OpenNorth

Gaps and opportunities for standardization in OGP members' open data catalogs

For the Open Data Working Group of the Open Government Partnership

James McKinney
Stéphane Guidoin
Paulina Marczak

The objective of the Standards stream is to promote the use of open data standards to improve transparency, create social and economic value, and increase the interoperability of open data activities across multiple jurisdictions. Its first deliverable is to complete an inventory of open data standards by type to develop a global view and to identify gaps and overlaps. Its final deliverable is an OGP document outlining baseline standards and best practices for open data, along with guidance for adoption and implementation.

As such, the Standards stream's focus is on OGP members' open data initiatives, which largely consist of the implementation of an open data catalog. The co-leads wrote automated scripts to automatically collect, normalize and analyze data from catalogs, in order to both set a baseline and identify gaps and opportunities for standardization. The bulk of this document reports on this analysis. The analysis simply states the choices that OGP members have made with respect each area for standardization; it makes no judgment as to whether these choices are best practices.

References to specific standards are largely contained in the appendices. Rather than pursue a comprehensive inventory of data standards, this document focuses on those that are most relevant to OGP members' catalogs. The great majority of data standards are specific to one type of data; for example, *GTFS* is specifically for public transit schedules. The Standards stream may investigate such standards once key datasets are recommended by the Principles stream.

40 OGP members have catalogs, most of which can be automatically analyzed since they offer APIs or machine-readable data. For each section below, we specify which catalogs are included in the analysis. An important caveat is that catalogs change constantly; aberrations and errors can be short-lived. If a national catalog aggregates datasets from subnational catalogs, we omit those datasets; at time of writing, this filter has been applied to only the catalogs of Italy and the US. This document is also supported by reading prior work and corresponding with governments and civil society organizations.

The Open Data Working Group's *work plan* listed possible areas for standardization, including: licensing, metadata, file formats, key datasets and domain-specific standards. Domain-specific standards will be addressed in future deliverables once key datasets are identified. This inventory considers only the areas below.

The automated scripts are available at *https://github.com/opennorth/inventory*. A dump of all data collected is available on request. Please email *info@opennorth.ca*. A *reference spreadsheet* with high-level metadata about each OGP member is available.

Table of Contents

1. Structure **4**
1.1. Catalog structures 4
1.1.1. Qualitative analysis of catalog structures. 6
1.2. Data access methods 7
1.2.1. Direct download 7
1.2.2. Indirect downloads 8
1.3. URL structures 8

2. Metadata **10**
2.1. Controlled vocabularies 10
2.2. Federation 11
2.3. Metadata elements 12
2.4. Metadata values. 16

3. Data **18**
3.1. File formats 18
3.1.1. CSV 24
3.2. Character encodings. 26

4. Licensing and rights **28**
4.1. Licenses and public domain dedications. 28
4.2. Licensing and rights metadata 30

Acknowledgements **31**

Resources **32**

Additional resources **37**

1. Structure

1.1. Catalog structures

A catalog's structure is the way in which it organizes its data. For example, CKAN allows publishers to organize data into "packages," each of which may have multiple "resources." A resource is either a file download (e.g. a CSV file) or an external link (e.g. to a listing of CSV files). Both the package and the resource have metadata. In this way, a catalog's software determines, to a great extent, a catalog's structure, by limiting the possibilities. Throughout this document, we will use the terms "dataset" and "distribution" as defined by the *Data Catalog Vocabulary (DCAT)*, which roughly map to CKAN's "package" and "resource." The CKAN example is merely provided to illustrate an implementation of DCAT's specification.

The primary finding is that catalog structure varies greatly between OGP catalogs, with no particular structure emerging as most common. Possible structures include, among others:

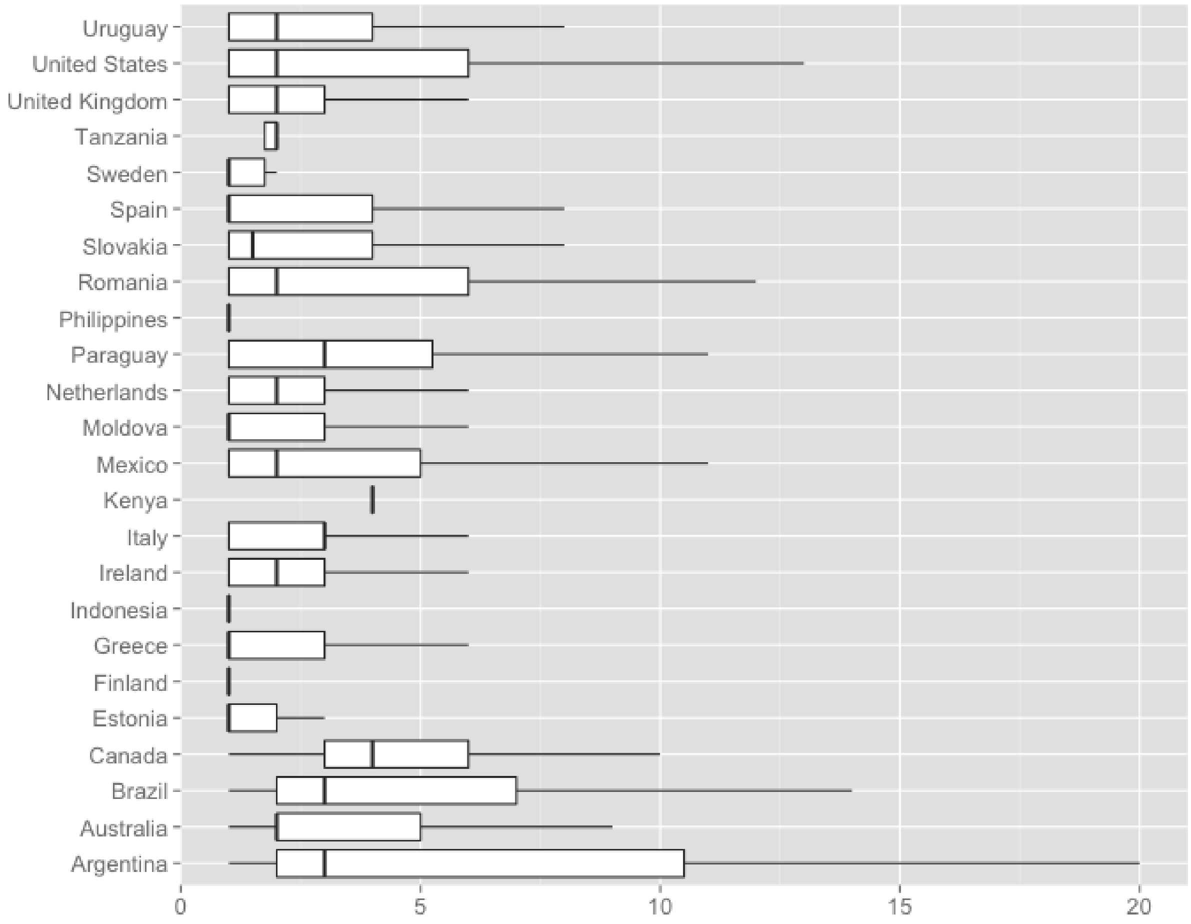
1. Grouping different formats of the same data: for example, grouping the CSV and Excel formats of a budget table. Offering multiple formats allows data users to use the format that is most accessible to them. Socrata's software generally follows this structure; for each dataset, all distributions are different formats of the same data: CSV, Excel, JSON and RDF for tabular data and Shapefile and GeoJSON for geospatial data.

2. Grouping the same type of data for different periods (longitudinal data): for example, grouping contract disclosures by year. In this structure, the different files could be merged into one file covering all periods. This structure may be used for any classification scheme, e.g. grouping the same type of data for different geographies.

3. Grouping different types of data about the same thing: for example, grouping the different tables that make up a national budget. In such cases, the different files could not be merged into one file, as each file has a different structure.

4. Grouping an analysis with its supporting data: for example, grouping an environmental impact assessment with its tabular and geospatial data, or grouping a geospatial file with its documentation.

To get a sense of the relationship between datasets (containers) and distributions (files) in OGP catalogs, a box plot was made with data from 23 catalogs showing the number of files per container. The average was five files per container, with a standard deviation of five. Some catalogs had a majority of containers with the same number of files, in which case the box plot appears as a vertical line.



The maximum number of distributions per dataset was very large in some cases (into the hundreds). Most datasets with a large number of files follow a simple, consistent structure. For example, the UK's *GP practice prescribing data–Presentation level* has four files for each month since August 2010, an example both of longitudinal data and of grouping different types of data about the same thing. Brazil's *Lista de Eleitores Filiados aos Partidos Políticos* has one file for each pair of political party and federal unit, an example of grouping the same type of data for different variables. On the other hand, Mexico's *México Próspero–Estadísticas Nacionales* groups different types of data about the same thing but with no evident structure. Canada's *East Coast Basin Atlas Series* and the US' *Kimama Well–Photos* list hundreds of photos with no way to distinguish one file from another.

If a dataset contains a large number of files but lacks a simple structure, the files are less easily discoverable by data users, especially if the files are very different from one another as in the structural patterns 3 and 4 above, or if the files are indistinguishable from one another before downloading. Poor discoverability negatively impacts a catalog's utility. It is therefore important for catalogs to **monitor the number of distributions per dataset**.

At the other end of the spectrum, 9 of the 23 catalogs have some datasets with no distributions, i.e. datasets without data. In some catalogs using CKAN, instead of having the link to the data appear as a distribution of the dataset, the link sometimes appears as part of the dataset's description, as in Australia, Finland and the Netherlands; in Brazil and Ireland, the link sometimes appears as part of the dataset's metadata. While many datasets can be similarly explained, the Philippines and US catalogs have several datasets without any data. The UK catalog explicitly lists several thousand "unpublished" datasets as a way to make users aware of the existence of data that has not yet been published; it's unclear whether these other catalogs are adopting a similar strategy for increasing awareness. That said, the UK has a few "published" datasets without data.

It is more difficult for users to access data if the links to the data are not consistently located (and it is impossible if there are no links to the data). It is therefore important for catalogs to **monitor the datasets with no distributions**.

1.1.1. Qualitative analysis of catalog structures

In this subsection, we will look at one dataset from each of four catalogs, to provide concrete examples of different approaches to catalog structure. As described previously, a catalog's software determines, to a great extent, its structure. These examples show how the large flexibility that CKAN allows, in comparison to Socrata, leads to greater variability and inconsistency – and therefore lower predictability – in catalog structure.

In Australia's CKAN catalog, the *Budget 2014-2015 Tables and Data* dataset contains 184 files, including: 17 Budget Overview appendices and Budget Paper tables, all but one of which are available in both CSV and Excel formats; 149 Portfolio Budget Statements (PBS) from different government departments in Excel format; one Excel file with each Budget Paper table on a different sheet, and one CSV file which is a "best effort machine readable list that includes line items extracted from the PBS tables." The BPS files use different column and row headers, making automated analysis and aggregation challenging; fortunately, the provided CSV file aggregates the most important information from the 149 PBS files. This dataset is an example of grouping different types of data about the same thing.

In Canada's CKAN catalog, the *Overview of Government Spending and Performance* dataset contains six files: three files for each of the two fiscal years 2011-12 and 2012-13. For each pair of files, the same format (CSV or XML) and the same CSV headers or XML schema are used, making analysis across years easier. This dataset is an example both of longitudinal data and of grouping different types of data about the same thing.

In the US' CKAN catalog, the *Economic Summaries (2005-2009)* dataset contains four identical ZIP files along with metadata records in XML and HTML formats following the ISO-19115:2003 and FGDC-STD-001-1998 standards. The ZIP file contains 13 tables in both CSV and Excel formats, along with documentation in plain text and Word formats. Distributing related tables in one ZIP file makes it easier to download all tables, but harder to see what is in the ZIP file before downloading. The duplicate ZIP files are an error.

The *Land Registry Monthly Price Paid Data* dataset in the UK's CKAN catalog contains files for each year since 1995, grouped by year, but also contains an aggregate file for all years. Each year contains a complete CSV file, a CSV file serializing the same data but with a ".txt" extension, and two halves of the complete CSV file. The complete file is available in parts both to decrease the size of each download and to make the file easier to load into software like Excel, which is slow to load large CSV files.

