European Public Sector Information Platform

Topic Report No. 2013 / 01

# Europe's Data Catalogues

Author: Daniel Dietrich

Published: January 2013

## Keywords

Data Catalogue, Open Data Portal, Open Data, Public Sector, Public Sector Information, PSI, PSI Directive, PSI reuse, Technology

# 1 Executive Summary

Over the past years there has been an explosion of interest in opening up Public Sector Information (PSI) for the public to reuse. At a national level, numerous countries from around the world have established central data platforms. Several European Union member states have set up official national data catalogues as single-entry points to simplify public access to PSI of the country. Examples range from the UK's data.gov.uk site to the French Government's data.gouv.fr. There are countless city level initiatives across Europe as well – from Helsinki to Berlin, Paris to Zaragoza. As many open data catalogues as exist now, still even more initiatives will emerge in the next years.

This topic report explains the basic concepts of data catalogues and introduces the technologies used. It showcases recent examples of data catalogues in European member states and pan-European initiatives to ultimately improve access to Europe's PSI. The topic report explores the role and impacts of data catalogues in accelerating PSI reuse.

# 2 Introduction

Information is the gold of the 21-century. The European Commission estimates the potential revenue from making Public Sector Information (PSI) available for reuse to deliver a €40 billion boost to the EU's economy each year.[1] In order to unlock the potential of digital PSI prospective reusers must be able to "discover and access" datasets they might be interested in reusing.

Publishing PSI following Open Data Principles[2] on the Internet is the first step, but it doesn't mean that people can actually find the data easily. There are several measures to improve the discoverability of PSI, these include:

- Use of formats than can be processed by machines (index)
- Proper use of metadata
- Search engine optimization
- Create dedicated central open data catalogues

1 European Commission, Digital Agenda: Turning government data into gold, 2011, http://europa.eu/rapid/press-release_IP-11-1524_en.htm See also recent economic analysis, http://ec.europa.eu/information_society/policy/psi/facilitating_reuse/economic_analysis/index_en.htm

2 See for example the Ten Principles For Opening Up Government Information, Sunlight Foundation, 2010, http://sunlightfoundation.com/policy/documents/ten-open-data-principles/

Open data catalogues play a key role in Open Government Data (OGD) initiatives around the world. An open data catalogue provides a central location for listing or storing data released by a government for use by outside consumers, making such data more easily discoverable for reuse. As of today 281 data catalogues have been established on the national, subnational and local levels[3] plus specialized catalogues for geo data or environmental data. In EU27 member states alone there are more than 70 official data catalogues.[4]

Source: Open Data Census Dashboard, http://census.okfn.org/catalogs/

# 3 Features of open data catalogues

Open data catalogues usually display datasets in lists than can be searched and sorted by topics and categories, or by data-holders (Public Sector Bodies - PSBs), location, date, license, format, etc. Each Dataset is displayed with additional information in form of a description and metadata like: formats, license, publisher, publication date, geographic coverage, publication date and temporal coverage, last update date and update frequency.

3 See the List and map of the Open Data Census Project: http://census.okfn.org/catalogs/

4 A list of official data catalogues in EU27 member states can be found at http://datacatalogs.org/group/eu-official

Datasets can be directly accessed via bulk download or via an Application Programming Interface (API).[5]

Some data catalogues provide extra features like APIs or SPARQL endpoints[6], lists of applications built with the certain dataset, discussion and comment features, and built-in tools to analyse and visualize data. These should be offered in addition to access to download datasets, rather than instead of. A good practice is to add contact information of the data-holder to each dataset and to offer the public a way to request for additional datasets that are not yet published in the catalogue. This can be done by sending a request to the central body responsible for the data portal or by filing a Freedom of Information request to the PSB responsible for the requested dataset.

Central data catalogues have become very popular in recent Open Government Data (OGD) initiatives to provide a single entry point to all the governments' data. Central catalogues usually harvest (automatically pull together) a multitude of datasets from different decentralized Public Sector Bodies into a central place: a one-stop-shop. While most national OGD initiatives establish and run their own data catalogue using specialist software, although at sub-national level and at municipality level, in a country where there is no central catalogue, it might be a better option to curate a list of datasets in a common public data portal like www.datahub.io or www.datamarket.com instead of investing resources in own data catalogue infrastructure.

