Open Data Ireland:

Data Audit Report

Authors: Richard Cyganiak, Maulik Kamdar, Fadi Maali, Deirdre Lee, Stefan Decker

Insight Centre for Data Analytics, NUI Galway

Contents

Glossary

API   Application Programming Interface

CKAN   Comprehensive Knowledge Archive Network

CRO   Companies Registration Office

CSO   Central Statistics Office of Ireland

DAFM   Department of Agriculture, Food and the Marine

DCENR   Department of Communications, Energy and Natural Resources

DECLG   Department of the Environment, Community and Local Government

DES   Department of Education and Skills

DPER   Department of Public Expenditure and Reform

DRI   Digital Repositories Ireland

DSP   Department of Social Protection

EPA   Environmental Protection Agency, Ireland

FSAI   Food Safety Authority of Ireland

GSI   Geological Survey of Ireland

HIQA   Health Information and Quality Authority

NRA   National Roads Authority

NTA   National Transport Authority

OSi   Ordnance Survey Ireland

PPSN   Personal Public Service Number

RIA   Royal Irish Academy

RSA   Road Safety Authority

SEAI   Sustainable Energy Authority of Ireland

1   Introduction

Before deciding what data to publish as Open Data, public bodies need to have an overview of what data they currently manage, and could therefore potentially release as Open Data. This can be a challenging task, as data in large organisations is typically dispersed over multiple websites, databases, shared-storage, and personal-computers. A data audit is carried out to assess to establish an inventory of what datasets currently exist.

Within the Open Data Ireland project, Insight-NUIG were tasked with (i) identifying existing published datasets for consideration for inclusion on the portal and (ii) providing an assessment of sectors where new datasets should be audited in order of priority for potential inclusion on the portal on the basis that their publication in Open Data format has the potential for creating a significant economic and/or societal impact.

To achieve this, we carried out a data audit to assess to establish an overview of Irish public-sector datasets currently available online. We then aligned our findings with the "common high-value datasets" based on the G8 Open Data Charter categories, which we outlined in the Open Data Ireland Best Practice Handbook, Section 6.2.4. The methodology and results of this data audit are described in this document, as well as recommendations for priority areas for the publication of datasets in the future.

2   Data Audit Methodology

To identify existing published datasets, the following activities were performed:

    a) A web crawl of the websites of government departments and state bodies;
    b) A review of relevant existing data catalogues;
    c) Web searches and web site reviews for each of the G8 Open Data Charter categories.

2.1   Web Crawl of Public Sector Websites

The web crawl included the websites of all government departments and state bodies listed on the Re-use of Public Sector Information (PSI) website[1]. Local authority websites and state-sponsored bodies were not included. The web crawl was performed using the Scrapy tool[2]. Web pages were downloaded from the sites, and the following information was extracted, using a combination of Scrapy, Xidel[3] and custom Ruby scripts:

- HTML links to machine-readable structured data formats
- Pages containing relevant keywords

The list of formats detected:

- Microsoft Excel (XLS, XLSX)
- Comma-Separated Values (CSV, TSV)
- Extensible Markup Language (XML)
- Keyhole Markup Language and Geography Markup Language (KML, GML)
- Resource Description Framework (RDF)
- ESRI Shapefiles (SHP)
- Archived forms of the formats above (ZIP, GZ, BZ2, etc.)

The list of keywords used includes 'data', 'database', 'download', 'API', 'export', 'statistics'. A manual review of the generated lists was performed to remove false positives and to identify interesting sources of data for further consideration.

2.2   Review of Relevant Existing Data Catalogues

There is a number of existing Irish data catalogues that list datasets, as described in the Open Data Ireland: Best Practice Handbook. These catalogues are valuable sources of information, as they already provide organised lists of datasets from specific publishers or on specific topics, with consistent metadata and in a searchable way. Most of these catalogues do not focus specifically on Open Data, so they may contain resources that are not in machine-readable form, not openly licensed, or from sources other than central government departments and state bodies.

