



Defining and measuring Open Data

Transparency International Georgia

Authors:
Etienne Baqué
Derek Dohler

Date: 28th of April 2014

Table of Contents

Introduction	3
Open Data Guidelines	5
1. Complete	5
2. Primary	5
3. Timely	6
4. Accessible	6
5. Machine-readable	6
6. Non-proprietary	6
7. Freely usable	7
8. Reviewable	7
9. Discoverable	7
10. Permanent	7
Open Data Usability Index	9
Applying the Index	9
1. Cost	10
2. Legal Framework	11
3. Accessibility	12
4. Readability	12
5. Timeliness	13
6. Completeness	14
7. Reviewability	15
Application of the ODUI	16
General observations	17
Data Set Evaluations	18
Evaluation of Judiciary-related data sets	18
Evaluation of Public Order related data sets	19
Evaluation of “Public Data” data sets	20
Evaluation of “Legislative Data” data sets	22
Evaluation of Financial Information related data sets	23
Evaluation of Public procurement related data sets	25
Evaluation of “Property” data sets	25
Evaluation of Elections related data sets	27
Evaluation of Online books/Catalogues related data sets	29
Conclusion	31

Introduction

As part of the efforts led by different government agencies regarding the Open Government Data (OGD) initiative, several NGOs decided to participate in the process of creating an Open Data portal.

Over 2 workshops that took place in January 2014 then in March 2014, recommendations about how to create such website were given by Johann Höchtl and Bernhard Krabina, advisers of the [EU Twinning project](#), to the different government agencies and NGOs. They also highlighted the process that would leads the different interested parties to release as much “Open Data” as possible.

A strong component that defines this initiative is Open Data, and Transparency International Georgia decided to help out, in defining what Open Data is, what it means, how it can be measured upon different data sets, in order to preserve the quality of the information that will be published.

In order to guarantee the success of the Open Data portal, that will offer various public data sets, it is indeed critical to serve data that is as “open as possible”. The final intent is to have data that is as open as possible, provided by different government agencies, freely usable by third with as few restrictions as possible, for different purposes.

We, at Transparency International Georgia, stand with the recommendations that were given over the course of these 2 workshops, and through this present document, we aim to release publications about what Open Data is, and how to measure it. This document is divided in 2 sections, that outline how we define and measure Open Data:

1st section: The Open Data Guidelines.

This section focuses on **what** defines Open Data. In other words, it answers the questions “what is open data?”

10 guidelines are provided, each of them presenting an aspect that characterized open data.

2nd section: Open Data Usability Index (ODUI)

This second section highlights **how** we can assess the “openness” of data. Because Open Data is defined through different aspects outlined by the guidelines, the openness of a given data set can vary. Because these guidelines are quite different from one to another, the evaluation of the data sets openness can turn out to be quite subjective.

This is why we created an Open Data Usability Index. This index is to be used as a tool, as a referential to apply on a data set, in order to define its degree of openness (via a score going from 0 to 100

This section is divided in 2 sub section: the first sub-section of this document will present more in detail the ODUI, and its different aspects. Explanations about how to apply this index will also be given.

The second sub-section presents applications of the ODUI, on several data sets. Over the past year, data has been released by different public agencies in Georgia. We first chose different data sets based on their public utility, and we applied the ODUI on them.

By doing so, we were able to measure the openness of each of them, and we wrote the different scores in a matrix (See second part of this document) . We're hoping that by releasing the different scores, and the recommendations we make to the different public agencies in order for them to improve the openness of their data, we will foster an environment in which organizations can seek to be as Open as possible.

Open Data Guidelines

(“The 10 Open Data Guidelines” presented below is a copy of what has been published on the Transparency International Georgia website, at that page)

The Open Data Guidelines are designed as a guide to help agency heads, IT managers, and web developers create open data websites. However, it is not meant to cover all situations - “data” is a broad term, and some data may require disclosure methods not discussed here in order to be fully open. In addition, there are other issues, such as accessibility for the disabled, which are not discussed here but which are key components of any good website.

Public data is data that is not restricted from disclosure due to privacy, security, or other valid concerns. These guidelines do not address in detail what information should be considered public, but in general, government data should be assumed to be public unless there is a specific reason to restrict its disclosure that is legitimate under international law, and the decision to withhold is based on the public interest. If data must be legitimately withheld, it should be redacted, the redaction labelled clearly, and the remainder of the data set released along with an explanation of the reasons for any redaction.

Open Data is:

1. **Complete**

All data comprising a particular public data set should be published, including non-digital archival data and data used to generate aggregate or derived figures.

- For data sets spanning many years, digitizing paper-based archives is often a difficult task. Ideally, this data should be made available, but given limited resources, the focus should be on creating systems that make collection and release of public data in a digital format efficient going forward. If possible, new digital systems should be designed so that non-digital archival data can be added as it is digitized.
- Aggregate and derived figures should be distributed along with the source data used to create them and explanations of the methods used to do so. For instance, an inflation figure should include the prices of the basket of goods used to measure inflation.

