Harvesting dataset records into data.gov.uk

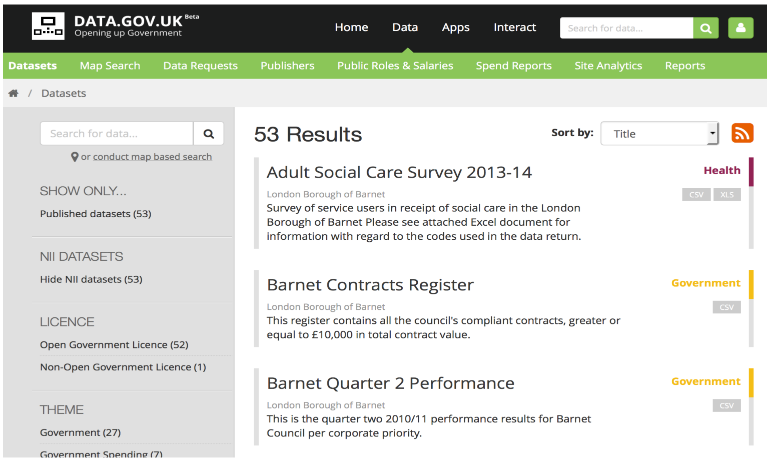
Technical Guide

Version: 1.2 – 17th Dec 2014
Author: David Read

Table of Contents

Introduction

This guide explains how a public body can transfer its dataset listings in bulk into data.gov.uk. The dataset records (metadata) are listed by the public body on their open data website and they then operate data.gov.uk's “harvester” to transfer them into data.gov.uk.



53 Results

Adult Social Care Survey 2013-14
London Borough of Barnet
Survey of service users in receipt of social care in the London Borough of Barnet Please see attached Excel document for information with regard to the codes used in the data return.
Health
CSV XLS

Barnet Contracts Register
London Borough of Barnet
This register contains all the council's compliant contracts, greater or equal to £10,000 in total contract value.
Government
CSV

Barnet Quarter 2 Performance
London Borough of Barnet
This is the quarter two 2010/11 performance results for Barnet Council per corporate priority.
Government
CSV

Public bodies have long been able to add datasets one at a time to data.gov.uk by using the web form[1]. In addition, bulk addition has been possible for location/INSPIRE data, by using the data.gov.uk harvester to collect datasets stored in the public body's GIS system. In November 2014 the data.gov.uk harvester has been extended to work for non-location datasets. This means that sets of data records (sometimes called 'inventories') can be added to data.gov.uk in bulk.

To work with the harvester, a public body's set of records are published on the internet in any one of the recognized formats. These formats represent the most common ones produced by 'open data websites' such as CKAN, DKAN, Socrata or DataShare.

The launch of the new harvesters are timed to support the Local Authorities to publish their metadata on data.gov.uk about the datasets that local authorities are required to publish locally due to the Local Government (Transparency Requirements) Regulations[2]. Local authorities are also encouraged to publish their open datasets on data.gov.uk through the Local Open Data Incentive Scheme[3]. Through a local government sector led approach, additional fields were requested and added by the LGA, such as function & service categories - these are imported by all of these harvesters. data.gov.uk is being enhanced to make good use of this extra metadata, to aid navigation by dataset type and automatically validate data against schemas.

NB The new harvesters and this guide are only for datasets that are **not** covered by the INSPIRE legislation. For more about this, see Appendix A – Location/INSPIRE datasets.

[1] See: http://data.gov.uk/library/user-guide-editors-and-administrators
[2] http://www.legislation.gov.uk/uksi/2014/2680/introduction/made
[3] http://incentive.opendata.esd.org.uk/

How to Publish Datasets

A public body may or may not have a open data web site (see Appendix C - Rationale for Using a Data Portal), but to be harvested into data.gov.uk it needs to ensure it is published in a suitable machine-readable way.

Each dataset should be described in a “metadata record”, giving details about the dataset, such as: title, description and web link to download the data file. The metadata records need to be published:

* in a machine-readable format that data.gov.uk recognizes - see section "Supported Formats"

* in a suitable place on the public Internet - probably the body's website or data portal

Supported Formats

Harvester	Suitability	Example Harvest URL
DCAT	Triple-stores	http://opendatacommunities.org/data.ttl
CKAN	CKAN	https://open.barnet.gov.uk/
DKAN	DKAN	http://opendata.cambridgeshireinsight.org.uk/
Inventory	DataShare	http://data.bracknell-forest.gov.uk/api/esdInventory
data.json	Socrata, custom systems	https://nycopendata.socrata.com/data.json

DCAT

DCAT has the major advantage that it was designed in an open process by practitioners from around the world of data publishing and has become a W3C recommendation. It has been selected by the EU for the harvesting and republication of data from the 28 EU national data portals including data.gov.uk – the EU-wide portal is in beta and will launch in Autumn 2015. DCAT is also supported by the majority of data portals around the world, including the national government portals in US, Canada, Australia and the large numbers of others that use CKAN and other software.

DCAT is a 'vocabulary' in the RDF Linked Data world, and it is normal to use fields/predicates from other vocabularies where appropriate. Whilst this adds expressiveness, data.gov.uk cannot read every eventuality, so this document publishes the fields/predicates that data.gov.uk expects in Appendix A – DCAT/data.json fields.

DCAT can be seen as somewhat verbose. Whilst it is ideal for linked data systems, it is often preferable to use a simplified version called data.json. The data.json format has the fields we need from DCAT, but removes the namespace prefixes and uses the well-known JSON syntax. This has the benefits of DCAT but is generally much more easily produced – see the section on data.json.

Although CKAN supports DCAT for the core fields, it is recommended to harvest from a CKAN using the CKAN harvester. This is because custom fields often do not map well to DCAT fields and can vary from portal to portal.

The DCAT harvester needs to be given the URL that returns the RDF for all the datasets. Optionally the datasets can be split into a number of pages, accessed using the 'page' parameter. i.e. page 2 would be accessed by appending to the URL: ?page=2

data.json

The data.json format was designed as having the same fields as DCAT, but expressed more simply. It is used extensively for harvesting the American public bodies into data.gov and is gaining popularity elsewhere.

The fields are documented in Appendix A.

The US government provide various tools http://project-open-data.github.io/ and full details of their implementation (it is similar to the UK definitions): http://project-open-data.github.io/v1.1/metadata-resources/

The data.json harvester needs to be given the URL that returns a JSON list containing the datasets. Optionally the datasets can be split into a number of pages, accessed using the 'page' parameter. i.e. page 2 would be accessed by appending to the URL: ?page=2

CKAN

The CKAN software is the most popular data portal software, powering the largest government portals in UK (including data.gov.uk), USA, Canada, Australia and much of Europe. It is open source, meaning there is no organization that can control functionality or usage. Numerous organizations and contractors offer services to administer, customize or host CKAN. The CKAN Association founded in 2014 provides a world-wide community.

Although most of the fields are core to all CKAN sites, it is common to customize fields. data.gov.uk has some customizations, so some translation of records occurs during harvest, and optionally extra metadata fields can be supplied that data.gov.uk will display. The fields are detailed in Appendix B.

The CKAN harvester needs the URL of the CKAN home page, from where it can find its API functions.

DKAN

DKAN is developed by a US-based consultancy and gives basic CKAN functionality. Although it aims for compatibility with CKAN APIs, as it stands (October 2014) the normal APIs for harvesting are not there, and some fields are expressed differently, so you'll need to use this custom DKAN harvester.

Follow the CKAN field guidance.

The DKAN harvester needs to be given the URL of the DKAN home page, from where it can find its API functions.