For the purpose of this data audit, two catalogues were identified as particularly relevant, because of their broad coverage across the Irish public sector:

---

[1] http://psi.gov.ie/
[2] http://scrapy.org/
[3] http://videlibri.sourceforge.net/xidel.html

a) **StatCentral**[4] is a portal that provides standardised documentation and metadata on recurring official statistics. The portal is maintained by the Central Statistics Office (CSO), but also lists statistics produced by other government departments and state bodies.

b) The **Irish Spatial Data Exchange (ISDE)**[5] is a data discovery tool which allows you to find spatial data and services that are hosted across multiple government and academic organisations. It was developed by the Marine Institute, with its partners the Geological Survey of Ireland (GSI), the Environmental Protection Agency (EPA), the Department of the Environment, Community and Local Government (DECLG) and the Coastal and Marine Research Centre UCC, as a step towards meeting the requirements of the INSPIRE Directive.

Metadata was screen-scraped from both catalogues, and brought into a unified format. Datasets that are not accessible in a machine-readable form, or are not published by organizations in scope for the data audit, were excluded.

2.3   Searches for Specific High-Value Datasets

A key goal of the data audit was to establish the status quo regarding the public availability of the high-value datasets identified in the G8 Open Data Charter. To augment the results of the web crawl, specific Web searches for these datasets were conducted. For each high-value dataset, the website of the agency or agencies responsible for the area was reviewed to identify any data downloads, statistical reports, APIs, etc. In addition, general web searches were performed using the Google search engine, to double-check that the most relevant sources have indeed been found. The keywords used in these activities were the same as in the web crawl, plus the names of the datasets, synonyms thereof, or other related terms. The results are outlined in detail in section 3.

2.4   Limitations of the Methodology

This section discusses the limitations of the chosen methodology for the data audit.

Publishing agencies. The audit was limited to central government departments and public sector bodies. Local governments, state-sponsored bodies (e.g., state-owned companies), and academic sources are not covered.

Machine-readability. The goal of the audit was to find machine-readable, re-usable datasets. Data published on the Web in a way that is not easily amenable to re-use was generally excluded. For example, the following categories of material were excluded: written PDF reports, Microsoft Word documents, tables in web pages, searchable web databases that do not offer data downloads.

Topics outside the G8 high-value data categories. Parts of the data audit required manual review of crawl results or websites. For these activities, attention was focused on the twelve categories of high-value data listed in the G8 Open Data Charter.

---

[4] http://statcentral.ie/
[5] http://catalogue.isde.ie/

Excel spreadsheets. Microsoft Excel is, despite the fact that it's a proprietary, a popular choice as a data publishing format. Excel spreadsheets can be found on almost any government website, and the web crawl has identified more than 2000 of such files. However, only a small part of these represent re-usable data, and that part is hard to identify:

- Many spreadsheets are not data, but forms, simple presentational tables, or calculators.
- Many spreadsheets are published on public sector websites without explicit metadata, so their purpose, context of creation, and value cannot be easily determined.
- Often there are multiple versions and copies of the same spreadsheet available, and determining their relationship (e.g., are they different versions of the same dataset?) often proved too time-consuming.

Data formats. The audit focused on specific, commonly used formats (see above). Other formats may be used by some data publishers, and would not have been caught by the web crawl.

Data deeply embedded in web applications. Some websites offer data downloads only through an "export" feature. A user first has to select the part of the data that is of interest through an interactive selection process, for example by creating a report or selecting a subcategory from a tree. Then they can export the result as a file in a machine-readable form. Such features cannot be discovered reliably by web crawls or search engines, and may or may not be readily apparent to a user.

Manual determination of metadata. Consistent and high-quality metadata is key to enabling data re-use, as it allows potential users to find data and judge its applicability for the user's purpose. Metadata should be captured close to the source of the data. Ideally, the creator or publisher of a dataset should provide the metadata. This data audit attempted to fill in metadata for discovered datasets based on already published information (e.g., from the web page where a dataset is offered for download), but this approach cannot fully replace publisher-provided metadata.

