WEBINAR FOR DATA PROVIDERS

Understanding open data: technical openness

data.europa academy

21 October 2022

10.00 – 11.30 CET

Introduction

Giuseppe Ascone Modica
data.europa.eu
Knowledge Management Assistant

Benjamin Dittwald
Fraunhofer FOKUS
Technical team

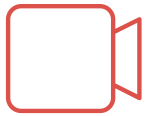
Lina Bruns
Fraunhofer FOKUS
Technical Team

Jakub Klímek
Ministry of the Interior of the Czech Republic
External Expert

Agenda of today

- Opening
- Introduction to open data
- Data format issues
- How to reach a better technical openness?
- Q&A
- Hands on: Best Practices from data provider
- Q&A and summary
- Closing

Rules of the game



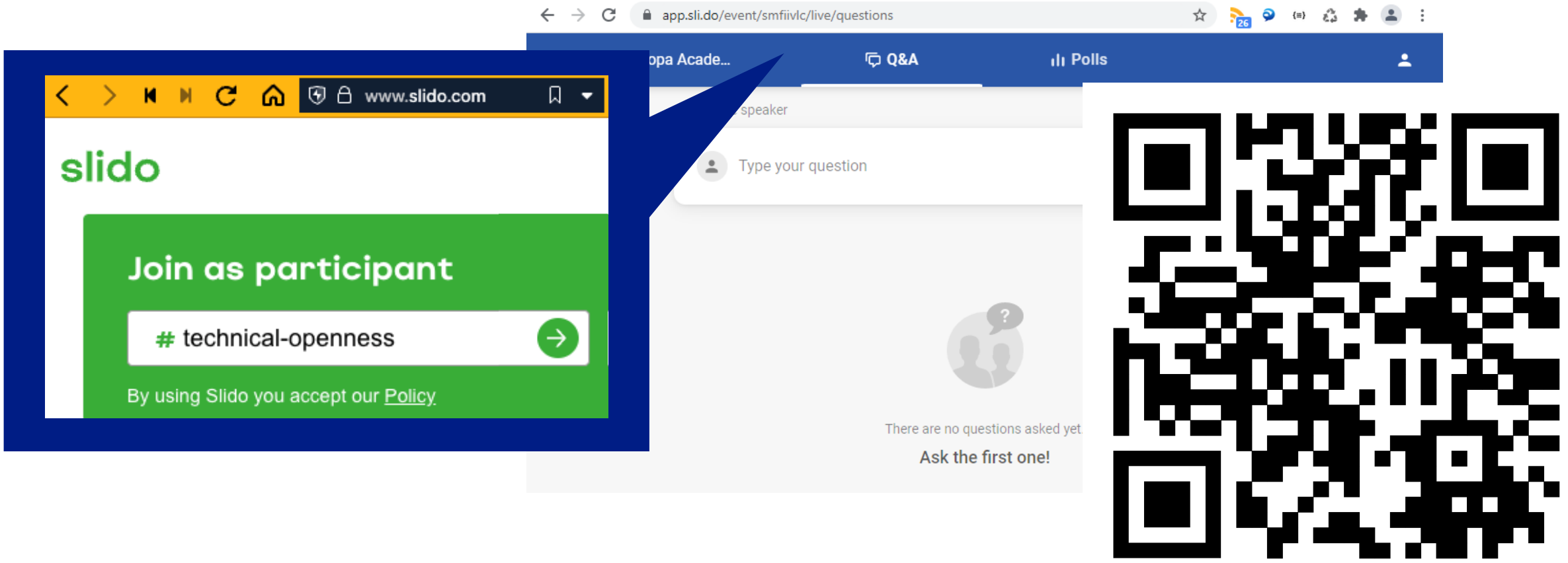
The webinar will be recorded

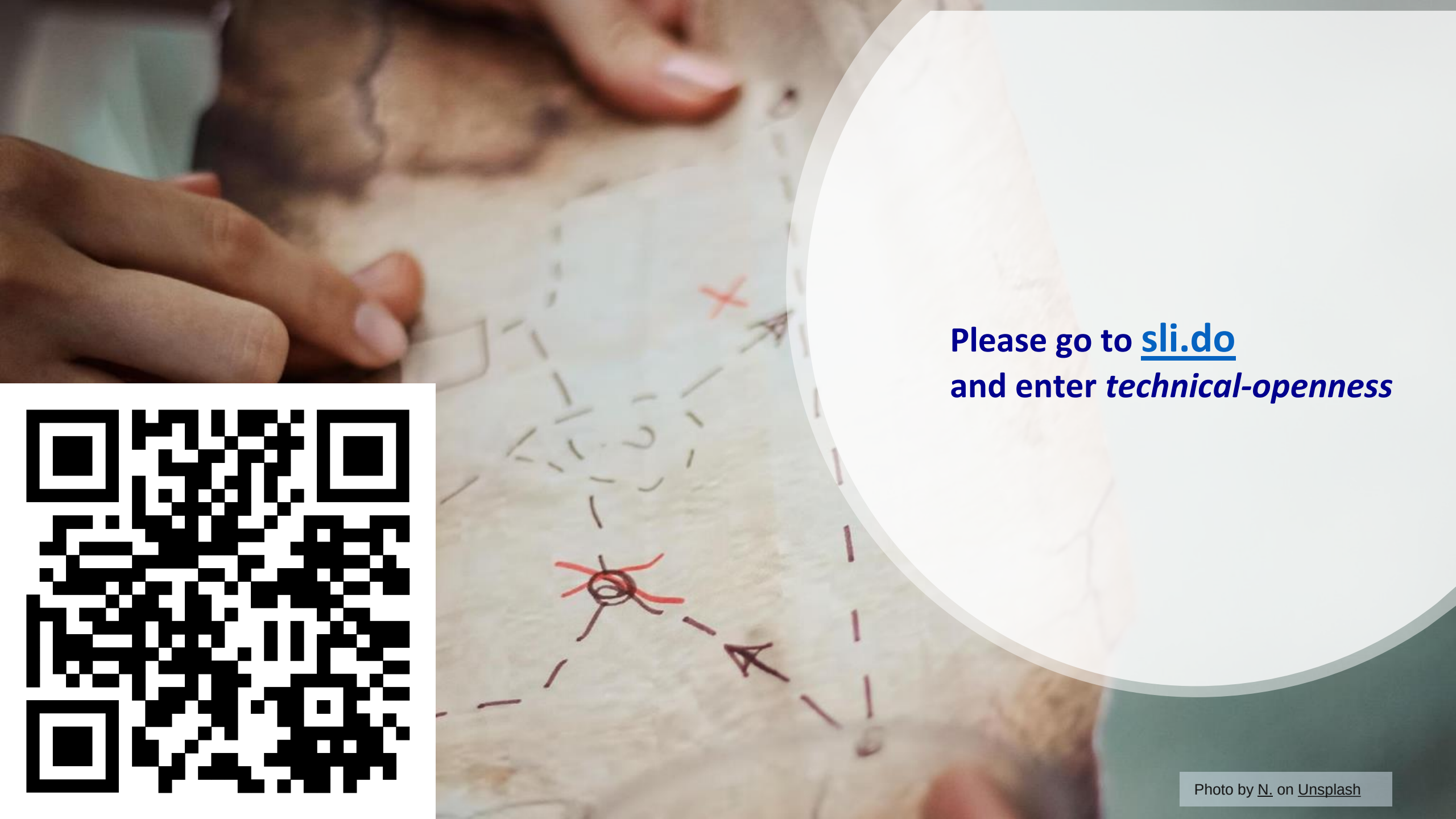
Please mute yourselves during the webinar

Please reserve 3 min after the webinar to help us improve by filling in our feedback form

For questions, please use sli.do. Vote for the questions that are of most interest to you, those will be discussed later

Ask your questions in sli.do using the code *technical-openness*





Please go to sli.do
and enter *technical-openness*

Photo by N. on Unsplash

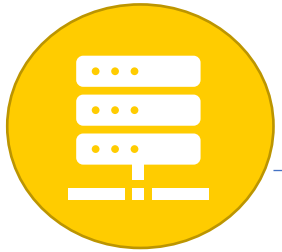
Introduction to open data

What data formats do you usually work with (e.g. Excel, PDF, …)?

What does ´open` mean?