Inventory

Developed for the LGA, this format is only suitable for local authority data and is implemented by DataShare. An XML schema is supplied by ESD for this harvester which checks the elements and basic types before being accepted by the harvester.

The CKAN harvester needs the full URL to the inventory XML file.

Full guidance for this format is here: http://schemas.esd.org.uk/inventory/InventoryGuidance.pdf

The schema is here: http://schemas.opendata.esd.org.uk/Inventory

How to Harvest

NB Steps 1 to 3 are covered in more detail in the User Guide for Editors and Administrators: http://data.gov.uk/library/user-guide-editors-and-administrators

1. Ensure your organization is listed at http://data.gov.uk/publisher

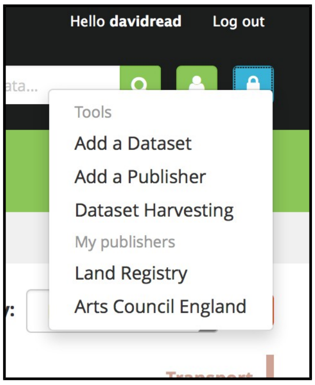
If it is not, then you'll need to request that it is added by using the contact form: http://data.gov.uk/contact

2. Ensure you have a user account on data.gov.uk

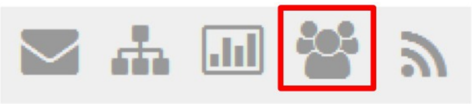
To register for an account, click on the green 'person' icon at the top-right or go to: http://data.gov.uk/user?destination=contact

3. Ensure your user account has been assigned permission for your organization, as either an admin or editor.

If you have this admin/editor permission then when you log-in you'll see a blue padlock icon and your organization listed underneath when you click it. For example in this screenshot, this user has admin or editor permission for Land Registry and the Arts Council:



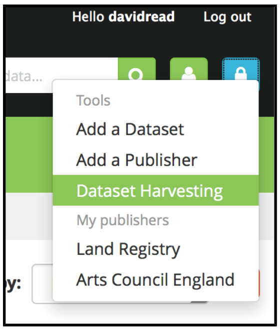
Permission can be granted by any existing admin for your organization, or an admin of any of its parent organizations (as shown in the tree at http://data.gov.uk/publisher). To request permission: log in, go to http://data.gov.uk/publisher , find your organization and go to its main page by clicking it.



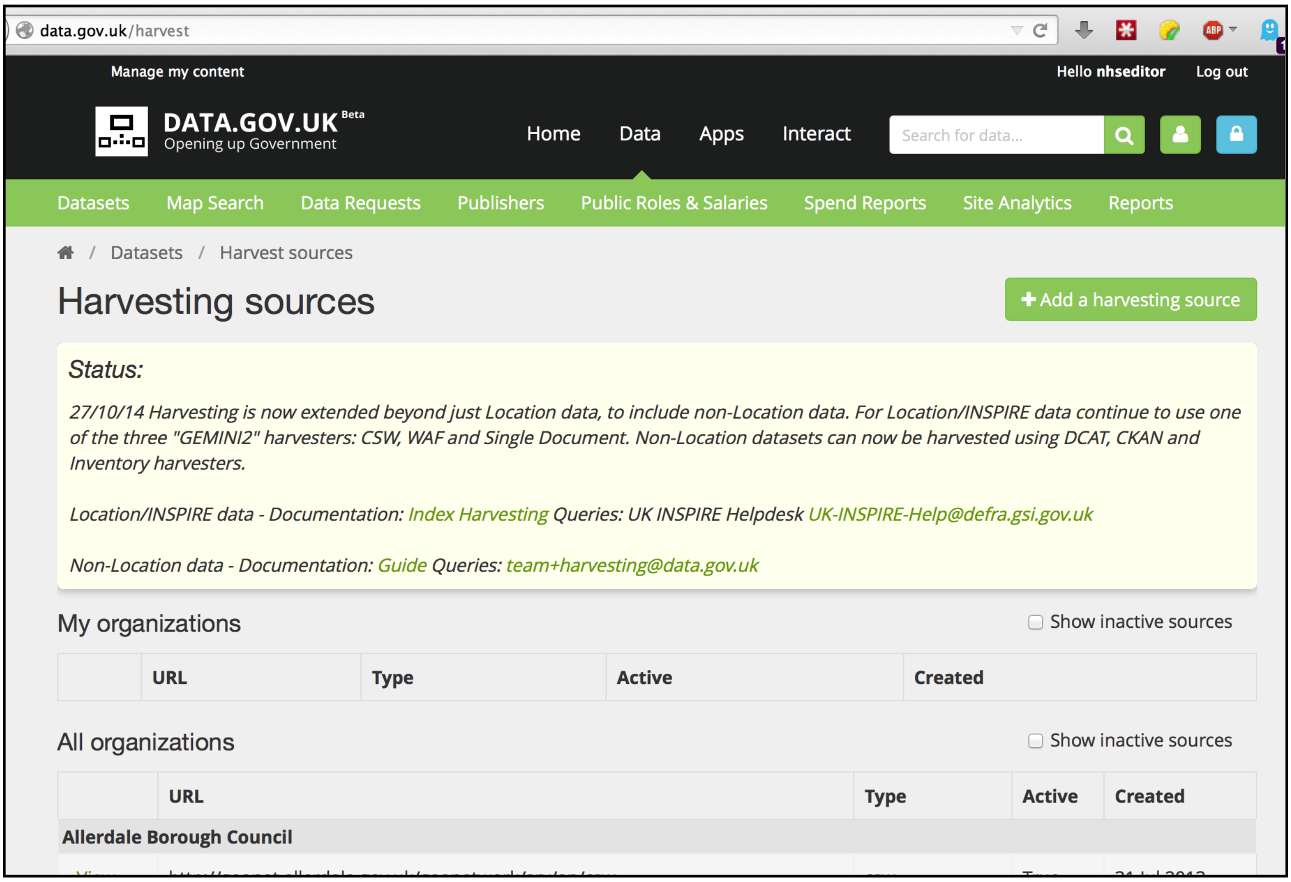
Click on the administrators icon on the icon bar on the top right of the page and select ‘request to become an editor’ and fill out the form. You should be notified by the publisher when you have been given publisher rights. If you don't hear back for a couple of days and permission has not be granted yet, contact the data.gov.uk team: http://data.gov.uk/contact

4. Create the harvester

When logged in as an editor/admin you'll see the blue padlock icon. Click on that and select Dataset Harvesting.



Now you should see the harvesting dashboard at http://data.gov.uk/harvest :



Harvesting sources

Status:

27/10/14 Harvesting is now extended beyond just Location data, to include non-Location data. For Location/INSPIRE data continue to use one of the three "GEMINI2" harvesters: CSW, WAF and Single Document. Non-Location datasets can now be harvested using DCAT, CKAN and Inventory harvesters.