**Interoperability**

With a multitude of catalogues emerging it becomes clear that exchange of data between the catalogues is essential. To improve the interoperability between data catalogues several stakeholders have initiated attempts to set standards for data exchange formats and data protocols.[7] The ultimate aim of such initiatives is to ensure that data catalogues don't end up being new data silos, but rather to allow for federated search over multiple data catalogues (indexes) and facilitate easy exchange of metadata and data.

# 4 State of Play

**data.gov.uk**, the central catalogue of the Government of the United Kingdom, surely is Europe's most famous data open data site. Launched in 2009 it was the first of its kind in Europe and has served as role-model and inspiration to other governments ever since. But this portal was not only the first one in Europe, but also a great example of the 'open

5 Application Programming Interfaces (API) have become very popular because they allow programmers to select specific portions of the data, rather than providing all of the data in bulk as a large file.

6 SPARQL Protocol and RDF Query Language) is a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format.

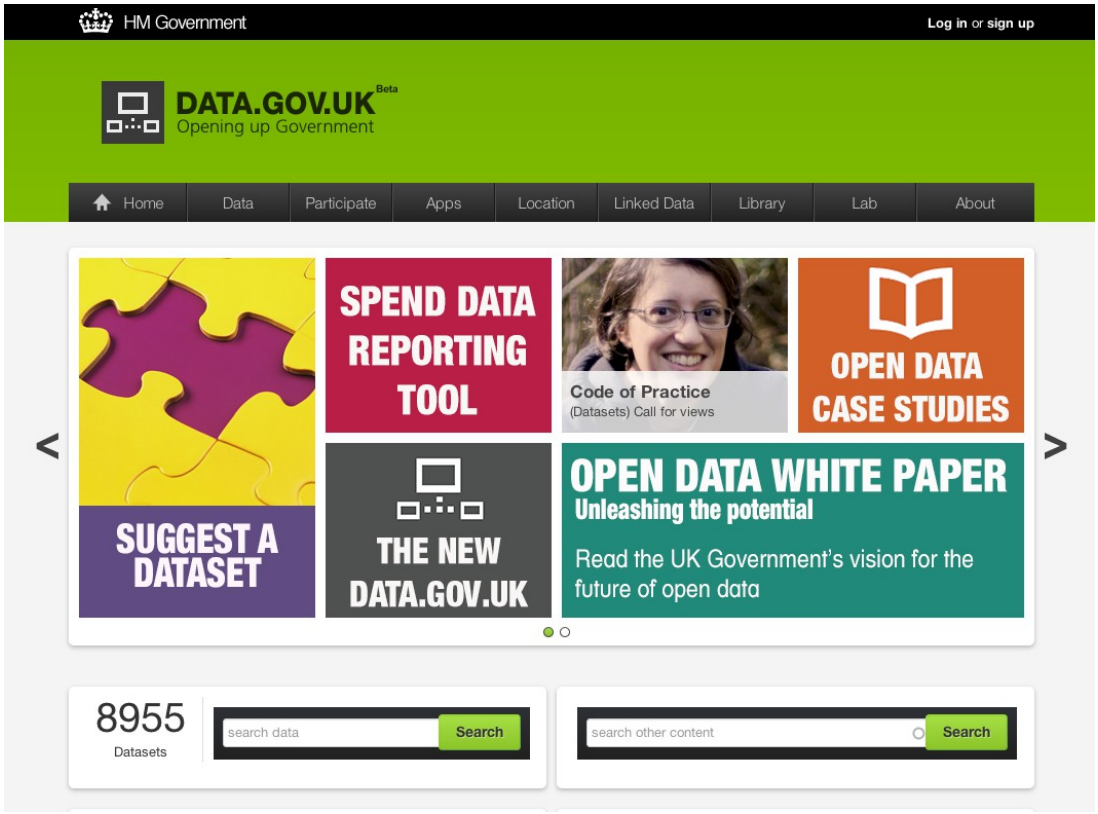
7 While most technical standards are being established by standards setting organizations like the W3C there are also community-driven efforts to develop simple, light-weight protocols and formats, like Data Protocols, http://www.dataprotocols.org/ or industry initiatives such as Microsoft's Open Data Protocol (OData), http://www.odata.org/