3 Data Audit Results

In this section we detail the current Irish status of each of the common high-value datasets identified in the Open Data Ireland Best Practice Handbook, categorised according to the G8 Open Data Charter categories. This content was gathered over the course of the data audit. All datasets described in this section that are available in machine-readable form have been uploaded to the National Open Data Platform data.gov.ie. However, most are not associated with an Open License, and some of the datasets are not available on the Web at all. These should be prioritised for release as Open Data by the Steering and Implementation Group.

3.1 Companies

Company registration information is available from the Companies Registration Office (CRO) through a free-of-charge API. Bulk downloads of the data are only available for paying subscribers. Insolvency and bankruptcy records do not appear to be published separately on the Web, but are partially covered by the CRO API.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Companies register	Companies register	CRO	API (free of charge) Bulk download (fee)	https://services.cro.ie/overview.aspx	https://services.cro.ie/downloads/CRO_WEBSERVICE_TC_2011.pdf	
Insolvency and bankruptcy records	-	CRO	-	-	-	Partially covered by the CRO API

3.2 Crime and Justice

Key statistics are available from the Central Statistics Office (CSO). Data on justice spending is available only as high-level annual totals through the export function of the Department of Public Expenditure and Reform (DPER) Databank. No detailed drill-down data is published.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Crime statistics	Quarterly recorded crime	CSO	Bulk download	http://www.cso.ie/px/pxeirestat/DATABASE/Eirestat/Recorded%20Crime/Recorded%20Crime_statbank.asp?sp=Recorded%20Crime&Planguage=0	http://www.cso.ie/en/aboutus/copyrightpolicy/	
Justice statistics	Persons Convicted of Relevant Offences by Sex, Type of Offence, Age Group and Year	CSO	Bulk download	http://www.cso.ie/px/pxeirestat/database/eirestat/Detections%20and%20Proceedings/Detections%20and%20Proceedings_statbank.asp?sp=Detections%20and%20Proceedings&Planguage=0	http://www.cso.ie/en/aboutus/copyrightpolicy/	
Justice spending	Net Expenditure Analysis by Vote	DPER	Export function	http://databank.per.gov.ie/Expenditure.aspx?rep=NetVA	No explicit license	The dataset only contains high-level totals. No detailed drill-downs are available.

3.3   Earth Observation

Comprehensive meteorological data is available from Met Éireann, but only for a fee, and not as an online download. The Sea Fisheries Protection Authority publishes relevant data on fishing levels on a web page, but not in a machine-readable form. Related geospatial data is available from the Marine Institute in machine-readable form, under a restrictive license. No data on Inland Fisheries or hunting appears to be available on the Web. Comprehensive statistics on agriculture are available in machine-readable form from the Department of Agriculture, Food and the Marine (DAFM), and the CSO.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Meteorological	Met Éireann data products	Met Éireann	Various	http://www.met.ie/climate/climate-products.asp		Data is available for a fee. Delivery via email/fax/post only, no Web downloads.
Fishing/ Hunting levels	Fisheries Landings	Sea Fisheries Protection Authority	Web page only	http://www.sfpa.ie/Sea-FisheriesConservation/FisheriesQuotas/FishLandings.aspx		
	Marine Data Online: Fisheries	Marine Institute	Mostly bulk download	http://data.marine.ie/Category/Index/13	https://www.marine.ie/NR/rdonlyres/6F56279C-631D-42AC-B495-74C76CE93A8B/0/MIDataLicenseMar2013.pdf	
Agriculture	Statistics on Agriculture and Fishing	CSO; DAFM	Bulk download	http://www.statcentral.ie/viewstats.asp?type=Agriculture%20and%20Fishing	http://www.cso.ie/en/aboutus/copyrightpolicy/	

3.4 Education

Comprehensive aggregate statistics, as well as data on individual schools, is available from the Department of Education and Skills (DES) for download in Excel format, although without explicit licensing information. Data on school locations appears to be not available on the Web in a re-usable form.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
School attendance	Education and Skills: Statistical Reports	DES	Bulk download	http://www.education.ie/en/Publications/Statistics/Statistical-Reports/	No explicit license	
	Data on Individual Schools	DES	Bulk download	http://www.education.ie/en/Publications/Statistics/Data-on-Individual-Schools/	No explicit license	
	Education Statistics	DES	Bulk download	http://www.cso.ie/px/pxeirestat/pssn/des/homepagefiles/des_statbank.asp	No explicit license	

