How to automate Open Data preservation

Martin Rechtorik

DLM FORUM

National Archives
Czech Republic

FAV

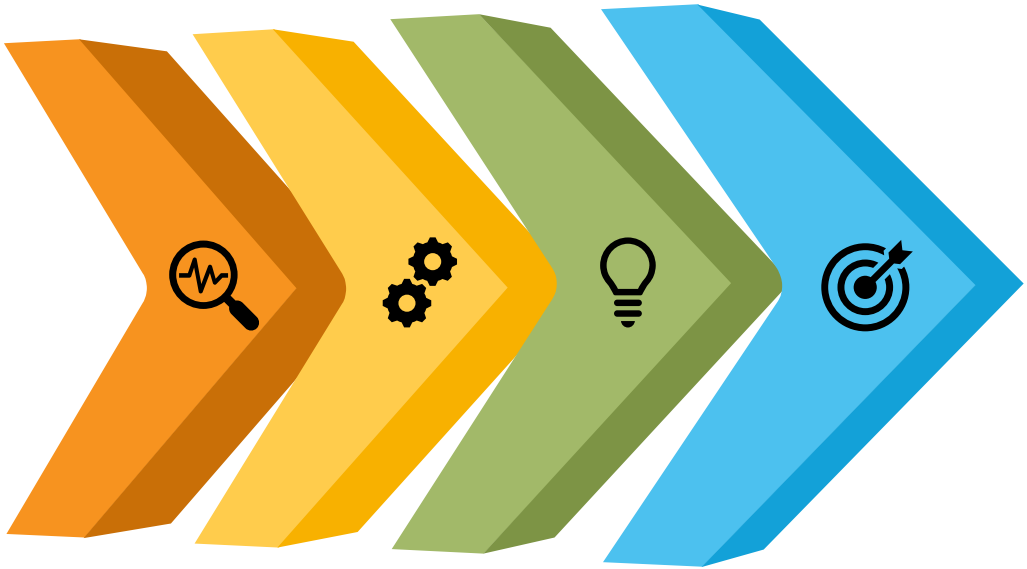
Agenda

1. Open Data, Metadata

Brief revision of several important facts from Salamanca.

2. DCAT, RDF…

Metadata is “must have” part of digital preservation and Open Data has it.



3. Archives and Open Data

Reasons to put effort and resources into it and not let it go.

4. Automated Preservation

What it takes to establish such workflow.

National Archives
Czech Republic

Open Data

- Open data refers to data that is freely available for anyone to access, use, and share without restrictions or limitations, except for minimal requirements such as attribution or the obligation to share derived data in a similar open manner.
- Open data is typically made available by governments, organizations, or individuals and is often released in a machine-readable and easily accessible format.
- When it is published the producer´s work is usually over and nobody cares and if somebody cares then only in short term.

Key characteristics

1. Accessibility: Open data should be easily accessible to the public through the internet or other means, without requiring special permissions, registrations, or fees.
2. Usability: Open data should be provided in a format that allows for easy analysis and use. Common formats include CSV, JSON, XML, or other machine-readable formats.
3. Reusability: Users should be able to use and republish open data without restrictions, except for basic attribution or sharing requirements.
4. Licensing: Open data is typically released under open licenses like Creative Commons or similar frameworks, which specify the terms and conditions for use.
5. Transparency: Open data initiatives often aim to promote transparency and accountability by making government data, research findings, and other information accessible to the public.

