Identifying recommended standards and best practices for open data

By Stéphane Guidoin (Open North), Paulina Marczak (Open North), Juan Pane (ILDA) and James McKinney (Open North)

Table of contents

1. Introduction

While reports suggest that open data can add trillions of economic value as well as support developing countries, the same sources describe significant barriers that must be overcome. The lack of standardization across jurisdictions is one major barrier; it makes discovering, accessing, using and integrating data cumbersome and expensive, above the expected return. A lack of knowledge about existing standards and a lack of guidance for their adoption and implementation contribute to this situation. This report seeks to address these lacks by outlining baseline standards and best practices for open data catalogs, while taking into account the differences between jurisdictions that make the global adoption and implementation of standards challenging.

This project is designed to support and reinforce the ongoing work of the Standards stream of the Open Data Working Group of the Open Government Partnership, whose ultimate deliverable is a similar document for OGP member countries along with guidance for the adoption and implementation of standards.

1.1. Research question

The research question follows from the Work Plan of the Standards stream of the OGP Open Data Working Group: “What baseline standards and best practices for open data should OGP members adopt?” To answer this question, this project investigates:

- What standards exist and what is their level of adoption and implementation by OGP countries?

- What challenges facing low- and middle-income countries impact the adoption or implementation of standards?

In this report, the term “standard” is understood in a broad sense to include many aspects of the publication, access and use of open data. In line with the Work Plan of the Standards stream of the OGP Open Data Working Group, this project evaluates standards in nine areas:

1. Licenses, dedications and metadata about licenses (e.g. Creative Commons)
2. Catalog and dataset metadata (e.g. DCAT)
3. Character encodings (e.g. UTF-8)
4. Data formats and serializations (e.g. CSV, Linked Data)
5. URL structures
6. Data delivery (e.g. API standards)
7. Definition and organization of datasets and distributions within data catalogs
8. General-purpose data standards (e.g. ISO 8601)
9. Domain-specific data standards (e.g. IATI)

2. Methodology

2.1. Overview

The research team completed the following steps:

- Read literature about: the needs of data consumers and data publishers as they relate to standards; the practices of data publishers which may qualify as *de facto* standards; the adoption and implementation of standards; and existing efforts to adopt the same standards across jurisdictions.
- Completed an inventory of *de jure* and *de facto* standards for the nine areas above, to develop a global view and identify gaps and overlaps in standardization.
- Evaluated the level of adoption and implementation of the standards by OGP countries, principally through the automated harvesting of information from open data catalogs. This evaluation aided in both the identification of interview candidates and the selection of recommended standards.
- Selected a diverse group of candidate interviewees from OGP governments, both with and without open data initiatives.
- Collected more specific information through interviews, which sought to understand interviewees’ choices with respect to standards as well as potential and real barriers to adoption and implementation. A questionnaire was designed and tested by the research team to guide this interview process.

2.2. Interviews

2.2.1. Interviewee selection

In order to understand both the supply and demand sides of open data, we interviewed two profiles in each country: national governments and data consumers (whether for-profit,

not-for-profit, or grassroots). For governments, the target interviewee was the open data "lead." For consumers, the target interviewee had experience using data from the government open data catalog: for example, civil society nonprofit organizations or software developers. In some cases, multiple members of the same organization were interviewed at the same time, given the wide range of interview questions.

A bias in interviewee selection is that all but two of the data consumers were members of nonprofit organizations, who were less interested in the commercial use of open data.

2.2.2. Country selection

Of the 65 OGP members as of December 2014, the 22 members classified as advanced economies by the IMF or as high income by the World Bank were removed from consideration in order to focus on the challenges facing low- and middle-income countries. Of the remaining 43, 20 had a national open data catalog. Their geographic distribution was:

- 3 Africa
- 4 Asia
- 5 Europe
- 8 Latin America

The 11 countries selected for interviews were:

- Africa
    - Ghana (only government interviewed)
    - Tunisia
- Asia
    - Georgia
    - Philippines
- Europe
    - Moldova
    - Ukraine
- Latin America
    - Chile
    - Guatemala (only government interviewed)
    - Panama
    - Paraguay
    - Peru (only consumer interviewed)

In Africa, three OGP countries have open data catalogs: Ghana, Kenya and Tunisia. A government contact in Kenya and a data consumer in Ghana could not be identified in time. Ghana and Tunisia were selected for interviews.

In Asia, four OGP countries have open data catalogs: Georgia, Indonesia, Jordan and the Philippines. Jordan did not respond to requests for interviews. Indonesia's open data initiative's

operations were interrupted by the country's 2014 presidential elections. Thus, Georgia and the Philippines were selected for interviews.

In Europe, five OGP countries have open data catalogs: Lithuania, Macedonia, Moldova, Romania and Ukraine. Moldova and Ukraine were selected for interviews, because the research team already had government and consumer contacts in these countries, and because identifying new contacts in other countries would risk delays.

Given that IDLA had better access in Latin America, we chose to expand the candidate pool for Latin America to include countries with planned open data catalogs, raising the number of candidates to 11. The three OGP countries in Latin America with a commitment in their most recent OGP Action Plan to create an open data catalog were: Guatemala, Panama and Peru. A government contact in Peru and a data consumer in Guatemala could not be identified in time. All three were selected for interviews. The eight OGP countries in Latin America with open data catalogs were: Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, Paraguay and Uruguay. Among Latin American countries, Chile's open data initiative was among the most advanced, while Paraguay's initiative was in its early stages. These two countries were therefore selected to collect responses from a broader range of experiences.

In terms of representativeness, this report's goal is primarily to identify the gaps and challenges with respect to the adoption and implementation of standards for open data, and to make recommendations for the most *urgent* gaps and challenges. The goal is not to make recommendations for the most *common* gaps and challenges. Thus, the comprehensiveness of the identified gaps and challenges is more important than their representativeness.

2.2.3. Interview process

Two test interviews were performed with the governments of Canada and Paraguay to test the interview process: the interview with Paraguay was retained in the results. Interviewees were sent an abbreviated version of the interview questions in advance. The complete version of the questions are provided as an appendix and discussed in the following section. The questions for governments and consumers were similar, but tailored to their role as publisher and consumer.

The interviews were semi-structured and conducted by VOIP and in-person in a few cases. Interviews in Latin America were conducted in Spanish and interviews in Tunisia were conducted in French; these interviews were later translated to English. All other interviews were conducted in English, with one country using an interpreter to translate. If permission was given, a recording of the interview was made. English interviews were conducted by two interviewers, with one primarily asking questions and the other primarily taking notes of the answers. Interviews took 1 to 1.5 hours in general. Some interviewees were sent follow-up questions after the interview, so that they may follow-up with an internal expert to answer the question from the interview.

3. Interview results

The structure of the following subsections follows the structure of the interviews, in which questions were grouped by subject.

3.1. Government interviews

The interviews gave an indication of a government's perception of a subject's importance, how advanced they were on the subject, whether they were aware of issues specific to the subject, whether they faced any challenges or limitations with respect to the subject, and whether they planned to make any changes in their handling of the subject.

3.1.1. Licenses, dedications and metadata about licenses

It should be noted that the interviewees were not their government's expert on data licensing; the following responses are indicative but not authoritative.

3.1.1.1. Is there an access to information law? Does it regulate how information should be published?

Eight interviewed countries had some form of access to information law. The two countries without such a law were in the process of introducing such a law. Respondents indicated that the law described **what** to publish but not **how** to publish. Most laws merely described that the information should be published proactively and regularly. One respondent indicated that their government was currently working on a guide to describe how to publish the information.

3.1.1.2. What does the default legal framework imply in terms of the consumer's requirements and permissions? *(For example, is there a law specifically about public sector information (PSI)? If not, is data (as compared to creative works) clearly protected by copyright, clearly not protected, or is it unclear? Are consumers, by default, prohibited from copying, modifying, and redistributing data? Is attribution, by default, required?)*

Six respondents indicated that consumers were, by default, not prohibited from copying, modifying or redistributing data, because data was not subject to copyright; in some cases, however, these statements were contradicted by the interview with the data consumer from the same country. Of the three respondents who indicated that data was not open by default, two said that copyright on data was a legal gray area, and one said that consumers were, by default, prohibited from copying, modifying or redistributing data.

Overall it appeared that both government and consumer interviewees had an unclear understanding of the legal framework for user rights relative to data reuse.

3.1.1.3. Do you consider there to be value in having a single, common license for all data on your open data catalog? *(What are the pros and cons? What are the challenges to overcome?)*

Most respondents indicated that there was value in having a single, common license, although some expressed a desire or need to use more restrictive licenses for some specific or sensitive datasets. Respondents reported difficulty in achieving agreement on the choice of license across departments, as departments had differing opinions on what was best. Other concerns were (1) ensuring that the license respected the law and (2) changing policies or laws – in cooperation with the legislature – to enable the open licensing of data. Varied organizational changes were also reported as being needed to achieve a single license.

3.1.1.4. Who selects the license for a dataset? *(Are there restrictions on the choice of license? Must all data be licensed? Is there a primary, preferred or recommended license? Were Creative Commons or Open Data Commons licenses considered?)*

This question, and most follow-up questions, were not relevant to countries whose national catalog used no licenses or whose catalog used one license. For the other five, each government department adding datasets to the catalog chose the licenses for its datasets.

In terms of license choice restrictions, one respondent indicated that departments must select from a list of licenses. Another government had a specific license for salary information. One respondent indicated that publishers must specify a license in order to publish a dataset.

Four of the respondents indicated that their policies required all datasets to be licensed.

Three respondents indicated that their catalog's primary or preferred license was an international license, such as a Creative Commons or Open Data Commons license. Two respondents reported using a single license for all datasets.

When developing their licensing policies, five respondents had considered, at a minimum, a Creative Commons license as a primary or preferred license. One respondent suggested an international license would be appropriate for software, but not data. Another was not aware of Creative Commons, and two had not considered any international licenses.

The follow-up question about whether international licenses were considered for use was relevant to all catalogs using licenses. Of the five respondents whose governments had considered or are considering Creative Commons or Open Data Commons licenses, two indicated that these licenses sufficiently fulfilled their licensing requirements. In one case, international licenses were considered due to a lack of resources to author a country-specific license. One respondent's primary concern about Creative Commons licenses was whether it would comply with the country's legal framework.

3.1.1.5. What are the challenges to adopting an international license?

Challenges included:

- the license's compliance with the country's legal framework;
- the availability of the license in the country's language(s);
- whether the license has unclear or misleading terms of use; and
- a desire for ownership of a custom, country-specific license.

3.1.1.6. Are departments given guidance in the choice of license? *(What guidance? What license-related issues require more guidance? Is there any central oversight?)*

For the three catalogs reported as using multiple licenses, two did not provide guidance, but one is working on guidance. The third respondent referred to a specific public government document providing guidance. No respondent reported central oversight of license selection.

