TRANSPARENCY IN HEALTH ENGAGEMENT PROJECT

Training on Data Usability

Dr. -Ing. Julia Hoxha

28-29 March 2018

Part I. Introduction of Open Data and Data Usability

1. What is Open Data?
2. Improving data usability
3. Choosing the right format
4. Achieving impact with your data
5. Initiatives of Open Data

What is Open Data?

- Open data is **data** that **anyone** can **access**, **use** and **share**.
- Governments, businesses and individuals can use open data to bring about social, economic and environmental benefits.
- Open data must be licensed. Its license must permit people to use the data in any way they want, including transforming, combining and sharing it with others, even commercially.

What makes data open?

Data is open if anyone can access, use and share it.

- Simple things to consider when defining openness:
  - + Limitations
  - + Cost
  - + Reuse

1. Introduction on Open Data

- What is open data?
- What makes data open?
- Why do we need open data?
- From data to knowledge

What makes data open?

Data is open if anyone can access, use and share it.

- Simple things to consider when defining openness:

- Limitations

- For data to be open, it should have no limitations that prevent it from being used in any particular way
- Anyone should be free to use, modify, combine and share the data, even commercially

+ Cost

+ Reuse

What makes data open?

Data is open if anyone can access, use and share it.

- Simple things to consider when defining openness:

+ Limitations

- Cost

- Open data must be free to use, but this does not mean that it must be free to access.
- There is often a cost to creating, maintaining and publishing usable data.
- Any fee for accessing open data should be ideally *no more* than the reasonable reproduction cost of the unit of data that is requested.

+ Reuse

What makes data open?

Data is open if anyone can access, use and share it.

- Simple things to consider when defining openness:

+ Limitations

+ Cost

- Reuse

- Once the user has the data, they are free to use, reuse and redistribute it – even commercially.
- Open data is measured by what it can be used for, not by how it is made available.
- Aspects like format, structure and machine readability all make data more usable, and should all be carefully considered.

Why do we need open data?

- Transforming government
- Building new business opportunities
- Protecting society and environment

Why do we need open data?

Transforming government

- Open data can help make governments more transparent. It can provide the evidence that public money is being well spent and policies are being implemented.
- Examples
  - Reduce government costs
  - Ensure proper spending of public funds
  - Transparency on law cases

Why do we need open data?

Building new business opportunities

- Open data can open up new opportunities for businesses to connect with customers
  - Transportation apps
  - Private companies connect to customers to improve services
  - Startup scene, new jobs

Why do we need open data?

Protecting society and environment

- Open data is also helping consumers to understand their personal impacts on the environment
- Warning systems for environmental disaster based on open data (UN case)
- Example: Green City Solutions

From Data to Knowledge

DATA

INFORMATION

KNOWLEDGE

Without data we can't build information,

without information there is no new knowledge.

From Data to Knowledge

DATA

Data is the raw material from which information and knowledge can be derived. Think of data as those locations, images, descriptions, reviews and prices that form the basis of information that can help you plan a trip, for example.

INFORMATION

Data becomes information when it is *given a context*.
E.g. locations, images, descriptions and prices can all help provide information related to a tourist attraction.
The collection and presentation of data helps to form information.

KNOWLEDGE

Knowledge is what is derived from information, and personalized for your needs.

Building knowledge is a process of turning information into choices.
E.g. make decision on your itinerary

From Data to Action

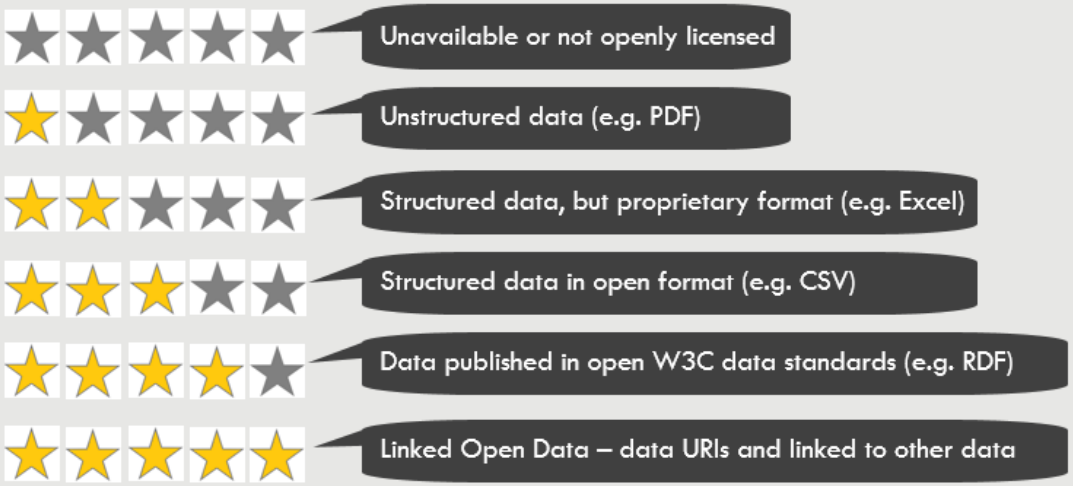
To unlock the power of *data*, just making *data open* isn't enough. It's critical to make *data* easier to find and use - to provide information and tools that make *data* accessible and actionable for all users.

