Importance of data quality of High Value Datasets (HVDs)

Non-paper of the Czech Republic

This document is intended for European Commission, CNECT Unit G.1 Data Policy and Innovation, author of impact assessment study, Open Data Committee, national contact points and their technical teams in charge of technical implementation of upcoming concept of high-value datasets.

Summary

The publication of the so called high-value datasets, a term introduced by the EU open data directive (2019/1024), will come up short unless their data quality is properly addressed along with their list. It is imperative that for each high-value dataset on the list, a technical specification exists along with a set of required data quality attributes. Compliance with the specification and the specified data quality attributes by publishers of the dataset in individual member states must be enforced, otherwise the published datasets will not be interoperable, and their value will be unexploitable. In this document the Czech Republic identify relevant data quality attributes and suggest how they could be defined and evaluated. The Czech Republic requests that each upcoming high-value dataset definition is accompanied by its technical specification, a set of required data quality attributes, and a mechanism of monitoring and enforcement of their compliance.

The requirements in this document come into play when the high-value datasets are identified. To ensure cross-border interoperability of the published data coming from individual member states representing a single high-value dataset, it is completely insufficient to just identify it. It needs to be ensured that for each such dataset, a proper technical specification is created and a required set of data quality attributes defined so that the member states can start publishing the data in a compliant, interoperable way.

The Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information introduces so-called *high-value datasets*, which, when used repeatedly by various data processors, will have interesting benefits for the European economy and society. The list of those datasets is currently being prepared. Following approval, the listed datasets will have to be published in all EU member states as open data, i.e. in a machine-readable and open format, without restricting terms of use, available for free on the Internet.

The concept of high-value datasets pertains to referred thematic categories, which currently are as follows:

- Geospatial
- Earth observation and environment

- Meteorological
- Statistics
- Companies and company ownership
- Mobility

It is imperative for fulfilling the goals of the directive 2019/1024/EU and the entire concept of high-value datasets that the implementing act dictates not just the datasets which the member states ought to provide but also the minimal requirements on data quality properties of the published data so that the quality of the data with regards to the aforementioned categories is guaranteed. The requirements put on the published data by the directive itself (i.e. open and machine-readable format) are too generic to sufficiently guarantee compatibility of data sources originating from various member states. Emerging differences could in practice prevent effective usage of European data and therefore undermine the goals which the 2019/1024/EU directive is trying to achieve in the context of the unified digital market. According to Article 14 of Directive 2019/1024/EU, a part of the implementing acts must, therefore, be a data specification that assures satisfactory technical harmonization and therefore high quality of data provided Europe-wide.

Thanks to the directive, EU member states will take the first step towards creating a unified EU open data space, in which data processors can create services that are usable EU-wide. That would be interesting progress compared to the current state, where the creation of similar services is very difficult. It is often the case that a dataset provided in one EU state is not available at all in other states or is not available as open data (e.g. access is subject to fees). While the creation of a unified space can be done in other areas (e.g. consumer protection), it is necessary to complete many steps in order to unify the open data space. The current fragmentation of open data can be perceived as contrary to the principles of a unified EU market because of this.

The directive cannot contribute to creating a unified data space without a uniform standard of quality for high-value datasets. We would eventually see member states formally providing all mandatory datasets, but each in a different way. It would still be very difficult to create EU-wide services on top of the heterogeneous datasets since data processors would have to invest a majority of their resources towards data integration instead of creating valuable services.

Before considering the unification of quality standards for datasets, it is necessary to examine the term *data quality* first. Data quality is a broad term that can nevertheless be summarized with a simple definition. High quality data is data that is well prepared for processing by a data processor. Well prepared data is data that can be easily located and that the processor can count on the correctness of its contents and the immutability of the form of its provision in time and space. The form of data provision can have many attributes. Among them are data structure, semantics, completeness and granularity, data interconnectivity, the formats in which the data is provided, the data access technical interface, the form of the dataset catalogization record in the European Data Portal, etc. The immutability of the form of data provision in time entails that the provider does not change individual attributes over time and makes only the necessary backwards-compatible changes that aim to increase quality. The immutability of the form of data provision in space means that the attributes are the same across individual providers of the dataset.

Oftentimes, these attributes are summarized into 4 main principles as FAIR data (Findability, Accessibility, Interoperability, Reusability).

We, therefore, request that all EU member states and the European Commission put maximum effort into achieving a uniform definition of high-value dataset quality. Since the term is very broad, it is necessary that the definition provides a concise listing of attributes which are important in terms of creating a unified open data space within the EU. Furthermore, we request that the definition of quality be applied to every formulated high-value dataset; that is, determine the exact fulfillment of individual attributes of the definition of quality for each dataset.

We suggest that the definition of quality includes attributes listed in the following table and that they are fulfilled for each dataset in a way also mentioned in the table.