2. **Primary**

Data should be collected at the source, and should be published with the level of granularity with which it was collected.

- Data may not always be in a usable form when it is collected (e.g. sensor data). In these cases, post-collection processing may be performed. This processing, however, should result in data in a format that reflects the full capabilities and granularity of the original collection mechanism. Data should never be lost.

3. **Timely**

Data is made available as rapidly as possible in order to maximize its value to the public.

- There is no time period that is appropriate for all types of data, but in many cases, a well-designed system will enable data to be published in real-time, or nearly so. Thus, information should be published as soon as possible.
- Updates should be easy to locate from within the larger data set, through mechanisms such as RSS feeds, search functions that can filter by date, and archives which contain snapshots of the data set taken at regular intervals. A successful site will probably use some combination of these techniques.

4. **Accessible**

Data is available to as many users as possible, for the widest range of purposes possible.

- The data should be easy to share digitally. Every page and document published should be assigned a unique and easily obtained URI (URL) that can be distributed via email and social websites. Web development techniques such as cookies, Flash, and AJAX/AHAH should not be used to obscure the location of data or make it difficult to share direct links.
- URIs should be in a human-friendly format, e.g.
"www.domain.gov.ge/data/ministry/justice/complaints/2009/10/11/complaint.html" not
"www.domain.gov.ge/data.php?lang=GEO&searchcode=15&searchstr=10,11,2009&type=4&min=01A" .
- Full access to data should never require registration or payment.
- Bulk downloads should be made available via protocols such as FTP or rsync.
- Ideally, a well-documented API will make automated access to data possible.

5. **Machine-readable**

Data is stored using a format and structure that allow automated processing.

- Machine readability should not replace human readability; both formats should be presented (e.g., a transcript of a speech should be provided along with a recording of it).
- Data should never be made available only in the form of scanned images. Scanned images may be the best way of presenting certain documents, such as documents bearing a seal or signature, but they should always be accompanied by a machine-readable representation of the document's other content.
- Data should never be stored in non-Unicode fonts such as AcadNusx or LitNusx.
- Data should be presented in an easily-processed format, such as CSV, JSON, or XML.
- In all cases, the meaning of each field in the data should be well-documented and this documentation included along with the data.

6. **Non-proprietary**

Data is available in an open format over which no entity has exclusive control, to ensure that computer programs capable of processing the data will always be available.

- An open format is one which is not subject to intellectual property controls in any country, and for which documents defining the format's structure are freely available. HTML and XML are examples of open formats.
- Simple formats and formats supported by numerous programs should be preferred. For example, formats such as XML and JSON should be preferred over formats such as PDF or OOXML (OOXML is commonly known as the MS Office .docx, .pptx, and .xlsx formats).

7. **Freely usable**

Intellectual property protections, such as copyright or trademark, are not used to restrict potential users or uses of the data.

- Data should be made free for all types of use, including commercial use, without restriction.
- Even if the law is clear that public government data may be used freely, a notice to this effect should be included with every item of data that is published.

8. **Reviewable**

Every public or private body releasing data should designate a person to respond to questions and complaints about the data, and this person's contact information should be included with the data.

9. **Discoverable**

Data must be able to be found by those who are looking for it; it must be included in appropriate data catalogues, and data websites should be accessible to search engines.

- Ideally, all government agencies should adopt a unified means for displaying their data online, so that users can rapidly locate the data produced by any agency from its website.
- Agencies should ensure that the listings at government data portal websites are accurate and up-to-date.
- Agencies should generate comprehensive sitemaps and submit them to all major search engines.

10. **Permanent**

As data ages, it should be archived in ways that satisfy the above criteria.

These guidelines were adapted from guidelines published by the [Sunlight Foundation](#), [Access-Info Europe](#), [OpenGovData.org](#). This document and its contents are [Creative-Commons licensed](#) and may be freely reused, modified, and redistributed. Any omissions from or additions to the text of these

guidelines should be attributed to their authors and not Transparency International Georgia or Transparency International.

Open Data Usability Index

The Georgia Open Data Usability Index (gODUI) is based on the [Open Data Usability Index](#) (ODUI) developed by Open Data BC, with modifications by Transparency International Georgia.

The original ODUI focused more on "usability" than "openness". We have added indicators to the ODUI in order to shift the focus toward transparency generally, while maintaining the original focus on assessing the quality of the data sets that agencies have already released, rewarding progress achieved, and encouraging further improvements in the quality of open data efforts.

We have also updated and clarified the scoring methodology in cases where we felt it was necessary for application to the Georgian context.

Applying the Index

The index is applied in three stages.