Location/INSPIRE data - Documentation: Index Harvesting Queries: UK INSPIRE Helpdesk UK-INSPIRE-Help@defra.gsi.gov.uk

Non-Location data - Documentation: Guide Queries: team+harvesting@data.gov.uk

My organizations

□ Show inactive sources

	URL	Type	Active	Created

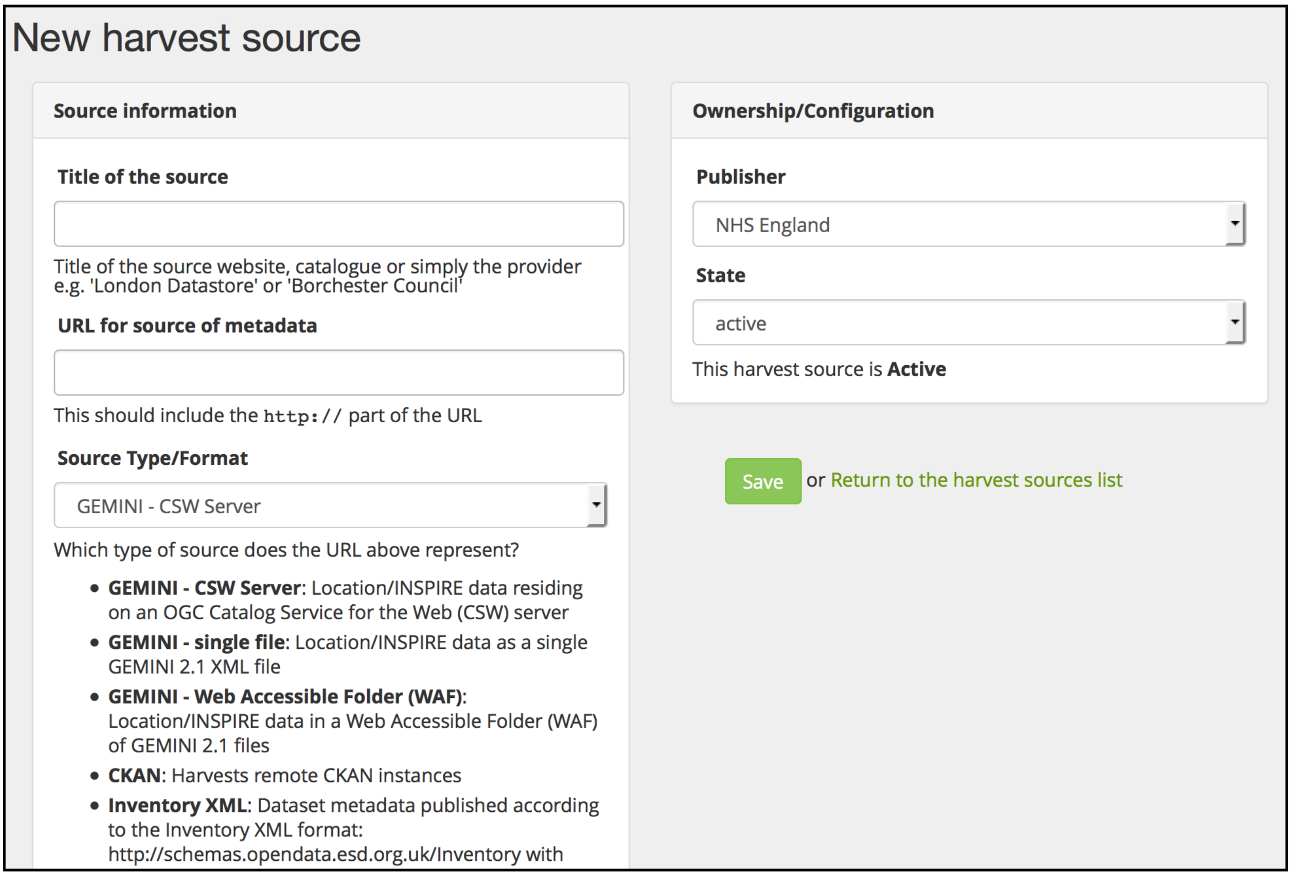
All organizations

□ Show inactive sources

	URL	Type	Active	Created
Allerdale Borough Council				

Note that the harvesters are public, but because you are logged in as an editor/admin you have the ability to create a harvester and see the status of your harvesters.

To create the harvester click "Add a harvesting source" and fill out the form and click 'Save':



New harvest source

Source information

Title of the source

Title of the source website, catalogue or simply the provider e.g. 'London Datastore' or 'Borchester Council'

URL for source of metadata

This should include the `http://` part of the URL

Source Type/Format

GEMINI - CSW Server

Which type of source does the URL above represent?

- **GEMINI - CSW Server**: Location/INSPIRE data residing on an OGC Catalog Service for the Web (CSW) server
- **GEMINI - single file**: Location/INSPIRE data as a single GEMINI 2.1 XML file
- **GEMINI - Web Accessible Folder (WAF)**: Location/INSPIRE data in a Web Accessible Folder (WAF) of GEMINI 2.1 files
- **CKAN**: Harvests remote CKAN instances
- **Inventory XML**: Dataset metadata published according to the Inventory XML format: http://schemas.opendata.esd.org.uk/Inventory with

Ownership/Configuration

Publisher

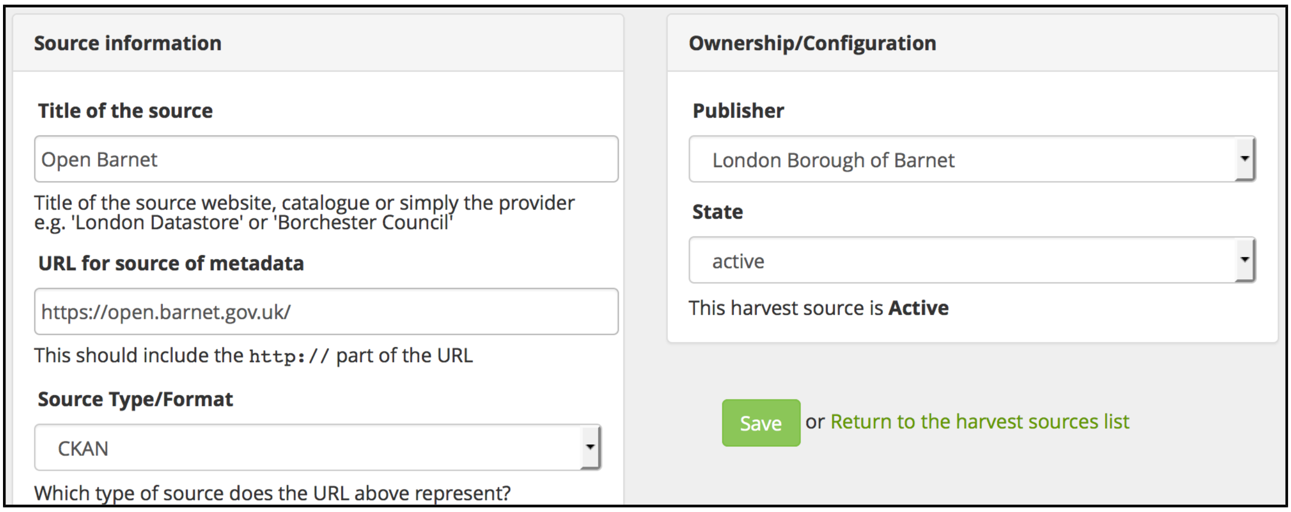
NHS England

State

active

This harvest source is **Active**

Here is an example for a harvesting CKAN server:



Source information

Title of the source

Open Barnet

Title of the source website, catalogue or simply the provider e.g. 'London Datastore' or 'Borchester Council'

URL for source of metadata

https://open.barnet.gov.uk/

This should include the `http://` part of the URL

Source Type/Format

CKAN

Which type of source does the URL above represent?