2. Improving data usability

- 5-stars of linked open data
- The first three stars
- How to recognize the stars in data

The 5 stars of open data

- The 5 Stars guideline for linked open data is a way to measure how well data is integrated into the Web.
- It examines the accessibility and technical usability of a dataset
  - ranging from being available online (1 star)
  - to being part of the web of data (5 stars).
- Each star must be awarded sequentially and none can be skipped



The first 3 stars

- 1 Star: Open License

- The first star is awarded to any data that is open at a basic level.
- The content, in any format, must be available under an open license.
- Regardless of the quality of a dataset, it cannot qualify for the first star unless it is available under an open license.
- A PDF file on a website available under an open license is enough to fulfil the first star.

+ 2 Stars: Re-usable Format

+ 3 Stars: Open Format

The first 3 stars

+ 1 Star: Open License

- 2 Stars: Re-usable Format

- awarded to any data that allows for simple re-use
- data must be available in a 'highly-reusable, structured format' that can be read by a machine and understood by a human.
- key to achieving 2 stars is to select the most re-usable format. In some cases, the most re-usable format may be a closed or proprietary one, such as an Excel file.

- *Making data available in any format is better than none at all.*

+ 3 Stars: Open Format

The first 3 stars

+ 1 Star: Open License

+ 2 Stars: Re-usable Format

- 3 Stars: Open Format

- Data must be available in a structured, machine-readable format which is not tied to a specific software package.
- **An example of a dataset that would be awarded 3 stars is a CSV file with an appropriate open license.**

Web of Data: Star 4 and 5

- 4 Stars: Open Identifiers

Web of Data: a set of interlinked data resources in the Web

- Current web is configured as a series of pages or 'documents'. These documents are designed by humans to be visually attractive for other humans to read and to provide them with information.
- Web of data is a set of deeply linked data resources that can be instantly read and understood by both humans and machines.
- 4th star refers to the ability to reference things on the Web with an identifier

+ 5 Stars: Linked Data

Web of Data: Star 4 and 5

+ 4 Stars: Open Identifiers

- 5 Stars: Linked Data

- This star introduces the concept of linking data points together.
- These links enable humans and machines to navigate between data sources on the Web.
- Linking data points directly to one another allows you to direct the person or machine to a definitive record on that subject, rather than each person maintaining their own record of the data.

Openness score

Unavailable or not openly licensed

Unstructured data (e.g. PDF)

Structured data, but proprietary format (e.g. Excel)

Assessing data usability

- **Star 1 - Checking the licence**
- Finding the licence for an open dataset can be one of the hardest steps.
- An open licence is an explicit permission to use the data for both commercial and non-commercial purposes.
- The metadata record about the data
- Terms and conditions of the webpage or site
- Click on the data file, there might be a popup
- Check inside the data file itself
- Websites feature licensing information in different ways, you may have to search a little on each site to find them.

Assessing data usability

- **Star 2 - Checking (re-)usability**
- When you discover open data on the Web, there are several ways in which you can assess how usable it is. Ask yourself:
- Do I think that this is the most user-friendly format in which to publish this particular data?
- Does the file appear to be the same as the one the publisher uses?
- Is there any other information that has been added by a human?

Assessing data usability

- **Star 3 - Checking openness**
- Open formats are sometimes difficult to assess. The following questions will help you understand if the data is available in an open format:
- Can I open the file in more than one programme on my computer and still see the full functionality of the file?
- When I look up the file format (eg CSV) online, does it say I need to download a specific programme to open it?
- CSV is the most popular open format and is easy to identify. CSV qualifies for 3 stars as long as the other 2 are achieved.

Ensuring data usability

- Open data becomes usable when a *human can understand it* and a *machine can manipulate it*.
- Users of open data need the permission of its publisher, granted by an open license. But the open license alone is not enough to guarantee the usability of data.
- It is unlikely that people will be able to engage with data to derive insights and demonstrate benefits unless it is usable.