Every dataset in Kenya's Socrata catalog contains different formats for the same data. Taking *Embu County Piped Water Schemes*, for example, the data can be downloaded as CSV, JSON, PDF, RDF, RSS, XLS, XLSX and XML.

While in the previous section we presented catalog structures as alternatives of one another, catalogs frequently pursue multiple structures at once, though this can result in a loss of discoverability if the resulting structure becomes complex.

This brief analysis shows that, besides the influence of the catalog's software, a catalog's structure is largely determined by the nature of its datasets. Indeed, the differences in structure within one catalog can be as great as the differences in structure between two catalogs. Except in cases like Socrata catalogs, which impose a consistent structure across all datasets, data users must learn a dataset's structure on a case-by-case basis. This situation is likely due to a lack of guidance to governments on how to structure a dataset. The inconsistencies may create challenges for data users to understand, discover and use the data within a dataset, especially as the number of files increases.

1.2. Data access methods

A data access method is a way to retrieve a copy of a file (distribution). OGP catalogs offer several access methods:

- **Direct download:** A person clicks a link on the catalog's website, which initiates a download of a data file. We distinguish two types of direct downloads: whether the data file is hosted locally by the catalog or hosted externally by another website. For example, a publisher may upload a spreadsheet to the catalog, in which case it is hosted by the catalog. Alternately, a publisher may already have a spreadsheet hosted on another website; instead of uploading the same spreadsheet to the catalog, they link to the spreadsheet hosted on the other website, thereby avoiding duplication.
- **Indirect download:** Indirect downloads require an additional step, such as visiting an intermediate web page, creating a user account or filling out a form. In most cases, a person clicks a link on the catalog's website and is brought to another web page, which may be part of another catalog. From this page, they click a link, which initiates a download of a data file.
- **API:** In one case, catalogs offer an API that gives programmatic access to data. All Socrata catalogs use the *Socrata Open Data API* (SODA) and all Open Government Data Initiative (OGDI) catalogs use *ODATA*. CKAN catalogs using the *DataStore* extension use the DataStore API; of the 21 CKAN catalogs, 11 use this extension.
- In another case, a person clicks a link on the catalog's website, which either loads an API's documentation or submits an API request, returning a JSON response for example. Given the difficulty in identifying such APIs programmatically, an analysis is future work.

1.2.1. Direct download

The direct download access method is a ubiquitous access method across catalogs.

The domain name of a download URL is an indication of its provenance. If a file is hosted on a government-controlled domain, such as `.gov` in the US, users recognize that the government controls the file. If the domain name is a subdomain of a generic service like `cloudapp.net`, as is the case for some catalogs using the Open Government Data Initiative (OGDI) software, it is less clear whether the government controls the file. Vendor-specific URLs are also less likely to persist over time, if for example a government changes vendors. Most vendors allow governments to use custom domains using CNAME DNS records; it is therefore the government's prerogative to use this option. Other vendor domains include `arcgis.com`, `amazonaws.com`, `box.com`, `docs.google.com` and `dropbox.com`.

With respect to URL persistence, a few catalogs use the Digital Object Identifier (DOI) System Proxy Server `doi.org`, which acts as a persistent uniform resource locator (PURL) service.

1.2.2. Indirect downloads

If a distribution has a media type of `text/html` (media types are discussed in detail in section 3), then it's possible that the distribution is a link to an intermediate web page that then links to a data file. It is difficult to determine whether an HTML file is a data file itself or whether it is an intermediate web page without manually accessing the file. As such, an analysis is future work.

1.3. URL structures

URL structure is a hot topic on the web; however, recommendations frequently conflict, and the actual value of following recommendations is infrequently measured. It is therefore unsurprising that no specific URL structure has achieved significant adoption, though some general practices like brevity, hierarchy and readability have. We collect recommendations under *Annex 1: Resources* for reference.

In most cases, a catalog's software determines, to a great extent, its URL structure. The URL structures of common software are:

	http://{catalog}/dataset/{reference}
CKAN	*http://data.gov.uk/dataset/index-of-multiple-deprivation*
OGPL	*http://data.gov.gh/dataset/2015-budget-appendix-final*
OpenColibri	*http://data.gov.gr/dataset/37*
	http://{catalog}/dataset/{name}/{reference}
Socrata	*http://www.opendata.go.ke/dataset/Economic-Survey-2014/4ygs-w9sr*
	http://{catalog}/datasets/ver/{reference}
Junar	*http://datos.gob.cl/datasets/ver/9404*
	http://{catalog}/DataBrowser/{container}/{reference}
	OGDI (all OGP members using OGDI use proxies)

Each URL structure has pros and cons. OpenColibri and Junar use opaque URLs. Opaque URLs are more resilient to change and more likely to persist over time – an important criteria to all recommendations for URL structure; however, opaque URIs are not human-readable (by definition). On the other hand, CKAN and OGPL use a unique, human-readable string to identify the dataset in the URL. Human-readable URLs are less resilient to change and less likely to persist over time. Socrata combines a human-readable string with an opaque identifier; in this combination, the opaque identifier will persist, but the URL as a whole will not persist if the human-readable part changes.

In terms of URL scheme (e.g. `http`, `https`, `ftp`), accessing data over the HTTPS protocol ensures that the data is authentic and not subject to a man-in-the-middle attack. Many catalogs link to data using HTTPS, but many targets produced SSL errors. Future work may measure the quality of publishers' SSL implementations.

In order for data users to make regular use of open data, the data must remain accessible at a stable URL. Moving data to a new URL can cause interruptions in services and projects using the data. Governments should therefore prefer URL structures and domain names that promote the persistence of URLs.

2. Metadata

Metadata is data about data: for example, a dataset's title and description or a distribution's format and download URL.

2.1. Controlled vocabularies

A *controlled vocabulary* is a deliberately selected list of terms used to tag units of information, in which exactly one term is used to refer to any given concept. For example, a vocabulary for people's names may select "given name" as one of its terms; users of the vocabulary would only use "given name", and not synonyms like "first name" or "forename", to tag a person's given name. Controlled vocabularies thus ensure consistency and reduce ambiguity.

Within the context of open data catalogs, controlled vocabularies are used to tag properties of catalogs, datasets and distributions, also known as metadata elements. The vocabularies used by OGP catalogs are:

- The *Dublin Core Metadata Initiative* (DCMI) *Metadata Terms*, which have the most widespread adoption of all.
- *ISO 19139 Geographic information – Metadata*, an XML schema implementation derived from *ISO 19115*, which has a *North American Profile*.
- The *World Wide Web Consortium*'s (W3C) *Data Catalog Vocabulary (DCAT)*, an RDF vocabulary that uses DCMI Metadata Terms and that has many formats, including *RDFa* in HTML.
- The US *Project Open Data* (POD) *Metadata Schema*, a *JSON-LD* format of DCAT with additional terms.
- *Schema.org*'s *DataCatalog*, *Dataset* and *DataDownload* schema, which are based on DCAT and which are primarily used as *microdata* in HTML.
- The US *Federal Geographic Data Committee* (FGDC) *Content Standard for Digital Geospatial Metadata* (CS-DGM), though agencies are encouraged to transition to ISO.
- The *British Standards Institution*'s *UK GEMINI* (GEo-spatial Metadata INteroperability INitiative) specification.
- The *DataShare* platform's *inventory schema* (UK).
- The *Infrastructure for Spatial Information in the European Community* (INSPIRE).

A catalog typically uses a single, primary controlled vocabulary to describe all its datasets and distributions. A catalog's software determines, to a great extent, this vocabulary. CKAN offers RDF/XML and N3 formats of DCAT[1] for dataset and distribution metadata in addition to its API's custom vocabulary; CKAN extensions can add support for the Project Open Data Metadata Schema. Socrata uses the Project Open Data Metadata Schema. Other software use their own, custom vocabulary, if any.

Looking at all 22 CKAN and Socrata catalogs, plus Spain's catalog, we determined the adoption of some of the above vocabularies. Five CKAN catalogs do not enable the bundled DCAT support. As future work, the adoption of other vocabularies may be determined.

[1] - Note however that at time of writing CKAN 2.2's implementation of DCAT is incomplete and non-compliant.

Controlled vocabulary	Number of implementations
CKAN API's custom vocabulary	21
DCAT (excluding Project Open Data Metadata Schema)	17
Project Open Data Metadata Schema	2
Schema.org	1

2.2. Federation

Controlled metadata vocabularies make it easier to federate (redistribute) datasets across catalogs. If a central catalog harvests data from other catalogs, data users can access the central catalog and browse the datasets of all linked catalogs, instead of having to access individual catalogs.

The technology used to federate datasets may be a dynamic API; a static file using any of the vocabularies above; or a Web Accessible Folder (WAF), which is essentially a public directory containing static files. CKAN's API uses its own, custom vocabulary. The *Open Geospatial Consortium*'s (OGC) *Catalog Service for the Web* (CSW) API may use Dublin Core, ISO 19139 or FGDC vocabularies.

Of the 22 CKAN and Socrata catalogs, dynamic APIs are used by eight catalogs:

- Catalog Service for the Web (CSW) is used by six catalogs to federate datasets from 39 sources.
- CKAN's API is used by five catalogs for 7 sources.
- *ISO 23950 Information and documentation – Information retrieval (Z39.50)* is used by the US catalog for 9 sources.
- The *ArcGIS Server REST API* is used by the US catalog for 4 sources.

Static files are used by the Mexico, UK and US catalogs.

- Mexico uses the Project Open Data Metadata Schema v1.0 for its 13 sources.
- The vocabularies used in the UK for its 34 sources are GEMINI (30), DataShare (2), DCAT as JSON (1) and DCAT as XML/RDF (1).
- The vocabularies used in the US for its 78 sources are Project Open Data Metadata Schema (56), FGDC (21) and ISO 19139 (1).

Web Accessible Folders (WAF) are used by the Australia, UK and US catalogs.

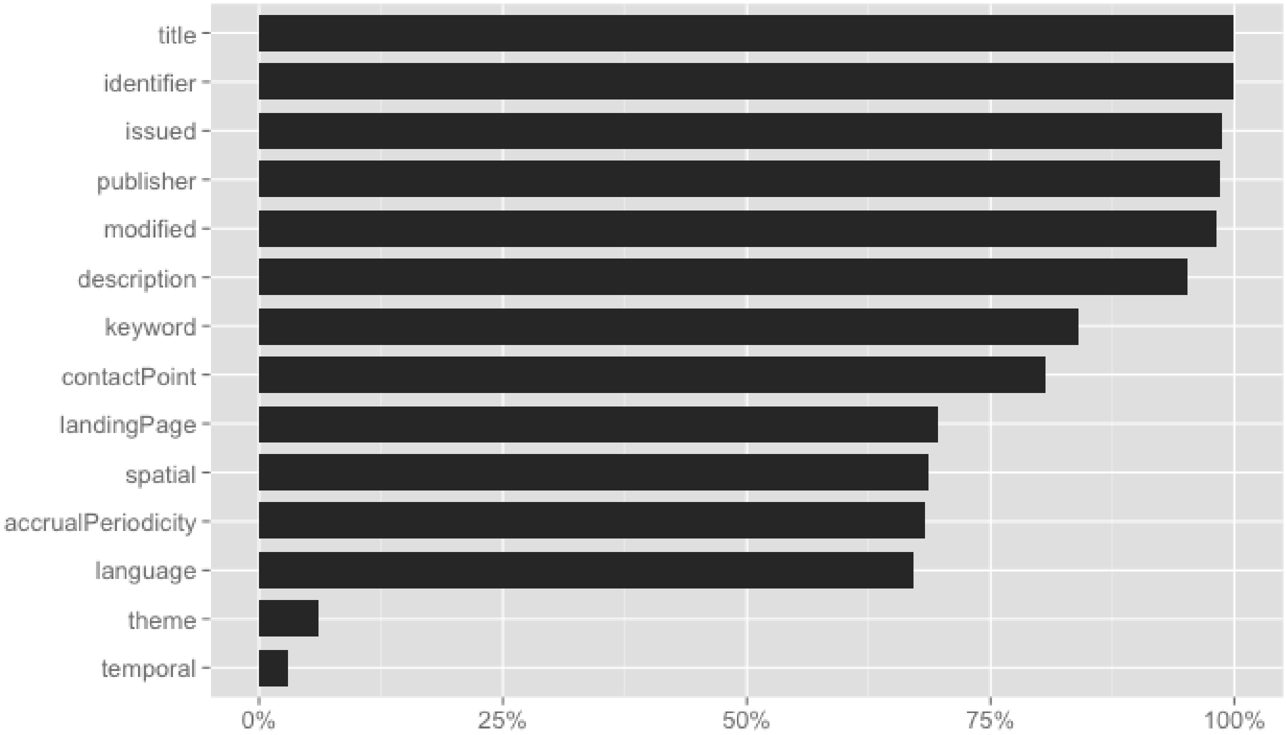
- Australia uses ISO 19139 (1).
- The UK uses GEMINI (200).
- The US uses ISO 19139 (153) and FGDC (76). The vocabularies of 232 folders could not be determined automatically, although sampling shows a prevalence of FGDC.