Legal Perspective

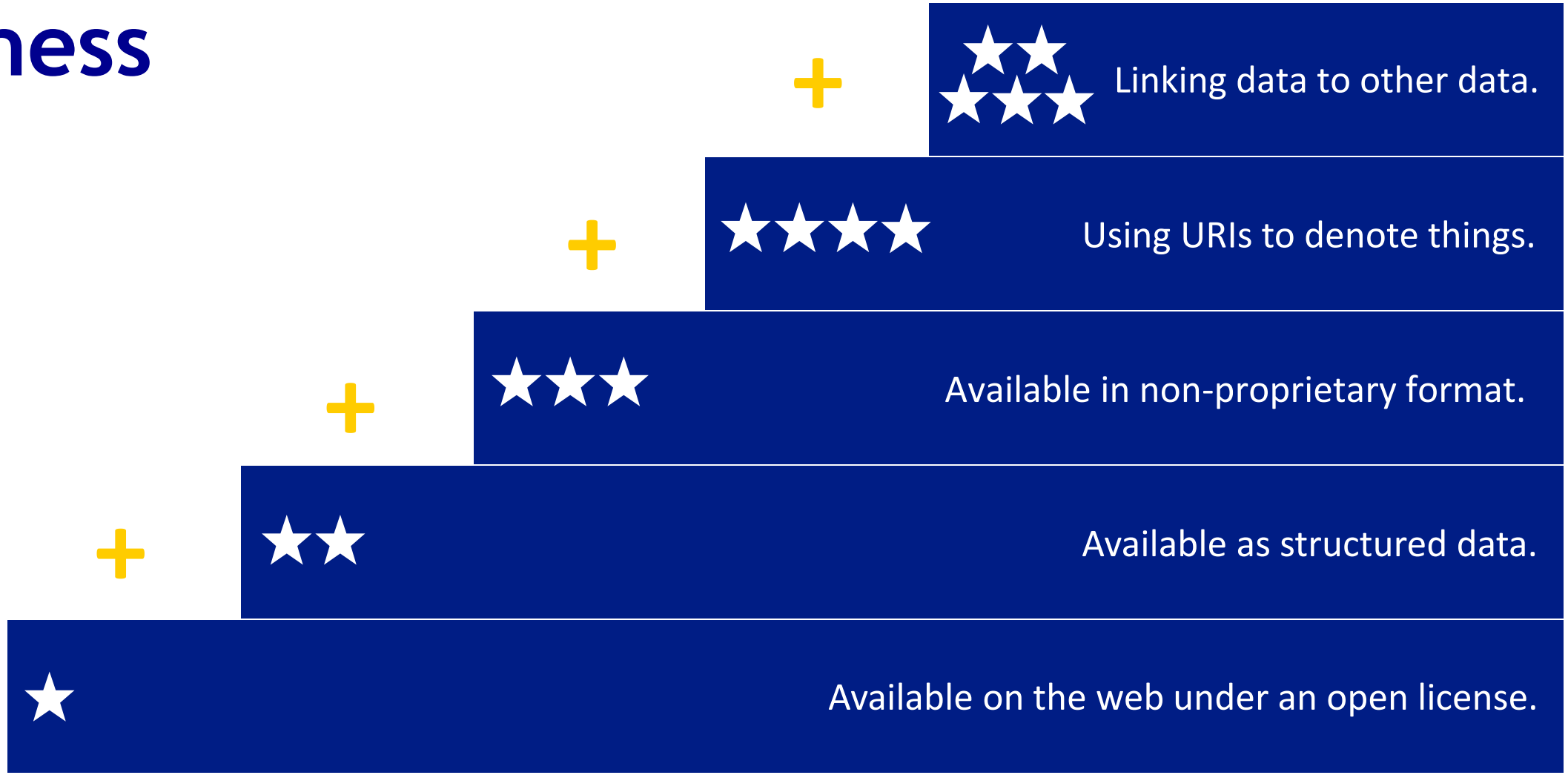
- Data can be freely accessed
- Data can be used, edited, and shared by all individuals and entities without legal restrictions