Quality attribute	**Fulfillment for a certain dataset**
structure and semantics	specified by an ontology expressed in all EU languages based on the ISA^2 Core Vocabularies and on other existing ontologies and vocabularies as much as possible.
completeness	a requirement to publish the entire list of all existing instances of all elements of the ontology determining the structure and semantics
granularity	a requirement to publish data on the granularity level according to the ontology without any aggregation of instances of specified ontology elements
data linking	usage of prescribed code lists and thesauri from the EU Vocabularies, which will be actively maintained for the needs of high-value datasets; semantic links to other datasets are expressed both on the ontology level and the level of particular entities expressed in the dataset
formats	definition of one or more formats using logical schemas (e.g. RDF, XML, JSON schema, etc.) such that the dataset is made available in at least one of these formats; usage of Linked Data principles for data linking, the concept of IRIs in particular; elements of the logical schema take advantage of or are mapped to ontology elements, which define their semantics
data access interface	prescribe the required interface, which is bulk download, API, or both; in the case of bulk download, the mandatory distribution of the content of the dataset to data files is prescribed, such that the files follow the specified ontology (see the structure and semantics attribute) and logical schema (see the format attribute); in the case of API, a mandatory technical specification of API is prescribed, which

	adheres to the ontology (see the structure and semantics attribute) and the logical schema (see the format attribute). Furthermore, the API must, in addition, provide a way to conveniently download the complete dataset
documentation	prepare unified documentation in all EU languages applicable for datasets from all countries publishing that dataset. The documentation must contain an example guide for users teaching how to use the datasets, starting from their discovery in the European Data Portal and ending in an example Proof of Concept application.
catalogization	prescribe a catalogization record according to the current DCAT-AP specification for the purposes of mandatory catalogization in the European Data Portal
terms of use	prescribe open terms of use in all EU languages guaranteeing their compatibility across law systems of publishing member states

The created definition specifies requirements on the minimal form of a dataset. Each provider will be obligated to adhere to this definition for the publication of the dataset. They can, however, extend it in any way that is backwards-compatible. Backwards-compatibility means that the provider has to make sure that any correct way of software dataset processing according to the uniform definition also works correctly with the extended definition specified by the provider.

Furthermore, we require that clearly defined, well documented and replicable automated and manual mechanisms for quality control of provided datasets defined by individual attributes mentioned above are created on the EU level. Quality testing should be performed routinely and its results should be made public on a website created for this purpose, perhaps even on the European Data Portal. Publishing results publically then puts pressure on providers who do not satisfy the required quality of publication and creates positive feedback for providers who publish high quality data. Manual testing must follow a clearly defined methodology. Automated tests must be conducted using open-source tools only and must be documented in such a way that anybody can verify the results; that is, deploy the tests themselves.

We propose the ways in which the attributes could be tested in the following table:

Quality attribute	**Way of testing a given dataset**
structure and semantics	The structure can be tested with automated data validators (SHACL, XML Schema, CSV on the Web validator). Semantics can be tested manually by evaluating data visualizations from individual countries in a Proof of Concept (PoC) application. The PoC application would be created along with the technical specification of the dataset by the authors of the specification. It would serve the publishers as a validation tool for their published data.
completeness	For the monitoring of data completeness, it would be appropriate to know how many records each dataset ought to have. Then, one can automatically check whether the provided data contains that number of records.
granularity	A set of database queries over the provided data, which would measure the number of occurrences of individual classes and properties across datasets provided by individual providers, can be used. Afterwards, manual evaluation of results is required, for example in the context of other datasets, whether any unexplained anomalies are present.
data linking	An automated tool measuring counts, types, and availability of outgoing links across datasets from individual providers can be created. Afterwards, manual evaluation of results is required, for

	example in the context of other datasets, to determine whether any unexplained anomalies are present.
formats	The provided formats of the datasetscan be automatically monitored and compared to their metadata records in the European Data Portal. This may include validation of the individual distributions.
data access interface	The availability of the API and the bulk data downloads, HTTP header correctness and the availablilty and validity of API documentation in the OpenAPI standard can be automatically monitored according to the metadata records of datasets in the European Data Portal.
documentation	Whether the datasets link to a unified documentation of the dataset can be checked automatically according to the metadata records of datasets in the European Data Portal.
catalogization	The validity according to DCAT-AP v2.0.0 or higher can be checked automatically according to the metadata records of datasets in the European Data Portal.
terms of use	Whether the record links to the specified terms of use and their availability can be checked automatically according to the metadata records of datasets in the European Data Portal.

In addition, it is necessary to routinely perform data usability studies for high-value datasets. Those can be conducted for example by questionnaires filled by known users of these datasets. These users could also voluntarily declare that they are using the datasets. Any involuntary dataset user tracking by requiring registration must be out of the question.

A position paper by the team of the National Open Data Coordinator of the Czech Republic, Department of the Chief Architect of the eGovernment, Ministry of the Interior of the Czech Republic. (michal.kuban@mvcr.cz)