government' spirit: "*pioneering the rapid development of a new digital part of government using open source code, and developed through fluid collaboration between government staff, academics, open source developers, and open data activists from outside government.*" as Tim Davies points out in an article on the relaunch of the platform in 2012.[8] The overall aim of the relaunch was to simplify search and improve the user experience for ordinary citizens and developers the like. As Tim Davies continues:

> *"The site includes a clear hierarchy of publishing organisations (over 700 of them) and somewhere in each hierarchy there is a named contact to be found. That means that when you're looking at any dataset it's now easier to find out who you can contact to ask questions about it, or, if the data doesn't tell you what you want, the new data.gov.uk lets you exercise your Right to Information (and hopefully soon Right to Data) and points you to how you can submit a Freedom of Information request. Whilst at first most of these enquiries will go off to the lead person in each publishing organisation who updates their records on data.gov.uk, the site allows contact details to be set at the dataset level, moving towards the idea of data catalogues not as a firewall sitting between government and citizens, but as the starting point of a conversation between data owners/data stewards and citizens with an interest in the data. Using data to generate conversation, and more citizen-state collaboration, is one of the key ideas in the 5 stars for open data engagement".*[9]

8 Tim Davies, Laying the foundations for open data engagement, 2012, http://data.gov.uk/blog/laying-the-foundations-for-open-data-engagement

9 See also Tim Davies, 5 stars for open data engagement, 2012, http://www.opendataimpacts.net/engagement/




Source: Central data catalogue of the Government of the United Kingdom, http://data.gov.uk/

**European Commission Open Data Portal**. In December 2012 the European Commission has launched its open data portal, as announced in December 2011.[10] At launch the EC data portal provides access to more than 5800 datasets. While the majority of datasets comes from Eurostat, it also includes datasets from the European Environment Agency, the Joint Research Centre and the Directorate-General for Health and Consumers, Education and Culture and other Union institutions, bodies, offices and agencies at their request. http://open-data.europa.eu/