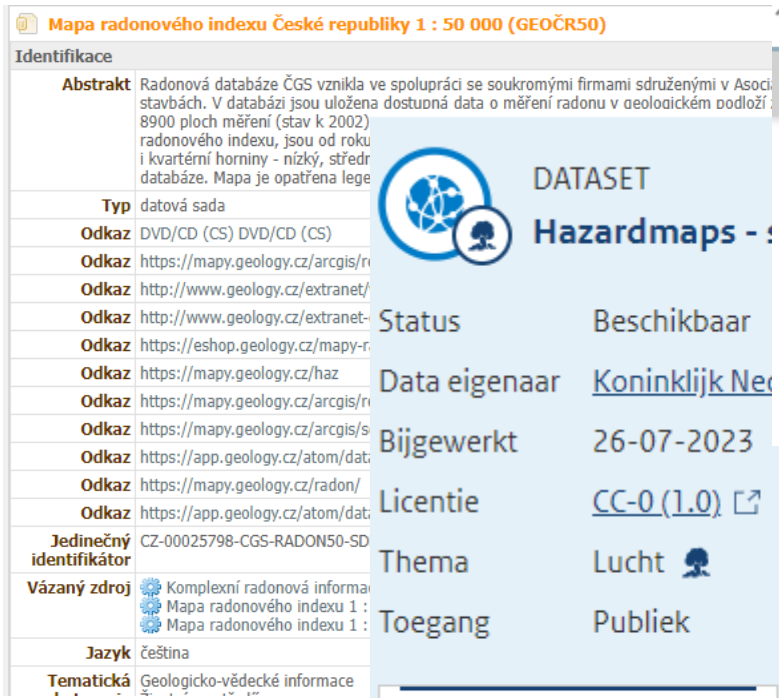
Metadata Issue

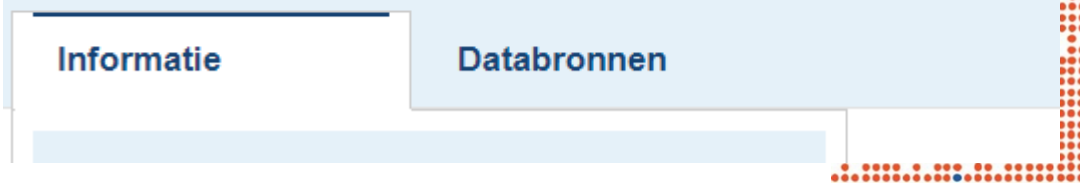
- No dataset is an island and important is quality of metadata
- DCAT = Data Catalog Vocabulary (now in v.3)
- Various standards (W3C DCAT standard (implementations: EU DCAT standard, Federal Governments of the USA, etc.)
- DCAT 3: An RDF vocabulary for describing data catalogs on the Web, which enables interoperability and federated search across catalogs.
- New features: DCAT 3 extends DCAT 2 with new classes and properties for versioning, dataset series, inverse properties, and more.
- Backward compatibility: DCAT 3 preserves the DCAT namespace and does not break the definition of previous terms. Existing DCAT 2 deployments do not need to upgrade unless they want to use the new features.

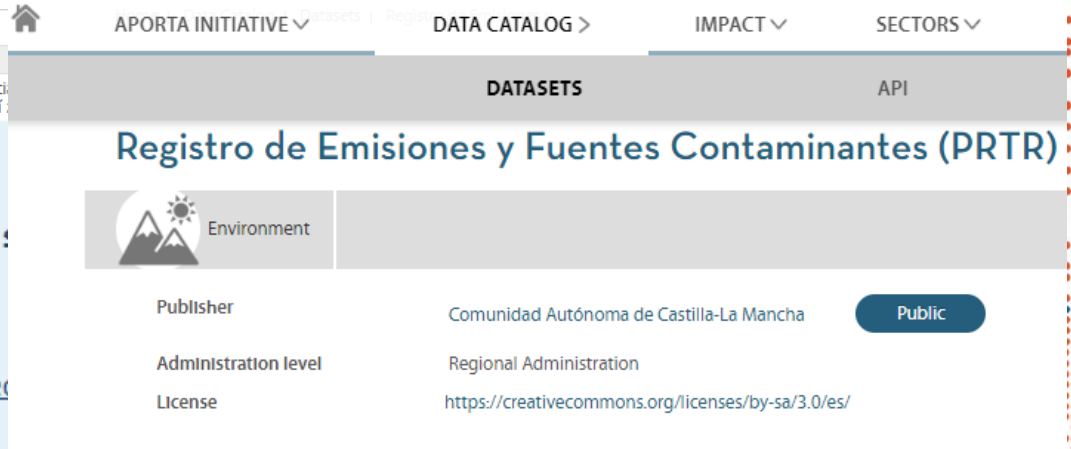
DCAT

- The original DCAT vocabulary was developed and hosted at the Digital Enterprise Research Institute (DERI), then refined by the eGov Interest Group, and finally standardized in 2014 [VOCAB-DCAT-1] by the Government Linked Data (GLD) Working Group.
- A second recommended revision of DCAT, DCAT 2 [VOCAB-DCAT-2], was developed by the Dataset Exchange Working Group in response to a new set of Use Cases and Requirements [DCAT-UCR] gathered from peoples' experience with the DCAT vocabulary from the time of the original version, and new applications that were not considered in the first version.
- This version of DCAT, DCAT 3, was developed by the Dataset Exchange Working Group, considering some of the more pressing use cases and requests among those left unaddressed in the previous standardization round. A summary of the changes from [VOCAB-DCAT-2] is provided in Change history.

Various DCATs examples







National Archives
Czech Republic

Let´s look closer

Abstract

Hazard maps published by the KNMI. The last version (v4) has been published on June 15, 2017. Previous versions were originally published in June 2016 (v2), October 2015 (v1) and December 2013 (v0). The hazard map shows surface peak ground acceleration (PGA, period T = 0.01 s) in the unit of [g] (9.82 m/s^2). The hazard map is calculated for the province of Groningen in the northern part of the Netherlands. The technical reports explain the specifications of the different versions of the hazard maps: http://www.knmi.nl/kennis-en-datacentrum/publicatie/probabilistic-seismic-hazard-analysis-for-induced-earthquakes-in-groningen-2013, http://www.knmi.nl/kennis-en-datacentrum/publicatie/probabilistic-seismic-hazard-analysis-for-induced-earthquakes-in-groningen-update-2015, http://www.knmi.nl/kennis-en-datacentrum/publicatie/probabilistic-seismic-hazard-analysis-for-induced-earthquakes-in-groningen-update-june-2016, http://cdn.knmi.nl/system/readmore_links/files/000/000/408/original/20170615_Technisch_rapport_hazardkaart_Groningen_2017.pdf Seismic hazard map 2017 (v4) which provides the spectral accelerations for specific locations and return-periods: http://rdsa.knmi.nl/hazard/