In the first stage, the user must select and carefully define the “data sets” that are to be assessed. There is no fully objective way to do this, but when doing so, the following factors should be considered, and the data set’s selection and definition justified.

- **Public utility:** This index is based on the belief that data collected by government entities is often a public good with few or no negative externalities related to overconsumption. In other words, the optimal production and consumption levels for such data are achieved when production is mandated by the government, and no restrictions are made on access to the data. Data sets which are of high public interest, for commercial, political/public accountability, anti-corruption, civic engagement, cultural, or other reasons, should be preferentially selected.
- **Proactive collection:** This index is designed to target data sets which are already being collected proactively by government entities, either as part of an established routine or procedure, or in an ad-hoc manner. If an important data set is not being collected, other advocacy tools should be used to support the collection and publishing of this data, and it can then be assessed using this index.
- **Countervailing considerations:** Not all data is appropriate for the type of publication envisaged by this index, which is designed to encourage the broad dissemination and usage of the data. Some data may contain, for example, private personal information. Data sets should be selected for which such countervailing considerations are minimal, nonexistent, or clearly outweighed by the public utility of the data.
- **Advocacy value:** This index is primarily designed to measure the quality of data which is already being collected and for which some steps to publish the data in an open way have already been taken. Therefore, the greatest advocacy value of the index will be achieved when it is applied to such data sets. Applying the index to data sets which are entirely unavailable or not being collected at all will have limited utility; in most cases, such data sets will receive

extremely low scores which will have limited advocacy value. Other advocacy tools should be used to push for expanded collection and disclosure of such data; once initial successes have been achieved, the newly published data sets can be assessed using this index.

- **Practical application:** This index is designed to apply to the publication of data in practice. The fact that agencies are often not legally required to comply with the definition of usability envisaged by this index should not present a barrier to applying this index to the data sets which they do publish. The goal of this index is to improve open data usability; the results from this index can be used to push for reformation of laws and regulations that are identified as shortcomings.

In the second stage the user assigns a value to each of the listed aspects of usability; each aspect may be composed of one or more components, which are summed to determine the final score (out of 100) for each aspect.

In the third stage, an weighting algorithm is applied to calculate a score out of 100 which can then be used to display the score visually.

The following section explains the aspects of usability that will be assessed.

1. Cost

In this section the cost of obtaining the data is assessed. Data provided for free receives the highest score. Nonetheless, if data is not offered for free, lower costs are better because they reduce barriers to using the data.

Data sets for which the cost to “access” the data are actually the costs associated with *collecting* the data may not be good candidates for inclusion into this index (e.g. paying to perform lab tests on food).

Scoring note: The cost should be calculated as the cost necessary for one individual to maintain access to an up-to-date version of the data set for one year. Because different data sets are updated with differing frequencies, ranging from microseconds to months, it is up to the researcher to determine how often a data set must be accessed in order to stay up-to-date. This determination should be supported.

Component 1 (weight: 1.0)

100	Free	The data is provided for no charge
75	Under GEL 25	The data is provided for a fee less than GEL 25 per year
50	Under GEL 250	The data is provided for a fee of GEL 25 or more but less than GEL 250 per year
25	Under GEL 2500	The data is provided for a fee of GEL 250 or more but less than GEL 2500 per year

0	GEL 2500 or more	The data is provided for a fee of GEL 2500 or more per year
---	------------------	---

2. Legal Framework

Organizations releasing data as open data use a variety of legal mechanisms to express the terms and conditions under which they grant permission to use the data being released. For the purposes of the gODUI we view these terms and conditions as a set of restrictions that the publisher has placed upon the use of the data.

Publishers often release data with custom text describing the restrictions that apply to the use, even when they place few restrictions. “Custom” in the context means that the text used by the publisher to describe the set of restrictions is unique to that jurisdiction. The use of custom licenses means that users of the data need to review and interpret each individual license to determine if the set of restrictions imposed on the data by the license allows them to do what they want to do with the data and imposes a level of risk that they can tolerate.

It is important that data be released with as few restrictions as possible. The Open Knowledge Foundation (<http://okfn.org/>) developed and maintains a tool called the open definition (<http://opendefinition.org/okd/>) which provides users with an easy way to determine whether a given set of restrictions placed upon data meets the minimum standard for the definition of open knowledge and open data.

It is also important that data is clearly labeled with the licensing restrictions that it bears. For example, each data file which makes up a data set might include a copy of the license which applies to it, or a link to the license might be available on the web page from which the data is downloaded. Data that is not labeled might require prospective users to consult other web pages, agency regulations, or local, state, or national law in order to determine the usage restrictions that apply to the data.

Component 1 (weight 0.10):

100	Labelled	The data is distributed alongside an easily accessed description of the usage restrictions that apply to it.
0	Not labelled	Irrespective of the usage restrictions that actually apply, they are not distributed alongside the data itself.