10 http://epsiplatform.eu/content/open-data-strategy-europe-here

Source: European Commission Open Data Portal, http://open-data.europa.eu/

**Need for a pan-European Catalogue**

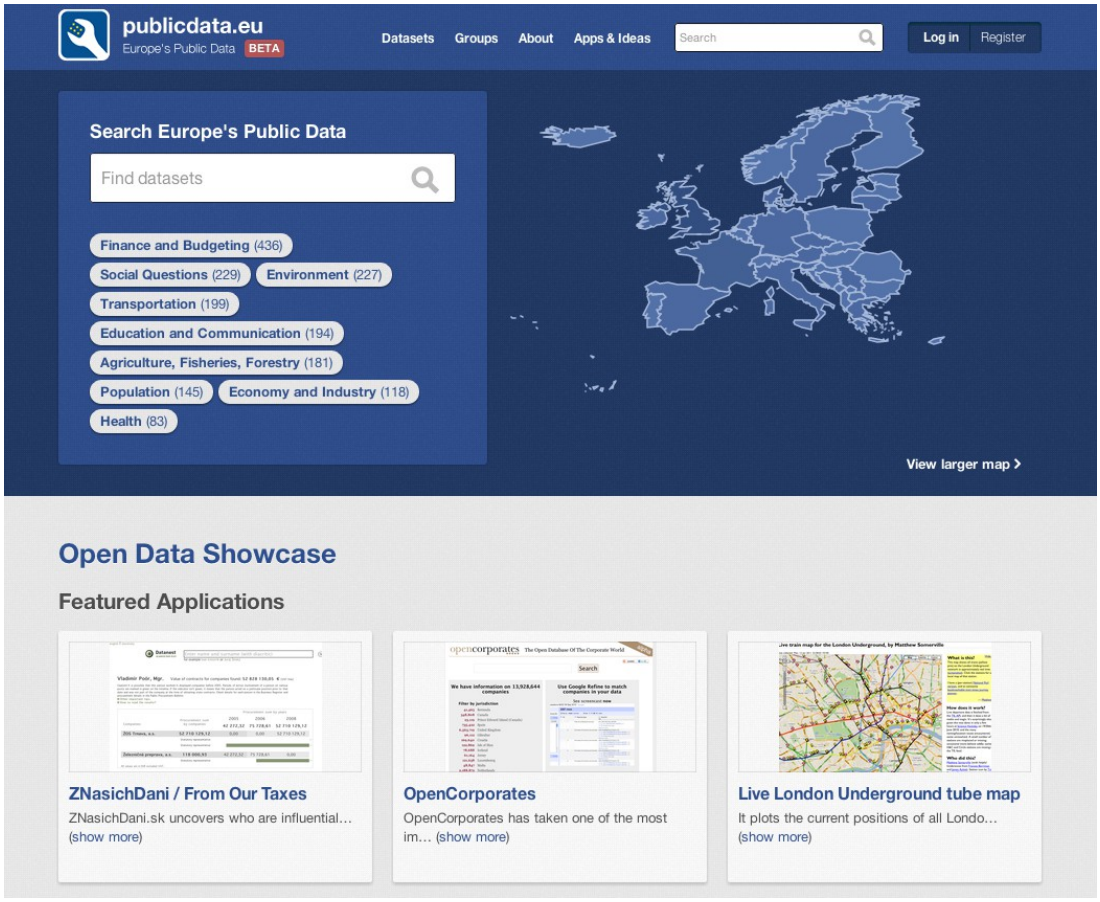
Information about European public datasets is currently scattered across many different data catalogues, portals and websites in many different languages, implemented using many different technologies. The kinds of information stored about public datasets may vary from country to country, and from catalogue to catalogue. To further simplify access to all European PSI the European Commission has announced the development and implementation of a central pan-European data catalogue that will aggregate huge amounts of datasets from EU Member States data catalogues.[11]

**Public Data EU** can be seen as a prototype and proof-of-concept for such an integrated pan-European data catalogue. It was developed as part of the LOD2 project in 2011.[12] Public Data EU aims to provide a single point of access to open, freely reusable datasets from numerous national, regional and local public bodies throughout Europe. The site harvests and federates this data to enable users to search, query and process the data from a single place. This helps to solve the "discoverability problem" of finding interesting data

11 See also Nigel Shadbolt, Towards a pan EU data portal – data.gov.eu, 2010, http://ec.europa.eu/information_society/policy/psi/docs/pdfs/towards_an_eu_psi_portals_v4_final.pdf

12 LOD2 is an ICT research project financed under the European Commission's 7th Framework Programme. http://lod2.eu/

across many different government websites, at many different levels of government, and across the many governments in Europe. http://publicdata.eu/