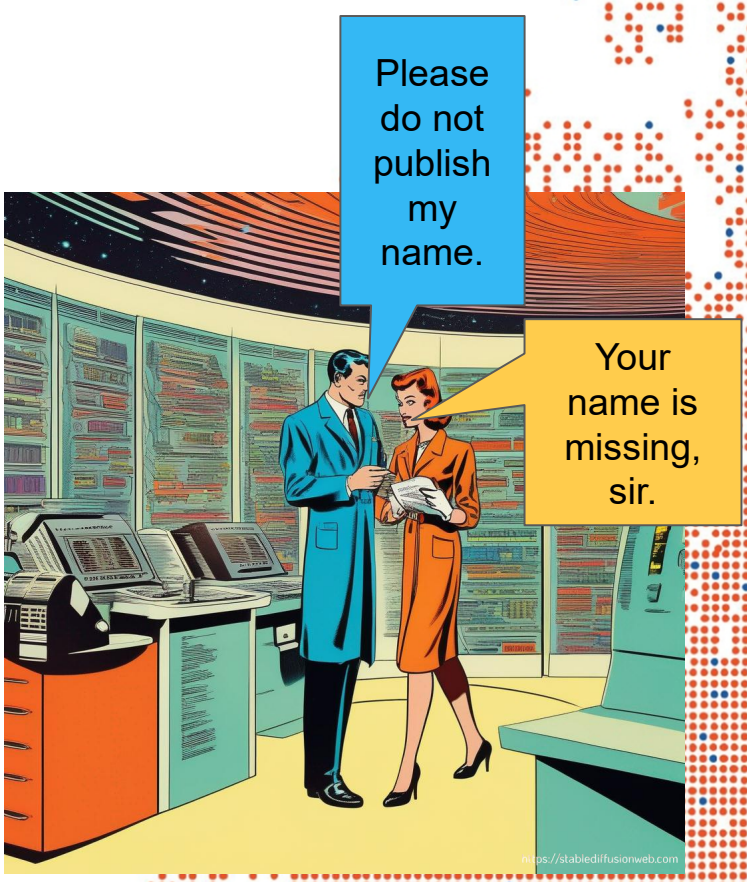
Dynamic | Earthquake | Earthquake hazard | Groningen | Ground accelerations | Ground motion | Hazard | Hazardmap | Open data | Pga | Seismic hazard | Seismicity | Seismology | Strong motion

National Archives
Czech Republic

Even closer…

Metadata

Dataset name	seismic_hazardmaps
Status	onGoing
Last metadata update	March 28, 2022, 2:44 PM (UTC+02:00)
Update frequency	asNeeded
License	https://creativecommons.org/publicdomain/zero/1.0/
North bound latitude	53.609034
East bound longitude	7.202048
South bound latitude	53.012058
West bound longitude	6.279166
Dataset version	1
Dataset edition	9
Dataset manager	Jesper Spetzler
Maintainer	KNMI Data Services
Publication timestamp	2016-10-13T13:41:19Z
Reference system identifier	EPSG28992
Dataset start time	2013-12-01
Dataset end time	2022-06-15



National Archives
Czech Republic

Resource Description Framework and GraphDB

- RDF is a standard model for data interchange on the Web. It was developed and standardized by the World Wide Web Consortium (W3C). RDF is used for representing highly interconnected data. Each RDF statement is a three-part structure consisting of resources where every resource is identified by a URI.
- RDF allows describing anything: persons, animals, objects, and concepts of any kind. They are considered resources. RDF represents meaningful information for software applications. We represent information by statements in the following format: **<subject> <predicate> <object>**. Those statements express a relation between the subject and the object. Both, the subject, and the object are resources.
- RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF enables effective data integration from multiple sources, detaching data from its schema. This allows multiple schemas to be applied, interlinked, queried as one and modified without changing the data instances.

SPARQL

- pronounced “sparkle” means SPARQL Protocol and RDF Query Language.
- It’s essentially a semantic query language for databases.
- Using SPARQL, you can extract any kind of data, with a query composed of logical combinations of triples.
- It’s recognized as one of the key technologies of the semantic web.
- SPARQL sees your data as a directed, labeled graph, that is internally expressed as triples consisting of subject, predicate, and object
- In contrast to SQL, SPARQL queries are not constrained to working within one database: federated queries can access multiple data stores (endpoints). This is technically possible because SPARQL is more than just a query language