2.3. Metadata elements

Some metadata is fundamental to the operation of a catalog: for example, a dataset's title and a distribution's download URL. Metadata, like subject, keyword and spatial and temporal extent make it easier for data users to filter datasets and discover datasets of interest to them; publisher and contact point allow a data user to contact the publisher about questions or corrections; modification date allows a data user to determine whether a dataset is up-to-date and fit for use; format allows a data user to select an appropriate file to download.

The metadata elements of 27 catalogs were mapped to DCAT to determine the use of each element. The mapping is imperfect as it doesn't account for all ad-hoc terms used by catalogs for properties.[2] It also could also not account for two differences: (1) CKAN and the Project Open Data Metadata Schema track the dataset's license rather than the distribution's license like DCAT; (2) OpenColibri, the software used by Greece, tracks the distribution's language rather than the dataset's language like DCAT. Licensing is analyzed in detail in section 4.

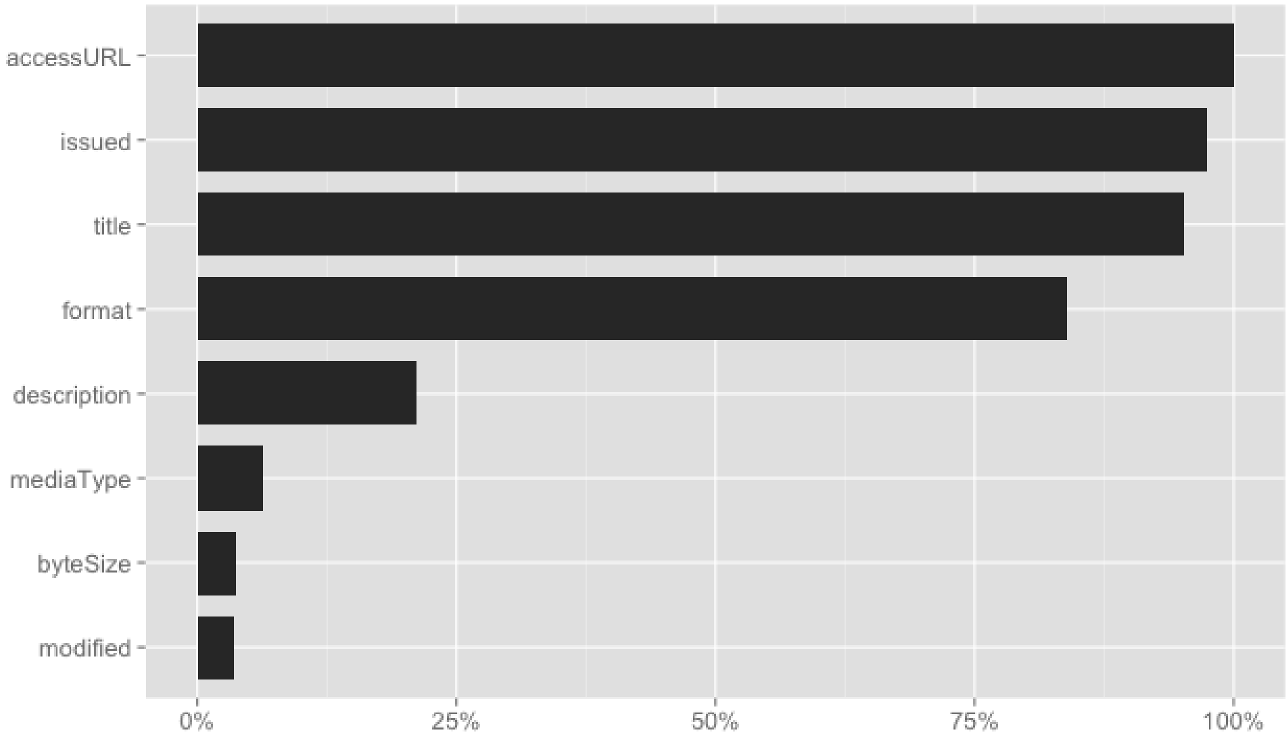
The usage of ***dataset*** elements across all catalogs is below; however, usage varies greatly between catalogs, so a later graph shows usage within each catalog.



Homepage (landingPage) is less frequency used, because in most catalogs, the homepage of a dataset is its webpage in the catalog. Language is not required for monolingual catalogs.

The usage of ***distribution*** elements across all catalogs is below; since usage varies greatly between catalogs, a later graph shows usage within each catalog.

[2] Specifically, we mapped all properties at the root of CKAN's JSON dataset objects, but neither the properties inside each dataset's "extra" object nor the ad-hoc properties at the root of distribution objects.

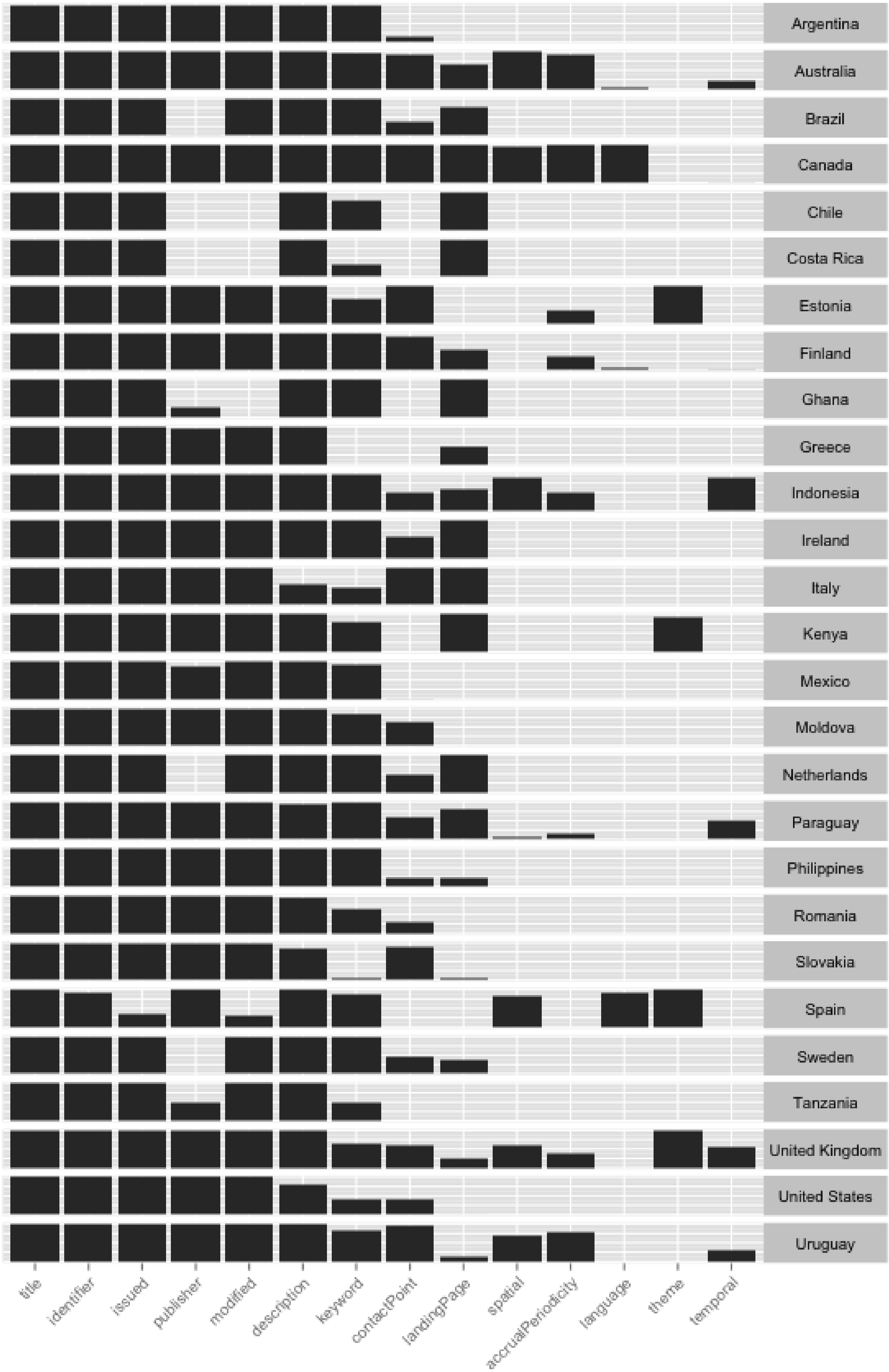


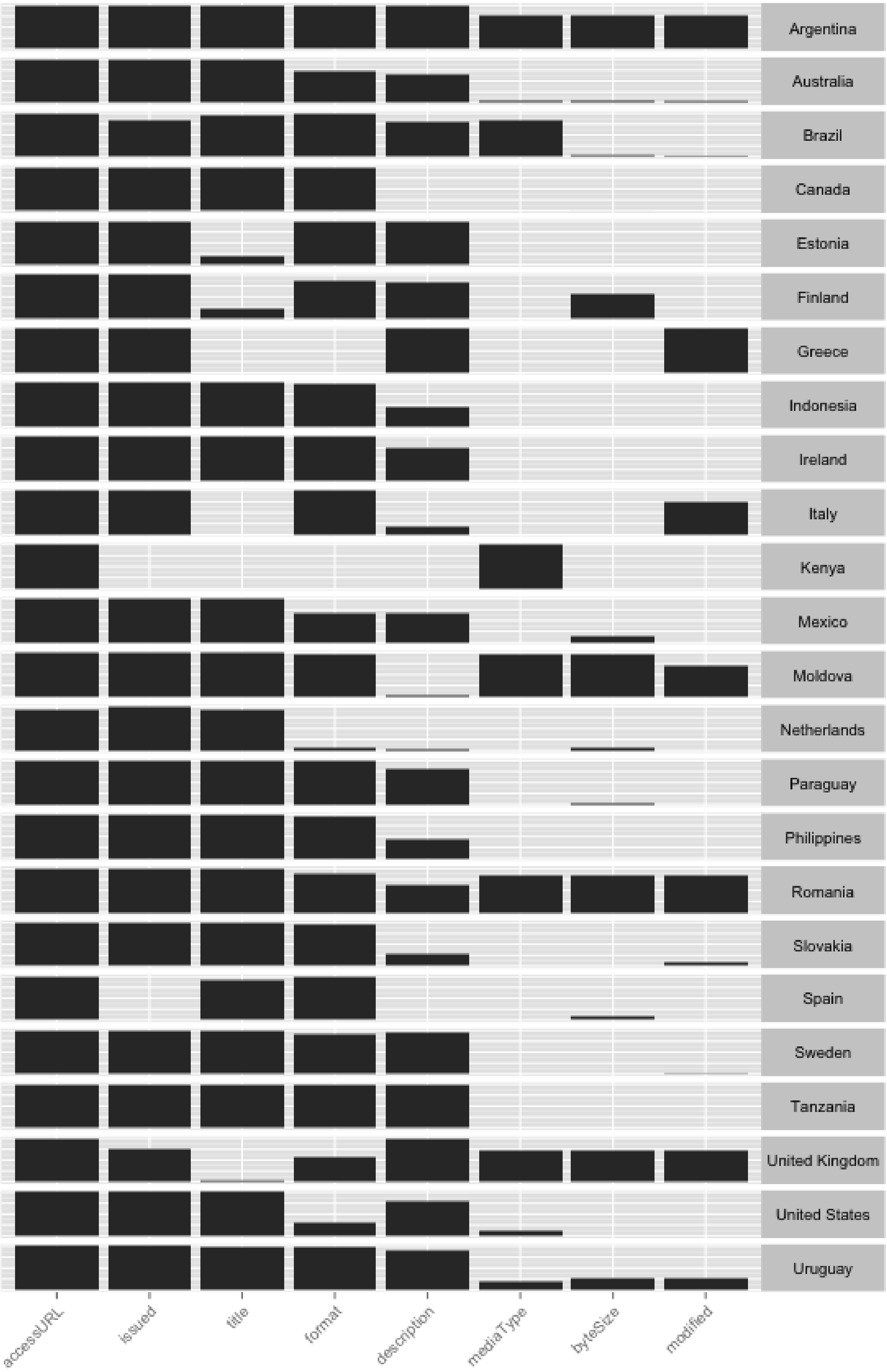
Description is infrequently used, because most distributions, in the context of their datasets, are sufficiently described by their title. Media type is infrequently used, because publishers cannot set it via CKAN's web forms. Byte size is less likely to be assigned a value when the data is hosted externally.

