Annex A. - Guide for identification of datasets – Inventory of datasets

When public sector organization make data inventory and identify data and datasets that they collect and/or produce the following definition of data and dataset must have in mind:

- Data – are things known or assumed as basic values or facts, making the basis of reasoning or calculation.[1] According to the Law on Use of Public Sector Data, data are a qualitative or quantitative value, or a separate part of an information.[2] Data become information when analysed to extract meaning and to provide context. The meaning of data can vary, depending on its context.
- Dataset - any organized collection of data.[3] The most basic representation of a dataset is the combination of data elements presented in tabular form. Each column represents a particular variable and each row corresponds to a given value of that column's variable. A dataset may also present information in a variety of nontabular formats, such as an extended markup language (XML) file, a geospatial data file, an image file, etc.
- Metadata – are (descriptive) data that explain the meaning of data.[4] Metadata provide relevant description about the data and datasets, such as: responsible access coordinator, associated laws and regulations (if any), structure, data elements, interrelationships with other data and/or datasets, and other characteristics of data/dataset such as: its creation, disposition, access and handling controls, formats, content, and context, as well as related audit trails.

The creation of dataset inventory is an iterative process and has the following four steps:[5]

1. Analysis of laws and regulation related to the work of public sector organization with focus on data/datasets production and/or collection.
2. Identification of all data sources.
3. Identification of all datasets from all data sources.
4. Creation of dataset inventory.

Step 1: Analysis of laws and regulation with focus on data/datasets

An analysis of the laws, bylaws and other regulation related to the work of the public sector organization should be done with aim to identify all data/databases/datasets/registers/records that the organization is obliged to produce and/or collect. The analysis should also provide the datasets structure if it is prescribed by any regulation. For example, in the Law on Health Care, in the Article 9-a it is written that the public healthcare institutions are obliged to keep records of the medical equipment they have and with which they perform the healthcare activity. The records of the medical equipment shall contain in particular data on: type of equipment, description of the equipment, medical specialty or subspecialty in which the equipment is used, whether the use of equipment requires scheduling through the electronic

[1] https://en.oxforddictionaries.com/definition/data (29.12.2017).

[2] Law on Public Sector Data Use (2014), Official Gazette of Republic of Macedonia, no. 27. 3.2.2014. p.2.

[3] https://stats.oecd.org/glossary/detail.asp?ID=542 (30.12.2017).

[4] Ibid.

[5] http://open-data-manual.readthedocs.io/en/latest/inventory.html (2.1.2018)

list of scheduled examinations and interventions, the year of production, the name of the manufacturer, the year of purchase, the date of conclusion of the procurement contract, i.e. the contract for donation and the archival number under which the contract is registered, the purchase price of the medical equipment, the date of signing the contract for maintenance of the medical equipment and the archival number under which the contract is entered, depreciation rate, number of repairs carried out, parts replaced and degree of utilization.

The following words/terms/phrases could help to identify the data/datasets that should be produced and/or collected by public sector organization according to some regulation: date, data, list, electronic list, lists, network, registry, registries, record, system for electronic records, records, notifications, report.

Step 2: Identification of all data sources

Some of identified datasets might already exist (in analogue/paper or digital/computer form), and some of them have yet to be established. Both existing and non-existing (but prescribed by the law), in digital or analogue form should be listed in the inventory.

Existing data may be housed in a variety of places from inside information systems or databases stored on individual and/or shared drives and folders, or in paper form in some registrars, registers or other records. This step is about identifying the major data sources in the public sector organizations.

The following questions could be used as a guide how to identify data sources in each public sector organization (including all sectors and departments):

1. What paper-form registrars, registers or other records does the organization use?
2. What information systems does the organization use?
3. What databases does the organization use?
4. What applications capture information or are used in organization’s business processes?
5. Are some data resources kept in spreadsheets (on shared or individual drives)?
6. What data and/or information do the organization already publish? Where did that data/information come from?

Step 3: Identification of all datasets from all data sources

Some of organization’s data/information sources may be fairly straightforward. Some frequently generated reports or spreadsheets would be good examples of a dataset.

The following questions could be used as a guide how to identify datasets in each public sector organization (including all sectors and departments):

1. What datasets are used for reports?
2. What datasets are publicly available online or elsewhere?
3. What datasets are used internally?
4. What data/information is published as a performance metric?
5. What data is reported to other public service organizations?
6. What data do other departments ask for?