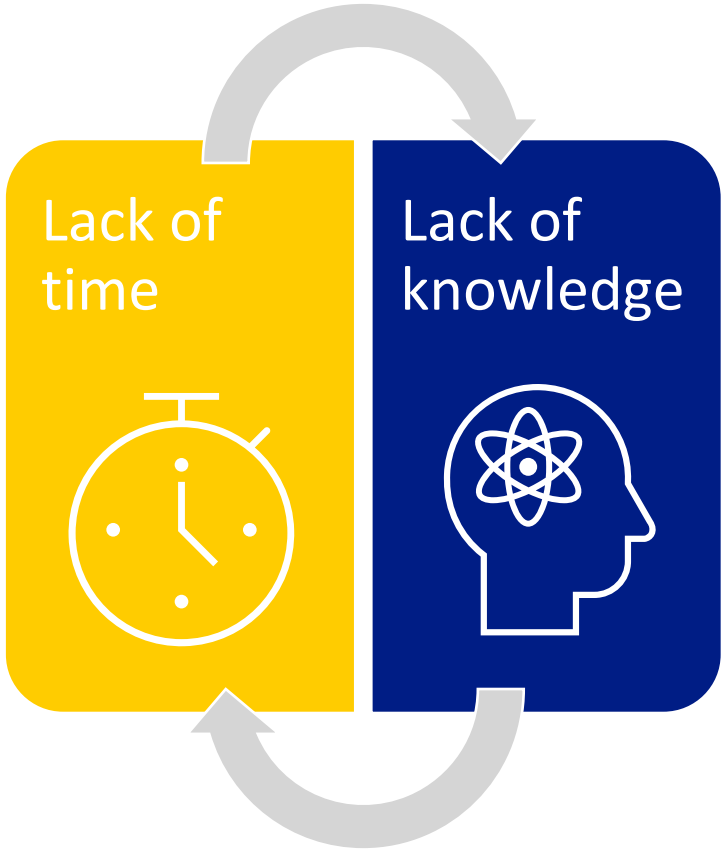
Technical Perspective

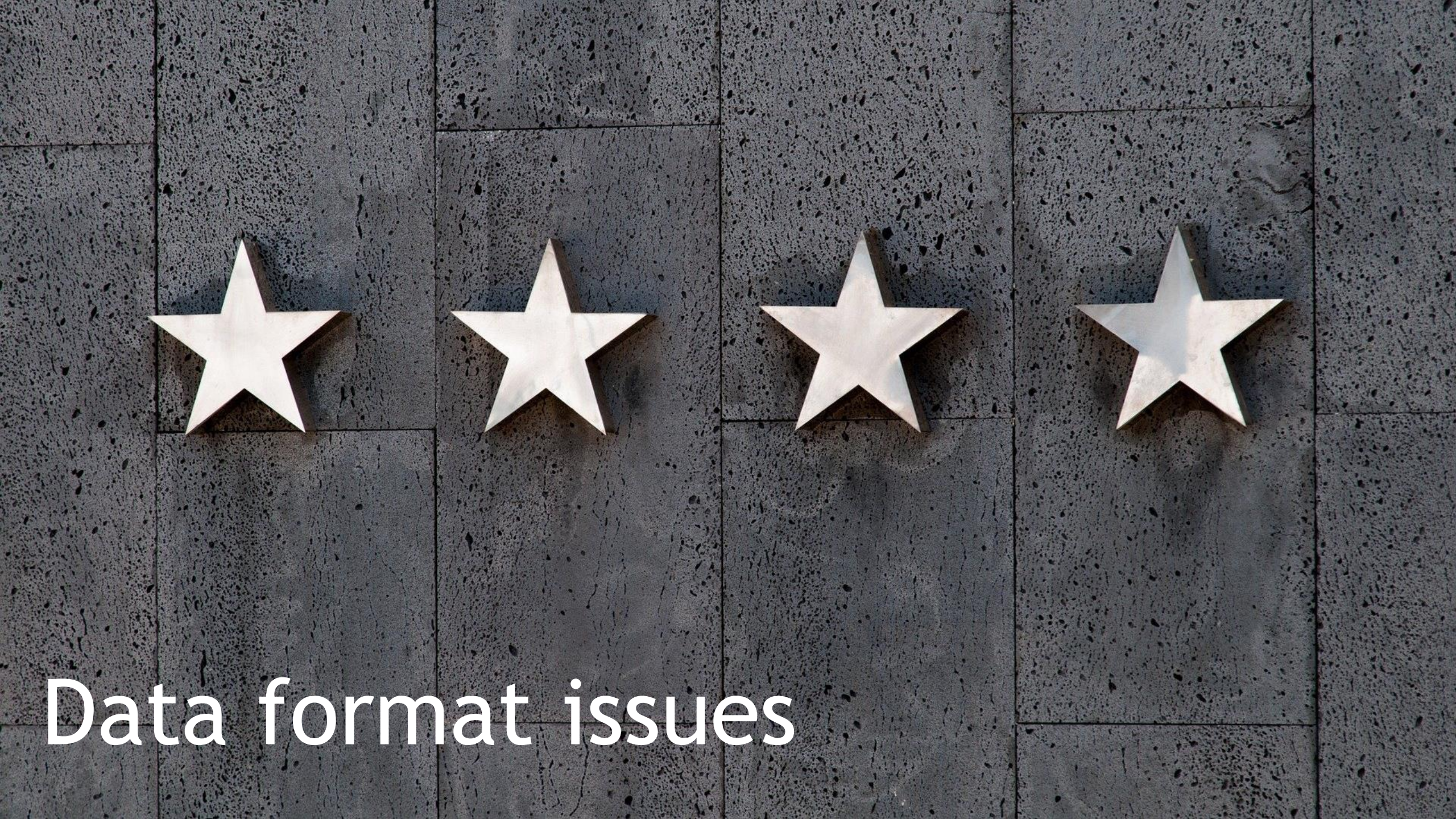
- Data is available in non-proprietary formats
- Data is machine-readable
- Data can be accessed without restrictions (e.g. no authorisation necessary)