3.5 Energy and Environment

Data on pollution, water quality and air quality is available from the Environmental Protection Agency (EPA), although not in a unified form. The Secure Archive for Environmental Research Data offers the results of many individual studies. The EPA Geoportal offers geospatial data files under an open license. Spatial data on natural resources is available from the Department of Communications, Energy and Natural Resources (DCENR), with a license restricted to non-commercial use. The EPA and the National Waste Collection Permit Office offer searchable databases and various written reports on waste management, but no relevant machine-readable data (bulk downloads, APIs) appears to be available on the Web. A number of specific statistics on energy usage and fuel efficiency are available from the Sustainable Energy Authority of Ireland (SEAI), but no comprehensive statistics or other machine-readable data are available from a public body. (The state-owned companies ESB and EirGrid were not covered by this data audit, and may publish additional data.)

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Pollution Water quality Air quality	Secure Archive for Environmental Research Data	EPA	Mostly bulk download	http://erc.epa.ie/safer/index.jsp	Mostly no explicit license	Many individual datasets, mostly the result of individual studies, available for download.
	EPA Geoportal	EPA	Bulk download	http://gis.epa.ie/GetData/Download	Open Data (except where based on OSi data) http://gis.epa.ie/ContactUs/Policy	Requires registration and confirmation email before download
Natural resources	DCENR Spatial Data	DCENR	Bulk download	http://www.dcenr.gov.ie/Spatial+Data/	Not open	Various spatial datasets on mining, petroleum, etc.
Waste	-	EPA, National Waste Collection Permit Office	-	-	-	Not available as machine-readable data

Energy consumption	Average Fuel Consumption and Distance Travelled for Private Cars by Engine Capacity cc, Year, Statistic and Type of Fuel	SEAI	Bulk download	http://www.cso.ie/px/pxeire stat/pssn/sei/Database/SEI/T ransport%20Statistics/Transp ort%20Statistics.asp	Not open	
	Energy Balance Statistics	SEAI	Bulk download	http://www.cso.ie/px/pxeire stat/pssn/sei/Database/SEI/E nergy%20Balance%20Statisti cs/Energy%20Balance%20Sta tistics.asp	Not open	

3.6 Finance and Contracts

Data on government budgets and government spending is available through the export function of the DPER data bank. Spending data on a transaction level appears not to be available on the Web. Information on tenders and procurement is available on the Web through the searchable eTenders database and the website of the Office of Government Procurement, but not in machine-readable form.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Government budgets	Voted Public Expenditure Databank	DPER	Export function	http://databank.per.gov.ie/Expenditure.aspx?rep=Welcome	No explicit license	
Government spending	Open Data Estimates Expenditure Source	DPER	Export function	http://databank.per.gov.ie/Expenditure.aspx?rep=OpenData	No explicit license	
Tenders/ Procurement	-	eTenders, Office of Government Procurement	-	-	-	Not available as machine-readable data

3.7 Geospatial

National maps, administrative boundaries, and topographical geo-information are available for purchase from Ordnance Survey Ireland (OSi). Administrative boundaries as of 2011 can be downloaded free of charge for non-commercial use only from the CSO. A range of environmental geo-information can be downloaded under an open license from the EPA. Data on postcodes and addresses should become available in 2015 with the introduction of the Eircode system, although information about machine-readable access, APIs, and licenses is not yet available.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
National maps; Local/ administrative boundaries; Topographical geo-information	OSi Digital Products	OSi	Bulk download	http://www.osi.ie/Products/Digital-Products.aspx	Commercial license available	Data available for a subscription fee
Local/ administrative boundaries	Census 2011 Boundary Files	CSO	Bulk download	http://www.cso.ie/en/census/census2011reports/census2011boundaryfiles/	Non-commercial use only	
Environmental geo-information	EPA Database	EPA	Bulk download	http://gis.epa.ie/GetData/Download	Open Data (except where based on OSi data)	Requires registration and confirmation email before download
Postcodes and addresses	Eircode	DCENR		http://eircode.ie	Unknown	To come in 2015

3.8   Global Development

High-value datasets on Government Accountability and Democracy include:

- development aid
- international assistance

Reports on development aid and international assistance are available from the Irish Aid website (Department of Foreign Affairs and Trade), but no machine-readable or re-usable data seems to be available.