Ownership/Configuration

Publisher

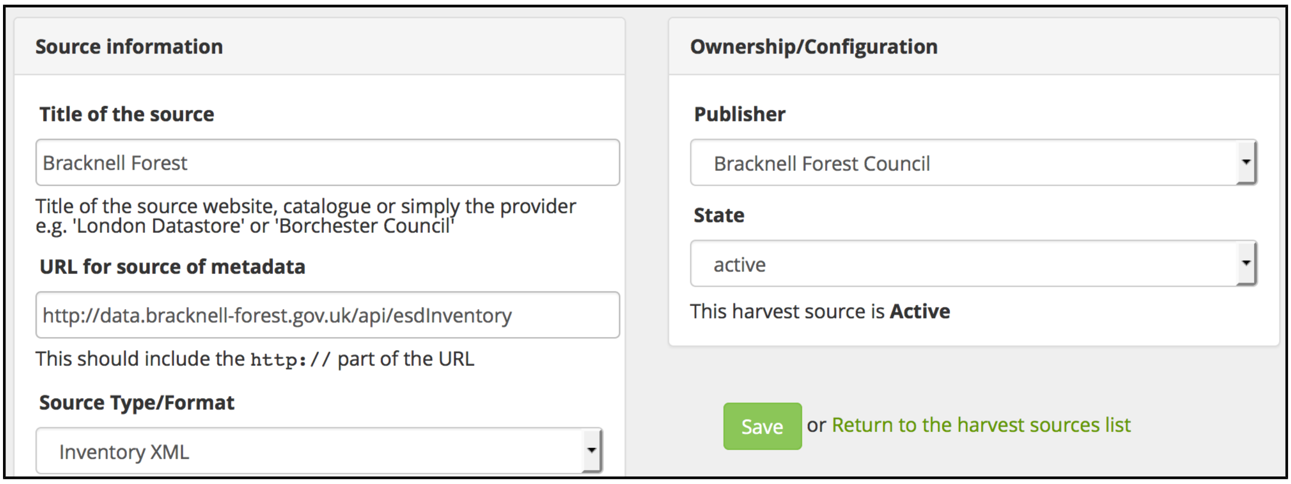
London Borough of Barnet

State

active

This harvest source is **Active**

And here is an example for harvesting datasets from a DataShare that publishes Inventory XML:



Source information

Title of the source

Bracknell Forest

Title of the source website, catalogue or simply the provider e.g. 'London Datastore' or 'Borchester Council'

URL for source of metadata

http://data.bracknell-forest.gov.uk/api/esdInventory

This should include the `http://` part of the URL

Source Type/Format

Inventory XML

Ownership/Configuration

Publisher

Bracknell Forest Council

State

active

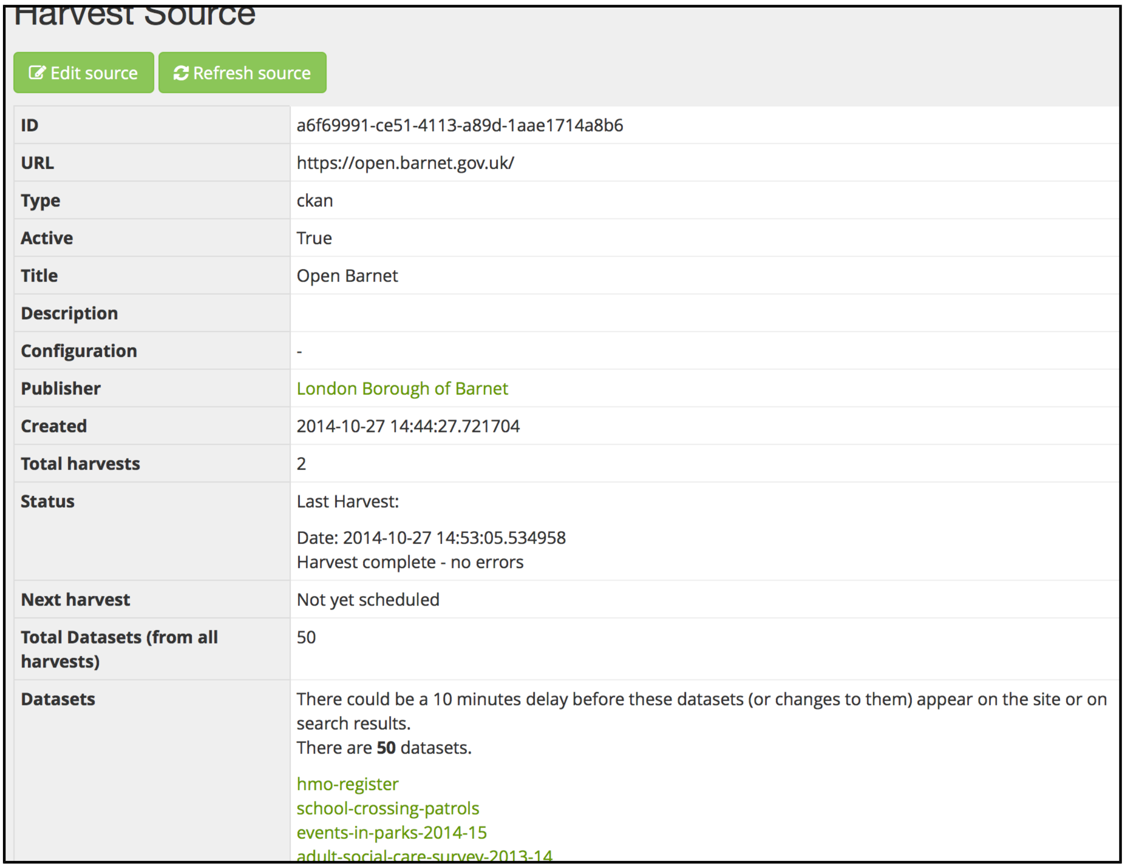
This harvest source is **Active**

When you click Save it will show you your harvester's details and tell you that the harvest has now been requested.

5. Wait for the harvest and check the results.

Harvests are started every 10 minutes, and take a few minutes more to complete. If you refresh the harvester's page it will tell you in the 'Status' and 'Last harvest' fields if the harvest is "scheduled" (i.e. waiting to start), "in progress" or when its complete it shows you the results of the "Last Harvest". (If you are not logged-in then status is not shown).

A successful harvest will look like this:



ID	a6f69991-ce51-4113-a89d-1aae1714a8b6
URL	https://open.barnet.gov.uk/
Type	ckan
Active	True
Title	Open Barnet
Description	
Configuration	-
Publisher	London Borough of Barnet
Created	2014-10-27 14:44:27.721704
Total harvests	2
Status	Last Harvest: Date: 2014-10-27 14:53:05.534958 Harvest complete - no errors
Next harvest	Not yet scheduled
Total Datasets (from all harvests)	50
Datasets	There could be a 10 minutes delay before these datasets (or changes to them) appear on the site or on search results. There are **50** datasets. hmo-register school-crossing-patrols events-in-parks-2014-15 adult-social-care-survey-2013-14

If you do get errors listed, then consult the section later on called "Harvest Errors".