Degrees of Openness



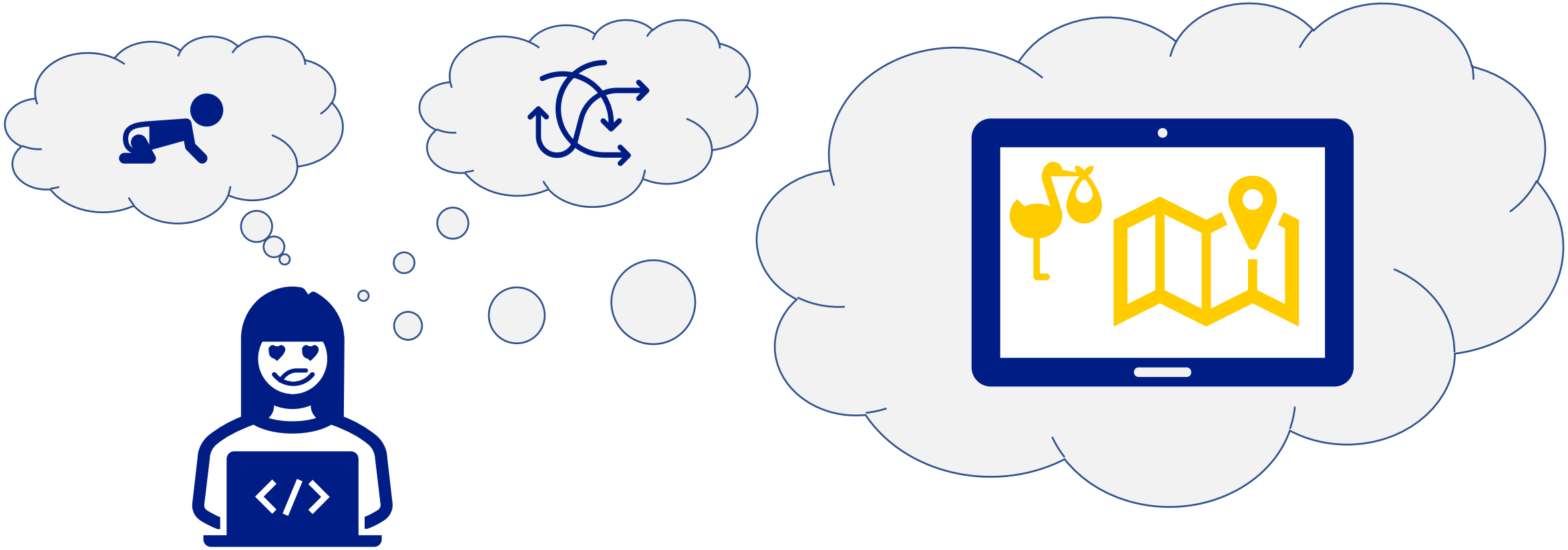
Frequent excuses for low technical openness