3.9   Government Accountability and Democracy

High-value datasets on Government Accountability and Democracy include:

- Government structures and contacts
- Government salaries and pay-scales
- Legislation (Irish Statute Book)
- Hospitality/gifts
- Election results

These datasets are not available in machine-readable or re-usable form on the Web, but only in the form of web pages or PDF reports, or not at all.

3.10 Health

The Health Well is an all-island (Ireland and Northern Ireland) health information website that promotes health and well-being by supporting evidence-informed decision making. Health Well provides an online catalogue listing publicly available interventions, 385 policies, 746 Datasets, and data repositories available from 1943 health organizations. Some can be downloaded from the website itself, whereas other redirect to original publisher.

Another detailed catalogue of all national health and social care data collections (administrative sources, censuses, national routine surveys, and national patient registries) in Ireland was published in 2010 by the Health Information and Quality Authority of Ireland (HIQA). It is available as a PDF or can be accessed using an interactive searchable version. The Food Safety Authority of Ireland (FSAI) undertakes various enforcement, monitoring and auditing activities to ensure the food produced, distributed and marketed in Ireland is safe. This data is released in the public domain as PDF files. It also released Information regarding the EU investigation into the adulteration of foods, for instance recent mixing of beef with horse meat.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Health performance Drug/prescription	Health Well	Northern Ireland's PHO Team in Institute of Public Health in Ireland (IPH)	Various	http://www.thehealthwell.info/search-all	Various	746 Datasets, 385 policies etc. retrieved from 1943 organizations. Some can be downloaded from website free, whereas other redirect to original publisher
Health performance	National and Regional Health Quality Data	HIQA	Data Catalogue	http://www.hiqa.ie/healthcare/health-information/data-collections	Various	Catalogue provides metadata and links to websites for download
Restaurant hygiene	FSAI Enforcement Order Reports	FSAI	PDF only	http://www.fsai.ie/enforcement_audit.htm	No explicit license	Audit and Monitoring reports available for download

3.11 Science and Research

The Royal Irish Academy (RIA) Library Online Catalogue has a number of searchable databases pertaining to Irish literature, periodicals, manuscripts, photographs, artefacts and prints, which could be interactively searched for social sciences research.

Several datasets pertaining to Irish Energy regulation, Ageism, Health, political sciences and Diversity, is made available by the Irish Social Science Data Archive. These datasets could be downloaded as SPSS files, but the archive allows download in SAS and CSV formats on request. End User License requires to be filled by the researcher before access.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Science and Research	Energy, Ageism, Health and Diversity.	Irish Social Science Data Archive	Individual downloads	http://www.ucd.ie/issda/data/	Research only	Archive of research datasets deposited by various data creators
Social Sciences Research	Manuscripts, Prints, Periodicals, Artefacts, Photographs	RIA		http://www.ria.ie/library/catalogue.aspx		

3.12 Social Mobility and Welfare

The DECLG publishes housing statistics pertaining to house prices, loans, social and affordable housing and housing needs assessment on an annual and quarterly level. These are available from 2000 to 2011, and could be downloaded as Excel or PDF files.