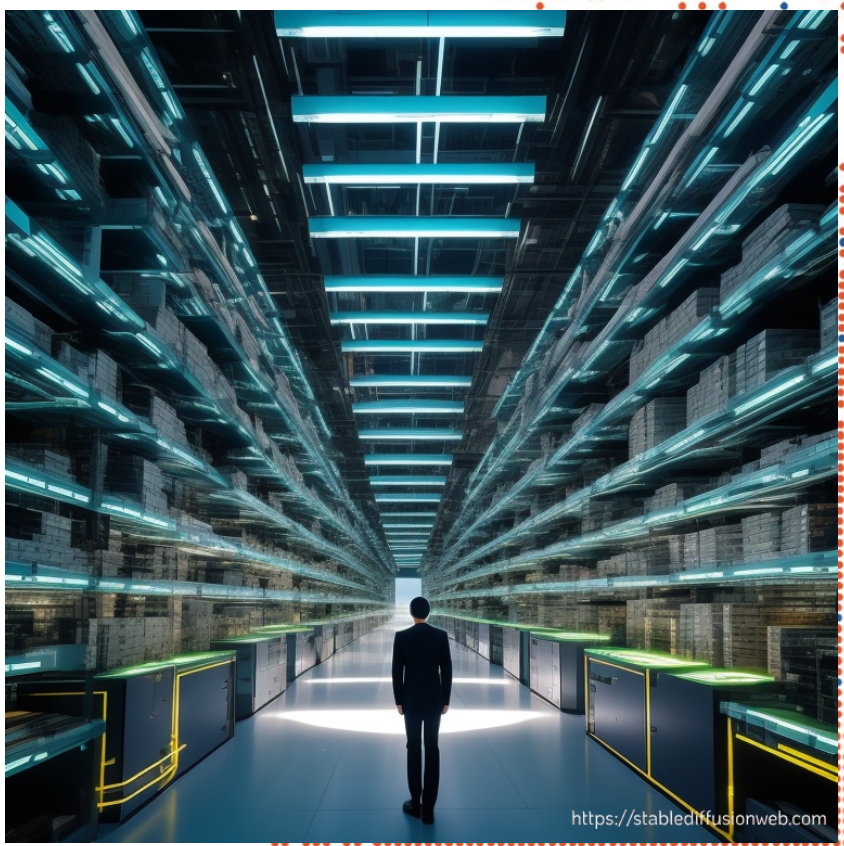
Graph or Relational database

- A graph database is a specialized, single-purpose platform which is used to create and manipulate data of an associative and contextual nature
- The relational model was designed for fast row-by-row access, but forming complex relationships between stored data can be challenging. This requires complex queries with many join operations over several tables, and consideration of foreign key constraints, which can cause additional overhead.
- Graph databases are often faster for associative data sets and map more directly to the structure of object-oriented applications. They can scale more naturally to large datasets as they typically do not need join operations. They are marketed as more suitable to manage ad hoc and changing data with evolving schemas due to their less dependence on a rigid schema.
- Relational database management systems are typically faster at performing the same operation on large numbers of data elements, allowing the manipulation of the data in its natural structure. Graph DB has different purpose. A graph database may become relevant if there is evidence for performance improvement by orders of magnitude and lower latency.

Why to preserve Open Data in Archives

Pros:

- FAIR Principles
- will not be easy to split Open Data from various results or products when republished
- trustworthiness of such processed Open Data
- integrity, reusability
- datasets are changing
- metadata and open formats
- various groups of users