Component 2 (weight 0.90):

100	Fully compliant	The license for the data complies with all requirements of open knowledge, as given here: http://opendefinition.org/okd/
90	1 non-compliant restriction	The license for the data contains no more than one restriction which does not comply with the requirements for open knowledge.

80	2 non-compliant restrictions	
70	etc.	

3. Accessibility

This aspect assesses the ease with which potential users can gain access to the data. There are three equally weighted components to this aspect.

Component 1 (weight 0.33):

100	Shareable URL	When the URL is shared via email, a chat window, social media, or other sharing mechanism, each user who clicks on the link receives the same data as the original user who shared the link.
0	Non-shareable URL	Users who click on a shared link do not always receive the same data as the original user who shared the link.

Component 2 (weight 0.33):

100	Anonymous access allowed	Users may access the full data set without logging in, registering, or authenticating in any way, including via tracking cookies.
0	Anonymous access not allowed	Users who do not log in or accept tracking cookies are not able to access the full data set.

Component 3 (weight 0.33):

100	Immediate	The data is immediately available at the URL without any human intervention (e.g. filling out a CAPTCHA).
0	Not immediate	Human intervention is required in order to receive the data.

4. Readability

This aspect assesses the ease with which users can make use of the data, after they have accessed it. There are three equally weighted components.

Component 1 (weight 0.33):

Scoring note: All digital formats are, at a certain level, parseable. For example, a scanned copy of a book page is stored in an image format which can be parsed by a computer to display the scanned image. However, assuming that the data of interest is the text printed on the page, such a format would not qualify as parseable because the text can only be reliably read from a scanned image by a human

being. In general, HTML documents are NOT parseable, because it is usually necessary to create a customized program to extract the data of interest from the surrounding web page.

100	Parseable	The data of interest can be parsed by standard parsers. A good rule-of-thumb test is to attempt to import the data into Google Refine, which can read many common formats. If the data relates to a specialized field, it may be necessary to consult a domain expert to determine whether the format is considered parseable in that field. Common examples of formats which satisfy this requirement are CSV, XML, JSON, and XLS.
0	Not parseable	The data of interest cannot be parsed or separated from the enclosing digital format by standard parsers.

Component 2 (weight 0.33):

100	Open format	Common examples of open formats: XML, JSON, CSV, SHP, SQL99, XHTML, PNG, SVG
0	Proprietary format	Common examples of proprietary formats: PDF, Word, XLS, FLASH, PSD

Component 3 (weight 0.33):

Scoring note: All data sets may contain errors; as long as the errors don't significantly impact the ability of users to accomplish useful work with the data, a small number of errors should not result in the data set being marked unstructured.

100	Structured	Importing the data using a standard parser results in data that can be used to accomplish useful analysis or work without significant further processing.
0	Unstructured	Using the data to accomplish useful analysis requires significant cleaning or processing. A good rule-of-thumb test is to try to sort the data based on a variety of fields. If the results are correct, then the data is probably well structured, but if they are not, or if it is impossible to define which fields should be used to sort, then the data is probably not well structured. For example, a data set in which dates are in multiple different formats (e.g. 21 December 2013 vs 12/21/13).

5. Timeliness

Data must be published soon enough after it is collected to make it useful; the usefulness of the data in relation to a hypothetical data set that was published immediately defines the timeliness of the data.

The timeliness of a data set depends on the type of data and the use to which a user wants to put it. For example, data on weather conditions would need to be published almost immediately in order to be useful for producing short-term weather forecasts. However, for a scientist investigating long-term

weather trends, such data could be published months or even years after its collection and still remain useful.

Thus, we can measure timeliness by the impact of the delay in releasing the data on the potential uses to which the data could be put. This calls for a degree of subjectivity, so the assessment should be supported. For specialized fields, a domain expert may need to be consulted to determine what data should be considered timely.

Component 1 (weight 1.0):

100	Immediate release	The data is published as soon as it is collected; this is usually accomplished through an automated publishing system with access to the data collection system.
90	No impact on usefulness	The data is not published as soon as it is collected, but the delay has no impact on the usefulness of the data.
50	Somewhat reduced usefulness	The data is not published as soon as it is collected, and the delay reduces the usefulness of the data somewhat.
25	Significantly reduced usefulness	The data is not published as soon as it is collected, and the delay significantly reduces the usefulness of the data.
0	Unusable	The data is not published as soon as it is collected, and the delay renders the data unusable.

6. Completeness

Data must be complete in order to be considered usable; incomplete data can render many types of analysis less useful.

Completeness must be judged relative to a “complete” data set. For the purposes of this assessment, the complete data set is all the data collected by a particular agency relating to a particular regulatory task. The complete data set does not include data that could be collected but is not, nor does it include data collected by other agencies (even if such data is related to a similar regulatory task).