Step 4: Creation of dataset inventory

Each dataset identified in Step 3, should be added to the dataset inventory. The dataset inventory will contain basic information about each dataset.

More precisely, the dataset inventory must have the following data about each dataset:

1. Unique Identifier - A unique identifier for the dataset.
2. Title – The name of the dataset. Should be in plain Macedonian and include sufficient detail to facilitate search and discovery.
3. Description – Description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the dataset is of interest.
4. Contact Email - Contact person's email for the dataset.
5. Format - What is the primary state or file format for containing this dataset? (i.e. paper, database, excel, CSV, JSON, other).
6. Keyword – Keyword or tag that describe the dataset
7. Public sector organization's name – The formal name of the organizations that release the dataset.
8. Theme – Main thematic category of the dataset.
9. Documentation - URL to the documentation describing the data set (if it is published).
10. Frequency - The frequency with which dataset is published.
11. Landing page - URL where the dataset is located (if it is published).
12. Language – The language of the dataset.
13. Spatial - The range of spatial applicability of a dataset. Could include a spatial region like a bounding box or a named place.
14. Temporal coverage - The range of temporal applicability of a dataset (i.e., a start and end date of applicability for the data).
15. License - The license with which the dataset has been published.

For the users of open data, it is very helpful if every dataset inventory is accompanied with the dataset structure for each dataset with the following data:

1. Unique Identifier - A unique identifier for the dataset.
2. Title – The name of the dataset. Should be in plain Macedonian and include sufficient detail to facilitate search and discovery.
3. Name - Human-readable name of the column.
4. Column description - Human-readable description of the column's contents.

An individual or group should be charged with oversight of the inventory to ensure its ongoing maintenance and accuracy.

Annex B. – Main data quality dimensions

There are several different dimensions for obtaining and improving open data quality:[6,7]

- **Accuracy** – is the extent to which it correctly represents the characteristics of the real-world objects, situation or event.
- **Availability** – is the extent to which it can be accessed. This also includes the long-term existence of data.
    - *Example:*
        - *A dataset that is identified by a certain URL that resolves persistently to the right resource (and does not give back 404 Not found).*
    - *Recommendation:*
        - *Responsibility for the maintenance of data should be clearly assigned in the organization.*
- **Completeness** – is the extent to which it includes the data items or data points that are necessary to support the application for which it is intended.
    - *Example:*
        - *A dataset that includes spending data for all ministries enables a complete overview of government spending.*
    - *Recommendation:*
        - *In order to include all the necessary data points a capture and publication process should be designed and detailed procedures should be developed that will check if completeness is fulfilled.*
- **Conformance** - is the extent to which it follows a set of explicit rules or standards for capture, publication and description.
    - *Example:*
        - *A description of a dataset (metadata) according to the DCAT-AP standard.*
        - *Publishing open data based on W3C standards.*
    - *Recommendation:*
        - *The most relevant and most used standards in the domain should be applied.*
- **Consistency** – is the extent to which it does not contain contradictions that would make its use difficult or impossible.
    - *Example:*
        - *A description of a dataset where the data of last modification is not before the creation date.*
        - *Dataset that contains data on the name of a municipality which is entered as free text with a possibility of a mistake while entering the name. Thus, some records will contain the exact name of the municipality (Skopje), while others will contain misspellings (Skojpe).*

[6] PwC (2014). *Open Data & Metadata Quality – Training Module 2.2,* EC, Open Data Support.

[7] https://www.w3.org/2013/share-psi/bp/eqa/ (8.2.2018)