3.1.1.7. Are there any challenges with respect to licensing?

Challenges include:

- making an international license comply with the country's legal framework;
- convincing the administration on the value or importance of licensing;
- licensing sensitive data;
- developing a custom, country-specific license;
- achieving consistent licensing across national government departments; and
- achieving consistent licensing across sub-national governments.

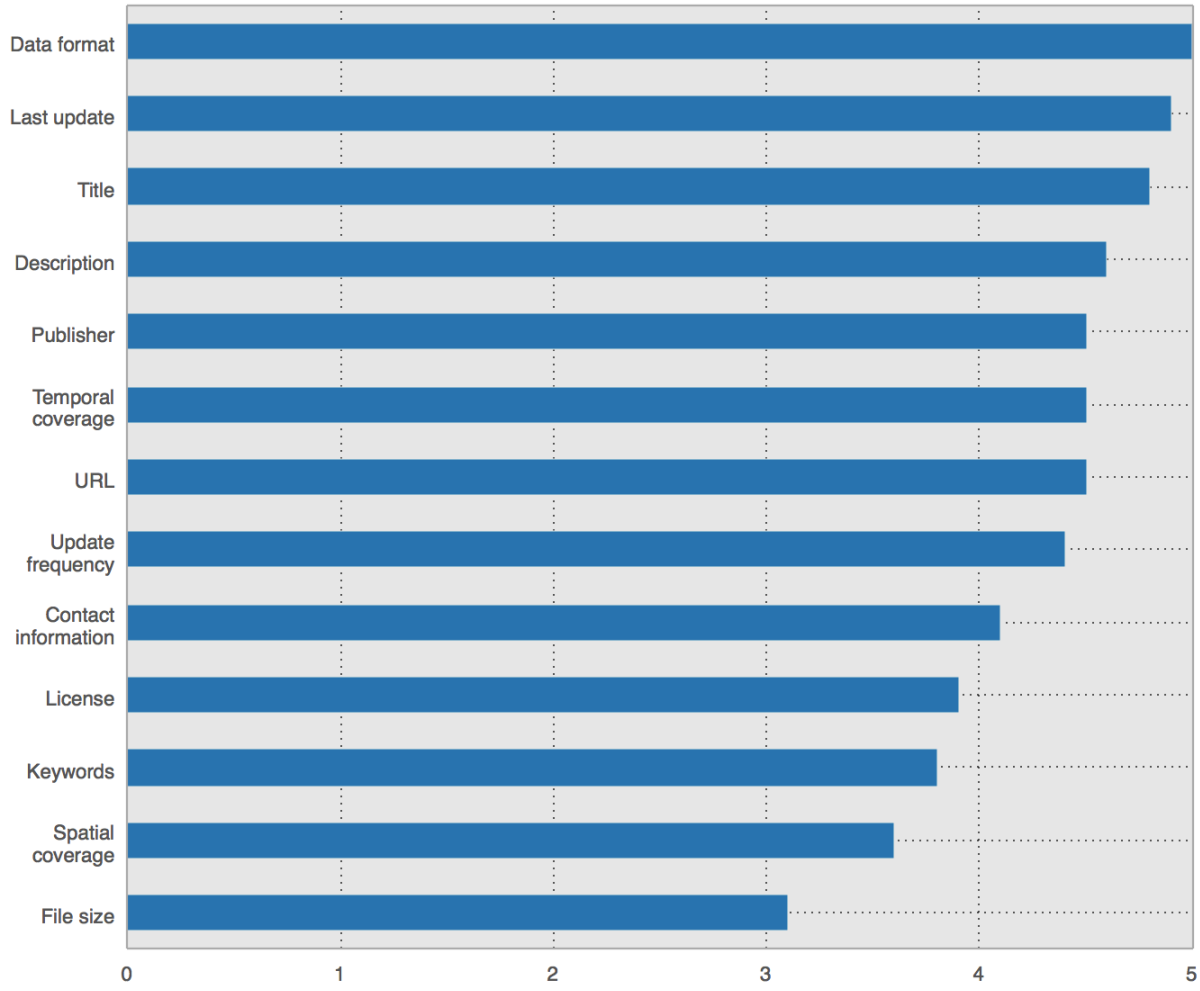
3.1.1.8. How is the metadata licensed?

Four respondents reported licensing metadata, and another four reported not licensing metadata. One respondent was unsure. One respondent specifically indicated that metadata is licensed under the Creative Commons Attribution-ShareAlike 4.0 license (CC-BY-SA 4.0).

3.1.2. Catalog and dataset metadata

3.1.2.1. What metadata do you consider most important? *(Why do you prioritize areas this way?)*

On a scale from 1 to 5, with 5 being most important, the weighted average of responses for each metadata element was as follows:



Respondents reported that the metadata elements with highest importance were those that:

- were required by operating procedures;
- were considered fundamental to the effective use by consumers of the catalog;
- most improved the accessibility and discoverability of the data.

3.1.2.2. Do you consider there to be value in having a single, common metadata scheme?

All respondents indicated that there is value in having a single, common metadata scheme. Reasons included:

- a better user experience for data consumers, who can predictably find the same information in the same place across all datasets;
- a better user experience for data publishers, who only need to become familiar with one web form for publishing metadata;
- efficiencies produced by adopting the same scheme across departments;
- easier automation of the aggregation/federation/harvesting of datasets;

- easier use of metadata by software; and
- better interoperability in general.

One respondent, however, questioned the feasibility of adopting a single, common scheme.

3.1.2.3. Did you consider international standards like DCAT/ISO 19115? *(What are the challenges to adopting such a scheme?)*

Six governments considered using international standards for metadata. Some governments adopted the World Wide Web Consortium's (W3C) Data Catalog Vocabulary (DCAT) or adopted a scheme similar to or based on DCAT, like Schema.org's Dataset type or the US Project Open Data Metadata Schema. Others plan to adopt DCAT. One respondent explained that, if they use the CKAN catalog software, then they use CKAN's metadata scheme.

Three governments had not yet considered standard metadata schemes, because:
- they were first concerned with bringing awareness to open data;
- they had not yet set up an inter-departmental working group to address such issues;
- they follow the practices of other catalogs.

The primary challenge reported (if any challenge was reported) was a lack of knowledge of standards and tools for metadata schemes.

3.1.2.4. Did you consider using RDFa or microdata on the web pages of datasets?

Two catalogs used RDFa or microdata, and one is considering these for the future. The other respondents explained that they:
- were not far enough along in their open data initiative to consider these technologies;
- did not have enough users to make these worthwhile;
- did not have the necessary technical knowledge;
- needed to address baseline standards first.

3.1.2.5. Are certain metadata elements required? Are controlled vocabularies used for some metadata elements/keywords?

Six respondents indicated that their catalog had required metadata elements; the four others indicated that theirs did not.

Nine respondents reported that their catalog did not use controlled vocabularies. One said they would appreciate help in implementing controlled vocabularies.

3.1.2.6. Who fills in metadata? *(Is it the publisher? Are they given guidance? If so, what kind, and which metadata issues require more guidance? Is there central validation on the completeness and accuracy of the metadata?)*

Seven respondents indicated that metadata was the responsibility of the publisher. Two others entered the metadata after the publisher submitted the data. One was not yet focusing on metadata. Four indicated that their catalog had quality assurance steps to ensure the accuracy of the metadata.

Five respondents indicated that their government gave guidance in the form of training, user guides (sometimes on the website itself) and technical meetings.

Five respondents indicated that their catalog had central validation – for example, a catalog administrator who would check the metadata before publishing – while the others expected the publishers to validate the metadata of their datasets.

3.1.2.7. Any other challenges in terms of metadata worth mentioning? *(Shortcomings in software tools?)*

A reported challenge was finding or developing a flexible metadata scheme that could be used across multiple institutions with few compromises by those institutions.

One respondent explained that some agencies would forget to submit metadata or would provide too much metadata, and that time was wasted in the back and forth between publishers and quality assurance. Adding a large number of datasets was challenging, because metadata had to be manually entered for each.

One respondent indicated that changing the set of metadata elements required coding or adapting a software extension, whereas editing a configuration file would have been easier**.**

One respondent reported that producing quality metadata was a challenge given their resources, in particular the challenge of editing dataset descriptions to be clear and concise.

3.1.3. Character encodings

3.1.3.1. Is character encoding an issue you have addressed on your open data catalog? *(Do you consider there being value in a single character encoding for all datasets? What are the challenges to adopting a single character encoding?)*

Five respondents reported addressing character encoding. One respondent indicated that a few departments used different encodings than the standard UTF-8, which led to information loss when that data was incorrectly translated into other encodings by tools.

One respondent indicated that information is manually converted to the proper encoding when possible. All respondents having addressed encoding believed there was value in a single character encoding, specifically UTF-8.

Respondents reported preferring UTF-8, because it:

- encodes all possible characters in Unicode;
- better ensures the correct display of non-ASCII characters;
- makes integrating multiple datasets easier;
- improves the interoperability between software tools;
- simplifies the development of open data applications;
- makes it easier for consumers to use the data.

Challenges indicated by one respondent included:

- educating publishers about encoding issues;
- software that cannot export data as UTF-8;
- complications in data harvesting due to encoding issues.

3.1.3.2. Who selects the encoding for a dataset? *(Are they given any guidance in character encoding? What guidance is given? What encoding issues require more guidance? Is there any central oversight on the encodings in use?)*

Three respondents reported that the publisher determined the encoding of a dataset, in which case publishers were generally encouraged to use UTF-8. Two other respondents reported that they set the encoding themselves, which can require re-encoding the data. Multiple respondents planned to give guidance on encoding. No respondent reported central oversight of encoding.

Catalogs were not equally affected by encoding issues: The Georgian alphabet, according to the respondent, can only be corrected displayed in UTF-8; for other alphabets, 4-5 different encodings were available, thus increasing the rate of encoding issues.

3.1.3.3. Are there any other challenges in terms of encoding worth mentioning? *(Have any tools made it difficult to control the encoding?)*

One respondent indicated that the primary encoding issue was due to different databases using different encodings by default. Another reported encoding issues due to the lack of metadata about data's encoding, and reported CP1125 as the most troublesome encoding. Another reported encoding issues due to the lack of support by proprietary tools for some encodings, and the challenge of developing and implementing an interoperability framework in which encoding would no longer be an issue. One respondent stressed the need for technical training and the need to achieve consensus on which encoding to use.

3.1.4. Data formats and serializations

3.1.4.1. How are the data formats for a dataset selected?

One respondent indicated that specific formats were encouraged in training sessions, but that the publisher made the choice of format. Seven respondents reported that datasets were generally published in the same format used by publishers internally (e.g. Microsoft Excel). Others base the choice of format on international practices. One respondent indicated that they urged publishers to publish the data in a “raw” format. One respondent had data format guidance which must be followed.