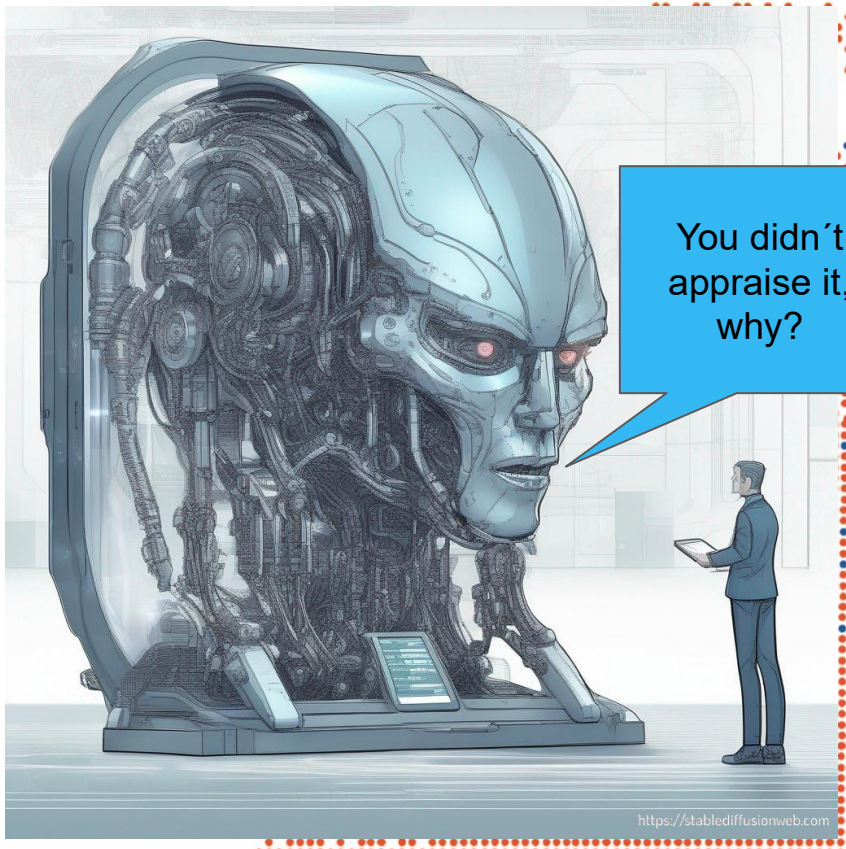


National Archives
Czech Republic

Why not to preserve Open Data in Archives

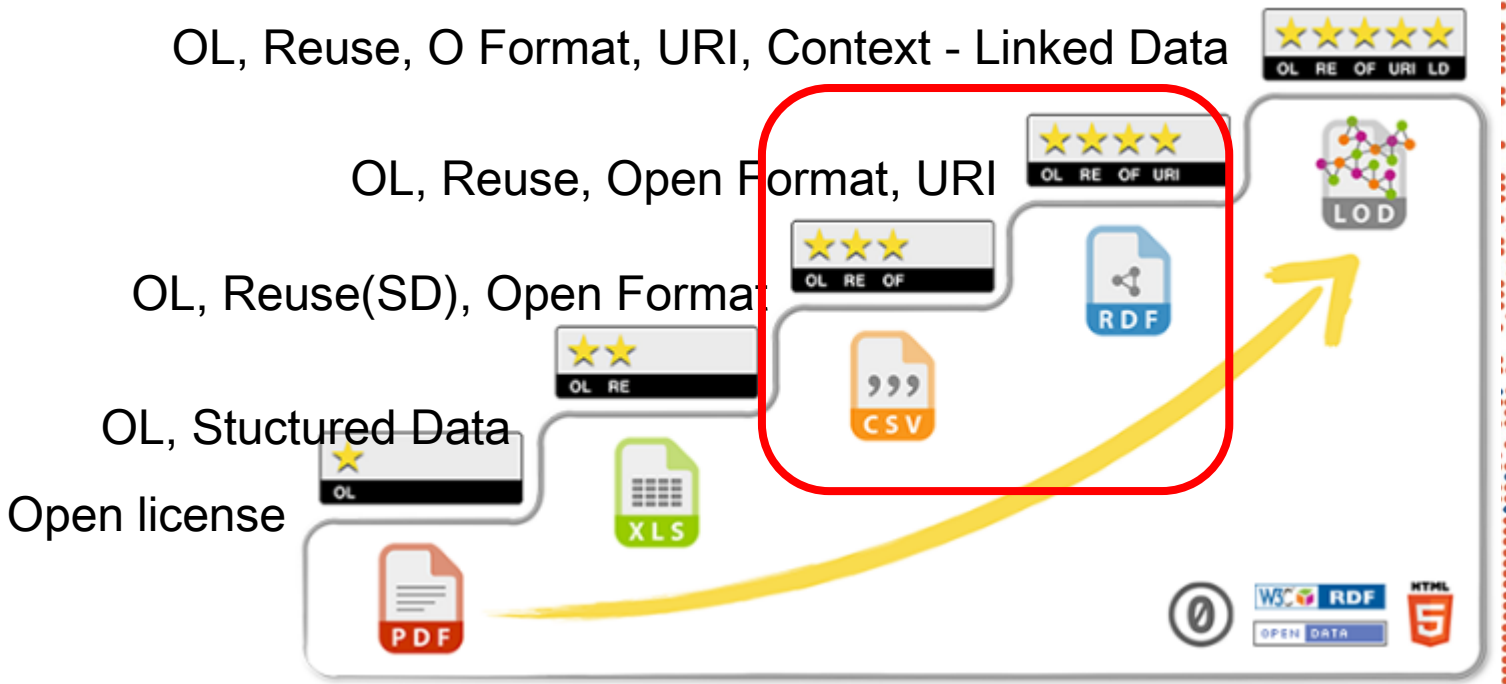
Cons:

- get it as soon as possible
- provide access immediately
- short time for archival processing and administration
- various types of data
- large and rapid growth of data volume over time