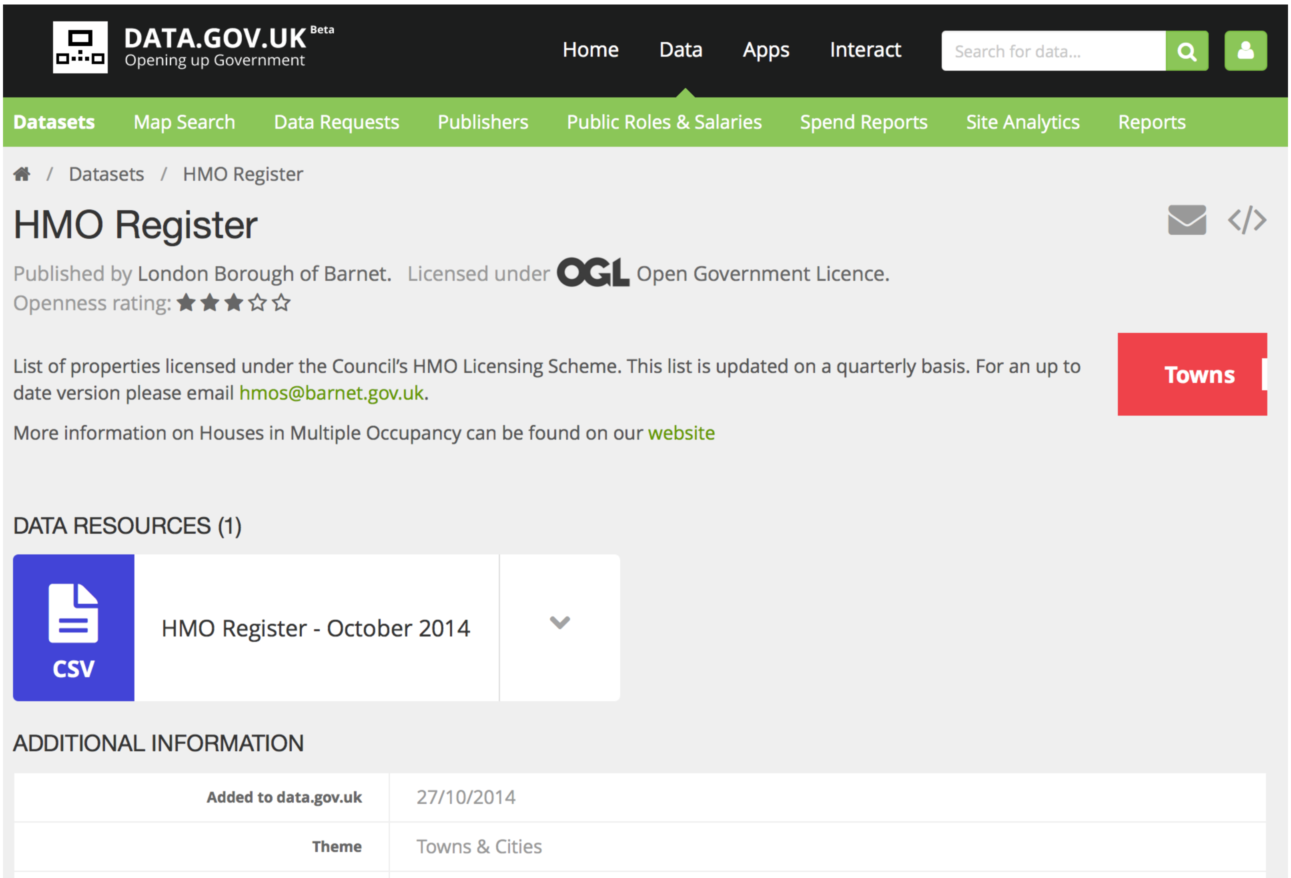
6. Harvest again when your datasets need updating

You can do another harvest by clicking the "Refresh source" button on the harvester page (remember to be logged-in to be able to see the button)

data.gov.uk will shortly add a feature to automatically schedule harvests on a weekly basis.

Harvested datasets

A harvested dataset looks mostly like other datasets on data.gov.uk:



HMO Register

Published by London Borough of Barnet. Licensed under OGL Open Government Licence.
Openness rating: ★★★☆☆

List of properties licensed under the Council's HMO Licensing Scheme. This list is updated on a quarterly basis. For an up to date version please email hmos@barnet.gov.uk.

More information on Houses in Multiple Occupancy can be found on our website

DATA RESOURCES (1)

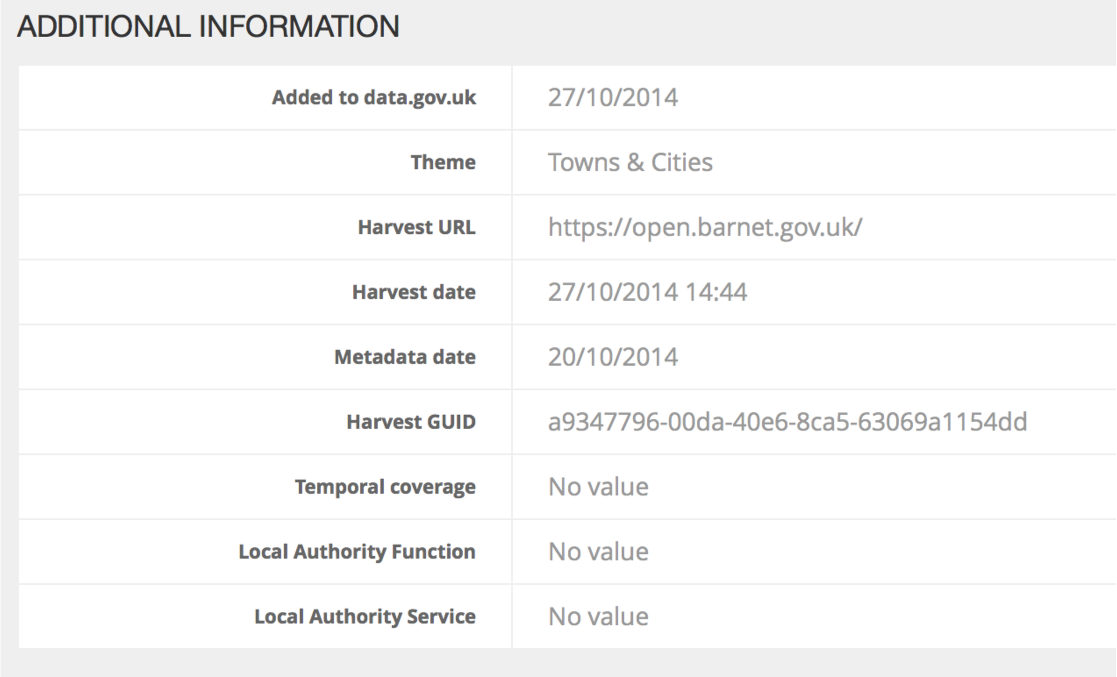
HMO Register - October 2014

ADDITIONAL INFORMATION

Added to data.gov.uk	27/10/2014
Theme	Towns & Cities

Details of the harvest

When someone views a dataset that have been harvested, they see a few extra fields listed towards the bottom of the page:



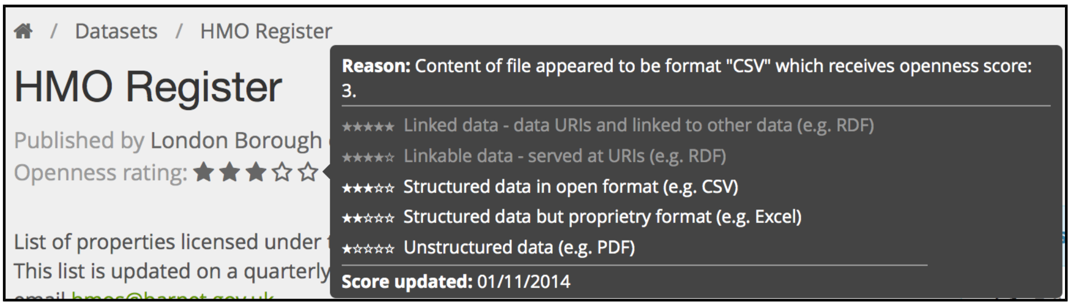
ADDITIONAL INFORMATION

Added to data.gov.uk	27/10/2014
Theme	Towns & Cities
Harvest URL	https://open.barnet.gov.uk/
Harvest date	27/10/2014 14:44
Metadata date	20/10/2014
Harvest GUID	a9347796-00da-40e6-8ca5-63069a1154dd
Temporal coverage	No value
Local Authority Function	No value
Local Authority Service	No value

The harvester has added: "Harvest URL", "Harvest date" and "Harvest GUID". The GUID is the dataset's ID as it was in the harvested system e.g. the CKAN "id", DCAT URI or the Inventory "inv:Identifier".