3.1.4.2. Does the open data initiative target specific formats for release, or are publishers free to use a variety of formats? And for geospatial data?

Seven respondents reported giving guidance, specifically encouraging the use of machine-readable and/or open formats like CSV or XML. About half of the respondents reported that they targeted a specific geospatial format for release, like Esri shapefile or KML. For one respondent, the current priority was to simply publish datasets, in any format. Three respondents indicated that there was, or will be, a document describing and recommending appropriate data formats.

3.1.4.3. Who selects the data format?

In most cases, the publisher selected the format. One respondent reported plans for the catalog to convert datasets to additional formats on-demand.

3.1.4.4. Are departments and agencies given any guidance in data formats? *(What guidance is given? What data format issues require more guidance?)*

Six governments had guidance on data formats. One respondent indicated that their guidance was based on a broader data standards document, which was to be consulted before the dataset was published. Three were planning on providing guidance.

3.1.4.5. Is there any central oversight of the data formats in use?

About half of the respondents reported central oversight. One government monitored and counted the number of datasets using each format, while three reviewed each new dataset and asked publishers to change the format if the submitted format was not acceptable.

3.1.4.6. Any other challenges in terms of data formats worth mentioning?

Two respondents wanted more feedback from users on what formats to use. Five described the challenge of the source data not being available in a machine-readable format, e.g. paper documents that were scanned to PDF without optical character recognition, or data being shared internally as PDF only. Another described the difficulty of balancing the accessibility of formats (e.g. CSV) with their expressiveness (e.g. XML).

3.1.4.7. Is your support for different formats limited by the tools used? *(Do tools cause problems with respect to data formats?)*

Eight respondents reported no such limitations. Of the other two, one reported that if the catalog software were capable of converting between formats, then they would only need to be responsible for a single, master format, saving time and resources. Another respondent reported issues due to the different implementations of the CSV format by different software, e.g. using a semicolon as the field delimiter instead of a comma, and due to a lack of technical assistance and budget capacity for training on data formats.

3.1.5. URL structures and data delivery

3.1.5.1. Is there a URL design scheme or guideline followed? *(What is it? Are the schemes followed for catalogs, datasets, data files, and/or metadata?)*

Seven respondents reported following a URL design scheme. In some cases, the scheme as determined by the catalog software (e.g. CKAN). In others, a standards document described the URL design scheme. The scheme for a dataset URL was either based on the dataset's title, a unique key generated by CKAN, or on the guidance in a standards document.

3.1.5.2. Does the catalog offer additional data delivery options besides direct download (e.g. APIs)? *(Are these more likely to be available for certain data (e.g. geospatial data)?)*

Seven catalogs offered an API; three reported using CKAN and offering its default API and datastore API. One respondent reported using Junar and offering its default API. For the catalogs not offering an API, the respondents were interested in implementing an API, and in some cases developing a standard for an API. Most respondents whose catalogs offered an API reported that API access was not more likely to be available for specific datasets.

3.1.6. Definition and organization of datasets and distributions within data catalogs

3.1.6.1. Could you provide examples of datasets with multiple files? *(How does the catalog typically structure datasets with multiple data files? Is the structure consistently followed? Is there any oversight?)*

Four respondents reported one file per dataset. Six respondents reported multiple files per dataset. One respondent indicated that files were grouped by topic; if the number of files per group were still too large, files were grouped by agency and then again by political division. In another case, files were grouped by keywords or as time series.

In terms of respecting a unique structure for files in datasets, responses varied. One respondent indicated the structure respected the Data Catalog Vocabulary's (DCAT) structure. Another

respondent indicated that a consistent structure is manually enforced. Two respondents reported that a consistent structure was enforced by the software used. One respondent reported that the structure of a dataset was centrally validated before publication.

3.1.6.2. When you publish multiple files in one dataset, how do you decide to include those files in that dataset, instead of across multiple datasets?

Two respondents had not established a consistent methodology for grouping files and left the decision to the publisher. Of the four catalogs that follow a structure, three grouped files as time series; for example, each file in a dataset represented one month of data. One respondent reported that the catalog grouped multiple serializations of the same data within a given year.

3.1.6.3. Does your primary data catalog aggregate from other government catalogs? *(How do you aggregate the data?)*

Two respondents reported that the primary catalog did not aggregate. Three catalogs aggregated manually, either by manually adding a copy of the agencies' datasets to the central catalog or by manually linking to the agencies' datasets from the central catalog. One respondent reported aggregating using an automated harvester that crawled public institutions' websites for data.json files as specified by Project Open Data and for HTML annotations using the Schema.org types DataCatalog, Dataset or DataDownload.

3.1.7. General-purpose data standards

3.1.7.1. Does the government's open data initiative adopt any general-purpose standards for specific types of data, like ISO 8601? *(Are departments given guidance for such standards? If so, what guidance is given? What issues require more guidance? Are the standards consistently followed? Is there any oversight?)*

Seven respondents reported not using any general-purpose standards. One indicated that such standards were described in a document but had not yet been implemented. Another indicated that they were presently more focused on quantity than quality, but that standards would rise in priority as more datasets were published. One reported using ISO 8601 for date formats, given that CKAN used ISO 8601.

Most respondents did not give any guidance. One gave guidance but could not enforce the guidance. One was considering including guidance in an upcoming standards document.

3.1.8. Domain-specific data standards

3.1.8.1. Does the government's open data initiative adopt any domain-specific standards for specific datasets? *(Are departments given guidance for such standards? If so, what*

guidance is given? What issues require more guidance? Are the standards consistently followed? Is there any oversight?)

One respondent planned to convert aid data to the International Aid Transparency Initiative (IATI) standard, after having been recommended the standard by multilateral organizations. No other catalog was reported as using domain-specific standards, although one respondent indicated that each publisher evaluated whether to use domain-specific standards. One respondent was considering adopting domain-specific standards. One respondent indicated that domain-specific standards may be used by departments, but that the central catalog's maintainers were not made aware of such use.

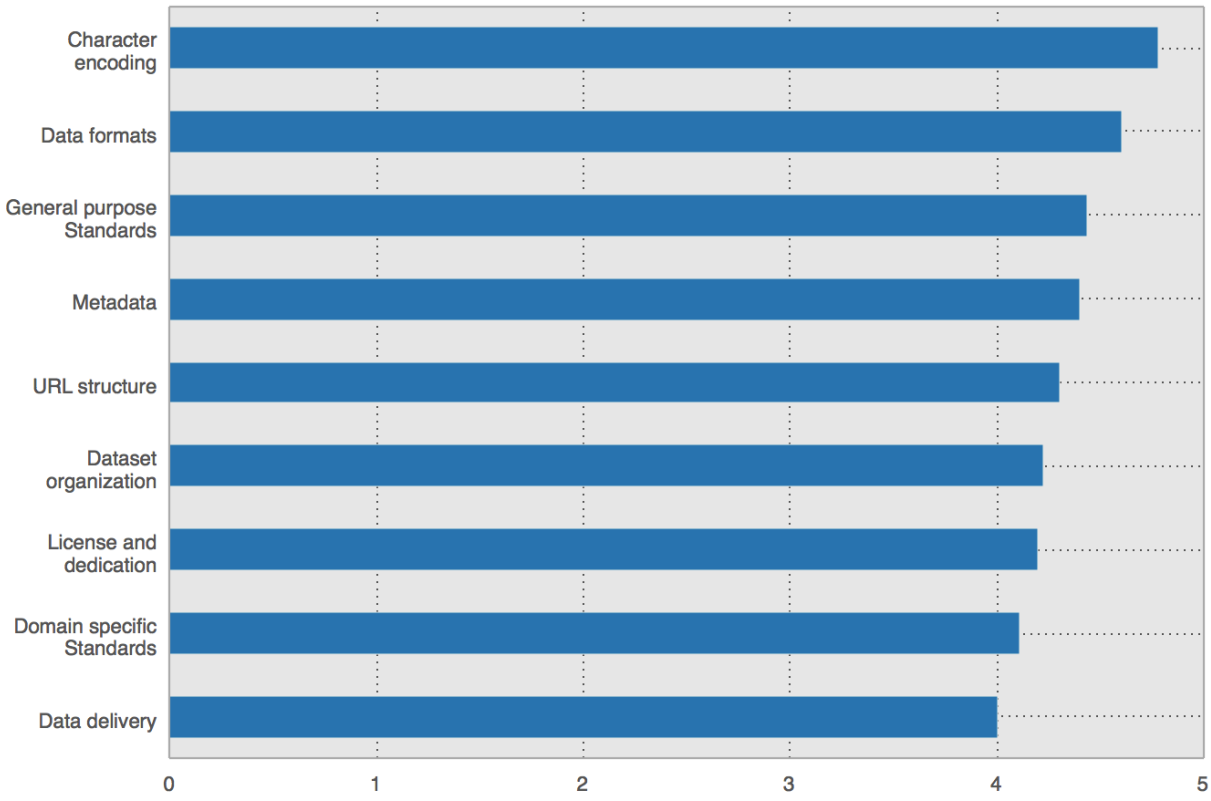
3.1.9. Process

3.1.9.1. Which government body, if any, is in charge of defining procedures for publishing open data (e.g. selection of license, format, encoding, etc.)?

Most have a specific department in charge. One respondent indicated that a specific federal government document described procedures and methodologies, and that this authoritative document was subsequently adapted by departments.

3.1.9.2. What area of the OD initiative do you think needs the most standardization? *(Why do you prioritize the areas this way?)*

On a scale from 1 to 5, with 5 being most important, the weighted average of responses for each area was as follows:



One respondent explained that the higher ranked areas were more important to catalyzing the growth of the open data initiative. Another explained that the more data-oriented areas were prioritized (e.g. data formats, character encodings, and data standards were prioritized over metadata, catalog structure, and data delivery), because data was ultimately the focus of an open data catalog; the only non-data-oriented area that was prioritized was licensing. Another prioritized areas according to how much each would contribute to the efficient management of the catalog. Another prioritized areas that would most ease the release and use of datasets.

3.1.9.3. If there is a standardization initiative, does it take place in coordination with other jurisdictions?

Five respondents were coordinated with other jurisdictions; the others were not. Of those that were, four were coordinating with the executive branch/central government. Two of these were also coordinating with autonomous institutions. Two were coordinating with local governments (e.g. municipalities). One was consulting with the EU on best practices to adopt. One had an extensive interoperability framework, developed in coordination with judicial, executive, and other government corporations, which included interoperability standards used by other countries.

Two respondents indicated that the open data initiative was just starting, and that they therefore had not consulted with other organizations.

3.2. Consumer

Nine data consumers were interviewed. The interviews gave an indication of a data consumer's perception of and justification for a subject's importance and collected general feedback on the subject.