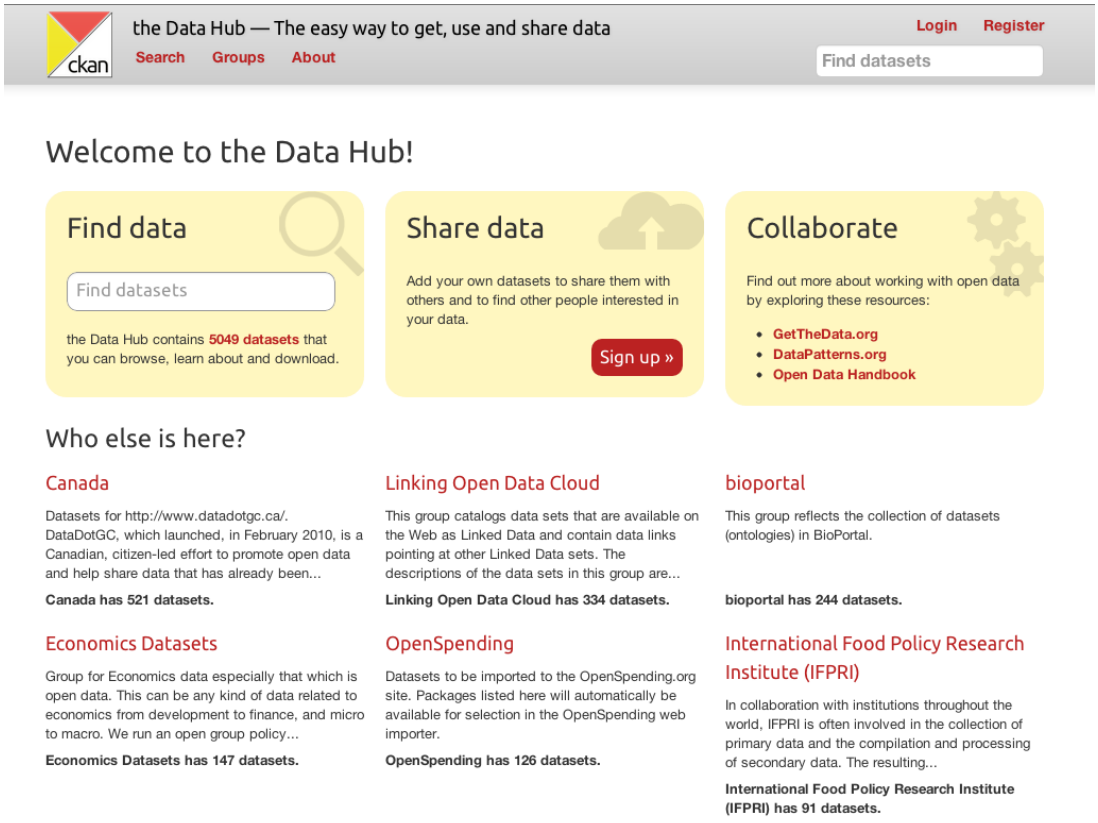



Source: Public Data EU, http://publicdata.eu/

# 5 Third-party platforms

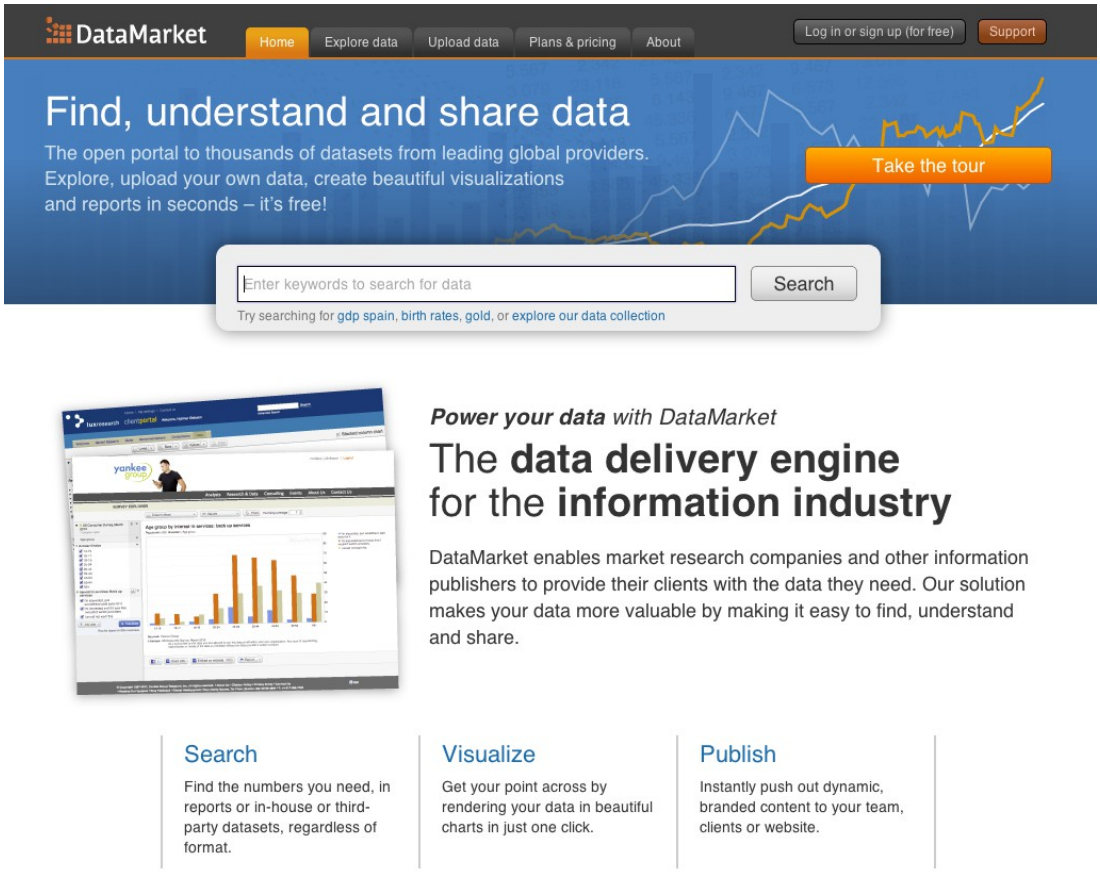
There are many platforms and public repositories that can be used to publish, share and access data on the Internet. For reasons of scope this report can only showcase some examples.