Not all of the publishing sources or harvesters cover all the fields, so for example Temporal Coverage operates differently in the Inventory format, and the DCAT harvester doesn't include it yet. We hope to fill in these gaps in time.

Five Stars of Openness



HMO Register

Published by London Borough

Openness rating: ★★★☆☆

List of properties licensed under

This list is updated on a quarterly

email hmos@barnet.gov.uk

Reason: Content of file appeared to be format "CSV" which receives openness score: 3.

★★★★★ Linked data - data URIs and linked to other data (e.g. RDF)

★★★★☆ Linkable data - served at URIs (e.g. RDF)

★★★☆☆ Structured data in open format (e.g. CSV)

★★☆☆☆ Structured data but proprietry format (e.g. Excel)

★☆☆☆☆ Unstructured data (e.g. PDF)

Score updated: 01/11/2014

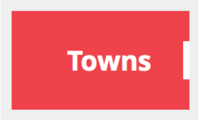
The 'Openness rating' is automatically calculated for a dataset shortly after harvest. It checks if the licence is open, if the resource links work ok and if the resource formats are open and linked. The resulting score is between 0 and 5 stars. If you hover the mouse pointer over the stars then an explanation of the scoring is provided. The overall score is the highest of all the resources, so if you have 3 PDFs (score 1) and a CSV (score 3), then the overall score is 3. There is more about the scheme at: http://5stardata.info/

Often this is a useful driver to improve the data or the metadata. For example releasing data as CSV instead of Excel files gets you to 3 stars instead of 2 (although it is good to provide both if you have them).

Also often this indicates problems with the dataset record - maybe the URL to the data was mistyped, or only goes to a web page about the data. Or maybe the harvester has not transferred the details accurately from your site - contact us in that instance. These are easily fixed and then you reharvest.

Occasionally the automatic calculation is wrong - it misidentified the format of your file. Do contact us and we can try and improve it.

Theme


Towns

The 'theme' of a dataset is automatically set on harvest. It is decided by data.gov.uk, based on the title, description and tags of a dataset. Local Authority datasets that have their service or function fields filled in will have their theme decided mainly on those.

This automatic system is generally a lot more consistent than asking humans to categorize datasets(!) And it makes it easy to recategorize them when the themes change as they occasionally do.

There are plenty of cases where a dataset could arguably be in another theme. Since there can often be disagreement about marginal topics, such as if 'planning permission' data should be in 'Environment' or 'Towns & Cities', or 'unemployment stats' goes in 'society' or 'economy'. We'll only consider changing a dataset's theme if it is completely wrong, such as if playground data ended up in 'Economy'. And even then, it would tend to be only we're generally only interested if there are a number of similar datasets. With 20000 datasets, getting a couple of lone datasets correctly categorized is not a priority. But if you think you have a case for a useful recategorization, do email us at team@data.gov.uk.

Harvest Errors

It is not unusual to see errors on the first harvest. Often it is simple to put right, either in the harvester configuration or it has revealed problems with how the datasets are published. Here are some error messages and tips for solving them:

Unable to get content for URL
There was some sort of connection error contacting the publishing site, or the publishing site gave an error for this URL. Check you've configured the harvester with the correct type and URL to go with that type. Some harvesters need the home page of the site, and some need a specific sub-url. Very occasionally these errors are just down to the internet being unreliable.

Failed to parse or validate the XML document
There is something wrong with the structure of the published Inventory content. Check that the URL is actually an Inventory XML file. Check it validates against the Inventory schema. Check the version of the schema is the same as the one data.gov.uk uses, which is given on the harvester edit page.

System error / Validation Error
Something has unexpectedly gone wrong internally data.gov.uk. Please contact the team to fix the problem.

Hopefully the detail of the error messages and these hints provide enough information to solve the issues. If something is still not clear, or data.gov.uk is not working as it should, please get in touch by emailing: team+harvesting@data.gov.uk

Appendix A – DCAT/data.json fields

Dataset:

data.json field	**DCAT predicate**	**Example value**	**Comments**
title	dct:title rdfs:label	Spend over £500	Mandatory
description	dct:description rdfs:comment	Spend transactions published monthly according to the Treasury transparency guidelines.	Mandatory
identifier	(equivalent to RDF object's URI or) dct:identifier	https://data.some.org/catalog/datasets/9df8df51-63db-37a8-e044-0003ba9b0d98 or http://dx.doi.org/10.7927/H4PZ56R2	Mandatory. Must stay the same, even if the dataset's title changes. Must be globally unique - not just unique to the publisher. A URI is highly recommended (preferably an HTTP URL).
license	dct:license	http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/	Mandatory. Either a license URI or a title (must be exact or may not be recognized as open)
keyword	dcat:keyword	data.json: "keyword": {"geochemistry", "geology"} DCAT: <dcat:keyword>geochemistry</dcat:keyword> <dcat:keyword>geology</dcat:keyword>	Not displayed by data.gov.uk but helps categorization
issued	dct:issued	2012-05-10T21:04	Date of formal issuance.
modified	dct:modified	2012-05-10T21:04	
publisher	dct:publisher	data.json: "publisher": {"name": "Geological Society", "mbox": "info@gs.org"} DCAT: <dct:publisher> <foaf:Organization> <foaf:name>Geological Society</foaf:name> <foaf:mbox>info@gs.org </foaf:mbox> </foaf:Organization> </dct:publisher>	
distribution	dcat:distribution	see "Distribution" table below	
language	dct:language	en or http://id.loc.gov/vocabulary/iso639-1/en	Language of the data
frequency	dct:accrualPeriodicity	"R/P1Y" (=annual) "R/P1W" (=weekly)	The frequency at which dataset is published. Format: ISO 8601 Repeating Duration (or "irregular") See: https://project-open-data.cio.gov/iso8601_guidance/#accrualperiodicity
temporal	dct:temporal	2000-01-15/2010-01-15	The date period that the data applies to. Formatted as two ISO 8601 dates (or datetimes) separated by a slash.
spatial	dct:spatial	{\"type\":\"Polygon\",\"coordinates\": [[[2.072, 49.943],[2.072, 55.816], [-6.236, 55.816], [-6.236, 49.943], [2.072, 49.943]]]}	The geographic location that the data applies to. If not specified, then it is inherited from the dataset's publisher. Formatted as a GeoJSON point, bounding box or polygon.

theme	dcat:theme	• "http://eurovoc.europa.eu/209416" - Police - Eurovoc • "COFOG/03.1.0" Police services - COFOG • "http://id.esd.org.uk/function/20" - Police services - ESD	Main thematic category of the dataset. Preferably expressed as a URI from a known vocabulary: • Eurovoc http://eurovoc.europa.eu/ • COFOG https://github.com/datasets/cofog/blob/master/data/cofog.csv • ESD Service/Function http://standards.esd.org.uk/?uri=list%2Fservices & http://standards.esd.org.uk/?uri=list%2Ffunctions More than one can be specified using a [] list.