3.2.1. Context

3.2.1.1. How would you describe your level of technical knowledge of open data?

Some rated their knowledge as expert-level. Some reported that they were software developers and were familiar with the technical aspects of open data. Others' knowledge was more focused on the principles and benefits of open data and on the needs of users.

3.2.1.2. How did open data start in your country? *(Was civil society involved?)*

For at least five respondents, civil society's involvement in open data initiatives was recent. According to four respondents, their country's open data initiatives were government-led. Two other respondents reported open data initiatives led by civil society or educational institutions, intended as advocacy tools to influence the government. One respondent reported that the national open data initiative was a combined effort between government and civil society.

3.2.1.3. What has been civil society's involvement in open data?

Four respondents described its involvement as fulfilling an advocacy or activist role, primarily promoting the availability of additional datasets. One respondent indicated that civil society requested data from the government, negotiated with the government on the prioritization of datasets, and provided feedback or criticism of the government's initiative. Some indicated that civil society had developed websites or applications using open data, organized open data hackathons, or organized data training workshops.

3.2.1.4. What is the government's open data initiative's attitude towards civil society?

In terms of the government's supportive attitude towards civil society, the answers were:

- 4 very supportive
- 2 supportive
- 1 neutral
- 2 unsupportive
- 0 very unsupportive

In term of its confrontational attitude, the answers were:

- 2 very non-confrontational
- 4 non-confrontational

- 2 neutral
- 0 confrontational
- 1 very confrontational

In terms of its collaborative attitude, answers were:

- 2 very collaborative
- 3 collaborative
- 3 neutral
- 0 uncollaborative
- 1 very uncollaborative

3.2.1.5. Are attitudes different at a personal level?

Six respondents indicated that attitudes were different at a personal level. Five respondents indicated that, at a personal level, attitudes were more supportive and collaborative, but one indicated that attitudes at a personal level were more negative in general.

3.2.2. Licenses and dedications

3.2.2.1. Would you consider the initiative's licensing to be simple and clear?

For the national catalogs using licenses, one respondent considered the licensing to be simple but unclear. Another respondent said the licensing was not explicit. Two respondents could not comment on licensing, one of whom was only concerned with whether the data could be used for free.

3.2.2.2. What are the major challenges as a consumer with respect to the initiative's licensing?

Respondents agreed with each challenge as follows:

- Some or all datasets are not licensed: 4
- The initiative uses too many licenses: 1
- The license does not grant sufficient rights: 1
- The license's obligations are too demanding: 1
- The license's warranty, liability, or indemnity clauses dissuade use: 1

Other challenges mentioned were:

- A lack of licensing;
- The license text is unavailable;
- The rights and obligations are unclear;
- Technical norms exist to regulate the licensing of some but not all government data.

3.2.2.3. Should the initiative adopt a common license for all datasets?

All but one respondent agreed that a single, common license should be used, because:

- it would follow the best practices of other countries;
- it would simplify the legal aspects of open data.

One respondent agreed as long as the licensing were not restrictive.

One respondent disagreed, explaining that they would rather have licenses that corresponded to the type of data, and that a fee-based license would be acceptable for APIs once users exceeded a threshold number of requests.

Two respondents explained that the challenge to implementing a single, common license was compliance with the country's legal framework.

3.2.2.4. Should the initiative adopt an international license for some or all datasets?

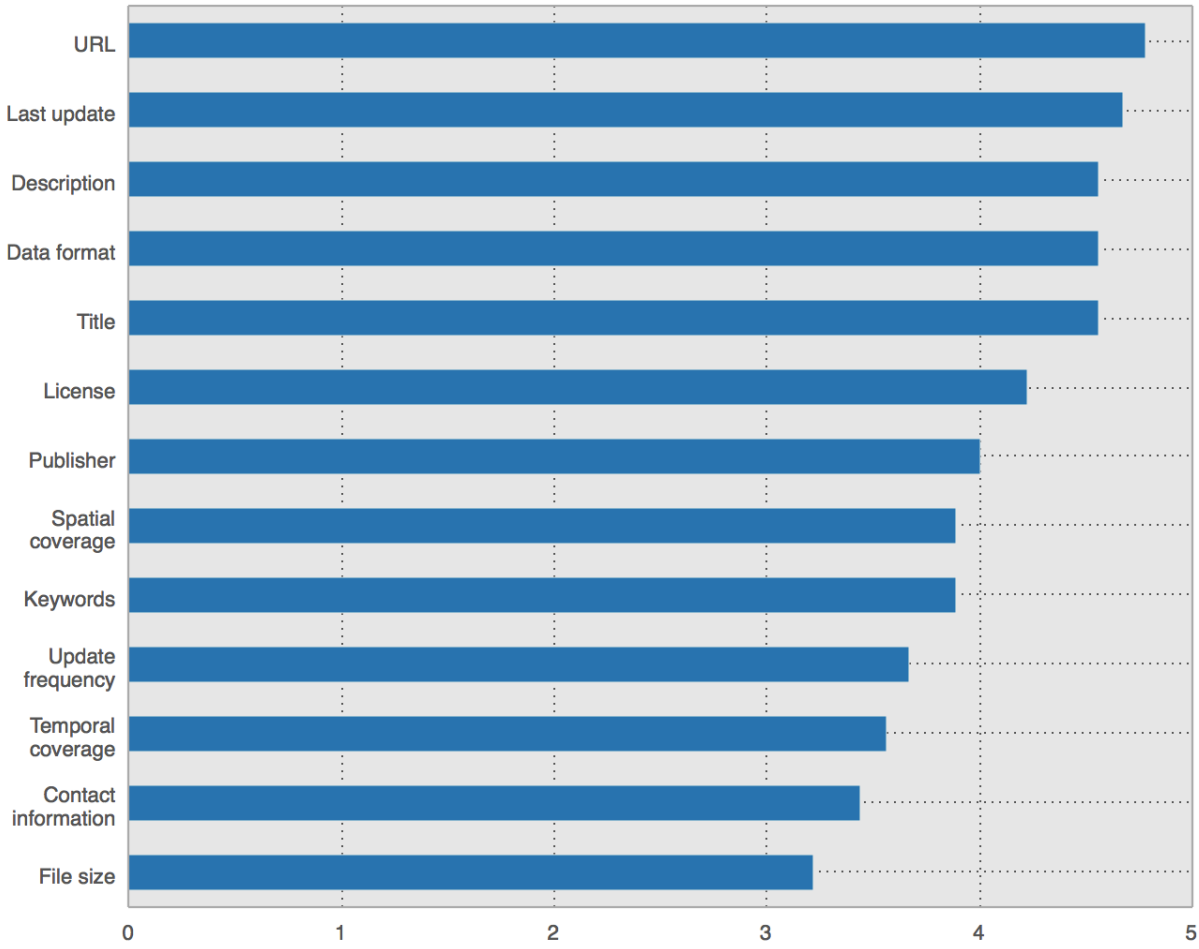
Almost all respondents agreed that an international license should be adopted for at least some datasets. One recommended Creative Commons licenses in particular, because users are more familiar with them and would readily use data licensed under them. Many respondents indicated that awareness of international licenses was high among users.

One respondent expressed concern that their country's legal framework would need to catch up to the policies of the EU in order to consider an international license.

3.2.3. Catalog and dataset metadata

3.2.3.1. What metadata do you consider to be most important? *(Why do you prioritize the areas this way?)*

On a scale from 1 to 5, with 5 being most important, the weighted average of responses for each metadata element was as follows:



Respondents reported that the metadata elements with highest importance were those that:

- were necessary to use the data (e.g. URL);
- described the dataset (e.g. title and description);
- made it easier to keep downloaded files up-to-date (e.g. last updated);
- described the file (e.g. data format);
- made it easier to filter datasets (e.g. spatial and temporal coverage).

Other strategies for prioritizing elements were according to:

- their their utility to both users who interact with the data via user interfaces and developers who interact with the data via software;
- how much each contributed to making the data easier to process automatically;
- how much each contributed to the dataset's discoverability or searchability;
- those which were frequently missing from the metadata records;
- those which made it easier to evaluate the data's quality, accuracy, or timeliness.

3.2.3.2. What are the major challenges as a consumer with respect to the initiative's metadata?

Seven respondents indicated that the metadata's completeness, quality and/or accuracy were lacking. One respondent further explained that they needed to be able to trust and refer to the metadata consistently; for example, some datasets with identical titles did not have identical files, making it hard to track which is which. One respondent raised the challenge of determining whether a dataset was being maintained, and pointed to the lack of last updated and update frequency metadata. Another concern was the lack of standards for naming and identifying entities. One respondent reported no challenges, explaining that the catalog's metadata was very good and exceeded requirements.

3.2.3.3. Should the initiative adopt a machine-readable format for all metadata?

All respondents agreed that the initiative should, if it were not already the case. Three explained that machine-readable metadata made it easier to analyze the metadata; made it easier to filter, discover and use the data; and helped software developers building tools using open data. One respondent added that the initiative should try to provide the most user-friendly metadata format.

3.2.3.4. Should the initiative adopt an international format for all metadata?

Three respondents were not previously aware of international formats for metadata, specifically DCAT, but all but one respondent were supportive of adopting an international format. Reasons for adopting an international format included making it easier to automate harvesting, making it easier to merge metadata from multiple catalogs, and promoting interoperability in general. One respondent who did not agree explained that adopting an international format could be limiting, when compared to creating a bespoke format.

3.2.4. Character encodings

3.2.4.1. What character encoding do you prefer in your work?

The preferred character encodings were:

- UTF-8: 6
- Windows-1252: 1 (specifically for Microsoft Excel files)
- ISO-8859-1: 0

No other encodings were mentioned.

One respondent had no preference, as long as a single encoding was used consistently and explicitly everywhere.

3.2.4.2. What are the major challenges as a consumer with respect to the initiative's encoding?

The major challenges were:

- Unknown encoding: 5
- Improper encoding: 2
- Inconsistent encoding: 1
- Unpopular encoding: 0

One respondent described the difficulty of working with CSVs in UTF-8 using Microsoft Excel. No other challenges were mentioned.

3.2.4.3. Should the initiative adopt a common encoding for all datasets?

Seven respondents reported that the encoding should be the same across all datasets. One respondent explained that this would make it easier to migrate data between databases. One said that the use of a common encoding should be enforced, if possible by law. Two respondents disagreed: one explained that Microsoft Excel has poor support for CSVs in UTF-8, and the other was concerned that achieving a common encoding would be difficult, and that a more achievable goal would be to specify the encoding in the metadata.

3.2.5. Data formats and serializations

3.2.5.1. What formats do you prefer for tabular data?

The preferred formats were:

- CSV: 6
- Microsoft Excel: 4
- OpenDocument Spreadsheet: 1

One respondent preferred Google Sheets. No other formats were mentioned.