    - *Recommendation:*
        - *All data should be processed before publication to identify all possible conflicting statements and other errors (in particular if data is collected and aggregated from different sources)*
- **Credibility** - is the extent to which it is based on trustworthy sources or delivered by trusted organizations.
    - *Examples:*
        - *A dataset that contains data from processes that can be independently verified, e.g. election results or parliamentary proceedings.*
    - *Recommendation:*
        - *Data should be based on sources that can be trusted.*
- **Processability** - is the extent to which it can be understood and handled by automated processes.
    - *Examples:*
        - *A dataset that contains coded information based on publicly available controlled vocabularies and code lists.*
        - *A description of a dataset that expresses dates in W3C Date and Time Format (e.g. 2013-06-01) rather than as text (e.g. 1 June 2013).*
    - *Recommendation:*
        - *To apply recommendations for syntax of data given in common standards and application profiles.*
        - *To Identify the source of terminology and codes used in the data in machine-readable manner.*
- **Relevance** - is the extent to which it contains the necessary information to support the application.
    - *Examples:*
        - *A Dataset that contains temperature measurements rounded to degrees Celsius for climate calculations; a dataset with precision of a thousandth of a degree for chemical reactions.*
    - *Recommendation:*
        - *To match coverage and granularity of data to its intended use within constraints of available time and money.*
        - *However, potential future usage of data should be also considered.*
- **Timeliness** - is the extent to which it correctly reflects the current state of the entity or event and the extent to which the data (in its latest version) is made available without unnecessary delay.
    - *Examples:*
        - *A dataset that contains real-time traffic data that is refreshed every few minutes.*
    - *Recommendation:*
        - *To adapt the update frequency of data to the nature of the data and its intended use*
        - *To make sure that processes and tools are in place to support the updating.*

Annex C. – Prioritization model for opening data

Prioritization will be made according to the following set of criteria:

- Institutions will open up data that is already in an open format.
- Institutions will open up datasets that are already publicly available, not in open, but in some other format.
- Institutions will open up data that is updated, well-structured and of high quality.
- Institutions will open up data that require minimal input of resources to prepare for opening.
- Institutions will open up the datasets for which there is a request for opening made by potential users (private sector, civil sector, other government institutions, etc.).
- Institutions will open up datasets that appeared to be useful given usage elsewhere (Open Data Index, Open Data Barometer).

Each of these criteria priority has equal weight in determining the final grade for and gets a value of 1 for each fulfilled criterion. The final grade is the sum of the individual grades of all 6 criteria. The highest priority for opening will have those datasets that will have the highest rating of all the criteria that the institution is having (the maximum possible score is 6). By completing the publication of the datasets with a priority score of 6, the process continues with the opening up the datasets with the grade 5 for priority, etc., until the last sets with priority grade of 1 are exhausted.

By creating new datasets due to new legislation or due to the need itself, the practice in the functioning of the institution should be given priority for their opening according to this model. Gradually (after 2-3 years from the beginning of the publication of open data on the national open data portal) the criterion for the requests from the users will get a bigger weight factor and will become increasingly dominant in determining the priority for the opening.

Annex D. - The 5-star model for OGD and set of open standards

Datasets will be published according to the 5-star model for Open Government Data as a minimum 2-star level, where:[8]

- 0 star – data is not available with an open licence.
- 1 star – data with documented metadata are available online with open license permitting re-use.
    - *Example:*
        - *PDF file containing table presented as a scanned image table with data for temperature forecast for Skopje: day, lowest temperature, highest temperature.*
    - *Benefits for the users and publishers:*
        - *User can look at the data, print it, store it locally, can enter the data into any other system, can change the data, and share them with anyone*
        - *For the public sector institution, it is simple to publish*
- 2 stars – data with documented metadata are available online in a machine readable format with open license permitting re-use.
    - *Example:*
        - *Excel (.xlsx) file instead of image scan table with data (in rows and columns) for temperature forecast for Skopje: day, lowest temperature, highest temperature.*
    - *Benefits for the users and publishers:*
        - *The users can do all as 1-star data and in addition they can directly process the data with some proprietary software to aggregate it, perform calculations, to make charts and diagrams, etc.*
        - *For the public sector institution, it is simple to publish*
- 3 stars – data with documented metadata are available online, in non-propriety machine readable format, with open license permitting re-use.
    - *Example:*
        - *Data for temperature forecast for Skopje in CSV format instead of Excel format.*
    - *Benefits for the users and publishers:*
        - *The users can do all as 2-stars data and in addition they can manipulate the data in any way they like, without any need to have any proprietary software*
        - *Publishers might need converters or plug-ins to export the data from the proprietary format*
- 4 stars - data are available online, in non-propriety machine readable format, with open license permitting re-use. Data are described in a standard way and uses unique reference indicators (URIs) to identify things, so that people can point to the data.
    - *Example:*

[8] KDZ, (2016). *Open Government Implementation Model – Implementation of Open Government, Ver. 3.0,* KDZ – Centre for Public Administration Research, p.28