Component 1 (weight 1.0):

Scoring note: This component assesses the percentage of the full data set which is released. It is permissible to score this question to any degree of precision which can be achieved. The calculation used to arrive at the percentage should be provided along with the score.

100	100% published
99	99% published
75	75% published
etc.	

7. Reviewability

Data sets should have assigned and public identified maintainers who are responsible for the maintaining and updating the data set, and for responding to inquiries regarding the data set.

Component 1 (weight 1.0):

Scoring note: A specific person **MUST** be identified in order to score 100 for this component. It is acceptable for inquiries and maintenance of a data set to be handled by several people or a department, but the person bearing ultimate responsibility for the data set must be identified. Thus, a generic “contact us” form, without any further information, is not sufficient.

100	Contact provided	The contact information of a person responsible for the data set is clearly identified alongside the data set, along with contact information.
0	No contact provided	No contact information is provided alongside the data set, or the contact information provided is a generic and does not identify a specific person responsible for the data set.

Table of Aspect Weighting

The following table shows the different weights that will be applied to each of the 7 scores defined by the index. Some aspects will have a greater impact than others on the general openness of the data. For example, if a data set is only available by paying a fee, that will reduce the “Cost” score, which counts 25% of the final score, and therefore will significantly decrease this final score.

Aspect	Weight
1. Cost	25%
2. Legal Framework	20%
3. Accessibility	12.5%
4. Readability	12.5%
5. Timeliness	12.5%
6. Completeness	12.5%
7. Reviewability	5%

Application of the ODUI

After creating the Open Data Usability Index, we decided to apply it on several data sets that already are publicly available.

Here's below the list of websites we evaluated the data sets from. and their respective type of service they belong to (eg Judiciary, Public order...). There are more data sets out there that could be evaluated using the ODUI. But in order to demonstrate how the ODUI can be applied, we chose to evaluate the data sets that we considered having a notable public utility.

Judiciary

Supreme Court of Georgia <http://www.supremecourt.ge/>
<http://www.supremecourt.ge/court-decisions/>
<http://prg.supremecourt.ge/>

Public order

Ministry of Interior <http://police.ge/>
<http://police.ge/ge/useful-information/statistics>

Public Data

National Statistics office of Georgia <http://geostat.ge/>
Asset declarations: <http://www.declaration.gov.ge>

Legislative Data

Legislative Herald of Georgia: <https://matsne.gov.ge/>
Parliament of Georgia <http://parliament.ge/>
<http://votes.parliament.ge/ka>

Financial Information

Ministry of Finance <http://mof.ge/>
<http://mof.ge/Budget>
Revenue Service <http://www.rs.ge/>
http://www.rs.ge/Default.aspx?sec_id=5099&lang=2
National Bank of Georgia <http://nbg.ge/>
<http://nbg.ge/index.php?m=304>

Public Procurement

Competition and State procurement Agency <http://tenders.procurement.gov.ge/>

Property

Public Register <http://reestri.gov.ge/>
<https://enreg.reestri.gov.ge>
<https://naprweb.reestri.gov.ge>
National Agency of State Property
Privatizations: <http://privatization.ge/>
Auctions: <https://www.eauction.ge/>

Elections

Central Election Commission <http://cec.gov.ge/>
2012 Election results: <http://results2012.cec.gov.ge/>
2013 Election results: http://results2013.cec.gov.ge

Online books/Catalogues

National Library <http://www.nplg.gov.ge/>
Digital library: <http://www.nplg.gov.ge/ec/ka/changedb.html>
Dictionaries (all use the same software, so can be evaluated as a group):
<http://www.nplg.gov.ge/gwdict/>

General observations

When going through the exercise of applying the ODUI on all these data sets, we consistently noticed a few points that are worth mentioning.

First, we noticed that while applying the index, each data set earned a score of 0 (out of 100), regarding the “Legal Framework” aspect. No website listed above make mention in any note, or on any webpage, that restrictions might legally be applied on the provided data sets.

Even if data seems to be easily accessible and freely usable, to avoid any doubt that users might have, it should be made clear, by having an explanatory text on the website, whether or not any potential restrictions are to be applied. (To avoid redundancy, this point about the Legal Framework won't be brought up in the individual observation of each data set)

It is also interesting to see that in, general, websites provide data in one of these 3 formats: HTML format, XLS (Microsoft Excel) and/or PDF files. As presented in the Open Data Guidelines, and explained more in details in the description of the ODUI, there are other file formats that are available to present data, and we recommend to use them: JSON, CSV, XML file formats. Because of the clear lack of these open file formats in our study group of data sets, one recommendation would be to use primarily one of these open formats, while generating new data sets in the future, not necessarily as a replacement for PDF documents (that are very human-readable), but as an additional resources. Having open format, parseable files will allow data processing for third parties, and therefore increase the openness of the data made available on the website.