Two respondents emphasized the importance of using a common CSV dialect (e.g. always using a comma as the field delimiter). One respondent explained that the preferred format depended on the particular dataset.

3.2.5.2. What formats do you prefer for geospatial data?

The preferred formats were:

- GeoJSON: 6
- ESRI shapefile: 2
- MapInfo: 1
- TopoJSON: 1
- KML: 0
- GML: 0

One respondent preferred OpenStreetMap. No other formats were mentioned.

Two respondents were unfamiliar with geospatial data and did not respond.

3.2.5.3. What are the major challenges as a consumer with respect to the initiative's formats?

The major challenges were:

- Non-machine-readable formats : 6
- Proprietary formats: 4
- Unpopular formats: 2

One respondent pointed to a disconnect between the government's open data catalog, which releases data online in machine-readable formats, and its access to information process, which releases information offline on paper.

One respondent explained that, given the amount of information that is stored on paper and the challenges of accessing that information, publishing datasets in non-machine-processable formats like PDF is acceptable, when compared to offline paper copies.

3.2.5.4. Should the initiative focus on providing specific formats for all datasets? *(e.g. instead of having some datasets only available as CSV and others only available as Excel)*

Eight respondents agreed, in particular on the provision of CSV and Excel for tabular data and JSON for other types of data. One respondent explained that the provision of data in multiple formats required data consumers to have a higher degree of technical knowledge, which is not always present in civil society.

One respondent explained that their priority was the release of data, in any format. One respondent described that the choice of format depended on the type of data.

3.2.6. Data delivery

3.2.6.1. How important is API access compared to direct download?

API access is considered very important, with a weighted average of 4.63 out of 5.

One respondent preferred API access for frequently updated data. Another respondent described API access as enabling effective, continuous and automated monitoring of government activity. Another respondent felt that, given the amount of effort spent by the government on the API, API access must be more important than direct download.

3.2.7. Definition and organization of datasets and distributions within data catalogs

3.2.7.1. Would you consider the catalog to be well organized? *(What are the major challenges with respect to the catalog's organization?)*

Three respondents considered the catalog to be reasonably well organized, but none considered it to be very well organized. One respondent described the catalog's simple structure as meeting basic functional needs.

Four respondents considered the catalog to be unorganized and expressed difficulty in finding and accessing data; for example, some datasets' files linked to the home pages of statistics agencies, instead of to specific pages or files. The lack of structure was reported as creating user experience issues.

3.2.7.2. Does the catalog link to its original source? Is it important that it does?

Two respondents reported that it did, and three reported that it did not.

The reported importance of linking to original sources was:

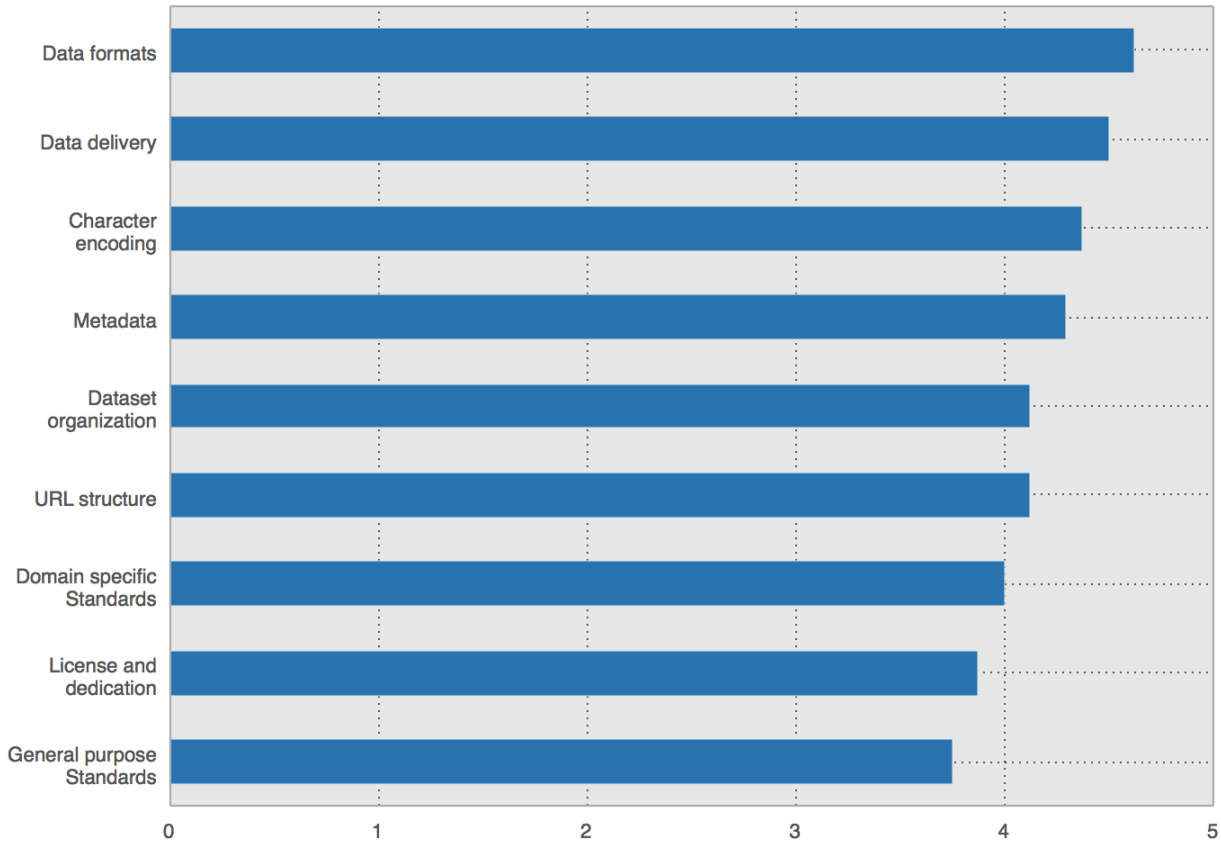
- Very important: 2
- Fairly important: 1
- Important: 1
- Slightly important: 1

One respondent reported that files were manually uploaded to the catalog, so the link to the original source was lost in the process.

3.2.8. Wrap-up

3.2.8.1. What area of the open data initiative do you think most needs standardization? *(Why do you prioritize the areas this way?)*

On a scale from 1 to 5, with 5 being most important, the weighted average of responses for each area was as follows:



Many respondents had a common concern for application developers, prioritizing areas that most affected applications developers, e.g. areas that made it easier to extract, transform and load the data into databases. One respondent specifically mentioned API access and machine-readable formats as being critical to application developers.

4. Draft recommendations

The research team proposes draft recommendations below for validation by stakeholders. The purpose of the recommendations is to improve open data initiatives by solving the most important challenges identified by this project. The recommendations are based on the above interview results and on the "Gaps and opportunities for standardization in OGP members' open data catalogs" report and appendices, published as the first deliverable of the Standards stream of the Open Government Partnership's Open Data Working Group.

The recommendations are intended to be realistic, by taking into account the needs and capacities of publishers and consumers: for example, the common challenge of limited resources. The recommendations are not intended to describe some ideal world scenario. Although many interviewed governments have committed significant resources, including full-time staff, to open data – something that is difficult for many jurisdictions to achieve – these resources are still insufficient to achieve the results they desire.

As such, the recommendations intend to:

1. Provide clear and specific guidance in order to limit the additional effort that implementers must commit to understand, evaluate and implement a recommendation;
2. Set targets that are achievable by a large portion of publishers and with the maximum impact for consumers.

In designing the recommendations, we recognize that governments are not solely responsible for the quality of their open data initiatives with respect to open data standards: software providers, whether they provide data catalog software or other data tools, have a responsibility to better implement and support standards.

Priority of recommendations

Each recommendation is given a priority:

- **Highly recommended:** These recommendations form the baseline that open data initiatives should meet early on, to ensure minimum accessibility to data. The recommendations tend to be easier to implement.
- **Recommended:** These recommendations are the most numerous and set targets for a well structured and standardized open data initiative that will make data easier for consumers to access and use.
- **Nice to have:** These recommendations are expected to be pursued once an open data initiative is advanced; are more demanding to implement; and/or meet the needs of a subset of users.

Recommendations at different levels of priority are compatible; an initiative may follow all recommendations concurrently. The recommendations are not mutually exclusive; for example, implementing a “highly recommended” item doesn’t mean changing the implementation of a “recommended” item.

It is possible to implement recommendations with lower priority before recommendations with higher priority, but doing so may benefit a subset of users at the expense of others.

Implementers of recommendations

As described above, governments are not solely responsible for implementing the recommendations: software authors share some responsibility. Each recommendation identifies types of stakeholders that should be involved in implementing a recommendation:

- **Government:** This project focused on national governments, but recommendations are likely relevant to any public body with an open data initiative.
- **Catalog software:** The authors of open data catalog software, like CKAN and Socrata.
- **Data tools:** The authors of software used in the authoring, editing, management and analysis of data, used by both publishers and consumers. Programs range from spreadsheet applications, to databases, to visualization tools.
- **Standardization body:** A national or international body responsible for developing and/or certifying specifications or processes as standards.

In some cases, multiple types of stakeholders are listed, in the order of greatest to least responsibility. Governments bear some responsibility for all recommendations' implementation (e.g. procuring appropriate catalog software and data tools), but if the responsibility is small, then government is not listed.

Beneficiaries of recommendations

The beneficiaries of a recommendation's implementation are listed to make clear who benefits. The labels for the beneficiaries are "personae," i.e. typical user profiles. A persona should not be interpreted as representing all entities associated with its label, but as an indication of a common set of needs and capacities among data consumers:

- **Civil society organizations** (CSO): has low or no technical capacities, is searching for data on specific topics, is concerned by the accessibility and interpretability of data.
- **Company:** has medium to high technical capacities, is searching for datasets with high value to some market, is concerned with the legal aspects of open data (e.g. liability clauses, commercial use), is concerned with scalability (e.g. ability to find similar data in other jurisdictions).
- **Researcher:** has medium to high technical capacities, wants to perform a large-scale analysis of the data and metadata, is searching for data on specific topics.
- **Government:** wants to aggregate data and/or metadata from multiple jurisdictions, wants to merge datasets from other jurisdictions.

Structure of recommendations

Each recommendation is described in one sentence, followed by its priority, implementers and beneficiaries. Implementation details are provided, along with a discussion of the motivations for the recommendation and its prioritization, based on this project and prior work.

4.1. Licenses, dedications, and metadata about licenses

In this section, it is understood that the relevant licenses and dedications are those that conform to the Open Definition.