A catalog's software determines, to a great extent, the terms used to tag the properties of datasets and distributions. If the software lacks a built-in term for a property, catalogs tend to omit the property or to adopt an ad-hoc term that is inconsistent between catalogs. Specifically, CKAN lacks terms for language, theme, frequency (accrualPeriodicity), spatial and temporal extent, though *a CKAN extension* adds support for spatial extent. We made efforts to identify all ad-hoc terms; however, more terms exist. With the exception of language, the properties without built-in terms are the least used.

Even in cases where software provides a built-in term, some catalogs use different terms. For example, the CKAN catalogs of Australia, Estonia and the UK use custom terms for the contact point instead of CKAN's `maintainer_email`. To maximize interoperability, **catalogs should not use a custom term if a built-in term would suffice.**

Studying the graphs below for dataset and distribution metadata per catalog shows: which terms are used by all; which are used by a subset; and which are used less frequently than others within the same catalog. Catalogs without distribution metadata are omitted from the second graph: Chile, Costa Rica and Ghana.


Argentina
Australia
Brazil
Canada
Chile
Costa Rica
Estonia
Finland
Ghana
Greece
Indonesia
Ireland
Italy
Kenya
Mexico
Moldova
Netherlands
Paraguay
Philippines
Romania
Slovakia
Spain
Sweden
Tanzania
United Kingdom
United States
Uruguay
title
identifier
issued
publisher
modified
description
keyword
contactPoint
landingPage
spatial
accrualPeriodicity
language
theme
temporal


Argentina
Australia
Brazil
Canada
Estonia
Finland
Greece
Indonesia
Ireland
Italy
Kenya
Mexico
Moldova
Netherlands
Paraguay
Philippines
Romania
Slovakia
Spain
Sweden
Tanzania
United Kingdom
United States
Uruguay
accessURL
issued
title
format
description
mediaType
byteSize
modified

Only the Netherlands' catalog has distributions without access URLs; the datasets of these distributions either have other distributions with access URLs, or provide a link to the data in the dataset's description. The omissions may therefore be errors.

2.4. Metadata values

The values of most metadata elements are simple data types. Most are text. Release date (issued) and modification date (modified) are dates, though many catalogs enter dates as text. Homepage (landingPage) and access URL are URLs. Byte size is an integer.

Contact point and spatial and temporal extent, on the other hand, may use structured values. DCAT structures *contact point* as a *vCard*, which the Project Open Data Metadata Schema (PODMS) inherits. On the other hand, DCAT does not specify structures for spatial and temporal extent. PODMS structures *temporal extent* as an ISO 8601 time interval and *spatial extent* as one of: a bounding box, a point, *Geography Markup Language* (GML), or a name from the *GeoNames* controlled vocabulary. The European Commission also *specifies* how to implement DCAT, though in less detail.

Most OGP catalogs, however, use a combination of metadata elements instead of structured values: for example, Estonia combines `contact-name`, `contact-email` and `contact-phone` for contact point, and Paraguay combines `valid_from` and `valid_until` for temporal extent. This is likely due to it being easier in CKAN to add a simple text field than to add a field containing subfields, which would introduce more structure.

For temporal extent, although *ISO 8601 Data elements and interchange formats – Information interchange – Representation of dates and times* is decades old, only the catalogs of Canada and Finland consistently use it. For spatial extent:

- Australia's catalog uses either a GeoJSON polygon, a GeoJSON point or free text.
- Canada's catalog uses a GeoJSON polygon and a vocabulary.
- Spain's catalog uses a linked data taxonomy.[3]
- The UK's catalog uses a vocabulary.

The values of the text[4] elements may also be controlled:

- Language may use *ISO 639* language codes, like the catalogs of Finland and Spain.
- Publisher may use the official names of departments. CKAN encourages the reuse of publisher values ("organizations" in CKAN) by design.
- Theme may use a controlled vocabulary. Spain's catalog uses a linked data taxonomy as recommended by DCAT.[3] Canada's catalog uses ISO 19115 Topic Categories and the Government of Canada Core Subject Thesaurus. The UK uses a vocabulary. Socrata encourages the reuse of theme values ("categories" in Socrata) by design.

[3] - However, at time of writing, the links do not resolve.

[4] - In DCAT, many elements would have URIs instead of text as values. See DCAT's `documentation`.

- Keyword may use a controlled vocabulary. Ghana uses URLs.
- Frequency (accrualPeriodicity) may use a controlled vocabulary. The catalogs of Canada, Paraguay and the UK use vocabularies. Uruguay's catalog uses an integer to represent the frequency in days.
- Identifier may use an identifier scheme, which controls the formatting of the identifier. CKAN uses universally unique identifiers (UUIDs).
- Media type may use *Internet Assigned Numbers Authority* (IANA) *media types*. Media types are discussed in *section 3*.
- License may use official, canonical license URLs. Licenses are discussed in *section 4*.

The lack of consistent structures for metadata values both within and across catalogs make it more difficult to interpret and use the metadata. Thus, structured values and controlled vocabularies for metadata elements are important areas for standardization.

Turning from the structure to the quality of metadata values, *Konrad Reiche* proposes a set of quality metrics. As future work, the catalogs' metadata could be measured by these metrics:

- completeness (does each metadata element have a value?)
- accuracy (are the values correct?)
- richness of information (do the values convey unique information?)
- accessibility (can a user easily interpret the values?)
- availability (are linked resources reachable?)
- intrinsic precision (are texts free of typos?)

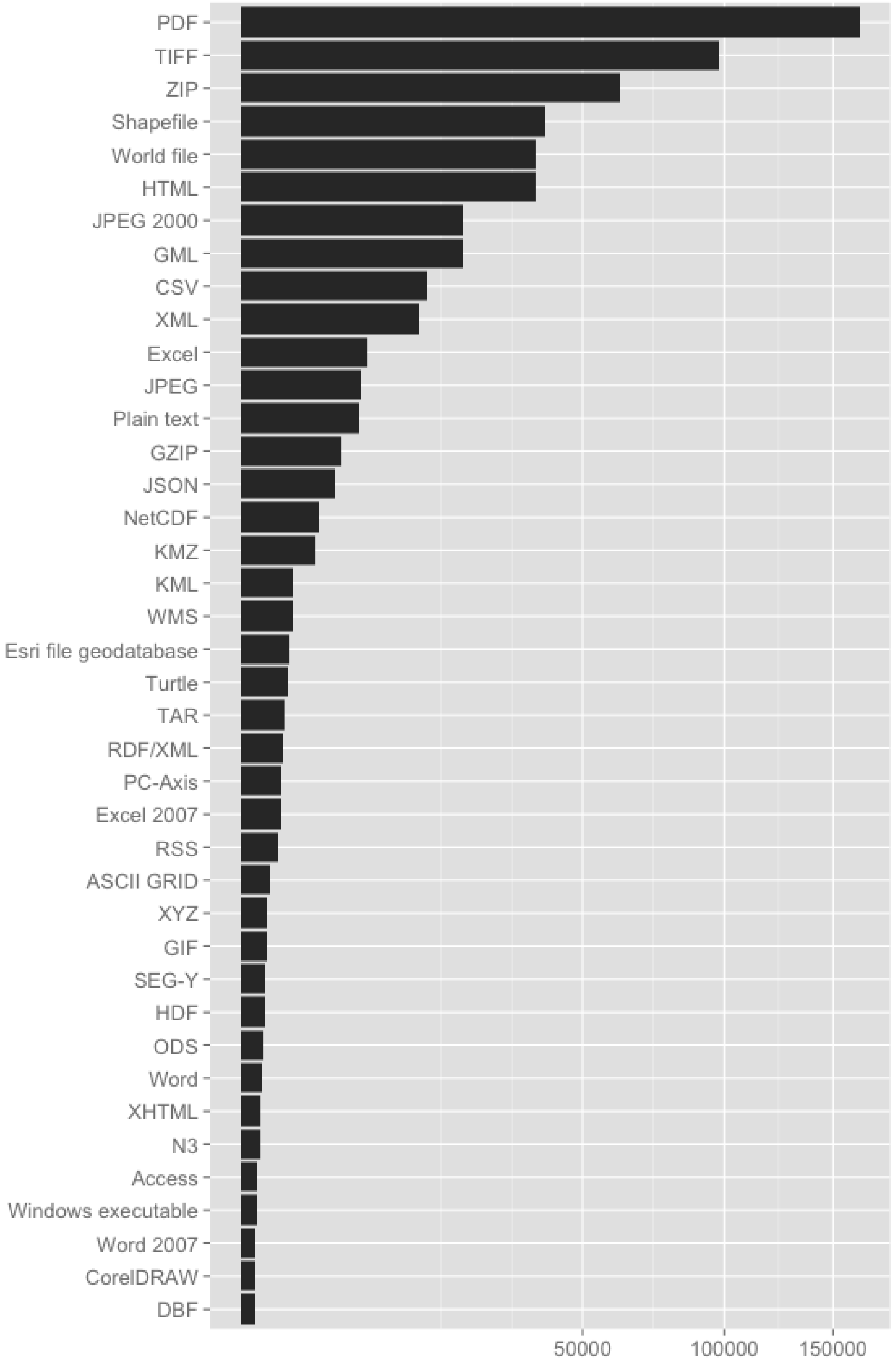
3. Data

3.1. File formats

A file format is a standard way of encoding data. A file format may be either proprietary or free and either unpublished or open. For example, HTML is a free, open format. Prior to 2007, Microsoft Excel was a proprietary, unpublished format. Free and open formats are generally preferred, because they generally have more support across applications, operating systems and platforms. The choice of file format for a distribution may depend on its openness, popularity, accessibility and consistency with other distributions.

A file format is identified by its media type, which is a standard identifier like “text/html”, composed of a type “text” and a subtype “html”. The *Internet Assigned Numbers Authority* (IANA) manages the *official registry of media types*.

Automated scripts were developed and used to collect the media types of all distributions in all 22 CKAN and Socrata catalogs, plus those of Spain and Greece. Socrata catalogs always provide a media type. CKAN catalogs, on the other hand, provide a media type for 6% of distributions, because publishers cannot set it via CKAN’s web forms. The scripts therefore had to determine the media type based on the non-standardized format description and based on the access URL’s extension; if the two conflicted, no media type was set. As future work, the scripts may instead determine the media type by downloading the file. After collecting all media types, any non-IANA media types were normalized to a single form. By the end of this process, 79% of the 1.5 million distributions had normalized media types.


PDF
TIFF
ZIP
Shapefile
World file
HTML
JPEG 2000
GML
CSV
XML
Excel
JPEG
Plain text
GZIP
JSON
NetCDF
KMZ
KML
WMS
Esri file geodatabase
Turtle
TAR
RDF/XML
PC-Axis
Excel 2007
RSS
ASCII GRID
XYZ
GIF
SEG-Y
HDF
ODS
Word
XHTML
N3
Access
Windows executable
Word 2007
CorelDRAW
DBF
50000
100000
150000

The graph uses a **square root scale** and shows, for each of the top 40[5] media types,[6] the number of datasets with at least one file of that media type; this avoid overly weighting media types that frequently occur in multiples. We can identify groups of media types.