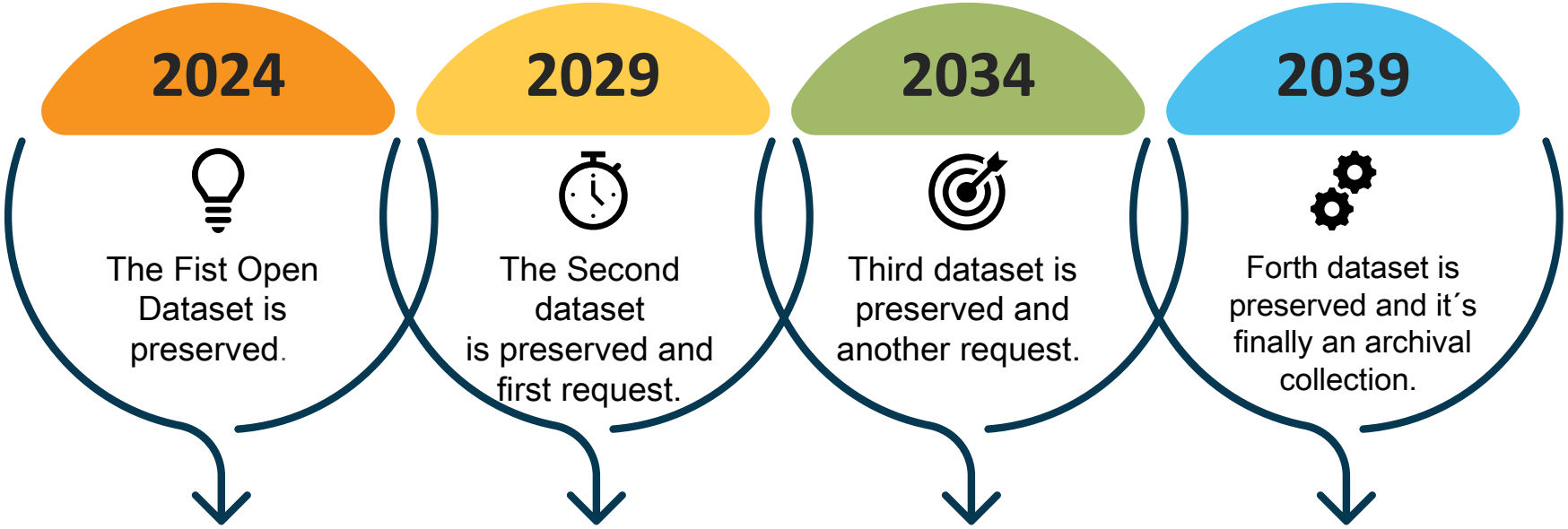
National Archives
Czech Republic

Open Data Levels



National Archives
Czech Republic

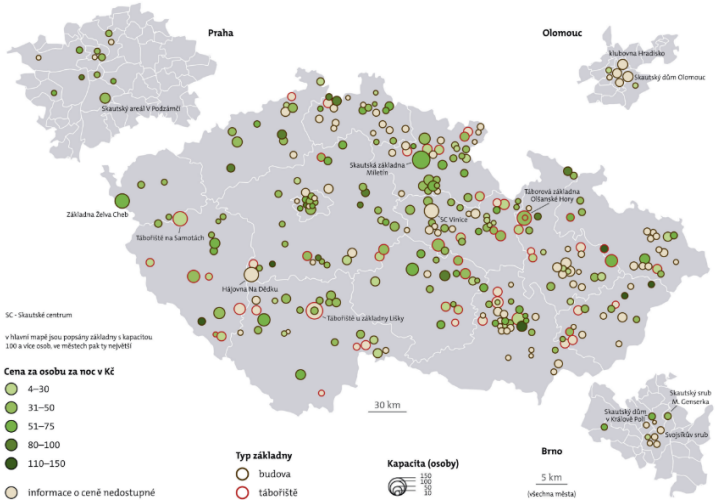
Expected Open Dataset´s Lifecycle in Archives

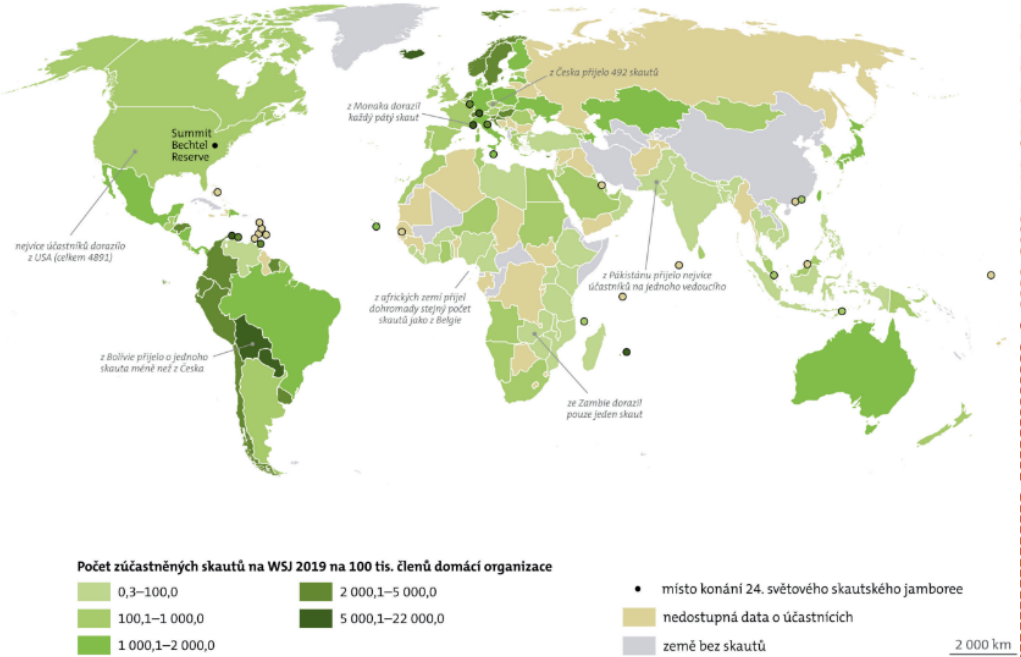


National Archives
Czech Republic

Example 1 - Combination

Open Map + Open Data





National Archives
Czech Republic

Example 2 - Data aggregation

- Data aggregation 2010 - 2022 Flu vaccination in CZ population - comes from Health Registers
- statistical data on the number of commuting per sons

  by districts - comes from Czech Statistical Office (data based on Census

Proočkovanost na chřipku

Záznamy (doklady o vykázané zdravotní péči) vstupující do analýzy jsou identifikovány pomocí vykázané

Zdroj dat: Národní registr hrazených zdravotních služeb (NRHZS)

Limitace: data NRHZS obsahují pouze data vykázaná v rámci veřejného zdravotního pojištění a tak jak b