Recommendation 1: If a catalog uses no licenses or dedications, the terms of use for the data must nonetheless be explained.	
Implemented by: **Government** Benefits: **Company, CSO**	✔✔✔ *Highly recommended*

Implementation

- The terms of use for the data should be accessible from the open data catalog's homepage, e.g. via a clearly-worded link in the primary or secondary navigation.
- If the choice of license or dedication is in progress, this work should be described along with the terms of use, and may link to other pages describing the work.

Discussion

Several government respondents were unsure whether licensing was necessary for data and were unsure of what the law implied when no license was used. Consumers shared the same uncertainties when using the data; companies and CSOs in particular may be discouraged from using open data without clear terms of use. To avoid uncertainty, the terms of use for the data should be made clear, even in the absence of explicit licensing or rights statements.

Several interviewed governments were working on their open data licensing and/or access to information laws; adding a page to the open data catalog to describe the terms of use can be done concurrently, without interfering with this work.

Recommendation 2: A human-readable statement of the applicable license, dedication or terms of use must be accessible from each dataset's landing page.	
Implemented by: **Government, Catalog software** Benefits: **CSO, Company, Researcher**	✔✔✔ *Highly recommended*

Implementation

- Each dataset's landing page should clearly state its license, dedication or terms of use, and should link to a page describing the license, dedication or terms of use. For example, a dataset's landing page may state that it is "licensed under the Creative Commons Public Domain Dedication (CC0 1.0 International)" and may link to the local language version of http://creativecommons.org/publicdomain/zero/1.0/.
- The open data catalog should have a page describing how the data in the catalog is licensed or dedicated in general and describing any relevant, general issues with respect to licensing and rights.

Discussion

Several interviewed consumers reported not finding any relevant information about rights and obligations on their national open data catalog and therefore were unsure of their rights and obligations. It should be easy for consumers to determine their rights and obligations with respect to using a dataset and to understand a catalog's general practices with respect to licensing and rights.

Many consumers have limited or no knowledge of copyright issues with respect to open data. It is therefore important for this information to be visible and accessible, so that consumers discover this information. Like with recommendation 1, a lack of clarity in terms of licensing and rights may discourage the use of open data.

Recommendation 3: For each dataset, a link to the applicable license, dedication or terms of use should be available in a machine-readable format.	
Implemented by: **Government, Catalog software** Benefits: **Company, CSO, Researcher**	✔✔ *Recommended*

Implementation

- Open data catalogs should provide for each dataset, in a machine-readable format, the name and official URL of the applicable license, dedication or terms of use.
- This information should be embedded in the dataset's machine-readable metadata: see recommendation 10.

Discussion

At a minimum, the URL of the applicable license, dedication or terms of use should be provided, to allow consumers and software to determine the rights and obligations. A name is more human-readable than a URL, so providing the name in addition to the URL is recommended. The name may be a translated.

The URL should be the official URL of the license or dedication, to make it easier for software to determine when the same license is used. For example, the link for the Creative Commons Public Domain Dedication (CC0 1.0 International) should be:

http://creativecommons.org/publicdomain/zero/1.0/

The link should not be for the dedication's page at opendefinition.org, clipol.org, or any other third-party website. Only its official page at creativecommons.org is authoritative. The link may be for an official translation of the dedication, for example:

http://creativecommons.org/publicdomain/zero/1.0/deed.pt_BR

Names and URLs are especially useful when the licenses or dedications are well-known (e.g. Creative Commons), because consumers can recognize the names or URLs and remember the associated rights and obligations.

Providing this information in a machine-readable format makes it easier for software to determine, monitor, analyze, federate and otherwise use this information.

Recommendation 4: The number of open data licenses and dedications used by a catalog should be limited, ideally to one.	
Implemented by: **Government, Catalog software** Benefits: **All**	✔✔ *Recommended*

Implementation

- The open data initiative should provide guidance to publishers on the choice of license /dedication and recommend a small number of preferred licenses/dedications.
- The number of open data licenses/dedications should be monitored, in order to limit their proliferation.
- The catalog software should restrict the choice of license/dedication to a list of approved licenses/dedications. (New licenses/dedications may still be approved.)
- The open data initiative should develop and implement a plan to reduce the number of licenses/dedications, preferably to one.

Discussion

Both government and consumer interviewees agreed that having a single, common license would significantly benefit the open data initiative. Limiting the number of licenses/dedications reduces the time and effort spent by consumers to read, understand, and evaluate the impact of legal texts on their use of open data.

Given the legal and administrative challenges to reducing the number of licenses/dedications to one – e.g. passing new legislation and harmonizing licensing policies across departments – it is acceptable to use more than one license/dedication. Reducing this number should not be pursued at the expense of reducing the rights granted to users or increasing the obligations required of users; for example, if adopting a single license would require introducing new obligations to a common license in order to satisfy one department's needs, it would be preferable to instead use two licenses, one for that department and another for the rest.

Recommendation 5: Governments may adopt an international license or dedication or may develop a common, regional license or dedication.	
Implemented by: **Government** Benefits: **All**	✔ *Nice to have*

Implementation options

- Governments should encourage publishers to use international licenses and dedications.
- Alternatively, governments should work with sub-national and/or neighboring governments to develop a common, regional license or dedication.

Discussion

Government and especially consumer interviewees agreed with the preference for international licenses/dedications. Licenses/dedications that are specific to one catalog contribute to the proliferation of licenses/dedications, which has negative consequences for the use of open data as discussed above.

International licenses/dedications like those of Creative Commons benefit from significant implementation experience, have supporting documentation that is accessible to users, have greater awareness among users, and are available in multiple languages, which benefits international users of a catalog's data.

If international licenses/dedications are not an option, governments should work to develop a regional license/dedication.

Recommendation 6: Metadata should be licensed or dedicated to the public domain.	
Implemented by: **Government, Catalog software** Benefits: **Researcher, Company**	✔✔ *Recommended*

Implementation

- The open data catalog should have a page describing how the metadata in the catalog is licensed or dedicated in general and describing any relevant, general issues with respect to licensing and rights. The page may be the same as in Recommendation 2.

Discussion

While the focus of open data initiatives is on data, metadata should also be clearly licensed or dedicated, since it is used in aggregation/federation/harvesting, catalog analysis, search tools, and product development. Achieving a single, common license/dedication for metadata is, in most cases, easier to achieve than for data, since government publishers are rarely concerned about the intellectual property rights in metadata.

Recommendation 7: <*Omitted from final report.*>

4.2. Catalog and dataset metadata

Recommendation 8: A common metadata scheme should be implemented and enforced.	
Implemented by: **Catalog software, Government** Benefits: **Researcher, CSO, Company**	✔✔✔ *Highly recommended*

Implementation

- Governments should adopt or define a metadata scheme for all metadata and document its implementation (which elements are required, what values are valid).
- The catalog software should allow the configuration of the metadata scheme and the definition of validation rules.

- Automated tools or manual validation should be used to enforce the correct implementation of the metadata scheme.

Discussion

Most government interviewees reported having a metadata scheme, but fewer reported having a common approach to its implementation with publishers, and even fewer reported monitoring or enforcing its implementation. As a result, most consumers described issues with the completeness, quality and accuracy of metadata – making it harder to discover, interpret and use data. Therefore, enforcement is critical. Validations to enforce may include:

- requiring values for basic elements (for example, requiring all datasets to set a value for the `title` metadata element)
- controlling values where possible (for example, restricting the values of the `language` metadata element to ISO 649 language codes)
- requiring publishers to use an element from the common metadata scheme where possible, instead of creating a new element (for example, using the existing `license` element instead of creating a new `licence` element)

In terms of controlling values where possible, the controlled vocabularies below may be used:

- For temporal extent: ISO 8601 Data elements and interchange formats.
- For media type: the Internet Assigned Numbers Authority (IANA) media types.
- For language: ISO 639 language codes.
- For character encoding: the IANA character sets.

An open data catalog may use multiple metadata schemes. For example, many governments already use domain-specific metadata schemes for geospatial data. The recommendation is not to abandon existing schemes, but rather to implement a common metadata scheme to serialize the metadata elements shared by all datasets, like title, description, etc.

Recommendation 9: Values for the data format, update frequency, spatial coverage and temporal coverage metadata elements should be provided and monitored.	
Implemented by: **Catalog software, Government** Benefits: **All**	✔✔ *Recommended*

Implementation

- The common metadata scheme should include elements for: data format, update frequency, spatial coverage and temporal coverage.
- Governments should document the elements' valid values and monitor the elements' values for completeness, quality and accuracy.
- Catalog software should assist the correct entry and active monitoring of the elements' values.

Discussion

The previous report determined that over 90% of datasets across all analyzed catalogs provided values for basic elements like a dataset's title and description or a file's URL and publication date. However, the four elements in this recommendation – data format, update frequency, spatial coverage and temporal coverage – were less frequently provided, despite these elements being important to several important use cases as described by both government and consumer interviewees in the interview results above.

Recommendation 10: Metadata should be available in a machine-readable format respecting an international metadata standard like DCAT or ISO 19115.	
Implemented by: **Catalog software** Benefits: **Government**	✔✔ *Recommended*

Implementation

- Catalog software should support the publication of metadata using international metadata standards.
- Support for the W3C Data Catalog Vocabulary (DCAT) and ISO 19115 Geographic information – Metadata is recommended.

Discussion

Most government and consumer interviewees supported the use of an international metadata standard, to avoid the need to define a new metadata scheme, to make it easier to automate harvesting, to make it easier to merge metadata from multiple catalogs, and to promote interoperability in general.

Most government interviewees proposed DCAT or ISO 19115 as appropriate international metadata standards. Government interviewees identified a lack of technical capacities as a barrier to the adoption of these standards; catalog software, therefore, should play a role in facilitating the adoption of these standards.

This recommendation benefits government in particular, as it greatly simplifies automated aggregation/federation/harvesting across departments and sub-national governments.

For an XML schema implementation of ISO 19115, see ISO 19139 Geographic information – Metadata. Note that DCAT uses many Dublin Core Metadata Initiative (DCMI) Metadata Terms, that the Projet Open Data Metadata Schema is based on DCAT, and that the the Open Geospatial Consortium's (OGC) Catalog Service for the Web (CSW) API may use ISO 19139.

This recommendation does not prevent catalogs from publishing metadata in multiple, additional formats.

Recommendation 11: *<Omitted from final report.>*

Recommendation 12: Metadata may be serialized in a catalog's HTML pages as microdata or RDFa.	
Implemented by: **Catalog software** Benefits: **CSO, Company**	✔ *Nice to have*

Implementation

- Catalog software may support the serialization of metadata in the catalog's HTML pages as microdata or RDFa.

Discussion

Many popular search engines extract structured data from the markup of web pages; in particular, many popular search engines support data structured using Schema.org's types and implemented using microdata or RDFa. This structured data is used to improve the relevance and display of search results. In other words, this recommendation is intended to increase the likelihood that a web user discovers a catalog's data when using a search engine. Support for microdata or RDFa is best implemented by the catalog software providers, considering the government interviewees almost all reported challenges to implementing these technologies themselves. The relevant Schema.org types for open data catalogs are DataCatalog, Dataset, and DataDownload.