Documents (PDF, HTML, Word, XHTML, Word 2007)

PDF is by far the most common format at over 163,000 datasets; the next most common is TIFF at over 97,000 files. The expectation of many open data advocates is for open data to be machine-processable, which most PDFs are not. However, PDFs are accessible to the greatest share of users, given their human-readability and given the broad availability of and familiarity with software capable of reading PDFs (including web browsers). When published alongside a primary machine-processable format, they can provide useful context to the data.

PDF files may also be georeferenced; since *geospatial PDF* and PDF have the same media type, it's unknown how many PDFs are geospatial. Note that the HTML and XHTML files require additional study as part of future work as discussed in *subsection 1.2*.

Images (TIFF, JPEG 2000, JPEG, GIF, CorelDRAW)

Image formats are frequently used for raster geospatial data, like aerial photos. *GeoTIFF* is a metadata standard for georeferencing TIFF files. The *Open Geospatial Consortium* (OGC)'s *GMLJP2* is a metadata standard for georeferencing JPEG 2000 files using *Geography Markup Language* (GML). Even if a user lacks the capacity to use the geospatial metadata, they may still view the image. Georeferenced and non-georeferenced files have the same media types; however, sampling suggests that most if not all images are georeferenced.

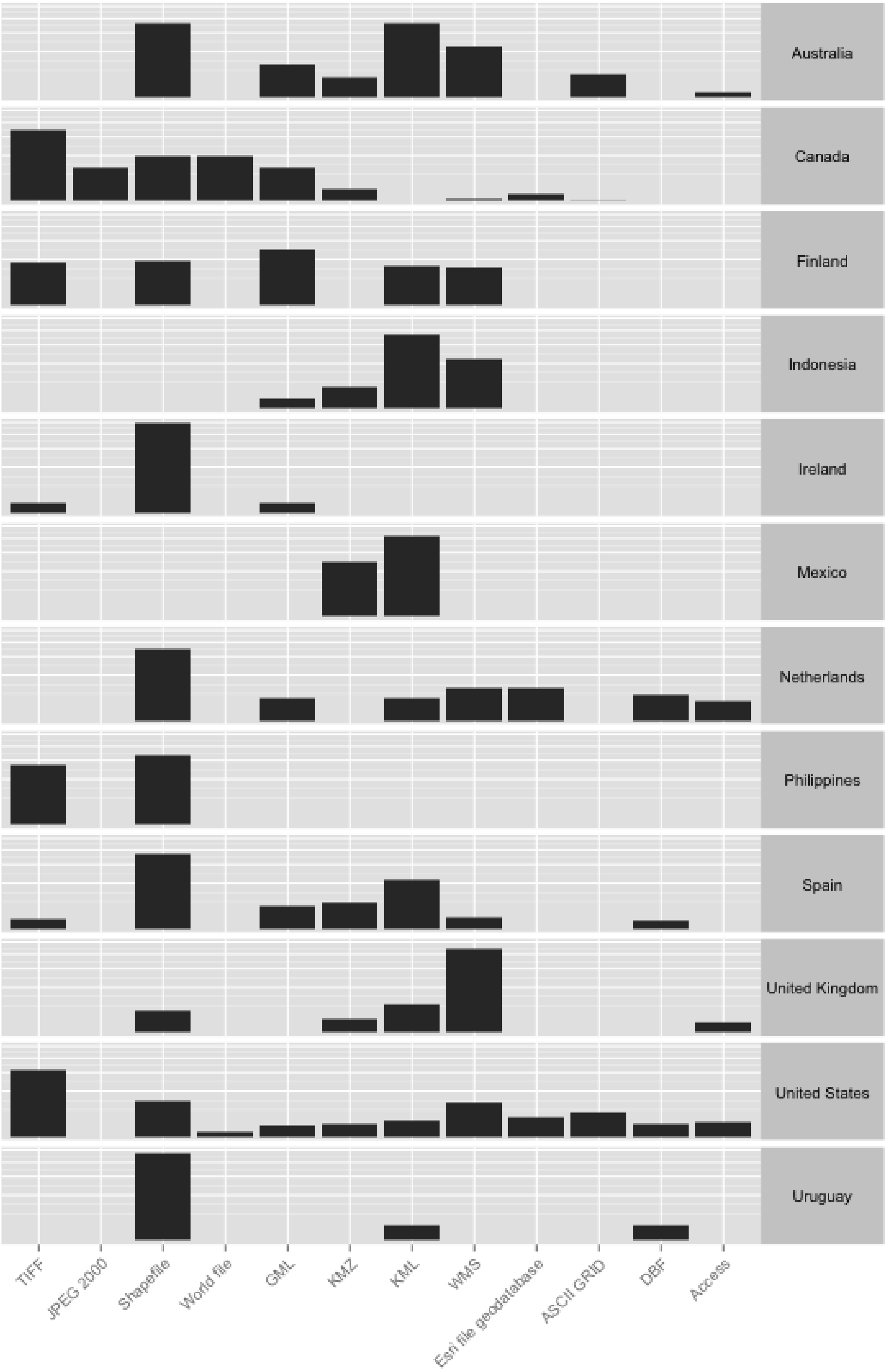
Geospatial (Shapefile, World file, GML, KMZ, KML, WMS, Esri FileGDB, ASCII GRID)

Geospatial formats can be divided into raster and vector formats. A raster format uses pixels as its building block; digital photographs are an everyday example. A vector format uses geometric points, lines and polygons as its building blocks; each geometry is linked to a row in a database that describes its attributes, such as its name. The most popular formats among OGP catalogs, and which are not already mentioned in other format groupings, are listed.

- Shapefile is a vector format.
- A World file is a plain text file for georeferencing raster data, as an alternative to embedding the georeferencing context in the raster file itself.
- *Geography Markup Language* (GML) is an XML-based format. Versions 1.0 and 2.0 support vector data only, but GML 3.0 adds support for raster data.
- Keyhole Markup Language (KML) is a XML-based vector format and is a popular format on the web along with GeoJSON. KMZ is a ZIP file containing KML files.
- Web Map Service (WMS) is an XML-based web service typically used to distribute raster data, though vector data is supported.
- Esri file geodatabase supports both raster and vector data.
- ASCII GRID is a raster format.

[5] - The top 40 media types each occur in at least 75 datasets. 51 media types do not meet this criteria.

[6] - Media types, like `application/vnd.openxmlformats-officedocument.spreadsheetml.sheet`, are abbreviated as vernacular names, like "Excel 2007".


Australia
Canada
Finland
Indonesia
Ireland
Mexico
Netherlands
Philippines
Spain
United Kingdom
United States
Uruguay
TIFF
JPEG 2000
Shapefile
World file
GML
KMZ
KML
WMS
Esri file geodatabase
ASCII GRID
DBF
Access

The above graph uses a **square root scale** to show per-catalog usage of geospatial formats. Catalogs publishing fewer than 20 datasets with geospatial formats are omitted: Romania (2) and Sweden (2); the remaining catalogs publish from 37 (Uruguay) to over 155,000 (Canada). dBase table files and Microsoft Access files are included, because, as discussed below, these tabular formats may be used for geospatial data.

The catalogs use 5 geospatial formats on average (minimum 2, maximum 11). Shapefile and KML/KMZ are the most ubiquitous formats, appearing on 10 of 12 catalogs. The next most common are GML and WMS, on 8 catalogs. On the other hand, the high rankings of JPEG 2000 and World file in the global graph above are primarily due to the large number of such files in Canada's catalog.

Archives (ZIP, GZIP, TAR, Windows executable)

Archives makes it easier and faster for users to download multiple files at once; they also save disk space and bandwidth for the publisher. On the other hand, they conceal the files they contain, making it harder for users to discover those files. Alternatives exist to avoid the drawbacks. To save disk space, file systems with transparent compression (like HFS or ZFS) can be used. To save bandwidth, HTTP compression can be used. To make it easier for users to download multiple files at once, implementing a web interface in which users can select the files to download can promote both easy discovery and easy downloading.

Non-standard media types exist for some common archives; for example, a Shapefile is generally distributed as a ZIP file containing the mandatory `.shp`, `.shx` and `.dbf` files and an optional `.prj` file. The media type `application/x-shapefile` is sometimes used, which is *recommended* by the European Commission. However, this convention is not widely established.

When a media type is not available, publishers sometimes improvise, for example: `application/zip+-text/csv`, `csv (zip)`, `zip (csv)`, `zip+csv`, etc. However, such constructs are not standardized. CKAN provides a "mimetype_inner" metadata element to describe the media type of the archived file; however, publishers cannot set it via CKAN's web forms, and it is only used on 264 files across all catalogs.

Note that most if not all of the Windows executables are self-extracting archives.

Generic (XML, Plain text, JSON, RSS)

XML, JSON and RSS[7] are common formats for encoding many different types of data. The structures of XML and JSON files can be described and standardized by technologies like XML Schema or JSON Schema; for example, eXtensible Business Reporting Language (XBRL) is a standard schema for business information, and GeoJSON is a standard schema for geospatial data. Many standard schemas do not have their own media types; future work may look into what schemas are used by XML and JSON files.

XML and JSON are preferred by programmers, because programming languages have strong support for the formats and because XML and JSON are often used to asynchronously exchange data between a web browser and a server to create web applications. On the other hand, XML and JSON are generally not accessible to non-programmers.

Tabular (CSV, Excel, Excel 2007, OpenDocument Spreadsheet (ODS), Access, DBF)

Tabular formats encode data as a grid of rows and columns, and are second to geospatial formats in prominence on catalogs. In terms of accessibility, tabular formats benefit from the wide use of tables in society

[7] - While RSS was designed to enable publishers to syndicate content, it has also been used to encode simple lists of items, and is thus included among the generic group of media types.

(as product comparison charts, for example) and given the broad availability of and familiarity with office productivity software capable of reading these formats. The CSV format is discussed in detail below.

Note that Esri personal geodatabases are distributed as Microsoft Access files and that Shapefile (and other Esri formats) uses dBase table files (DBF) to store attribute data.

Scientific (*NetCDF*, *PC-Axis*, *XYZ*, *SEG-Y*, *HDF*)

Scientific formats generally require familiarity with specialized software to use. SEG-Y, for examples, encodes geophysical data like seismic reflection data, which geophysics software can read. Many scientific formats are optimized for organizing and sharing large amounts of numerical data, like NetCDF and Hierarchical Data Format (HDF), for which less specialized formats are unsuitable.