Data usability: Legal Requirements

- In addition to the open licence, there are three legal requirements that need to be considered. You must:
- Protect sensitive information like personal data.
- Preserve the rights of data owners.
- Promote correct use of the data.

Data usability: Practical Requirements

- It is important that open data can be found and relied upon by users.
- This means that open data publishers must:
- Link to the data from their website.
- Update the data regularly if it changes.
- Commit to continue to make the data available.

Data usability: Technical Requirements

- There are three recommendations that define the technical aspects of open data:
- The format in which the data is published.
- The structure of the data.
- The channels through which the data is available.

Data usability: Social Requirements

- For data use to be sustainable, it is important to have an engaged community of users. The best datasets have:
- Active support channels.
- Discussion groups and forums.
- Published how-to guides on working with the data.

3. Choosing the right format

- Why formats matter to open data
- Choosing the correct structure
- Access open data formats
- Keeping it simple with CSV

Usability, management and access

- Open data must be available for anyone to access, use and share. In this section we consider what formats maximise the usability of data, ease of access for users and simple management for publishers.

Common Data Structures

- Tabular
  - The most common structure for data is tabular. Data is organized into rows and columns listing sequential values, such as expenditure.
  - If the data is based on separate entries that are not linked to each other then a tabular file structure in a format such as CSV is ideal.
- Hierarchical
  - Hierarchical data shows the relationships between data points, such as a family tree or municipalities in each country. If the dataset depends on the relationship between data points and follows a structure in which data points are linked in vertical 'trees', a hierarchical data structure in a format such as JSON is ideal.
- Network
  - Network structured data allows relationships to exist between any combination of elements in any direction.
  - A good example of a network data structure is a social network.

Finding the right format

Downloadable data

- Tabular data is the best suited for download; most government open data portals are predominantly tabular data.
- Managing millions of rows of data may need other considerations:
  - Should the data be split into smaller datasets?
  - How often should the data be updated?
  - How will changes in the way you publish affect the previous releases?

Live data and data feeds

- Some data is not suitable to be made available as a downloadable file.
- Much of this data is updated so regularly that the file downloads would be too large for most users. This type of open data structure may be made available by a machine interface, also know as an application programming interface (API).
- There are many services that make machine interfaces available over the Web. These services can be directly integrated into other Web applications.

Starting with CSV

- When it comes to open data formats, start with CSV.
- A comma separated values (CSV) file is simply lines of data, with each data point separated from the next by a comma. CSV is perfect for tabular data and can be easily loaded into and saved from applications like Excel, making it accessible to users.
- Although CSV doesn't maintain formatting and graphs like Excel formats, it is an open, machine-readable format. CSV represents the simplest format that still supports broad reuse of open data. In other words, CSV is the 'lowest common denominator' for open data – open data should be made available in this format wherever possible?

4. Achieving impact with your data

The goal of any open data project or initiative is to have some kind of impact, be it political transparency and accountability, social benefit or economic growth.

- Why is achieving impact is important?
- How can we achieve impact?
- How can we measure the impact we have achieved?
- Overcoming barriers to achieving impact

Achieving impact with your data

- Achieving impact, and measuring that impact, is one of the most challenging aspects of leading an open data project, programme or initiative. There are a number of techniques that you can use to help you to achieve impact.
- Identify shared goals
- Seek out and foster stories
- Reflect, iterate, reflect

Measuring impact

- There are a number of assessment tools that can help you assess the impact of an organisation or country's open data policy, programme or initiative. Each assessment tool uses slightly different methodologies and measurements, so find the one that works best for your context. Think about which criteria are important for your organisation.
- This module looks in detail at the European Data Portal's Open Data in Europe tool, which compares the open data initiatives of European Union countries.
- You can also look at Global Open Data Index and Open Data Barometer, which compare countries internationally.

Measuring impact

- Open Data in Europe looks at three areas of impact: **political, economic and social.**
- The factors that make up the political impact indicator are: impact on government efficiency and effectiveness, and impact on transparency and accountability.
- The economic impact indicator is based upon the implementation of multiple macro-economic studies assessing the market value of open data and studies regarding better service delivery or looking at related subjects.
- The social impact indicator includes impact on environmental sustainability and the inclusion of marginalized groups in policymaking and in accessing government services.

5. Open Data Initiatives

- ODI UK
- Open Data Gov US
- Open Government Partnership
- Albanian Law

Exercise Session No. 1

- Assessing overall data usability
- Exercises on the Badge Stars
- Detecting open data standards in the Albanian Law
- How useful is our data? Hands-on assessment exercise with examples publication/report from ALSAI