management system** (DMS) is a computer system (or set of computer programs) used to track and **store electronic documents**. It is usually also capable of keeping track of the different versions modified by different users (history tracking). The term has some overlap with the concepts of content management systems. It is often viewed as a component of enterprise content management (ECM) systems and related to **digital asset management**, document imaging, **workflow systems** and records management systems.”

Document management systems commonly provide storage, versioning, metadata, security, as well as indexing and retrieval capabilities. Here is a description of these components:

| Topic | Description |
|---|---|
| **Metadata** | Metadata is typically stored for each document. Metadata may, for example, include the date the document was stored and the identity of the user storing it. The DMS may also extract metadata from the document automatically or prompt the user to add metadata. Some systems also use optical character recognition on scanned images, or perform text extraction on electronic documents. The resulting extracted text can be |

[5] http://openid.net/

| | |
|---|---|
| | used to assist users in locating documents by identifying probable keywords or providing for full text search capability, or can be used on its own. Extracted text can also be stored as a component of metadata, stored with the image, or separately as a source for searching document collections. |
| **Integration** | Many document management systems attempt to integrate document management directly into other applications, so that users may retrieve existing documents directly from the document management system repository, make changes, and save the changed document back to the repository as a new version, all without leaving the application. Such integration is commonly available for office suites and e-mail or collaboration/groupware software. Integration often uses open standards such as ODMA, LDAP, WebDav and SOAP to allow integration with other software and compliance with internal controls. |
| **Capture** | Capture primarily involves accepting and processing images of paper documents from scanners or multifunction printers. Optical character recognition (OCR) software is often used, whether integrated into the hardware or as stand-alone software, in order to convert digital images into machine readable text. Optical mark recognition (OMR) software is sometimes used to extract values of check-boxes or bubbles. Capture may also involve accepting electronic documents and other computer-based files. |
| **Indexing** | Indexing tracks electronic documents. Indexing may be as simple as keeping track of unique document identifiers; but often it takes a more complex form, providing classification through the documents' metadata or even through word indexes extracted from the documents' contents. Indexing exists mainly to support retrieval. One area of critical importance for rapid retrieval is the creation of an index topology. |
| **Storage** | Store electronic documents. Storage of the documents often includes management of those same documents; where they are stored, for how long, migration of the documents from one storage media to another (hierarchical storage management) and eventual document destruction. |
| **Retrieval** | Retrieve the electronic documents from the storage. Although the notion of retrieving a particular document is simple, retrieval in the electronic context can be quite complex and powerful. Simple retrieval of individual documents can be supported by allowing the user to specify the unique document identifier, and having the system use the basic index (or a non-indexed query on its data store) to retrieve the document. More flexible retrieval allows the user to specify partial search terms involving the document identifier and/or parts of the expected metadata. This would typically return a list of documents which match the user's search terms. Some systems provide the capability to specify a Boolean expression containing multiple keywords or example |

| | |
|---|---|
| | phrases expected to exist within the documents' contents. The retrieval for this kind of query may be supported by previously built indexes, or may perform more time-consuming searches through the documents' contents to return a list of the potentially relevant documents. |
| **Distribution** | A published document for distribution has to be in a format that can not be easily altered. As a common practice in law regulated industries, an original master copy of the document is usually never used for distribution other than archiving. If a document is to be distributed electronically in a regulatory environment, then the equipment tasking the job has to be quality endorsed AND validated. Similarly quality endorsed electronic distribution carriers have to be used. This approach applies to both of the systems by which the document is to be inter-exchanged, if the integrity of the document is highly in demand. |
| **Security** | Document security is vital in many document management applications. Compliance requirements for certain documents can be quite complex depending on the type of documents. For instance, in the United States, the Health Insurance Portability and Accountability Act (HIPAA) requirements dictate that medical documents have certain security requirements. Some document management systems have a rights management module that allows an administrator to give access to documents based on type to only certain people or groups of people. Document marking at the time of printing or PDF-creation is an essential element to preclude alteration or unintended use. |
| **Workflow** | Workflow is a complex process and some document management systems have a built-in workflow module. There are different types of workflow. Usage depends on the environment to which the electronic document management system (EDMS) is applied. Manual workflow requires a user to view the document and decide whom to send it to. Rules-based workflow allows an administrator to create a rule that dictates the flow of the document through an organization: for instance, an invoice passes through an approval process and then is routed to the accounts-payable department. Dynamic rules allow for branches to be created in a workflow process. A simple example would be to enter an invoice amount and if the amount is lower than a certain set amount, it follows different routes through the organization. Advanced workflow mechanisms can manipulate content or signal external processes while these rules are in effect. |
| **Collaboration** | Collaboration should be inherent in an EDMS. In its basic form, a collaborative EDMS should allow documents to be retrieved and worked on by an authorized user. Access should be blocked to other users while work is being performed on the document. Other advanced forms of collaboration allow multiple users to view and modify (or markup) a document at the same time in a collaboration session. The resulting document should be viewable in its final shape, while also storing the markups done |