**The Data Hub** is a community-driven catalogue of useful sets of data on the Internet. The free service allows everyone to collect links to data from around the web or search for data that others have collected. The platform offers publication and management of metadata, search and an API. Depending on the type of data (and its conditions for reuse), the Data Hub may also store a copy of the data or host it in a database, and provide basic visualisation tools: http://datahub.io/




Source: the Data Hub, http://datahub.io

**DataMarket** is a data portal that provides access to thousands of data sets holding hundreds of millions of facts and figures from a wide range of public and private data providers including the United Nations, the World Bank, Eurostat, Economist Intelligence Unit and others. The portal allows all this data to be searched, visualized, compared and downloaded in a single place in a standard, unified manner. DataMarket's data publishing solutions allow data providers such as market research companies, financial institutions and analytics firms to easily publish their data on www.datamarket.com and on their existing websites through embedded content and branded versions of DataMarket's systems, enabling all the functionality of DataMarket on top of their own data collections. http://datamarket.com/




Source: Data Market, http://datamarket.com/

# 6 Technology solutions for data catalogues

There is a plenitude of open source and proprietary software solutions for data catalogues available on the market. While most of them are open source, there are also proprietary platforms and tools. For reasons of scope this report can only showcase some examples.

**Socrata** is a powerful set of software components that allows easy publishing and managing of data on the Internet. It offers advanced features like an API and visualizations as well as support and hosting solutions in the cloud with SLA. http://www.socrata.com/