Distribution:

downloadURL	dcat:downloadURL	http://site.gov.uk/river-levels/dec2012.csv	The direct URL that downloads a file
accessURL	dcat:accessURL	http://www.site.gov.uk/api/sparql http://site.gov.uk/river-level-data.html	If there is not a downloadURL, specify the accessURL, which is the URL of an API or web page about the data
title	dct:title	Spend transactions, Dec 2012	
description	dct:description		Not currently displayed on DGU
format	dcat:mediaType	text/csv	
conformsTo	dct:conformsTo	http://schemas.opendata.esd.org.uk/publictoilets/PublicToilets.json?v=0.41 or	URL of the machine-readable schema that the data conforms to. See: Appendix E - Local Authority data schemas
temporal	dct:temporal	2000-01-15/2010-01-15	The date period that the data applies to. Formatted as two ISO 8601 dates (or datetimes) separated by a slash.
spatial	dct:spatial	{\"type\":\"Polygon\",\"coordinates\": [[[2.072, 49.943],[2.072, 55.816], [-6.236, 55.816], [-6.236, 49.943], [2.072, 49.943]]]}	The geographic location that the data applies to. If not specified, then it is inherited from the dataset if not its publisher. Formatted as a GeoJSON point, bounding box or polygon.

Namespaces:

dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/
rdfs	http://www.w3.org/2000/01/rdf-schema#

Local Authorities should ensure they add an ESD service or function URI as the dataset's theme, to ensure good classification.

Comparison with USA data.json schema

The UK data.json format is based on the Project Open Data data.json schema used by the U.S. Federal Government and agencies:

https://project-open-data.cio.gov/v1.1/schema/

There are a few conscious differences that are listed here for reference:

Field	Change	Explanation
bureauCode, programCode, primaryITInvestmentUII, systemOfRecords, dataQuality	not required	They are codes specific to US Federal Government
temporal, spatial	Can be applied to not just a dataset but also distributions.	e.g. Spend data is split up by months
spatial	Formatted with GeoJSON	GeoJSON is analagous to GML but preferred. Place name strings are not preferred as they can be ambiguous.
theme	The values are URIs from known classification vocabularies, rather than simply strings.	A catalogue chooses its own classification vocabularies, so there is little value in simple strings determined by the data publisher. Strings related to the data's topic can go in the keywords field.

Appendix B – CKAN fields

Here is guidance for CKAN fields when harvesting into data.gov.uk:

Dataset fields - core

Key	**Value guidance**	**Example**
name	If there is a clash with an existing dataset in data.gov.uk then it will be modified to have a number suffixed, to make it unique.	river-levels-sussex
title	Name of the dataset. No jargon allowed - this is for general public to quickly understand what the data is about.	"River Levels in Sussex"
notes	Longer description of the data, how it was collected, etc.	Recorded daily from automated monitoring stations. Recorded at midday, published by 3pm the same day. Accurate to nearest centimetre.
tags	Optional. Key words/phrases that describe the subject of the data. Not displayed, but helps categorize it.	["transportation", "road safety", "accidents", "casualties"]
license_id	CKAN internal id for the data's licence, from CKAN's standard list of IDs. If the license isn't in CKAN's standard list then specify (in free-text) the name of the licence in the most standard way possible. If no standard licence is used then supply the full text of the terms of use.	uk-ogl or OS OpenData Licence
Ignored by DGU: owner_org, maintainer, maintainer_email, author, author_email, relationships	The owner_org/publisher is determined by the value given in the harvester's settings, not the harvested dataset.	

Dataset fields - extras

Key	**Value guidance**	**Example**
la_service	Service as listed here: http://standards.esd.org.uk/?uri=list%2Fservices More than one can be specified, separated by spaces.	http://id.esd.org.uk/service/190
la_function	Function as listed here:	http://id.esd.org.uk/service/437

	http://standards.esd.org.uk/?uri=list%2Ffunctions More than one can be specified, separated by spaces.	
contact-name, contact-phone, contact-email	Contact details for this dataset. Normally you shouldn't specify these - instead fill them in for the publisher in data.gov.uk and the publisher's datasets will have those details displayed.	"contact-name": "Information Handling Team", "contact-phone": "0208 445 3423", "contact-email": "informationhandling@dft.gsi.gov.uk"
foi-name, foi-phone, foi-email, foi-web	Freedom of Information (FOI) contact details for the dataset. Normally you shouldn't specify these - instead fill them in for the publisher in data.gov.uk and the publisher's datasets will have those details displayed.	
temporal_coverage-from, temporal_coverage-to	Gives the temporal coverage of the data. Each value can be YYYY, YYYY/MM or YYYY/MM/DD	"temporal_coverage-from": "1979-01-01", "temporal_coverage-to": "2013-12-31"
mandate	URL for the exact part of legislation that brought about the collection of this dataset (or if none, the public commitment or announcement, if available.)	http://www.legislation.gov.uk/ukpga/2014/26/section/168/enacted
update_frequency	Example values: never, discontinued, annual, quarterly, monthly	annual
unpublished	Allows publishers to report data that exists but is not published	false
geographic_coverage	No need to specify - a better field for spatial coverage will be coming soon.	
theme-primary, theme-secondary	Should not be included - data.gov.uk will automatically give it a theme based on keywords in the content.	

Dataset fields - resources

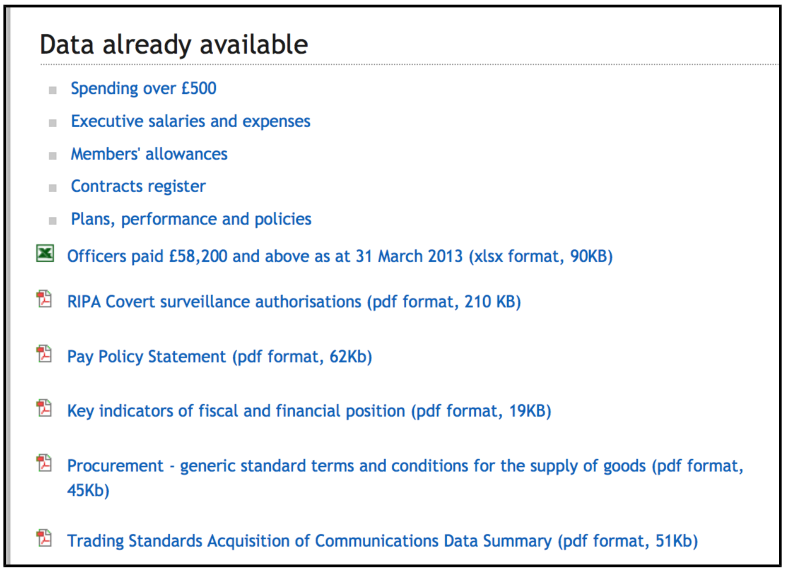
Key	**Value guidance**	**Example value**
name	Text to describe the resource.	Toilet spreadsheet
date	Date the resource applies to. Format: YYYY or YYYY/MM or YYYY/MM/DD. A dataset's resources should all have a date value or none should.	2014/12

url	URL to the file, webpage or API. It is not good enough to simply link to a page of links to the data - all the direct links to the data files must be supplied separately.	https://open.barnet.gov.uk/dataset/d7384c92-5828-41a3-9d12-36d6db5c192f/resource/3e9934ac-492a-4042-971c-fe49bd9ca7e7/download/publictoiletsabfv3.0.csv
format	File format. e.g. CSV, XLS, PDF, RDF, SHP, HTML, SPARQL, API. Don't mention if a file is zipped - just the format inside the zip.	CSV
resource_type	If it is data, then: file or api Otherwise: documentation	file
schema_url	URL of the machine-readable schema that the data conforms to. If it is not a standard format then leave blank. See: Appendix E - Local Authority data schemas	http://schemas.opendata.esd.org.uk/publictoilets/PublicToilets.json?v=0.41
schema_type	The type of schema specified in schema_url. Examples: csvlint or xsd	csvlint

Local Authorities in particular should make sure they fill in the la_service, la_function, schema_type and schema_url fields.