Recommendation 13: Governments may initiate or participate in the development of regional profiles of international metadata standards.	
Implemented by: **Government** Benefits: **Government**	✔ *Nice to have*

Discussion

International metadata standards tend to have fewer and looser validation rules for metadata elements' values, given that they must operate in a great diversity of contexts. For example, DCAT does not specify a structure for the value of the spatial extent element; as such, the same catalog may use the text "London, UK," the GeoName ID 2643743, GeoJSON, and Well-Known Text to describe different datasets' spatial extent as being London, UK. Such inconsistencies make the processing of metadata by software more difficult.

To address such issues, the European Union developed an application profile of DCAT, and Canada and the US developed a North American profile of ISO 19115. Other groups of

governments may refine these and other international standards, while maintaining compliance, by introducing new validation rules to improve the quality of metadata.

4.3. Character encoding

Recommendation 14: The default character encoding for data must be UTF-8.	
Implemented by: **Data tools, Government** Benefits: **Company, CSO**	✔✔✔ *Highly recommended*

Implementation

- Governments should publish data production guides for publishers, explaining how to set the character encoding to UTF-8 for common formats and software.
- Data tools should support the import and export of data in UTF-8.
- Automated tools or manual validation may be used to enforce the use of UTF-8.

Discussion

Among the consumer software developers interviewed, character encoding was a source of significant challenges; data could not be correctly read into software systems due to the encoding being unknown or due to improper or inconsistent encoding. As character encoding is an esoteric topic, many developers are ill-equipped to overcome these challenges.

Encoding issues do not affect all catalogs equally; in English-speaking jurisdictions, for example, many datasets use ASCII, which is rarely problematic. In other jurisdictions, many encodings may be in use, creating the challenges above. Government and consumer interviewees preferred to standardize on UTF-8, as described in the interview results above.

Furthermore, the W3C recommends UTF-8. UTF-8 is the default encoding of JSON and Shapefile, and the only encoding for Turtle and Notation3. RFC 2854 declares UTF-8 as the preferred encoding for HTML documents. As an example guidance document, the UK sets UTF-8 as its standard, default character encoding.

With respect to data tools for authoring and managing data, both government and consumer interviewees reported uneven support for UTF-8, in particular Microsoft Excel's poor support for CSV files in UTF-8. The authors of data tools must improve UTF-8 support and make it easier to set the encoding.

This recommendation does not require all data to be encoded in UTF-8. However, if another encoding is used, either the metadata, the protocol used to distribute the data, or the data itself should declare the character encoding in order to avoid issues.

Recommendation 15: A file's metadata should declare the character encoding, as should the protocol used to distribute the data if possible.	
Implemented by: **Catalog software, Standardization body** Benefits: **Company, Government**	✔✔ *Recommended*

Implementation

- Catalog software should support a character encoding metadata element for files.
- Metadata standards should provide a character encoding metadata element.
- If the encoding is not UTF-8, governments should set the character encoding metadata element.
- Catalog software should set the "charset" parameter in the Content-Type HTTP header, if the file's encoding is known.

Discussion

The European Union's application profile of DCAT uses the W3C `characterEncoding` RDF property for its character encoding metadata element. See Recommendation 14's discussion.

Recommendation 16: A file should declare its encoding, if the file format allows.	
Implemented by: **Data tools, Government** Benefits: **All**	✔✔ *Recommended*

Implementation

- XML uses the encoding declaration in the XML declaration.
- HTML5 uses the `meta` tag's `charset` attribute.
- HTML4 uses the `meta` tag's `http-equiv` and `content` attributes.
- Shapefile uses a `.cpg` file.

Discussion

Not all formats with support for encoding declarations are listed. All XML formats/notations/grammars use the same technique, including GML, KML, RSS, RDF/XML, XHTML, etc.

4.4. Data formats and serializations

Recommendation 17: A list of data formats to be used for publish open data should be defined and enforced.	
Implemented by: **Government** Benefits: **Company, CSO**	✔✔✔ *Highly recommended*

Implementation

- Government should define and publish to its agencies a list of acceptable data format for the open data portal with corresponding use cases (e.g usual situation where each format should be used).
- The government should enforce this list either by setting a pre-publication validation (e.g gatekeeper approach) or using automated validation tools to detect datasets not following the guidelines.

Discussion

Several government interviewees answered that there was formal or informal list of formats accepted on the open data portal but that it was usually not supported by clear document shared with agencies in department. The consequence it that consumer interviewees noted that too many data are using unknown data format or non-usable formats like PDF.

The implementation of this recommendation depends on the organization of the open data initiative. With a centralized organization where a given team funnel all the datasets published, it is easy to have a gatekeeper approach where dataset not meeting the requirement are rejected. Some government interviewed process this way, but it is difficult to handle large volume of data. With decentralized approach where each agency can directly publish data, it is can be useful to have a centralized, if possible automated, validation of the formats used.

Enforcing a list of clearly defined format makes it easier for companies and CSOs to use the data, and to develop technical skills that can be used in a recurrent manner when getting new data from the portal.

Recommendation 18: For tabular data, prioritize CSV file, and to a lesser extent Office Open XML workbook	
Implemented by: **Government, Data tools** Benefits: **CSO, Company**	✔✔ *Recommended*

Implementation

- As part of the supported format, government should specify that tabular data should be provided as much as possible using format dedicated for this structure with a priority to CSV format
- Providing additional widespread format, namely Office Open XML workbook

Discussion

While publishing data using structured machine readable format like JSON or RDF is powerful, many consumer interviewees mentioned that they were mainly proficient with tools like MS Excel. Consequently, in order to make data accessible to people not proficient in software

development or non-mainstream software, it is important that tabular data should always able available in simple tabular formats.

As a pure open format, CSV should be the prefered serialization. An additional option like Office Open XML workbook should be considered since many consumer interviewees put MS Excel as their prefered format for tabular data. Although the status of “open format” for Office Open XML workbook is frequently mentioned as contentious, this format is now recognized by ISO and ECMA and supported by several software besides the MS Office suite.

Recommendation 19: Produce valid CSV files as per IETF RFC 4180.	
Implemented by: **Data tools** Benefits: **All**	✔✔ *Recommended*

Implementation

- Data tools used to produce or read CSV files should be compliant with specification IETF RFC 4180

Discussion

Data authoring tool generate a large variety of CSV files with different field separators, line separator or text quoting. It can frequently generate incorrect import of CSV data that can only be solve by tedious modification work or development of small modification scripts/software.

These recommendation is targeted to data authoring software editors since several publisher interviewed mentioned they had difficult to get their software to generate valid CSVs.

The inventory demonstrated that CSV tend to be used as the main format for tabular data but that getting valid CSV was a significant issue.

Note: RFC 4180 does provide character encoding guideline but as per recommendation 14, the use of UTF-8 should also be considered as one of the criteria for CSV validity.

Recommendation 20: For vectorial geospatial data, prioritize GeoJSON or GML, and to a lesser extent shapefile format	
Implemented by: **Government, Data tools** Benefits: **CSO, company**	✔✔ *Recommended*

Implementation

- As part of the supported format, government should specify that file with vectorial geospatial data should be published using GeoJSON or GML.

- In order to make the data more accessible, the Shapefile can be used although it does rank as an open format.

Discussion
Vectorial geospatial represent a significant part of existing open data and is supported by numerous open, proprietary and in-between formats which makes it difficult to recommend one over the other.

GML is an open format well supported by GIS software and largely used in existing open data portal as demonstrated in the inventory. On the other hand consumers with a more technical development background preferred GeoJSON, another open format and an extension of the popular JSON format but with less adoption in existing open data portal. As consequence, those 2 formats can be considered as best candidates for vectorial geospatial data.

On top of this, Shapefile enjoys widespread adoption and can be considered as the *lingua franca* of vectorial geospatial data. Shapefile's specification is publicly available but under the unique control of it's owner, ESRI, so it does not qualify as a pure open format but remain accessible in quantity of software, including open source one. Consequently it could be considered as an acceptable secondary representation after GML and GeoJSON.

KML/KMZ format enjoy a large adoption but appear to be more structured for display (e.g attributes are provided in HTML tags), making it less relevant for data sharing.

Note: Overall it is difficult to choose among the variety of existing format. The key aspect here for a given publisher lies in recommendation 16, more specifically to chose an open format or a widespread non-proprietary format and use it as consistently as possible.

Recommendation 21: Propose linked data format with existing vocabularies	
Implemented by: **Government, Data tools** Benefits: **Researcher, company**	✔ *Nice to have*

Implementation
- When possible, government could provide data using linked data format like RDF or JSON-LD
- Data tools should support existing ontologies and vocabularies and adopt the relevant ones when possible

Discussion
On one hand, linked data and use of existing vocabularies is perceived a significant aspect of open data. With linked data, it is possible to significantly increase the automation capabilities since data attributes become much less ambiguous.

On the other hand, linked data almost did not appear in the answers of either government or consumer interviewees, and most of the government agencies appeared to already have difficulties to comply with more simple simple recommendations. Finally linked data is more difficult to use and require more advanced tools to be used that average citizens, CSOs and even developer do not always know.

As a consequence, use of linked data is presented as a “nice to have” recommendation and can be added to, but should not replace more accessible formats like tabular data.

Recommendation 22: Avoid file compression or provide better support for compressed files within open data catalogs	
Implemented by: **Government, Catalog software** Benefits: **All**	✔ *Nice to have*

Implementation

- Government should document and share with its agencies that file compression should be avoided when make data available on their open data portal
- In order to improve transfer speed, open data portal should support HTTP content negociation to compress file during downloads

Discussion

Currently catalog software and metadata standards do not provide satisfying support for compression since the resulting data format will frequently be “zip” thus hiding all information about the “internal” format. As a consequence, format metadata becomes opaque and useless. While improving portal and standards should be the prefered solution in the long run, the short term solution should be to avoid compression.

In order to improve transfer time, using standard HTTP content negociation -supported by all major browser and some HTTP libraries- should be considered by portal providers.

The recommendation should not be applied to formats where compression is part of the data structure (e.g Shapefile and GTFS). Compression might also be a good second best option when lots of file are to be provided (see recommendation 31)

Recommendation 23: Provide dedicated location within open data portal for non machine readable files	
Implemented by: **Catalog software** Benefits: **CSO**	✔ *Nice to have*

Implementation

- Open data portal should provide a mechanism to link non-machine-readable file to dataset, sometimes in place of existing data.

Discussion

In some situation, the only file accessible are scanned documents provided as PDF files or images. Interviewees from several government raised that digitization is far from complete everywhere and that several data only exist as printed document.