One last recommendation we can make to the interested parties would be to provide an API. An API (Application Programming Interface) is a great tool for third-party developers to freely tie, connect their program to the data provided by the website. Because this connection is made programmatically, every time the data changes on the original website, this update will be automatically taken into account on the third-party application. An API is a great tool one can provide in order to increase the openness of the provided data sets. Among the websites we applied the ODUI to, we only found one website that provides an API (<http://votes.parliament.ge/ka>)

Data Set Evaluations

Evaluation of Judiciary-related data sets

1) *Court Decisions* - Website: <http://www.supremecourt.ge/court-decisions/>

Scoring sheet:

Observations:

Here's what was noticed when assessing the data published by the Supreme court of Georgia. We will suggest also some actions to be taken, in order to increase the openness of the provided data.

- URLs on this website are very good: relatively short and explicit for the main pages, (<http://www.supremecourt.ge/court-decisions/criminal-cases/>), and doc urls have relevant information (eg date: <http://www.supremecourt.ge/files/upload-file/pdf/noemberi2008.pdf>)
- The provided documents are only PDF documents. By definition, they are non-structured, several of them are formatted in a different way. If the Supreme Court of Georgia was to transform these document in an open format files (JSON, CSV...), this would increase the openness of their data.
- Publications had been made available over the past several years, in a very consistent manner. These documents have been published until the end of 2013. (Publications related to 2014 cannot be uploaded until the end of this year). However, the decisions-related documents have not been published since 2011.
- Finally about reviewability, not a clear indication about who to contact about the data.

2) *Supreme Court of Georgia database* – Website: <http://prg.supremecourt.ge/>

Scoring sheet:

Observations:

Here's what was noticed when assessing the data found while browsing the Supreme court of Georgia database:

- The URL always remains the same (<http://prg.supremecourt.ge/DetailViewCrime.aspx>). Specific data is loaded within it. So it's not possible to use the URL as a way to reach data. As a result, users have to go through a form in order to send some information to the server, so that the relevant information can be displayed (HTTP POST request). This situation prevents an easy access to the data made available on this website.
- In addition to what the website already has to offer, one recommendation would be to group data logically (eg. grouped by grievance ("საჩივროს სახე ") or by Result ("შედეგი")), and provided as downloadable XML or JSON files.
- Data is provided online pretty efficiently. The most recent data is about appeals which date is 28th of February 2014
- Reviewability: There is not a clear indication about who to contact, should a user have a question about the data. There is a contact us section, but it is not explicitly said that a specific contact should be reached.

Evaluation of Public Order related data sets

1) *Crime Statistics* – Website: <http://police.ge/ge/useful-information/statistics>

Scoring sheet:

Observations:

In addition to the general observation (see above), here's what was noticed when assessing the data found while studying the Crime Statistics data sets

- It is very easy to access the data, no need to input anything in order to retrieve the search result list, as it is directly accessible via URL.
- Data is provided only in PDF format and JPG files, which impedes the openness of the published data. Although it is important to serve human-readable document, it is also just as critical to provide machine-readable file formats, like JSON and CSV, in order to allow further processing by third parties. Making this data available under this format, in addition to the current formats, is definitely recommended.
- Some data was published in February 2014 (in "Different types of statistics: <http://police.ge/ge/useful-information/statistics/skhvadaskhva-sakhis-statistika-kvlevebi>), but the most recent data regarding Crime Statistics is a year old. So unless this data doesn't relate to one another, we cannot confirm the completeness of the provided data.
- Reviewability: There is an extensive list of contacts on the "Contact us" page (<http://police.ge/ge/contact-us>) which might be seen as very useful. That said, there is no clear mention of someone being in charge (the person to contact) for the published data set.

Evaluation of "Public Data" data sets

1) *National Statistics office of Georgia* – Website: <http://geostat.ge/>

Scoring sheet:

Observations:

Here's what was noticed when assessing the data found while studying the Geostat data sets:

- Even though a lot of information is published on Geostat, which makes this website a fantastic resource of data, we found out that fees are applied on several data sets, restricting greatly the general openness of the data provided on this website. Here's the list of prices to be applied upon requests of data sets: http://geostat.ge/cms/files/September%202012_Eng.pdf
- All the main statistics are provided in an XLS format, which is considered a parseable format.

However, no open format is used to represent the data. Since many XLS files are published, a recommendation would be to convert automatically them into CSV files (open format).

- GeoStat published its data periodically: depending of its nature, data is published monthly, quarterly, or annually. The rules of these publications are defined by elaborated methodologies, which are available on the Geostat website, and have been written by different third-parties organizations. Therefore, since rules are regulating data releases, the timeliness aspect score is high. However, we cannot guarantee that immediate release of data (when available) takes place, so we cannot give the highest note for that aspect.
- The contact information of the regional offices is detailed on the website. In addition to this a hotline number is advertised on the home page: +995 32 2 60 11 60. The good visibility of the phone number makes it easier for the users to find someone, should they have any request.