Appendix C – Rationale for using a data portal

The first step to publishing open data tends to be listing a few datasets on a web page, for example:



Data already available

- Spending over £500
- Executive salaries and expenses
- Members' allowances
- Contracts register
- Plans, performance and policies

Officers paid £58,200 and above as at 31 March 2013 (xlsx format, 90KB)

RIPA Covert surveillance authorisations (pdf format, 210 KB)

Pay Policy Statement (pdf format, 62Kb)

Key indicators of fiscal and financial position (pdf format, 19KB)

Procurement - generic standard terms and conditions for the supply of goods (pdf format, 45Kb)

Trading Standards Acquisition of Communications Data Summary (pdf format, 51Kb)

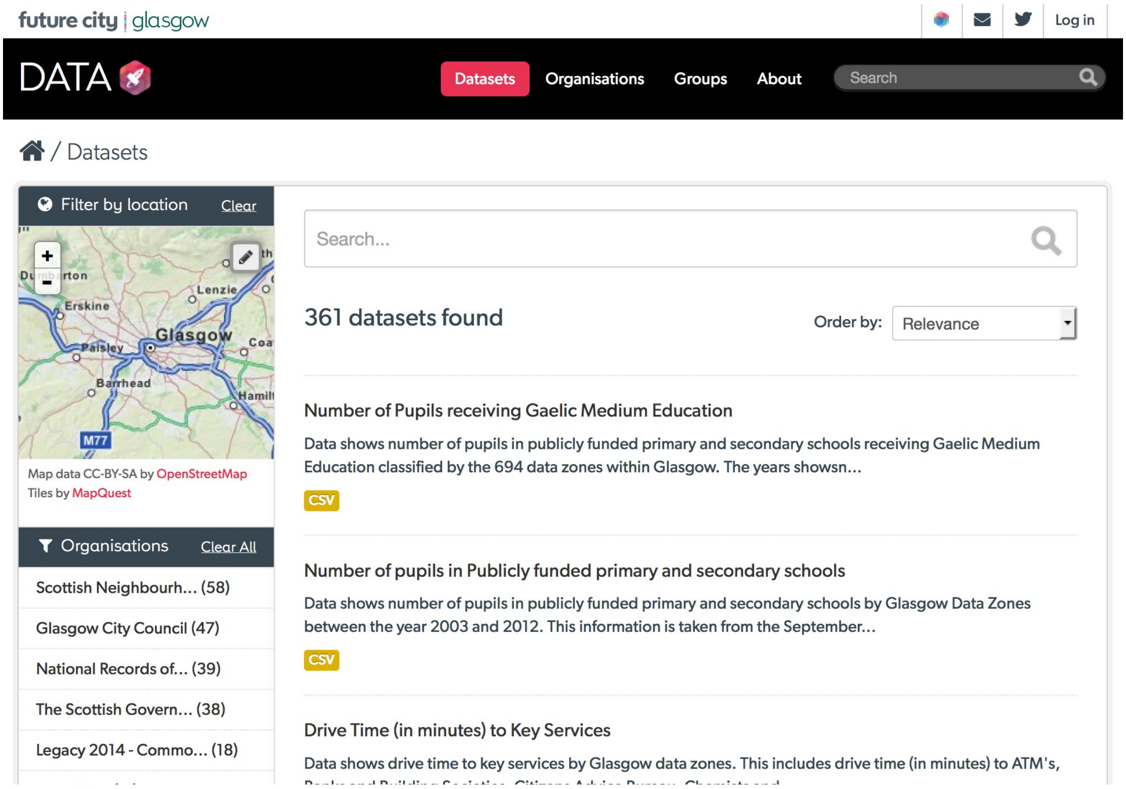
In a similar way you could produce the metadata for data.gov.uk by typing it into a suitable editor with the extra information needed. You'd save the file and publish it on the website for data.gov.uk to harvest.

This will certainly work, but the problem is you now have two lists of datasets to maintain and keep in sync. The administrator will get it right the first time, but the metadata file is less visible than the web page and so it is easily forgotten about. Having two lists to maintain is just not going to work long-term.

So what is essential is to have the web-page listing the datasets **automatically** generated from the metadata file. This could be a relatively small job for a web programmer to implement - to write a bit of PHP or JS perhaps to take the metadata (JSON) and produce the web page listing the dataset links.

Or a more comprehensive solution would be to publish data using an open data portal.

A data portal such as http://data.glasgow.gov.uk/ is an advanced way to publish metadata. Users can browse multiple pages of datasets, click on them for more detailed information and search for keywords or themes. So it makes clear sense to use one when you start to publish more than a few datasets, or have regular updates. And the good news is that they (nearly) all automatically publish the metadata in a machine-readable format that data.gov.uk can harvest.



/ Datasets

Filter by location Clear

Map data CC-BY-SA by OpenStreetMap
Tiles by MapQuest

Organisations Clear All

Scottish Neighbourh...	(58)
Glasgow City Council	(47)
National Records of...	(39)
The Scottish Govern...	(38)
Legacy 2014 - Commo...	(18)

Search...

361 datasets found

Order by: Relevance

Number of Pupils receiving Gaelic Medium Education

Data shows number of pupils in publicly funded primary and secondary schools receiving Gaelic Medium Education classified by the 694 data zones within Glasgow. The years showsn...

CSV

Number of pupils in Publicly funded primary and secondary schools

Data shows number of pupils in publicly funded primary and secondary schools by Glasgow Data Zones between the year 2003 and 2012. This information is taken from the September...

CSV

Drive Time (in minutes) to Key Services

Data shows drive time to key services by Glasgow data zones. This includes drive time (in minutes) to ATM's, Banks and Building Societies, Citizens Advice Bureau, Chemists and

Appendix D – Location/INSPIRE data

The INSPIRE legislation says how public bodies must publish particular sorts of location data, across 34 themes (Annex 1, 2 & 3). UK Location and data.gov.uk encourage all location data (i.e. if it can be plotted on a map) to be published in this way as well. The key requirement is to serve the location data on the internet using GIS servers, rather than just providing a downloadable file. Along with this it is required to provide metadata records to data.gov.uk in the GEMINI2 format, rather than any of the formats described in this document. Full details about publishing location data are published by UK Location here:

- http://data.gov.uk/location

Questions should be sent to the UK Location Helpdesk:

- UK-INSPIRE-Help@defra.gsi.gov.uk

Appendix E – Local Authority data schemas

The Local Government Association (LGA) has released schemas for key datasets, for example Public Toilets and LA Spend Transactions. When publishing metadata on these datasets, Local Authorities should include the URL for the schema that it uses. This allows users to understand the format of the data. It also allows them to do basic validation that the file is formatted correctly. Indeed data.gov.uk intends to provide an automatic checking service, to validate the data files against the schemas.

The schema should be referred to in the metadata by the URL which downloads the schema file. (Rather than the URL to a web page about the schema, which is not machine-readable.) ESD is looking to improve the process to get the URL, but currently you need to follow these instructions:

1. Browse to the schema page in Firefox or Chrome. e.g. for Spend data it is:

http://csvchecker.opendata.esd.org.uk/spend

2. Fill in the form saying what options you have chosen for the fields that have options

3. Open the developer toolbar and click on the 'Network' tab/ (i.e. enable it to record the network requests)

4. On the web page press "Validation File". (You should get a download dialog box - ignore it)

5. In the Network tab you'll get a list of URLs loaded - its the first one you need. Copy and paste it into the correct field, which is "conformsTo" (data.json/DCAT/Inventory) or "schema_url" (CKAN).

It should look a bit like:

> http://csvchecker.opendata.esd.org.uk/schema/downloadjsonschema?schemaId=spend&majorVersion=0&requiredFieldIds=