On the other hand, several consumer interviewees said that they prefer to have access to scanned document than nothing. Since scanned document cannot qualify as data, it would be recommended not to make them available using the same mechanisms are regular data (e.g CSV, etc.) but provide a mechanism to make take available.

4.5. URL structures

Recommendation 24: Define and enforce a clear URL structure.	
Implemented by: **Catalog software, Government** Benefits: **All**	✔ *Nice to have*

Implementation

- Catalog software providers or governments should develop and implement a clear and coherent URL scheme for all key resources including dataset and files.

Discussion

Predictable and clear URL structure help human and software find the data and access it. However, although both consumers and government judged URL structure as fairly important, no strong benefit was provided; as consequence this recommendation is set to “nice to have”.

Recommendation 25: Establish stable URLs with redirects if needed.	
Implemented by: **Government, Catalog software** Benefits: **All**	✔✔ *Recommended*

Implementation

- Open data portal should support stable URL for files and dataset and avoid dynamic URL that might change due to inpredictable events
- When URL cannot be maintained, effectivement redirection strategies should be put in place.

Discussion

As stated by Tim Berner-Lee, "cool URIs don't change". URLs to access data or metadata should not change over time. If, for a reason, it is not possible to maintain the same URLs (e.g software change coming with it own URL structure), redirection should be implemented, ideally using HTTP 301 or 303 code or other automated redirections.

Changing URL can break tools relying on the data or simple bookmarks used by data consumers. It can also temporarily cause issues with search engine.

4.6. Data delivery

Recommendation 26: Build and provide domain-specific API interfaces.	
Implemented by: **Data tools, Government** Benefits: **Company**	✔ *Nice to have*

Implementation

- Data tools editors or government can develop application programming interface (API) or Webservice for specific datasets

Discussion

While several consumers ranked API access as very important a limited number of them seemed to have the capacity to use them. Most government interviewees mentioned they had APIs but most were not able to say for which data or in which conditions such approach is being developed.

APIs are perceived an important step for open data progress mainly in order to make data easier to valorize for companies. However it require good to superior technical capacities to use them and it takes significant effort for governments to develop and maintain APIs. Finally, consumers who are able to use APIs are also usually able to build their own API based on downloaded files. As a consequence, this recommendation is set to "nice to have" and it is intended that software editor should be at the forefront to support domain specific APIs (e.g transit management systems should support GTFS realtime or SIRI formats).

A larger discussion would be needed around this topic since some types of data are more or less a good fit for APIs.

Recommendation 27: Automatically generate generic API based on files.	
Implemented by: **Catalog software** Benefits: **Company**	✔ *Nice to have*

Implementation

- Data portals could provide features that automatically transform tabular data in to APIs with filtering and sorting option on the data.

Discussion
As for domain specific APIs, there seem to have some demand but limited knowledge to use it. Generic APIs can be useful to query large dataset without downloading the complete file but it usually lacks features that would needed for a specific dataset.

Some interviewees from the government seemed to be aware that such existed on their portal but had limited knowledge about it.

Recommendation 28: Always propose download options.	
Implemented by: **Government** Benefits: **CSO, Researcher**	✔✔✔ *Highly recommended*

Implementation
- Government should clearly define and enforce that direct download has priority over APIs and web services.

Discussion
APIs require technical knowledge to be used, a knowledge not accessible to many of the potential users: journalists, CSOs, citizens. Providing direct download, if possible using commonly accepted formats (see recommendations 13) is an important step to make open data accessible to the majority.

For some data, the option of the API is the most relevant one (very large volumes, real time data) and file download might not always be a simple solution, however, where possible, a file dump or historical file should be available for download.

4.7. Definition and organization of datasets and distributions within catalogs

Recommendation 29: Define and enforce a clear a data structure in the portal.	
Implemented by: **Government, Catalog software** Benefits: **CSO, Researcher**	✔✔✔ *Highly recommended*

Implementation
- Governments should define and share with relevant agencies and department a clear structure of dataset and file within the data portal
- Open data portal should provide tools and options to effectively structure the data.

Discussion
Several consumers raised the fact that open data portal seems poorly organized and that it was frequently unclear how files were structured in datasets. Such situation discourage users and the bigger the portal, the more discouraged users are.

Along with the metadata structure, dataset and file structure should be shared with all the teams that can publish data. After that, depending on the published process (centralized or decentralized), either a prepublication validation can take place or a post publication validation, ideally automated.

Recommendation 30: Implement and promote multiple serializations of one data.	
Implemented by: **Government, Catalog software** Benefits: **CSO, Researcher, Company**	✔ *Nice to have*

Implementation
Open data portal should support multiple files per dataset
Government should promote as part of the open data portal structure, the publication of different serializations of the same data.

Discussion
Some open data portal allows to have multiple files attached to one dataset. As explained in recommendations 18 and 19, it might be a good approach to provide different formats or serializations for the same data in order to support different profiles of consumers.

Some portal also provide time series (same data on different period) or geographical series (same data for different regions) but it tends to break spatial or temporal metadata. Combining multiple formats with time or geographical tend to make it difficult to sort out the data.

As a consequence, the recommended structure is to only propose datasets with multiple file format.

Recommendation 31: Governments should monitor the number of files per dataset, and limit the number of files where appropriate.	
Implemented by: **Government** Benefits: **All**	✔✔ *Recommended*

Implementation

- If appropriate, split a dataset containing a large number of files into multiple datasets. For example, instead of using one dataset to collect all budget documents, use the keywords metadata element to collect all budget documents with one keyword.

Discussion

The previous report determined that the maximum number of files per dataset was very large in some cases (into the hundreds). 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. Poor discoverability negatively impacts a catalog's utility. Further, most catalog software does not provide a search feature to filter files within a dataset.

4.8. General-purpose data standards

Recommendation 32: Adopt general-purpose standards for highly used information, such as dates.	
Implemented by: **Government, Data tools** Benefits: **Company, Researcher**	✔✔ *Recommended*

Implementation

- Governments and data authoring software should adopt existing general-purpose format for commonly used information

Discussion

Specific type of data are frequently part of dataset: dates, locations, languages, etc. Many of these information are already supported by international standards; for example ISO 8601 provides extensive specification to format dates and intervals of date in a non-ambiguous way and such format are frequently supported by software development libraries.

Government should have guideline about general-purpose standards to be used and share these guideline with software providers.

Although this section was answered as one of the least important both by governments and consumers and it could significantly ease reusability of the data.

4.9. Domain-specific data standards

Recommendation 33: Domain-specific standards should be used where possible.	
Implemented by: **Government, Data tools** Benefits: **Company, Researcher**	✔✔ *Recommended*

Implementation

- Government and data tools editor should evaluate and adopt existing domain-specific standards

- Supported data standards should be declared in open data portal

Discussion
Domain-specific standards is perceived as an important trigger to obtain widespread use of open data. On the economic development side, such standard dramatically increase market opportunity. The example of the transit schedule format GTFS is frequently used since it opens the markets of more than 700 transit agencies. On the civil society and transparency side, it allows to compare different jurisdictions and to get a larger point of view.

At the same time, development and adoption of such standards is time consuming while government already have limited capacities. For most of the governments interviewed, domain-specific standards were not yet on their radar or at least not in the main priorities. And while some format like GTFS have obtained high adoption, several other standards struggle to follow the same path.

Consequently additional work would be required on domain-specific standard (how they created, by who, how they are adopted and governed) to come up with more specific recommendations. On the meantime, government should still evaluate relevant standards while software vendor are probably in the best situation to adopt such standards.

Recommendation 34: Governments may support the development or improvement of domain-specific standards.	
Implemented by: **Government, Data tools** Benefits: **All**	✔ *Nice to have*

Implementation
- Establish domains within governments and data tools editors where it would be relevant to have data standards
- Evaluate existing or nascent standards that could be supported
- Delegate human resource to author or comment technical specification

Discussion
Like all standards, data standard frequently face the chicken and egg dilemma where all potential adopters want to see existing adoption before joining. As a consequence proposed data standard take time to mature and to find adoption. Proactive support from government and data tools editors would could quickly improve the development and adoption curve of standard providing positive value for all the categories of consumers.

5. Conclusions

Recommendations

As described in the introduction to section 4, a goal was to set targets that are achievable by a large portion of publishers and with the maximum impact for consumers. In this regard, the interviews were critical to ensuring the recommendations were realistic, taking into account the needs and capacities of publishers and consumers.

We recognize that many discussions among experts within the open data community propose advanced, technical solutions to the challenges identified in this report, many of which the above recommendations either do not repeat or repeat with a low priority of “nice to have”; for example, API access and linked data do not figure prominently. The research team proposed the draft recommendations above for validation by stakeholders, and looks forward to engaging the community in a discussion of the recommendations.

Applicability

While the interviewees were from low- and middle-income countries, the challenges they faced were not dissimilar to those in high-income countries. Similarly, although this project studied the national initiatives of OGP members, the experiences were not dissimilar to those of sub-national initiatives in high-income countries, which are often similarly resource constrained. As such, we expect the results and recommendations to be of value to initiatives outside those of OGP members.

Regional trends

Interviewee selection was designed to allow regional trends to surface; however, geography did not seem to have a strong effect, with most countries reporting similar challenges and proposing similar solutions. Given that most of the open data initiatives were young, this could have been expected. As the initiatives mature, trends may appear.

Future work

As raised several times throughout the “Gaps and opportunities for standardization in OGP members’ open data catalogs” report, there is much future work to be done on open data standards, just in terms of measuring and understanding current practices. However, we limit our discussion to two areas for future work.

Domain-specific standards

The success of the General Transit Feed Specification (GTFS) for transit data had kindled a strong desire for domain-specific standards, and many attempts have been made to replicate its success since 2005. However, very few data standards achieve high levels of adoption. Future work should focus on better understanding what makes standards succeed or fail, by looking at multiple aspects of standardization, including stakeholder identification, development process, technical design, communications strategy, and governance. With this understanding, new

standards could be developed, with a greater likelihood to succeed, for types of data for which no popular standards exist.

Self-assessment tools

The government interviews highlighted a major gap in the operation of open data initiatives: the lack of tools to monitor the quality of the different areas of standardization. Indeed, few government interviewees reported central oversight for any area of standardization. It is difficult for a government to achieve a high quality standard for its open data initiative if it cannot measure its performance. Future work should therefore create the necessary tools for governments to measure their performance, in order to help them improve their operations.

6. Acknowledgements

The funding for this work has been 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 the Social Sciences and Humanities Research Council (SSHRC) Partnership Grant Program.

We would like to thank all participants for their interest and participation in assisting us complete this project. Their suggestions and ideas helped produce more well-rounded and useful recommendations.

We would like to especially thank Abhinav Bahl of Global Integrity for his role in assisting us to find and contact appropriate interviewees.