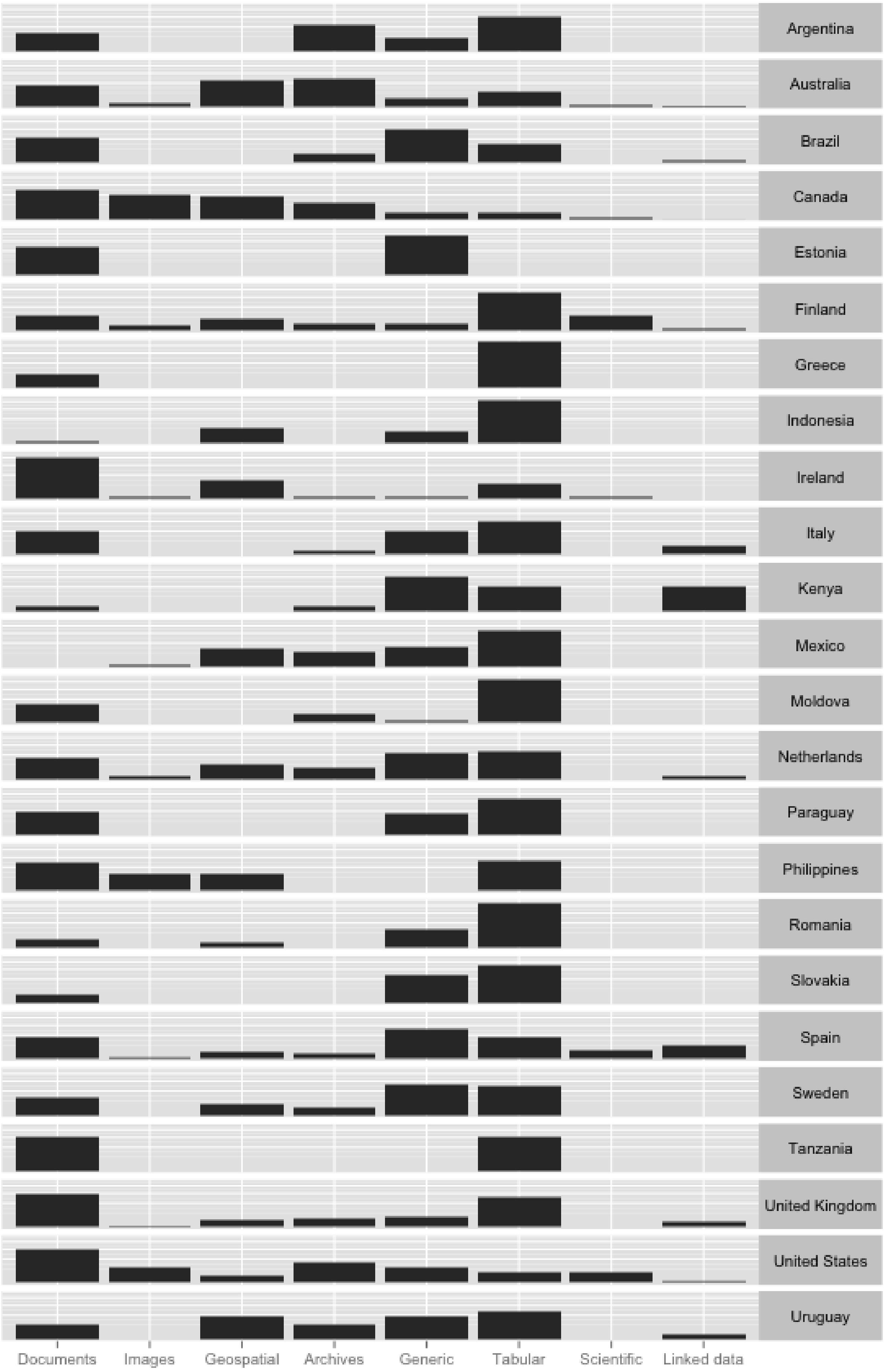
Note that the NetCDF format with *Climate and Forecast (CF) metadata conventions* can be used for raster geospatial data.

Linked data (Turtle, RDF/XML, N3)

Turtle, RDF/XML and N3 are all *Resource Description Framework* (RDF) formats. RDF is a general-purpose language for representing information on the web. The meaning of RDF terms can be described and standardized by technologies like the Web Ontology Language (OWL). Using linked data requires familiarity with RDF and with specialized software, such as graph databases.

The following graph uses a **square root scale** and shows, for each group of media types, the number of datasets per catalog with at least one file in that group.

The graph tells a different story than the first graph in this subsection. While there are only 22,864 datasets with tabular formats across all catalogs – compared to 201,270 datasets with document formats – their share of datasets per catalog is greatest. Generic formats similarly increase in rank from 5th to 3rd, overtaking geospatial and archive formats. Like in the analysis of geospatial formats, the high rankings of image formats in the first graph above are primarily due to the large number of such files in the two largest catalogs: Canada and the US. Note that the catalogs of Estonia and Tanzania had only just launched and had fewer than 10 datasets at time of writing. Otherwise, most catalogs offer a diversity of formats.


Argentina
Australia
Brazil
Canada
Estonia
Finland
Greece
Indonesia
Ireland
Italy
Kenya
Mexico
Moldova
Netherlands
Paraguay
Philippines
Romania
Slovakia
Spain
Sweden
Tanzania
United Kingdom
United States
Uruguay
Documents
Images
Geospatial
Archives
Generic
Tabular
Scientific
Linked data

3.1.1. CSV

The comma separated values (CSV) format had been used to exchange data between spreadsheet programs for decades before the format was first formally documented in *RFC 4180* by the *Internet Engineering Task Force* (IETF) in 2005. RFC 4180 remains a ***de facto*** standard, and no general standard exists. As such, while a wide range of programs can read CSV files, some programs write CSVs in idiosyncratic ways. For example, whereas RFC 4180 escapes a double quote with another double quote, some programs escape with a backslash.

RFC 4180 describes the structure of a CSV file, but not the substructure of a CSV field. For example, many countries use a decimal point while others use a decimal comma. Absent any standardization of number formats, software support for numerical data in CSV files may vary. RFC 4180 also allows any character encoding for CSV files; however, some versions of Microsoft Excel, for example, cannot read UTF-8 CSV files correctly. Some programs write a byte order mark (BOM) to set the encoding, but other programs will read the BOM as data.

The World Wide Web Consortium (W3C) *CSV on the Web Working Group*, whose charter ends August 31, 2015, is working to normalize CSV files on the web and address some of these issues.

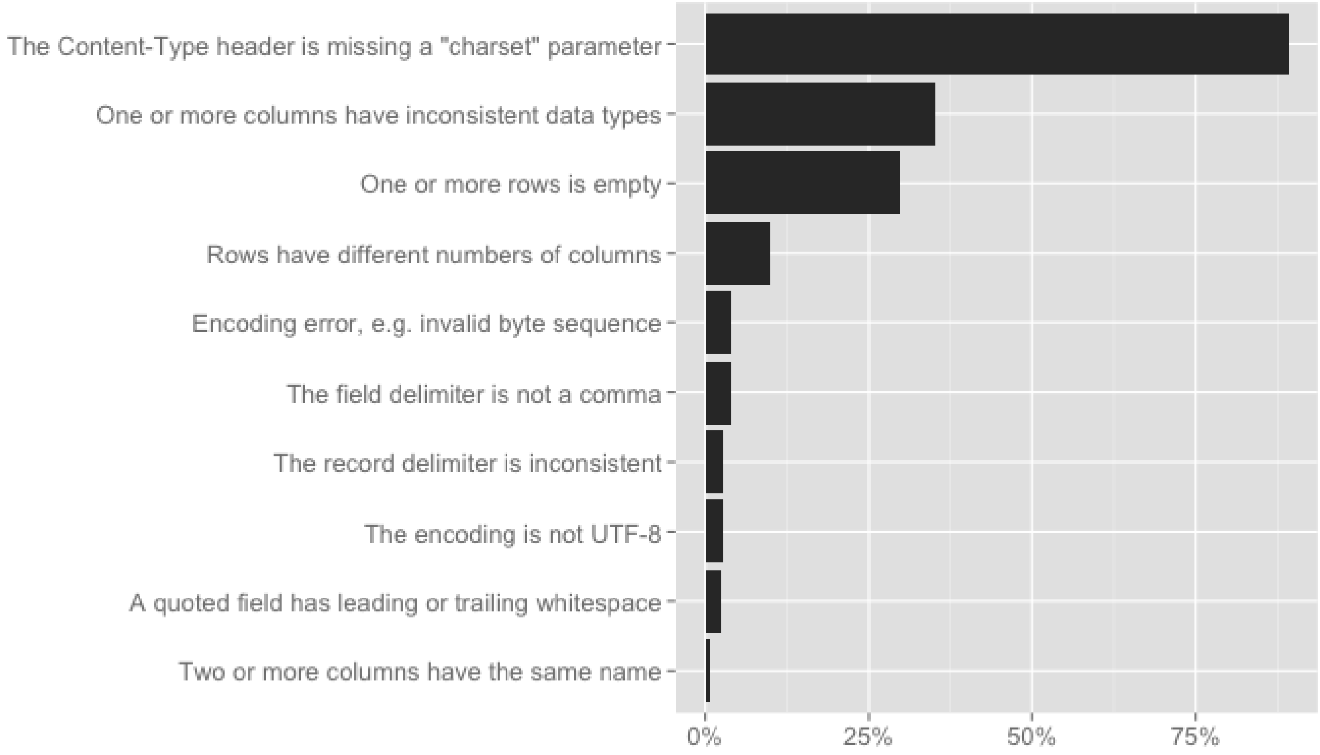
For our analysis, the *CSVLint* Ruby library was used to identify common inconsistencies and errors in CSV files, in order to identify areas for standardization. Of the 42,905 distributions with metadata indicating a media type of `text/csv`, only 14,572 have a Content-Type HTTP header of `text/csv`. 10,107 have a Content-Type of `application/octet-stream`, which according to *RFC 7231* is a generic media type for binary data; in other words, it says nothing about the file's format. 7,778 have a Content-Type of `application/zip`, which may be ZIP files containing CSV files. 5,524 have a Content-Type of `text/html`, which may be indirect downloads, discussed in subsection 1.2. Among many other incorrect Content-Type headers are several invalid media types like `text/x-comma-separated-values`.

To avoid reporting on non-CSV files, analysis was limited to files with a Content-Type header of `text/csv`. A CSV file may have multiple errors.

The most common errors and inconsistencies are:

- As discussed in the next subsection on *character encoding*, few CSV files declare a character encoding by setting a "charset" parameter in the Content-Type header, for which CSVLint issues a warning. That said, only 4% caused an encoding error.
- The next two most common errors are related. CSVLint reports an error if a CSV file contains empty rows between non-empty rows. Many CSV files have rows of data, followed by empty rows, followed by rows of explanations. A data user would generally have to remove or ignore the explanations to use the data.
- The CSV format doesn't have a mechanism to declare the data type of a cell. For example, cells with the values `12345` and `2014-12-31` are read as text, although a human would interpret them as an integer and a date. CSVLint tries to guess the data type of each cell and issues a warning if the types are inconsistent within a column.
- When CSVLint encounters the rows of explanations, it correctly determines that the data types in the rows of explanations are inconsistent with the data types in the rows of data above. Thus, these two errors often co-occur. In general, **publishers should omit rows of explanations from CSV files**, and instead publish them as metadata.

- CSV files should have the same number of columns as each row; in other words, CSV files should resemble a table. CSV files often violate this expectation by having header content, like titles and subtitles, above the table. **Publishers should put column names in the first row of CSV files,** and move any other content into the metadata.



The remaining errors occur less than 4% of the time. It should be noted that no CSV file had quoting errors – e.g. an unclosed quoted field like `“Jane”,”Doe”,”100` or a stray quote character like `“Gary “Kid” Carter”`. Other errors that would prevent a user from reading a CSV file occurred at a low rate.

3.2. Character encodings

To visually display a character to a human reader correctly, its encoding must be known. Many data formats, like HTML, can declare their encoding; if the declared encoding is incorrect or unsupported, however, some characters may be visually displayed as ▯ or □. Other data formats cannot declare their encoding. A lack of encoding information becomes an issue whenever data uses characters outside the 128 ASCII characters. If there is no convention for the encoding (that is, if a catalog doesn’t adopt a single, common encoding) or if software fails to detect the encoding (software may guess the encoding based on character frequencies), a data user is unlikely to be able to read the data, as few data users have the experience and skills required to determine the correct encoding.

CSV is a common format lacking a mechanism to declare encoding; our analysis therefore focuses on CSV files. The analysis can inform encoding issues in other formats lacking a mechanism to declare encoding.

RFC 2616 3.7.1 states that data received via HTTP with a media subtype of the "text" type but without a "charset" parameter has a default "charset" parameter of ISO-8859-1.[8] Of the 14,572 CSV files analyzed above, only 1,858 had an explicit "charset" parameter. Of the remaining, the default encoding of ISO-8859-1 caused an encoding error in 422 cases; that is, the file contained an invalid byte sequence in ISO-8859-1. That said, ISO-8859-1 may nonetheless be the incorrect encoding for those files, because byte sequences can be read as valid but incorrect characters. For example, if the correct encoding is UTF-8, then the single two-byte UTF-8 character é would be read as the two one-byte ISO-8859-1 characters Ã©. Future work may try to establish the actual encoding of files.

The software used to author data files may not make it easy or possible to select a desired encoding. For example, Microsoft Excel on OS X in North America only saves CSV files using the MacRoman encoding, while LibreOffice gives control over the encoding. The problems caused by poor software support for character encoding is *well documented*.

Even if the publisher knows the encoding, the software used to publish data files may not make it easy or possible to declare the encoding. As described above, few CSV files in the sample set a "charset" parameter in the Content-Type HTTP header. Furthermore, many metadata standards, including DCAT, do not allow publishers to declare the encoding. Indeed, no OGP catalog declares the encoding of files in its metadata.