The CSO publishes Employment/Unemployment Statistics, seasonally adjusted unemployment rates and labour disputes as HTML Tables or can be downloaded as CSV files through their Interactive Table Viewer. These statistics range from 2007 to 2013. CSO also publishes Housing and Households datasets. Allocation of Personal Public Service (PPS) numbers (depending on nationality), annual redundancy statistics and statistical information of social welfare services provided to different groups is made available in the Public domain by the Department of Social Protection.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Housing	Housing Prices, Allocations, Loans, Affordability	DECLG	Bulk download	http://www.environ.ie/en/Publications/StatisticsandRegularPublications/HousingStatistics/	No explicit license	
Employment/ unemployment	Employment Statistics, Seasonal Rates, Employment Disputes	CSO	Bulk download	http://www.cso.ie/en/statistics/labourmarket/	http://www.cso.ie/en/aboutus/copyrightpolicy/	
Social security/ welfare	PPS Allocation, Poverty Statistics	DSP	PDF and web pages only	http://www.welfare.ie/en/Pages/Statistics_holder.aspx	Personal use only	

3.13 Statistics

Data on National Statistics pertaining to population, health, science, tourism, construction, agriculture, industries etc. is made available by the CSO on their website. Most of this is published in the form of tables, however they provide an interactive data viewer which allows the download in the form of CSV or PX files.

CSO also publishes the Census 2011 and 2006 results as interactive tables, compiled reports, community profiles, population maps, grid datasets and boundary files. These results can also be visualized using their interactive mapping tool. While, most of these statistics are available for public download, the Place of Work Census for Anonymized Records (POWCAR) is only available to researchers approved by CSO. Results from older census are available as archives.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
National statistics	Population, Health, Tourism,Economy, etc.	CSO	Bulk download	http://www.cso.ie/en/statistics/	http://www.cso.ie/en/aboutus/copyrightpolicy/	
Census	Boundary Files, Population Maps	CSO	Bulk download	http://www.cso.ie/en/census/index.html	http://www.cso.ie/en/aboutus/copyrightpolicy/	

3.14 Transport and Infrastructure

The National Transport Authority (NTA) of Ireland has published information on the routes and schedules of transport services provide by various operators across Ireland. This can be downloaded as Zipped CSV files under the Transport for Ireland web portal. The National Roads Authority (NRA) has a primary function, under the Roads Act 1993, 'to secure the provision of a safe and efficient network of national roads'. OSi mapping is used by the NRA for planning, construction and maintenance of Ireland's national road infrastructure, research and development, and for the collection of traffic counts and analysis on a spatial basis. The Irish Road Network is available under as Geo-referenced TIFF images which can be bought from the OSi's Road Atlas of Ireland system. The Traffic Count Data is published on the NRA's website and can be downloaded as Excel files. The Road Safety Authority (RSA) of Ireland releases statistics on a yearly basis summarizing deaths, injuries and collisions on Irish Roads, penalty point statistics, as well as driving tests pass rates and waiting times. These statistics on Road Traffic Accidents are available as PDF files.

Dataset	Irish dataset(s)	Public body	Format	URL	License	Notes
Public transport schedules	Transport for Ireland	NTA	Bulk download	http://www.transportforireland.ie/transitData/PT_Data.html	Public Sector General License	
Public transport stops	Transport for Ireland	NTA	Bulk download	http://www.transportforireland.ie/transitData/PT_Data.html	Public Sector General License	
Road network	Digi Ireland & Road Atlas of Ireland	OSi	Bulk download	http://www.osi.ie/Products/Professional-Mapping/Ireland-Series.aspx	Commercial license available	Data available for a subscription fee
Road traffic accidents	Collision Statistics, Death Data	RSA	PDF and web page only	http://www.rsa.ie/en/RSA/Road-Safety/Our-Research/		
Traffic Count Data	Traffic Count Data	NRA	Web page and export function	http://www.nra.ie/network-monitoring-and-management/traffic-data/	No explicit license	

4 Conclusion

This document described the data audit carried out as part of the Open Data Ireland support project. A standard methodology was followed, as described in section 2, and the results were aligned with the common high-value datasets identified in the Open Data Ireland Best Practice Handbook, categorised according to the G8 Open Data Charter categories, as described in section 3. Machine-readable datasets described in this section have been added to the National Open Data Platform data.gov.ie. In addition all of the datasets described in StatCentral and the ISDE catalogue have also been uploaded to data.gov.ie.

However, most of the datasets identified during the data audit are not associated with an Open License, and many of the datasets are not available in a machine-readable form, or are not published on the Web at all. These should be prioritised for release as Open Data by the Steering and Implementation Group.