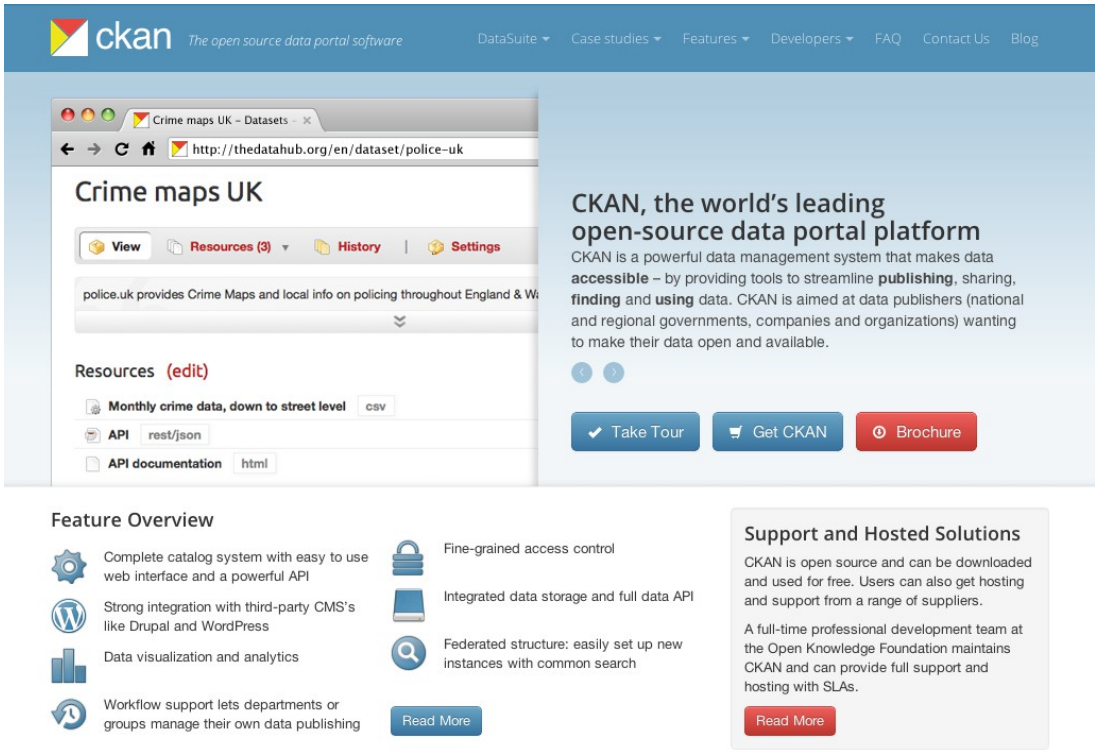
Source: Socrata, http://www.socrata.com/

**GeoNetwork** is a catalogue application to manage spatially referenced resources. It provides metadata editing and search functions as well as an embedded interactive web map viewer. It is currently used in numerous spatial data infrastructure initiatives across the world. The open source software provides an easy to use web interface to search geospatial data across multiple catalogues, combine distributed map services in the embedded map viewer, publish geospatial data using the online metadata editing tools and optionally the embedded GeoServer map server.[13] Administrators have the option to manage user and group accounts, configure the server through web based and desktop utilities and schedule metadata harvesting from other catalogues. http://geonetwork-opensource.org/

13 GeoServer map server, http://geoserver.org/

Source: Geo Network Opensource, http://geonetwork-opensource.org/

**CKAN** is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available. CKAN is open source and can be downloaded and used for free. Users can also get hosting and support from a range of suppliers. A full-time professional development team at the Open Knowledge Foundation maintains CKAN and can provide full support and hosting with SLAs. http://ckan.org/




Source: CKAN, http://ckan.org/

# 7 Data Catalogues and Data Usage

It is evident that making PSI accessible for reuse by setting up data catalogues is a key element of any OGD initiative. However one should not think that the story ends there. Publishing data into a central catalogue is just the first step of what is sometimes referred to as "building a sustainable open data ecosystem".[14] The success of a government program to make data available for reuse is that the data actually gets reused. But how to grow and sustain a community of commercial and non-commercial reusers? How to help citizens, civil society organisations, businesses, researchers and journalists to engage with open government data? One should not think of opening government data and accelerate its reuse as a technical problem that can be solved with the implementation of a technical infrastructure (e.g. providing access to the data in a data catalogue). In fact it is likely that reuse and engagement will not uptake automatically but need to be stimulated, promoted and supported with multiple measures.

**Measuring Impact and success**

Little evidence-based research is available on the actual impact of OGD initiatives. While some organizations have started to elaborate criteria to evaluate and compare the impacts of OGD initiatives[15], there is consensus within the research and Open Data communities that more research is needed to better understand the parameters and environmental

14 See Building the (Open) Data Ecosystem, Rufus Pollock, 2011, http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/

15 See the Open Data Research network, focussed on the impacts of open data in developing countries, http://www.opendataresearch.org/ and the

conditions that affect OGD initiatives. The same applies to data catalogues as key part of such initiatives.[16]