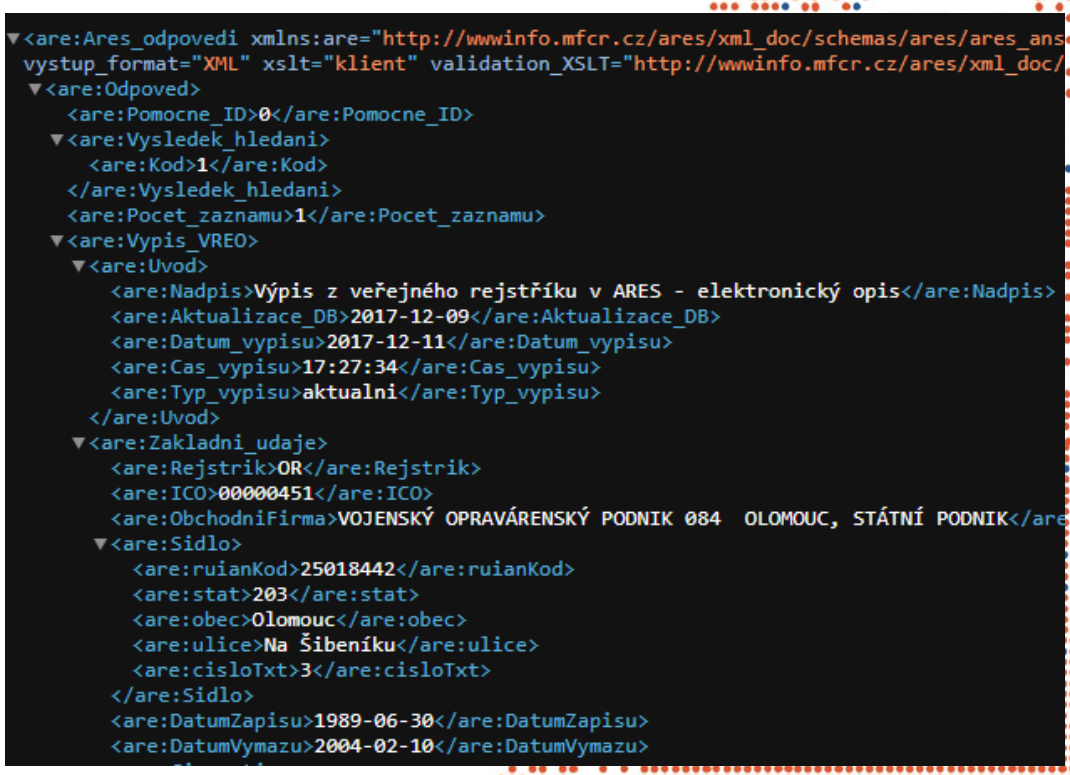
územní celek	rok 2010		
	demografie	počet vakcinovaných	proočkovanost vakcinovaných (%)
Česká republika	10 532 770	388 924	3,7%
Hl. m. Praha	1 257 158	46 286	3,7%
Středočeský kraj	1 264 978	39 932	3,2%
Benešov	94 652	2 847	3,0%
Beroun	85 081	2 495	2,9%
Kladno	160 742	4 936	3,1%
Kolín	95 764	3 714	3,9%
Kutná Hora	75 004	2 457	3,3%
Mělník	102 628	3 292	3,2%
Mladá Boleslav	122 816	3 360	2,7%

Example 3 - raw data

- Raw data - list of honorary citizens
- Usually results from Select statements exported to either CSV, JSON or XML
- JSONs and XMLs more suitable for structured data
- CSVs better for imports or flat data

ARES scenario

- Data on economic entities registered in the Czech Republic. It facilitates the display of data held in individual source registers of the state administration.
- XML file for each economic entity incl. its history
- daily updates



National Archives
Czech Republic

Dataset´s workflow

Business sector

Datasets incl. Open Data to generate profit, to sell new products etc. To predict future sales based Censuses, birth statistics,...

Producer

Usually produces datasets from which Open Data is just a version of their data and usually does not care what happens with published data. LTP is acceptable without any problem.

Scientist

Produces or uses Datasets for his or her research. It is inevitable to preserve various versions of datasets like Open Data and not Open just because of FAIR Principles and to reproduce various results during a long periods.

Archives

Goal is the preserve datasets for a long time, doing migrations etc. But can benefit and mine data too.