        - *Data for temperature forecast for Skopje in RDF format (Resource Description Framework). On one hand there are information where a resource can be found. On the other hand, there the first steps to a semantic approach. It defined what is to be talked of.*
    - *Benefits for the users and publishers:*
        - *The users can do all as 3-stars data and in addition it can link to it from any other place, can bookmark it, can re-use part of the data, can combine the data safely with other data, URIs are a global scheme so if two things have the same URI then it's intentional, and if so that's well on it is way to being 5-star data.*
        - *Publishers have fine-granular control over the data items and can optimize their access (load balancing, caching, etc.), the publishers can link to other publisher's data – upgrade it to a 5-star rate, need to assign URLs to data items and think about how to represent the data.*
- 5 stars – data are available online, in non-propriety machine readable format, with open license permitting re-use. Data uses unique references and links to other data to provide context.
    - *Example:*
        - *data for temperature forecast for Skopje in RDF format (data provided by the National Hydro-meteorological Service - Republic of Macedonia) and linked data for air quality in Skopje (data provided by the Ministry of environment and physical planning – Republic of Macedonia)*
    - *Benefits for the users and publishers:*
        - *The users can do all as 4-stars data and in addition they can discover more related data while consuming the open data, they can directly learn about the open schema.*
        - *Publishers can make their own data discoverable, can increase the value of their data, and their institution will gain the same benefits from the links as the users, but they will need to invest resources to link their data to other data on the web, and might need to repair broken or incorrect links.*

All public sector institutions will open up and publish their data according to a set of best practices to maximize re-usability e.g. the 5-star model and as a **minimum 2-star level**.

Annex E. – Possible scenarios for opening up source data by ETL (Extract, Transform, Load)

Four different scenarios are possible for opening up data:[9]

1. **Opening up data from existing publications/documents.**
   The public sector institutions collect data from an existing publication/document (graphs or tables in PDF or DOCX file) and publish them as open data. Here the data source from which the raw data are collected and processed in order to be included in a publication should be found. In this scenario, there is no extraction and transformation steps, only metadata for the dataset should be collected and the dataset should be published as open data.

2. **Opening up data from an existing dataset.**
   The public sector institutions already publish data and information on their web sites in a format such as XLSX available via download or viewer. Here, in the existing process, the IT department will extract (if the data from the publication do not satisfy the open data criteria), transform and publish the data in an open data format. Extraction will isolate data and filter them from the database in a uniform dataset. Transformation encompasses a thorough quality check of the data, as is the case in every dataset environment. For instance, using uniform names for fields and content - no cryptic abbreviations, storing addresses in a consistent manner, writing names in full and in the same format, etc. All this will be done only for the datasets with the highest priority to be open up. As final step metadata for the dataset should be collected and the dataset should be published as open data.

3. **Opening up data from a database.**
   In many cases the core data are in a database that has been created for an application to support a business process for the public sector institutions. The basis of this scenario is that the data must be extracted by the IT department from the application database first, before they are prepared for publication as Open Dataset. The assumption in this scenario is that application is developed internally on one of the existing internal environments (for instance, built in Java, .NET or other) and the database is one of the standard used by the public sector institutions (like Oracle, SQLServer, PostGre, etc.). Extraction will be done using the standard techniques for reading out tables from the database systems, and making them available as flat files.[10] For the more frequently changed data, more suitable is to read out the data through ODBC or JDBC drivers. Transformation encompasses a thorough quality check of the data, as is the case in every dataset environment. For instance, using uniform names for fields and content - no cryptic abbreviations, storing addresses in a consistent manner, writing names in full and in the same format, etc. As

[9] Government of Flanders in Belgium (2014). Open Data Manual: Practice-oriented manual for the publication and management of Open Data using the Flemish Open Data Platform, p.16-27.

[10] The method for extraction is to write SELECT queries that extract just the data needed into a flat, CSV file.

final step metadata for the dataset should be collected and the dataset should be published as open data.

4. **Opening up data from an existing source system or package.**

   Public sector institutions often use a commercial package that has its own database. These data can often not be accessed directly or they may be stored in a proprietary format that is determined by the supplier. In this scenario techniques are used for extracting these data from the package, transforming them and publishing them as Open Data. The difference with Scenario 3 is that packages often require the data to be opened up through other channels (like API or package-specific tools).