2) *Asset declarations* – Website: <http://www.declaration.gov.ge/>

Scoring sheet:

Observations:

Here's what was noticed when assessing the data found while studying the Asset Declaration data set:

- All the public officials' declarations can be found in PDF files only. As mentioned in the general recommendations, in addition to these PDF documents, it is important to provide data in a parseable/machine-readable open format.
- Timeliness aspect is well rated as some declarations are just a few days old.
- Like many other website providing data sets, reviewability is nonexistent, as there is no designated person to contact, should a user have any questions about the data.

Evaluation of “Legislative Data” data sets

1) *Matsne* – Website: <https://matsne.gov.ge/>

Scoring sheet:

Observations

Here's some observations we made about the Matsne website:

- In addition to the information shown on their website, Matsne provides a downloadable offline version. Having downloadable data (or data-related content) is definitely a very good feature on a website in terms of openness, as it increases information accessibility. That said, and in order to go further, Matsne should now make sure that this content is not tied to proprietary format, or proprietary software. This is the case here, the downloadable program runs on Windows only. It cannot be run on MAC OS X, or any Linux-based distributions. Instead, Matsne should provide, for example, downloadable zip/tar archives that contain open format files (JSON, CSV files...)

2) *Parliament of Georgia / Voting records* – Website: <http://votes.parliament.ge/ka>

Scoring sheet:

Observations:

Here are some observation we made about votes.parliament.ge

- The data on this website is very open, very accessible. It provides a great feature that improves

dramatically the openness of the data sets: it is indeed possible for a third party to get programmatically data via an API (<http://votes.parliament.ge/en/api>). An API guarantees that very accurate data is provided, in terms of timeliness and completeness, as a single change in the database is immediately reflected in the data provided by the API.

- The 2 areas that could be improved are Reviewability, and Legal Framework, as there is no mention of any text with regards to these 2 aspects.

Evaluation of Financial Information related data sets

1) *Budget* – Website: <http://mof.ge/Budget>

Scoring sheet:

Observations:

Here are some observations made about the Budget website:

- Data takes the form of graphs (budget analytical data). Therefore, the informational value is great, as graphs usually conveys more insights than plain text do. PDF files (yearly state budget) are also provided, which also increase the quality of the way data is served via this website. However, data on Budget Legislation page is rendered on HTML pages. So only a fraction of the information here is considered being displayed in an open format. In addition to all these ways the data is displayed, and as mentioned in the general observations, open format files (or even an API) should be preferred.
- Like many other website providing data sets, reviewability is nonexistent, as there is no designated person to contact, should a user have any questions about the data.

2) *Exports and imports* – Website: http://www.rs.ge/Default.aspx?sec_id=5099&lang=2

Scoring sheet:

Observations:

Here are some observations we made about the Export and Imports data, on the Revenue Service website:

- Data is really accessible, informational value is important, but data about export/import can only be found as charts. There is no open format file that allow third parties to use the data for further processing.
- Also, the accessibility of the data could be improved: it is not possible to get a specific sub-data sets right from the url, as refining the chart setting and consequently displaying results takes place on the same page (ie the url remains the same). In order to be able to access a very specific data set and to share it, it is important to have it identified by a unique url.

3) *National Bank of Georgia/Statistical Data* – Website: <http://nbg.ge/index.php?m=304>

Scoring sheet:

Observations:

Here are some observations made on the Statistical data available on the National Bank of Georgia:

- The data available on this website is shown through HTML, which is considered as an open format. It is also considered structured, as it is displayed in tables: parsing this data using a program would not need a lot of processing in order to make sense of it.
- That said, it is not considered as being parseable right away. As just mentioned, programs would need to be developed then used in order to use this data (see description of ODUI above, “Readability” aspect). To avoid having to use a program to re-use data, we're encouraging the

- use of open structured formats, like XML, JSON or CSV.
- As mentioned in the general observations, texts giving information about legal framework and person to contact (reviewability) are missing.

Evaluation of Public procurement related data sets

1) *Competition and State procurement Agency* – Website: <http://tenders.procurement.gov.ge/>

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on this website, in addition to the general recommendations made above:

- Even though it is possible to access every tender document via the URL, reaching the data available on this website is not as straightforward as it should be: as the home page states, “The registration at the Ge-GP system is obligatory only for the procuring entities and the suppliers”. While we understand that registration might be mandatory for some users because of some business restrictions, it should be noted that this additional step will impede the openness of the data.
- The raw data is only in PDF documents. In order to increase openness, it should be made available in an open format, such as CSV or JSON.