The literature review found only one country with official guidance on encoding: *the UK*. While our analysis methods need improvement to identify the encoding of a greater share of the CSV files, it is clear that a lack of encoding information and an inconsistent use of encodings are common issues across catalogs.

[8] - RFC 2616 (1999) is obsoleted by RFC 7231 (2014), among others, which removes the default charset of ISO-8859-1. However, Ruby still implements the RFC 2616 behavior.

4. Licensing and rights

4.1. Licenses and public domain dedications

Version 4.0 of the Creative Commons license suite is the most endorsed license suite in the literature, and the Creative Commons Public Domain Dedication (CC0) is the most endorsed public domain dedication. Notably, the *European Commission* highlights and the *Open Data Institute* endorses the open Creative Commons licenses and dedications: that is, the Public Domain Dedication (CC0), Attribution license (CC-BY) and Attribution-ShareAlike license (CC-BY-SA). Among OGP members, open Creative Commons licenses and dedications are the most popular. Several OGP members are in the process of determining which license to adopt, and all are favoring Creative Commons.

Automated scripts were developed and used to collect the license under which each dataset was licensed for 24 catalogs. License descriptors were normalized to canonical URLs where possible.[9] Some software, like Junar and the Open Government Platform (OGPL), provide no licensing metadata via API, preventing automated collection.

In 9 countries, the most common license is an open Creative Commons license.[10] In 8 countries, it is a government-authored, country-specific license. In 2 countries, it is an Open Data Commons license. In 5 countries, it is a unspecified or underspecified license.

In 13 countries, the most common ***internationally reusable*** license is an open Creative Commons license.[11] In 3 countries, it is an Open Data Commons license. 8 countries do not use any internationally reusable license.

Five countries use a single license for all datasets: Argentina, Indonesia, Kenya, Mexico and Moldova. Three countries use a single license for all but a small number of datasets using unspecified licenses, which may be easily corrected errors: the Philippines, Romania and Uruguay. It should be noted that all but two of Canada's over 200,000 datasets are licensed under the Open Government Licence – Canada.

While catalogs use a diversity of licenses, most have a single primary open data license. In nearly all cases, the number of licenses used is equal to the number of distinct documents describing licensing terms. Spain is an outlier, using over one hundred licensing documents, though 90% of datasets are described by one of 13 licensing documents. In 15 catalogs, 90% of datasets are licensed under one of two licenses. In the UK, while 97% of datasets are licensed under the UK Open Government Licence, others require unique attribution statements, creating a long tail of attribution requirements. The remaining 8 catalogs are discussed below. It is important for catalogs to **monitor the number of licenses, licensing documents and attribution statements in use to avoid their proliferation**, which would increase the burden on data users to read, understand and respect the licenses.

A significant problem is the quality of licensing metadata. In 8 out of 24 catalogs, the licenses of over 10% of datasets are either not specified or underspecified. The choice of catalog software affects the quality of

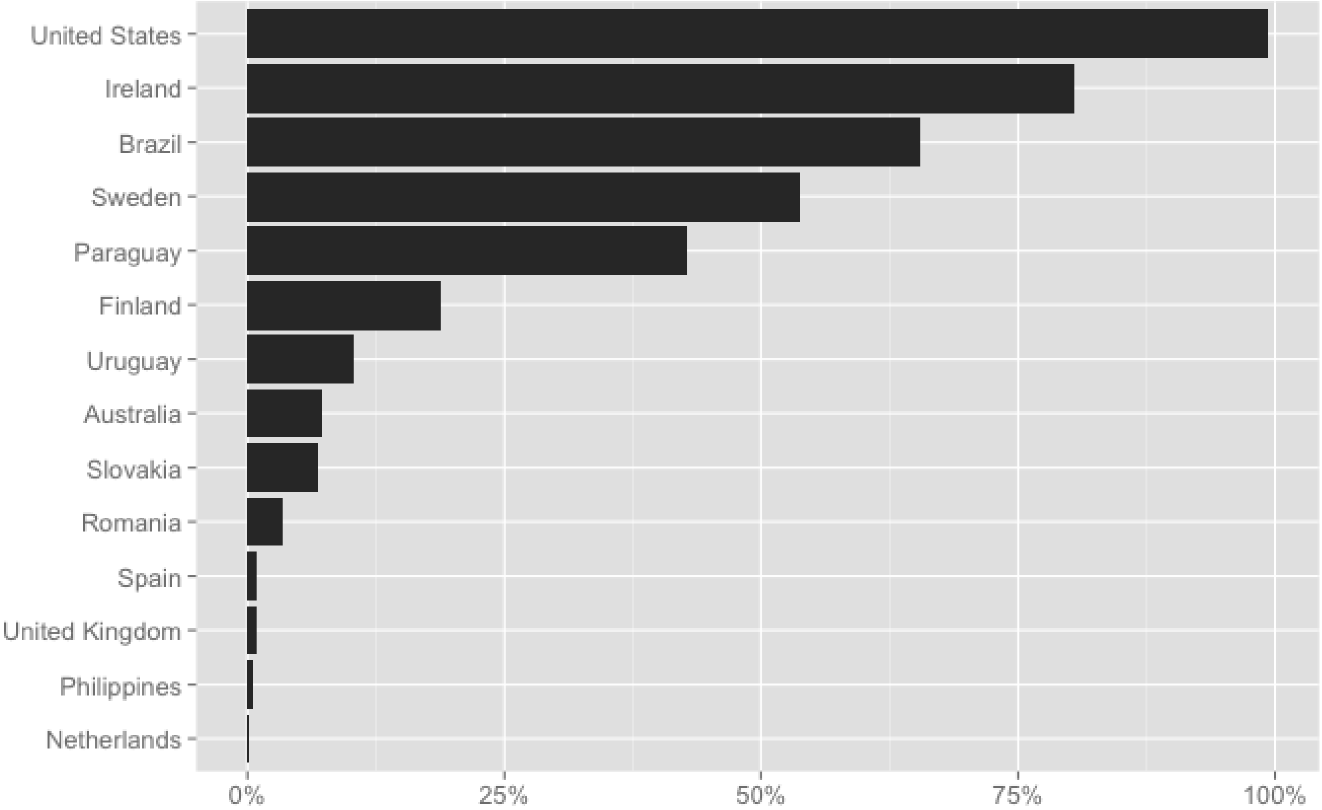
[9] - The CKAN data catalog software, by default, links its bundled licenses to http://opendefinition.org/. We instead use canonical URLs like https://creativecommons.org/licenses/by/4.0/.

[10] - The specific licenses are CC0-1.0, CC-BY-4.0, CC-BY-3.0-AU, CC-BY-3.0-EE, CC-BY-3.0-GR, CC-BY and CC-BY-SA (versions unspecified), all of which are Open Definition-conformant licenses.

[11] - In addition to the previous footnote, the specific licenses are CC-BY-3.0-ES and CC-BY-3.0-IT.

licensing metadata. CKAN *offers* several "generic" licenses, which are not licenses at all, such as "Other (Attribution)" or "License not specified." It also offers Creative Commons licenses without specifying version numbers, which are necessary to identify the licenses; we supplemented the licenses with version numbers where possible.[12] Catalogs should **avoid using the default CKAN licenses** as almost all are underspecified.

The following graph shows the proportion of datasets per catalog with an unspecified or underspecified license.



The graph bears a few important clarifications and caveats:

- Under US law, federal government works are not entitled to domestic copyright protection; therefore if no license is presented, U.S. public domain can be assumed. Project Open Data provides *guidance* to label US federal government data as U.S. public domain and, where possible, to use worldwide public domain dedications.
- At time of writing, Ireland is holding *a public consultation on licensing*, and Brazil is conducting a legal study on licensing.
- The catalogs of Sweden and Paraguay are among the five smallest – with under 60 datasets each – so licensing may be clarified as the catalogs mature.
- In Finland, 14% of datasets are subject to an unspecified Contributor Licence Agreement; the licenses of a further 1% are either unspecified or underspecified.
- Uruguay, Romania and the Philippines use a single license for over 89% of datasets; the remaining datasets aren't licensed, which may be easily corrected errors.

As future work, the license that applies to metadata for each catalog may be determined.

[12] - The version numbers were determined either by reading the catalogs' documentation or by corresponding with the catalogs' operators.

4.2. Licensing and rights metadata

It is possible to provide metadata about licenses and rights statements, such as whether the license requires attribution or whether the license permits commercial use. The Creative Commons licensing suite uses a composable nomenclature to communicate its licenses' requirements, e.g. "Attribution" (BY), "Attribution-NonCommercial" (BY-NC), and "Attribution- NonCommercial-ShareAlike" (BY-NC-SA). If licensing metadata is machine-readable, it is possible to determine automatically the permissions and requirements of a given license.

Data catalogs, however, provide little to no metadata about licenses and rights statements. The only common metadata are the title of the license and the URL to its full text. CKAN also provides a license identifier and an "isopen" boolean indicating whether the license is an open license according to the *Open Definition*. However, many catalogs misconfigure the "isopen" boolean, i.e. the metadata is unreliable.[13]

Several initiatives catalog licenses and provide metadata. For example, the Open Knowledge Foundation's Open Definition website catalogs *conformant licences* and provides *licensing metadata*. The Canadian Internet Policy and Public Interest Clinic's *Licensing Information Project for Open Licences* also catalogs open licenses and provides *licensing metadata*. Instead of replicating licensing metadata on each catalog, licensing metadata can be centralized on services such as these.

Acknowledgements

The funding for this work has been partially provided through the World Wide Web Foundation 'Open Data for Development Fund' to support the 'Open Government Partnership Open Data Working Group' work, through grant 107722 from Canada's International Development Research Centre (web.idrc.ca). Find out more at: *http://www.opengovpartnership.org/groups/opendata*

The funding for a research intern has been provided through the SSHRC partnership grant 'How the Geospatial Web 2.0 is Reshaping Government-Citizen Interactions' (Geothink) funded by Canada's Social Sciences and Humanities Research Council (SSHRC) Partnership Grant Program.

[13] - Australia ("cc-by-sa"), Italy ("cc-by-4"), the Netherlands ("cc-by", "cc-zero") and the Philippines ("cc-by", "cc-by-sa") misconfigure Creative Commons licenses, which are open, as being closed.

ANNEX 1

Resources

1. Data access methods

- Majewski, Christopher et al. *GC Web API Standard*. Government of Canada, 29 September 2014.
- Mill, Eric et al. *18F API Standards*. U.S. General Services Administration, 25 September 2014.
- Hirsch, Bryan et al. *White House Web API Standards*. The White House, 14 April 2014.
- Pizzo, Michael et al. *Open Data Protocol (OData)*. OASIS, 17 March 2014.

2. URL structures

- Leigh Dodds et al. *Creating Value with Identifiers in an Open Data World*. Open Data Institute and Thomson Reuters, October 2014.
- Hyland, Bernadette et al. "*The Role of "Good URIs" for Linked Data*." *Best Practices for Publishing Linked Data*. World Wide Web Consortium (W3C), 9 January 2014.
- Archer, Phil et al. *D7.1.3–Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC*. Interoperability Solutions for European Public Administrations (ISA), European Commission, 17 December 2012.
- Benitez, M.T. Carraso et al. *Best Practice for Web Data URI*. Data on the Web Best Practices Working Group, World Wide Web Consortium (W3C), 10 June 2014.
- Benitez, M.T. Carraso et al. *Compact Uniform Resource Identifier (COMURI)*. Data on the Web Best Practices Working Group, World Wide Web Consortium (W3C), 14 November 2014.
- Davidson, Paul. *Designing URI Sets for the UK Public Sector*. UK Chief Technology Officer Council, 9 October 2009.
- Williams, Stuart. *Revision of UK Public Sector URI Guidance–Summary of Survey Responses*. UK Government Linked Data Working Group (UKGovLD), 12 February 2013.
- Williams, Stuart. *Core URI Patterns for the UK Public Sector*. UK Government Linked Data Working Group (UKGovLD), 31 October 2014.
- Williams, Stuart. *Location URI Patterns for the UK Public Sector*. UK Government Linked Data Working Group (UKGovLD), 31 October 2014.
- Jennison, Teni. *Creating URIs*. DATA.GOV.UK, 28 March 2011.
- Nečaský, Martin et al. "Convention for URLs of catalogue entities." *Methodology for publishing datasets as open data*. Components Supporting the Open Data Exploitation (COMSODE), 31 July 2014.
- Nečaský, Martin et al. "Convention for URLs of entities in datasets published as 4* and 5* data." *Methodology for publishing datasets as open data*. Components Supporting the Open Data Exploitation (COMSODE), 31 July 2014.