National Archives
Czech Republic

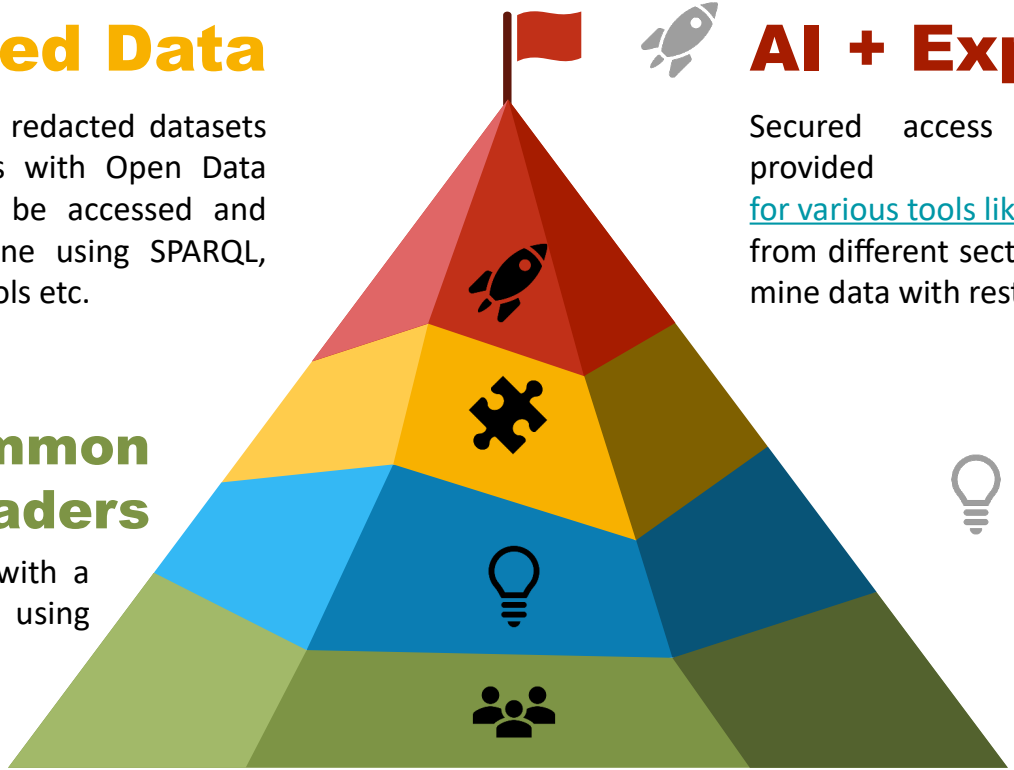
Designated Community

Linked Data

Storage for redacted datasets or datasets with Open Data which can be accessed and mined online using SPARQL, scraping tools etc.

AI + Expert

Secured access needs to be provided
for various tools like AI or Experts
from different sector to browse and mine data with restricted access.



Common Readers

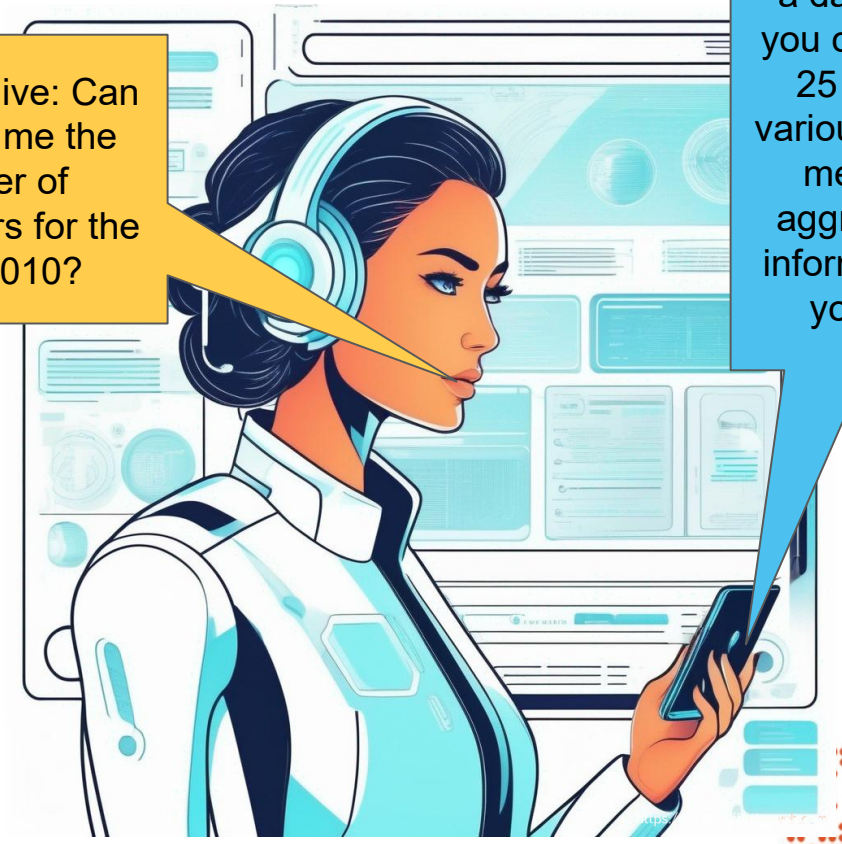
Usually are satisfied with a simple preview using Reading room´s UI.

Pioneers

Want to dive deeper, requires more info, metadata and downloads data to make up something or to carry out research.

National Archives
Czech Republic

How might it end?



Stable Diffusion Online

Prompt:

Female visitor of archives asks archival chatbot for data.

National Archives
Czech Republic

Automated Preservation

- define expected data structure and transformation
- define designated community
- map metadata to your standards/requirements
- define role of an archivist
- reduce/simplify administration
- develop or upgrade tools

National Archives
Czech Republic

Key questions

Is it better to:

- preserve open data in a single and independent SIP?
- put it together with non open data version?
- packages containing multiple related datasets?
- harvest everything at once?
- establish something like disposal schedule?

National Archives
Czech Republic

Conclusion

- various types of Open Data needs to be preserved for designated Community
- Geodata might be result from a redaction of sensitive information
- even simple flat files have high value
- Archives must participate in data market
- Archives can benefit and be beneficial

DLM FORUM

National Archives
Czech Republic

FAV

The End

- Let´s be positive
- Open data has much in common with Digital Preservation
- Some requirements can already be unchecked

Thanks!

Images generated by AI Stable Diffusion

DLM FORUM

National Archives
Czech Republic

FAV