| | |
|---|---|
| | by each individual user during the collaboration session. |
| **Versioning** | Versioning is a process by which documents are checked in or out of the document management system, allowing users to retrieve previous versions and to continue work from a selected point. Versioning is useful for documents that change over time and require updating, but it may be necessary to go back to or reference a previous copy. |
| **Searching** | Searching finds documents and folders using template attributes or full text search. Documents can be searched using various attributes and document content. |
| **Publishing** | Publishing a document involves the procedures of proofreading, peer or public reviewing, authorizing, printing and approving etc. Those steps ensure prudence and logical thinking. Any careless handling may result in the inaccuracy of the document and therefore mislead or upset its users and readers. In law regulated industries, some of the procedures have to be completed as evidenced by their corresponding signatures and the date(s) on which the document was signed.<br>The published document should be in a format that is not easily altered without a specific knowledge or tools, and yet it is read-only or portable. |
| **Reproduction** | Document/image reproduction is key when thinking about implementing a system. It's great to be able to put things in, but how are you going to get them out? An example of this is building plans. How will plans be scanned and scale be retained when printed? |

As we can see, features of DMS match well with features we have listed before to be a requirement for the workbench. Requirement to choose between available DMS on the market:

- ➔ open source
- ➔ well documented : Both administration and developer documentation must be provided
- ➔ large user community : The DMS should be already used by a large community to have a maximum feedback
- ➔ extensible (plug-in) : we must be able to add more feature ourselves to fulfil our needs
- ➔ preferred programming language : Python, We are using the python programming language
- ➔ Web Based User interface : A web browser based user interface should be provided for the user and a good API should be provided for the data/service providers

After some investigation, we have found a good candidate: the **Comprehensive Knowledge Archive Network**[6] (CKAN) because it offers nice features:

- ➔ open source, use of standards, well documented, python, easy use, plugin architecture, harvesting , Restful API, Q&A on dataset, comments …

[6] Ckan.org

### *The Comprehensive Knowledge Archive Network (CKAN)*

CKAN is a fully-featured, mature, open source data portal and data management solution. CKAN provides a streamlined way to make your data discoverable and presentable. Each dataset is given its own page with a rich collection of metadata, making it a valuable and easily searchable resource.

CKAN development is managed by the Open Knowledge Foundation[7], a non-profit organisation dedicated to finding solutions to the technical and social problems of opening up knowledge and data. The OKF has a team of full-time CKAN developers. In addition there are volunteer contributions to the code from around the world.

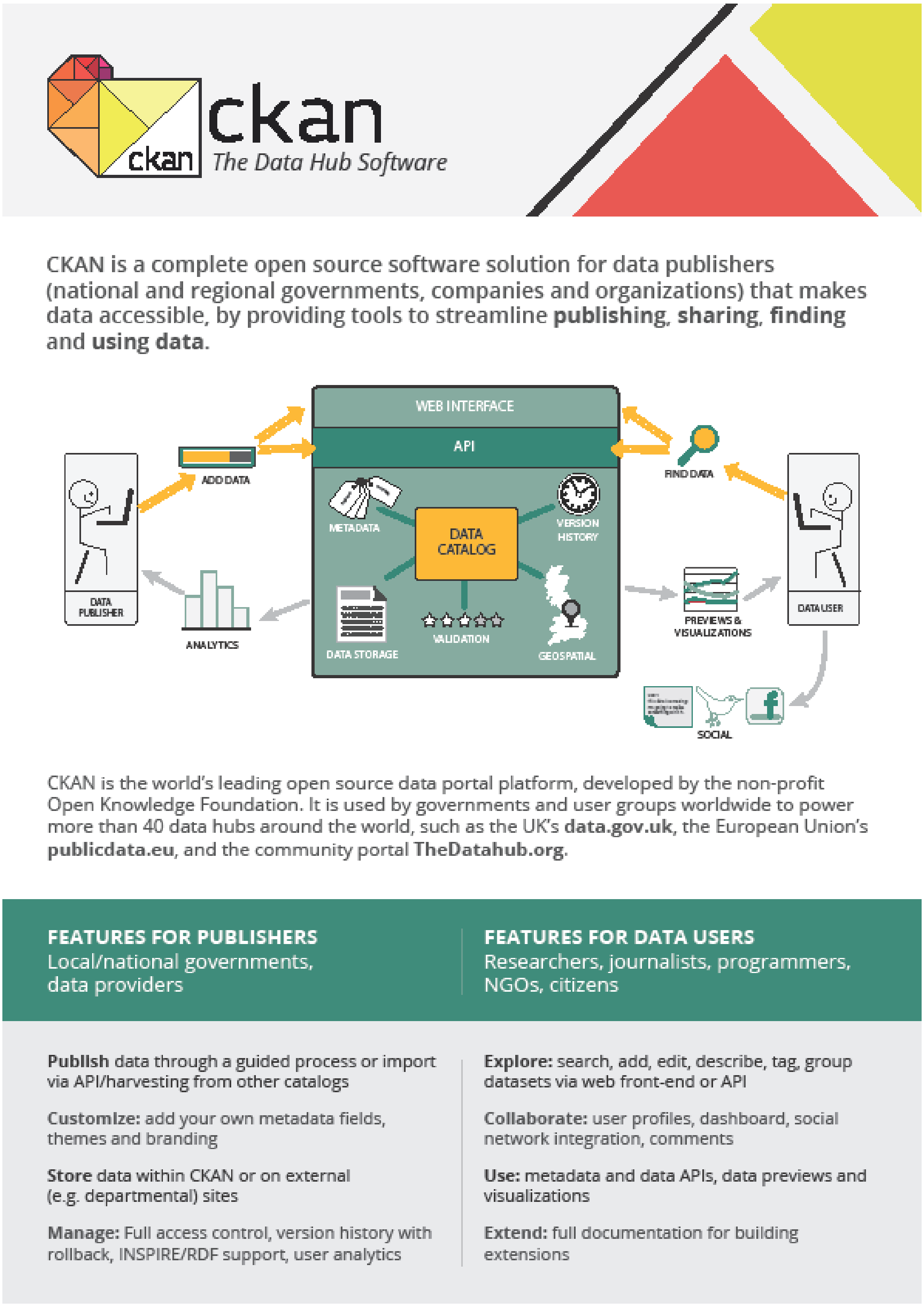
CKAN main features are:

- **Publish & find datasets**: Publish datasets via import or through a web interface. Search by keyword or filter by tags. See dataset information at a glance. Full change history lets you easily undo changes or view old versions.
- **Store & manage data**: Store the raw data and metadata. Visualise structured data with interactive tables, graphs and maps. Get statistics and usage metrics for your datasets. Search geospatial data on a map by area.
- **Engage with users & others**: Federate networks with other CKAN nodes. Theme with CSS or integrate with a CMS. Build a community with extensions that allow users to comment on and follow datasets.
- **Customise & extend**: Use the API‘s rich programming interface, and benefit from over 60 extensions including link checking, comments, and analytics. CKAN’s Open Source licence allows us to download and run it for free.

[7] Okfn.org

# Annexe

## *The CKAN brochure*



ckan
The Data Hub Software

CKAN is a complete open source software solution for data publishers (national and regional governments, companies and organizations) that makes data accessible, by providing tools to streamline **publishing**, **sharing**, **finding** and **using data**.

WEB INTERFACE
API
METADATA
DATA CATALOG
VERSION HISTORY
DATA STORAGE
VALIDATION
GEOSPATIAL
ADD DATA
DATA PUBLISHER
ANALYTICS
FIND DATA
DATA USER
PREVIEWS & VISUALIZATIONS
SOCIAL

CKAN is the world's leading open source data portal platform, developed by the non-profit Open Knowledge Foundation. It is used by governments and user groups worldwide to power more than 40 data hubs around the world, such as the UK's **data.gov.uk**, the European Union's **publicdata.eu**, and the community portal **TheDatahub.org**.

**FEATURES FOR PUBLISHERS**
Local/national governments, data providers

**Publish** data through a guided process or import via API/harvesting from other catalogs

**Customize:** add your own metadata fields, themes and branding

**Store** data within CKAN or on external (e.g. departmental) sites

**Manage:** Full access control, version history with rollback, INSPIRE/RDF support, user analytics

**FEATURES FOR DATA USERS**
Researchers, journalists, programmers, NGOs, citizens

**Explore:** search, add, edit, describe, tag, group datasets via web front-end or API

**Collaborate:** user profiles, dashboard, social network integration, comments

**Use:** metadata and data APIs, data previews and visualizations

**Extend:** full documentation for building extensions ckan

*the Open Source Data Portal Software*

# Using CKAN: storing data for re-use

*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.

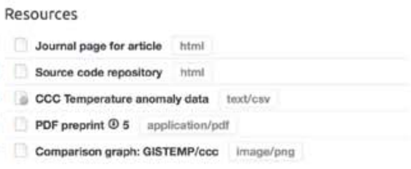
## Web, command line and API interfaces

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 ⓓ 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**

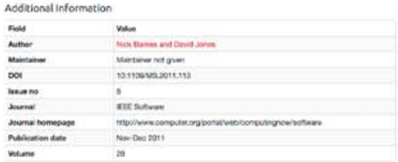
## Community repository: Personal publication lists

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.

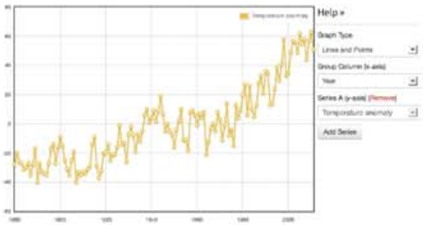
**publidata.eu harvests data from 18 European data catalogues**

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**