The Open Data Study[17] analyses the country specific conditions and the approach by both the US and the UK administration for their OGD initiatives (with the open data catalogues www.data.gov and www.data.gov.uk respectively). The report finds that in both the US and UK, a "three-tiered" drive was at play. The three groups of actors who were crucial to the projects' success were:

- *Civil society, and in particular a small and motivated group of 'civic hackers'*
- *An engaged and well-resourced 'middle layer' of skilled government bureaucrats; and*
- *A 'top-level' mandate, motivated by either an outside force (in the case of the U.K.) or a refreshed political administration hungry for change (in the U.S.).*

In an interview for the Open Data Study, Tim Berners-Lee observed, "*It has to start at the top, it has to start in the middle, and it has to start at the bottom.*" The conclusion to this report strengthens that assertion, and warns those attempting to mirror the successes of the UK and US projects not to neglect any of these three layers of influence. Other studies look into the feasibility of transferring such concepts and approaches from one country to another and especially from high-income to low and middle-income countries.[18] The bottom line is that beyond setting up the technical infrastructure of a data catalogue, governments need to be clear about the fact that other supporting measures and activities need to be taken in order to establish the data catalogue as part of a sustainable open data initiative of a country.

# 8 Conclusion

Open Data catalogues are one aspect in facilitating access to and reuse of PSI. They aim help citizens and business to identify what type of information and data exists and which public authority holds it. They can improve the discoverability of PSI and thus reduce the transaction-costs to a great extend. A number of countries, regions and municipalities have therefore created dedicated open data catalogues.

But open data catalogues can be much more than just directories or repositories – when implemented and managed successfully, they can also form the centrepiece for the community that generates value from publicly released open government data. The community around a catalogue is a direct contributor to the success of the OGD initiative. This community ideally consists both people inside government (data producers) and

16 See also Fournier-Tombs, Eleonore, Evaluating the Impact of Open Data Websites, 2011. Available at SSRN: http://ssrn.com/abstract=1926201 or http://dx.doi.org/10.2139/ssrn.1926201

17 Becky Hogge. Open Data Study, 2011, http://www.transparency-initiative.org/reports/open-data-study-new-technologies for the two following citations see page 4 and page 10.

18 Web Foundation, Open Government Data Feasibility Studies, 2012, http://www.webfoundation.org/projects/open-government-data-feasibility-studies/

outside (data consumers – developers, journalists, researchers, civic activists, etc.) and establishing a dialogue between these groups and around the data and its reuse. Growing and nursing this community should be a key element of OGD initiatives.

Understanding this approach to open data catalogues underlines some important considerations government officials should keep in mind when evaluating different approaches for an open data catalogue, and also highlights work that must be done beyond the publication of open data in a data catalogue to ensure the success of government transparency efforts and the increased reuse of PSI for innovations.

## About the Author

Daniel Dietrich was born in 1973 in Frankfurt, Germany. His academic work covers political science, computer science and communication science in Frankfurt and Berlin. He worked as Research Associate at Technical University Berlin, Department of Internet and Society until the end of 2011. He has been working for the Open Knowledge Foundation (OKFN), since 2009 and is Chairman of the German Chapter of the Open Knowledge Foundation. He is the Project Coordinator for the OKF Project Open Definition as well as the Coordinator of the Working Group on Open Government Data and the Working Group on Open Data in the EU. He is the co-founder of the Open Data Network, a non-profit advocacy organisation to promote Open Data, Open Government and Transparency in Germany, Europe and beyond. In 2011 he became Editor of the ePSI platform.

## Copyright information

© 2011 European PSI Platform - This document and all material therein has been compiled with great care. However, the author, editor and/or publisher and/or any party within the European PSI Platform or its predecessor projects the ePSIplus Network project or ePSINet consortium cannot be held liable in any way for the consequences of using the content of this document and/or any material referenced therein. This report has been published under the auspices of the European Public Sector Information Platform.

The report may be reproduced providing acknowledgement is made to the European Public Sector Information (PSI) Platform. The European Public Sector Information (PSI) Platform is funded under the European Commission eContentplus programme.