Evaluation of “Property” data sets

1) *Legal entities* – Website: https://enreg.reestri.gov.ge/main.php?m=new_index

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on this website, in addition to the general recommendations made above:

- The company registry website contains a lot of information, and is doing a good job to present over time the documents related to a company. That said, the entire data set related to a company is only available in PDF documents. As we browse through the website and we narrow down our path to the PDF document, it is possible to notice some information related to the company, but the full set is only available when we land on the PDF document itself. In order to improve the openness of the data available on this website, another file format than PDF, such as JSON and CSV formats, should be considered, in order to facilitate the reuse of the data by third parties.

2) *Real Estate* – Website: <https://naprweb.reestri.gov.ge/>

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on this website, in addition to the general recommendations made above:

- After applying the ODUI on the company registry website (analyzed in the previous point), we noticed that the Real Estate website seems to be own by the same entity: both website's url contains the domain name *reestri.gov.ge*. However, there is a notable difference between the 2 websites: after navigating through it, the Real Estate website provides data sets in HTML

format, and not PDF format. Even though data provided in HTML documents is defined as being not parseable, because it required custom program to do so (which is why we prefer having data in parseable open format such as CSV, or JSON files), it is provided as an open format, which is better than providing data in PDF files. Now, in addition to serve data in html format, the next step would be to provide it in one of the open parseable format mentioned above.

3) *Auctions* – Website: <https://www.eauction.ge/>

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on this eAuction website, in addition to the general recommendations made above:

- Every data sets (eg online auction) happens to have a very complete section called “Terms and Conditions”, which makes sense given the nature of the website. Should a user have questions or concerns about these data sets, it might be appropriate to add contact information into this section. That would increase the reviewability of this website. Another solution that might improve the reviewability is to always display the information about the seller: for every auction, there seems to be such a section, but it might happen to be completely empty (eg <https://www.eauction.ge/Home/EntityView/144907>)

Evaluation of Elections related data sets

1) *2012 Election Results* – Website: <http://results2012.cec.gov.ge/>

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on the 2012 Parliamentary elections, in addition to the general recommendations made above:

- The data on this website happens to be very accessible: right from the home page, different data sets are available. When browsing through the website, and looking at some more detailed informations, the url is kept short and is very explicit.
- Even though the data is not to be found in a more parseable open format (it only is available in html), it could easily be parsed with a custom program, as some data is very raw (no extra non-processable information – like styling).
- Graphs are often hard to parse using a program. The more graphs are used, the more we would recommend to serve data in a open parseable format (XML, JSON, CSV)

2) *2013 Election Results* – Website: <http://results2013.cec.gov.ge/>

Scoring sheet:

Observations:

The same recommendations we made for the 2012 elections data (see above case) are the same we're making for the 2013 elections data. Please refer to the 2012 elections data observation paragraph.

Evaluation of Online books/Catalogues related data sets

1) *Digital catalogues* – Website: <http://www.nplg.gov.ge/ec/ka/changedb.html>

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on this website, in addition to the general recommendations made above:

- Some data sets are available since 2007 ("Journal Chronicle") while other data sets are available since 1992 ("Newspaper 'Sakartvelos Respublika'"). Also, "Newspaper articles" seems that they have not been stored since 1990. Even though all these type of documents relate to the same section, they don't seem to relate to each other. But given the fact that we don't know why these data sets were not updated before (or after) a specific date, we cannot fully assert their completeness.

2) *Dictionaries* – Website: <http://www.nplg.gov.ge/gwdict/>

Scoring sheet:

Observations:

Here's what we noticed while applying the ODUI on this website, in addition to the general recommendations made above:

- We're dealing here with dictionaries, therefore it is very hard to judge the completeness of the

available data sets. That said, the online dictionaries are consistently being updated. We indeed noticed that the timeliness aspect was excellent: there are mentions of data that was updated over the past 2 weeks, or data that was created over the past week.

Conclusion

There are, of course, many other data sets available online, and it would be interesting to analyze them, in order to know if they could effectively be used (or linked to) on the final version of the Data Portal website. However, the number of data sets we applied the Open Data Usability Index on was great enough so that we were able to draw some conclusions about the Georgian data landscape, as well as the usefulness of the index.

As mentioned in the general observations, we've noticed the following:

- Among those we applied the ODUI on, no website had a legal framework text. The legal framework that would be applied upon a data set, should be written somewhere on a website, so that end users should be aware of any possible restrictions that could be linked to the usage of the data.
- Even though PDF files are very convenient for human-readability, open format files (like JSON or CSV...) should be preferred, or better, made available in addition to other formats.
- To serve data through an API will dramatically increase the transparency and the openness of the data a website is providing. In addition to providing existing open format files, we're recommending companies to create on their website an API, which will guarantee the accuracy and reusability of the data they provide.

Going through the exercise of defining these openness score allowed us to highlight what can be improved, in order to increase transparency and data-usability on these data sets. We're hoping that it also represents the proof that all the public agencies have now the tools to measure and publish open and accurate data, which will help the Georgian OGP initiative to be a successful project.