- Overbeek, Hans and Linda van den Brink. *Towards a national URI Strategy for Linked Data of the Dutch public sector*. Platform Linked Data Nederland, 19 September 2013.
- *URI Design Principles: Creating Persistent URIs for Government Linked Data*. Linking Open Government Data, Tetherless World Constellation, Rensselaer Polytechnic Institute, 23 October 2013.
- *Slash Data Catalog Requirements*. Project Open Data, 18 November 2014.
- *Persistent Identifiers*. Australian National Data Service, August 2011.
- *URIs: Best Practices*. Synthetic Biology Open Language.

3. Metadata

- Reiche, Konrad Johannes. *Implementation of Metadata Quality Metrics and Application on Public Government Data*. 2013.
- Reiche, Konrad Johannes. Assessment and Visualization of Metadata Quality for Open Government Data. Master's thesis. Freie Universität Berlin. 17 October 2013.
- Cyganiak, Richard and Fadi Maali. *Use Cases and Requirements for the Data Catalog Vocabulary*. W3C, 23 October 2014.
- Maali, Fadi and John Erickson. *Data Catalog Vocabulary (DCAT)*. W3C, 16 January 2014.
- *Marking up your dataset with DCAT*. Open Data Institute.
- *Dublin Core Collection Description Frequency Vocabulary*. Dublin Core Collection Description Task Group, 9 March 2013.
- Brown, Matthew et al. *G8 Metadata Mapping*. Project Open Data, 18 June 2013.
- Brooks, Gray et al. *Common Core Metadata Schema*. Project Open Data, 6 November 2014.
- Brickley, Dan and Ramanathan V. Guha. "*DataCatalog*." *Schema.org*. W3C, 12 September 2014.
- Lagoze, Carl and Herbert Van de Sompel. *The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*. Open Archives Initiative, 7 December 2008.
- Goedertier, Stijn. *DCAT application profile for data portals in Europe*. Interoperability Solutions for European Public Administrations (ISA), European Commission, 15 May 2014.
- *Alignment of INSPIRE metadata with DCAT-AP*. European Commission, 27 October 2014.
- Reuvers, Marcel et al. *INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119*. European Commission, 2013.
- *Government of Canada Open Data Metadata Element Set*. Treasury Board of Canada Secretariat, 18 June 2013.
- *Geopatial Metadata Standards*. Federal Geographic Metadata Committee. 12 September 2014.
- *AGLS Metadata Standard: Australian Government Implementation Manual*. National Archives of Australia, 30 September 2011.
- Weisberg, Peri. *San Francisco Metadata: Background and survey of options for San Francisco metadata standards that help users find data and allow for tracking and summarizing*. City of San Francisco, July 23 2014.

- *Data.cabq.gov Core Metadata Requirements*. City of Albuquerque.
- Simperl, Elena et al. “Availability of Different Metadata Attributes.” *Open data topologies, catalogues and metadata harmonisation*. OpenDataMonitor, 30 June 2014.
- Tirry, Diederik et al. “Publishing metadata of geospatial indicators as Linked Open Data: a policy-oriented approach.” *Proceedings Of The 17Th AGILE International Conference On Geographic Information Science*. June 2014.
- *Putting Things in Order: a Directory of Metadata Schemas and Related Standards*. JISC Digital Media.
- *Controlling your Language: a Directory of Metadata Vocabularies*. JISC Digital Media.
- Day, Michael. *Mapping between metadata formats*. UK Office for Library and Information Networking, University of Bath, 22 May 2002.
- *Metadata styles and standards*. Esri, 2012.

4. Data formats

- Humphries, Linda. *Sharing or collaborating with government documents: profile*. Open Standards Board, 12 June 2014.
- Ahuja, Anisha. *Exchange of Calendar Events Standards Profile*. Open Standards Board, 15 October 2014.
- Ahuja, Anisha. *Exchange of Contact Information Standards Profile*. Open Standards Board, 15 October 2014.
- Vinck, Sean. “Data Formats Compatible with the Open Data Portal.” *Open Data Standard and Technical Standards Manual*. State of Illinois, 26 March 2014.
- Tandy, Jeremy et al. “*Requirements*.” *CSV on the Web: Use Cases and Requirements*. CSV on the Web Working Group, World Wide Web Consortium (W3C), 1 July 2014.
- Atz, Ulrich. *What is a CSV? A case study of CSVs on data.gov.uk*. Open Data Institute, 18 February 2014.
- *SDMX 2.1 Technical Specification*. Statistical Data and Metadata eXchange, April 2013.
- Levine, Thomas. *What file formats are on the data portals?*. 20 October 2014.
- Cote, Paul. *Understanding GIS Data, Referencing Systems and Metadata*. Harvard University Graduate School of Design.
- Burrough, Peter A. and Rachael A. McDonnell. *Principles of geographical information systems*. Oxford University Press, 1998.

5. Character encodings

- Humphries, Linda. *Cross platform character encoding profile*. Open Standards Board, 25 September 2013.

6. Licenses and public domain dedications

- *Guidelines on recommended standard licences, datasets and charging for the re-use of documents*. European Commission, 24 July 2014.

- Nečaský, Martin et al. “Licensing.” *Methodology for publishing datasets as open data*. Components Supporting the Open Data Exploitation (COMSODE), 31 July 2014.
- Davey, Francis et al. *Publisher’s Guide to Open Data Licensing*. Open Data Institute, 2013.
- Nugroho, Rininta Putri. *A comparison of open data policies in different countries: Lessons learned for an open data policy in Indonesia*. Master’s thesis. Delft University of Technology, 2013.
- Williams, Rebecca. *US Open Data Licensing*. Sunlight Foundation, 22 September 2014.
- Levine, Thomas. *Open data licensing*. 20 October 2014.

7. Licensing and rights metadata

- DCMI Usage Board. *DCMI Metadata Terms*. Dublin Core Metadata Initiative, 14 June 2012.
- *Creative Commons Rights Expression Language*. Creative Commons.
- Dodds, Leigh. *Open Data Rights Statement (ODRS) Vocabulary*. Open Data Institute, 29 July 2013.
- *Publisher’s Guide to the Open Data Rights Statement Vocabulary*. Open Data Institute.
- *Re-user’s Guide to the Open Data Rights Statement Vocabulary*. Open Data Institute.
- Rodríguez, Víctor. *Linked Data Rights*. Ontology Engineering Group, Universidad Politécnica de Madrid, 1 September 2014.
- García, Roberto. *Copyright Ontology*. Universitat Pompeu Fabra, January 2014.
- García, Roberto. “*RELs Overview*.” *A Semantic Web Approach to Digital Rights Management*. Diss. Universitat Pompeu Fabra, November 2005.
- McRoberts, Mo and Víctor Rodríguez Doncel. *Open Digital Rights Language (ODRL) Ontology*. W3C ODRL Community Group, 10 November 2014.
- Mazzini, Silvia et al. *License Model (LIMO) Vocabulary Specification*. 23 February 2013.
- Davis, Ian. *WAIVER: A vocabulary for waivers of rights*. 6 July 2009.
- Villata, Serena and Fabrien Gandon. *Licenses for Linked Open Data (L4LOD) Vocabulary Specification*. Inria, 10 May 2013.
- Cover, Robin. *MPEG Rights Expression Language*. Organization for the Advancement of Structured Information Standards, 19 June 2004.
- *The PRISM Usage Rights Namespace*. International Digital Enterprise Alliance, June 15 2009.
- Hoebelheinrich, Nancy J. *METSRights*. Library of Congress, 2 June 2004.

8. Resources covering several of the above areas

- Lee, Deirdre et al. *Open Data Ireland: Best Practice Handbook*. Insight Centre for Data Analytics, NUI Galway, May 2014.
- *Open Government Guide: Open government data*. Transparency and Accountability Initiative, 2014.
- Lee, Deirdre et al. "*Requirements by Challenge*." *Data on the Web Best Practices Use Cases & Requirements*. Data on the Web Best Practices Working Group, World Wide Web Consortium (W3C), 14 October 2014.
- Braunschweig, Katrin et al. *The State of Open Data–Limits of Current Open Data Platforms*. Proceedings of the 21st World Wide Web Conference 2012, Web Science Track at WWW'12, Lyon, France, April 16-20, 2012. ACM, 2012.
- *What do you need to do to get an Open Data Certificate?*. Open Data Institute.
- Romain, Pascal et al. *Open Data best practices*. OpQuast Open Quality Standards, 20 April 2010.
- Nečaský, Martin et al. *Methodology for publishing datasets as open data*. Components Supporting the Open Data Exploitation (COMSODE), 31 July 2014.
- Bloomberg, Michael and Rahul Merchant. "*City Standards*." *NYC OpenData Technical Standards Manual*. City of New York, September 2012.
- *Open Data Handbook*. New York State, 6 November 2013.

ANNEX 2

Additional resources

1. General-purpose data standards

- Brickley, Dan and Ramanathan V. Guha. *Schema.org*. W3C, 12 September 2014.
- DCMI Usage Board. *DCMI Metadata Terms*. Dublin Core Metadata Initiative, 14 June 2012.
- Cyganiak, Richard and Dave Reynolds. *The RDF Data Cube Vocabulary*. W3C, 16 January 2014.
- Several International Organization for Standardization (ISO), Open Geospatial Consortium (OGC) and OpenStreetMap (OSM) standards.

2. Domain-specific data standards

- *Open Budget Survey Tracker*. International Budget Partnership, 30 September 2014.
- *IATI Standard*. International Aid Transparency Initiative, 18 August 2014.
- *IATI Registry*. International Aid Transparency Initiative, 2014.
- *Aid Transparency Index 2014*. Publish What You Fund, 8 October 2014.
- *Guidance Note 3: Developing an Implementation Plan*. Construction Sector Transparency Initiative, October 2013.
- *EITI Standard*. Extractive Industries Transparency Initiative. 11 July 2013.
- *2013 Resource Governance Index*. Natural Resource Governance Institute, 15 May 2013.
- *2010 Revenue Watch Index*. Revenue Watch Institute, 6 October 2010.
- Reynolds, Dave. *The Organization Ontology*. W3C, 16 January 2014.
- Gregory, Arofan et al. *DDI Specification*. Data Documentation Initiative, 12 March 2014.
- Sufi, Shoaib and Brian Mathews. *Core Scientific Meta-Data Model (CSMD)*. CCLRC, August 2004.
- Nečaský, Martin et al. *Summary report on user requirements and techniques for data transformation, quality assessment, cleansing, data integration and intended data consumption of the selected datasets*. Components Supporting the Open Data Exploitation (COMSODE), 29 July 2014.

3. Key datasets

- "*Release of high value data*." *G8 Open Data Charter and Technical Annex*. 18 June 2013.
- Davies, Tim. "Listing of data categories included in Barometer survey." *Open Data Barometer: 2013 Global Report*. Open Data Institute and Web Foundation, 31 October 2013.
- *Global Open Data Index 2014*. Open Knowledge.

- Lee, Deirdre. *G8 Open Data Action Plan Datasets Comparison*. 4 August 2014.
- Nečaský, Martin et al. *List of selected datasets*. Components Supporting the Open Data Exploitation (COMSODE), 30 May 2014.
- "High-Value Data From Three Perspectives." *2010 Open Government Data Benchmark Study*. Socrata, 4 January 2011.
- *Top 10 Datasets*. Socrata.

4. Open data commitments

a. Bahl, Abhinav. *So What's In Those New OGP Action Plans, Anyway? 2014 Edition*. Open Government Partnership, 13 November 2014.

b. *What's in the New OGP National Action Plans?*. Open Government Partnership, November 2014.

Gaps and opportunities for standardization in OGP members' open data catalogs

For the Open Data Working Group of the Open Government Partnership

James McKinney
Stéphane Guidoin
Paulina Marczak

OpenNorth