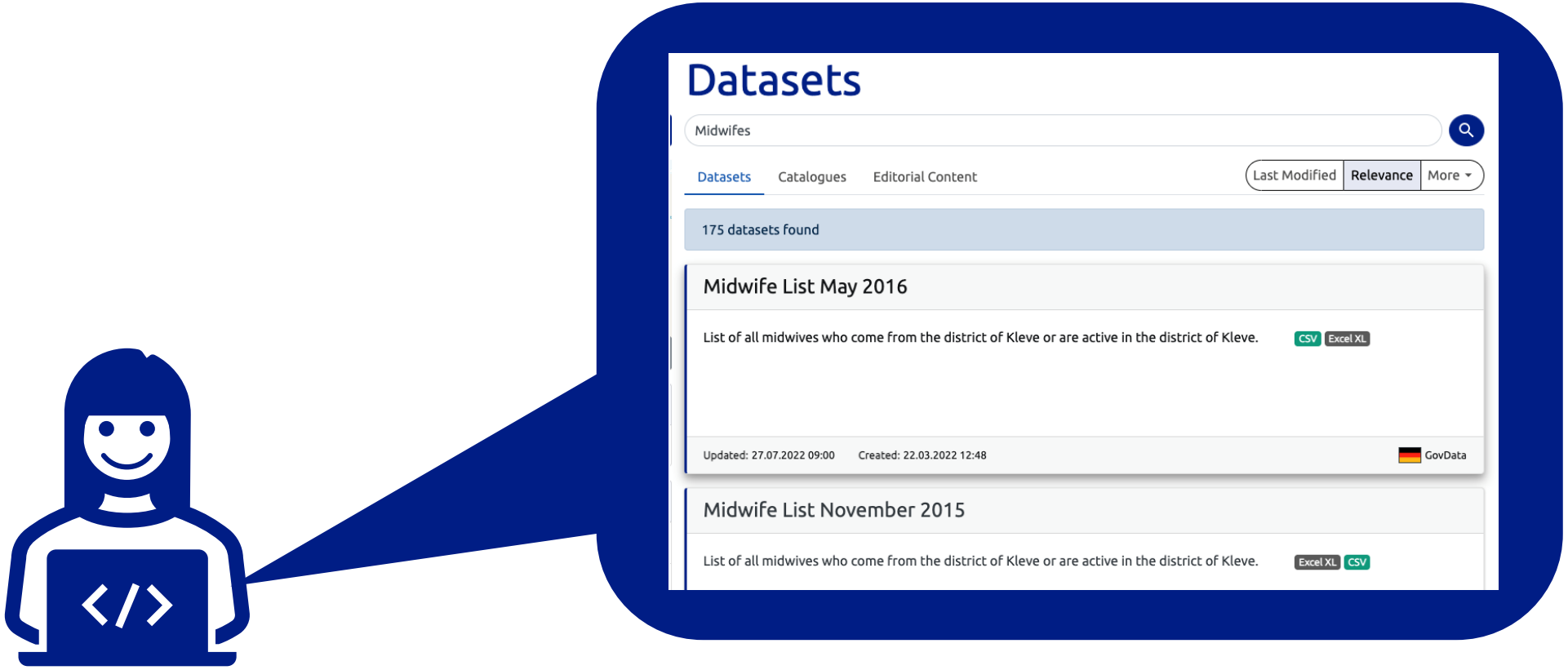


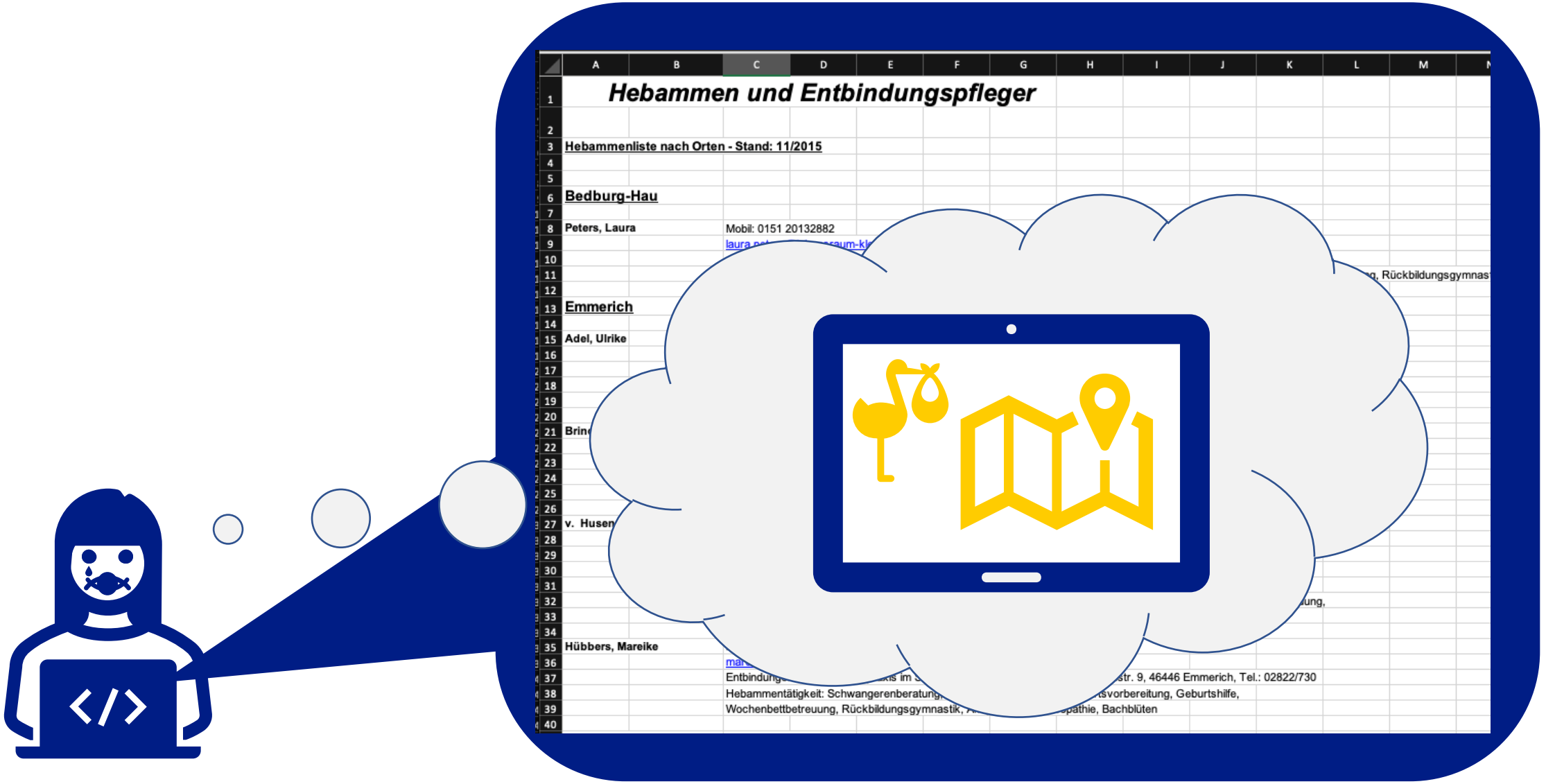
Data format issues

From a user perspective...



From a user perspective…




Hebammen und Entbindungspfleger
Hebammenliste nach Orten - Stand: 11/2015
Bedburg-Hau
Peters, Laura
Mobil: 0151 20132882
Emmerich
Adel, Ulrike
v. Husen
Hübbers, Mareike
Entbindung
Hebammentätigkeit: Schwangerenberatung,
Wochenbettbetreuung, Rückbildungsgymnastik,
Rückbildungsgymnas
str. 9, 46446 Emmerich, Tel.: 02822/730
svorbereitung, Geburtshilfe,
spathie, Bachblüten

How to improve technical openness?

- Introduce a data management in your organisation

Data Management

Data Governance

Specifications and rules for handling data considering standards and legal frameworks

Define tasks, roles and responsibilities

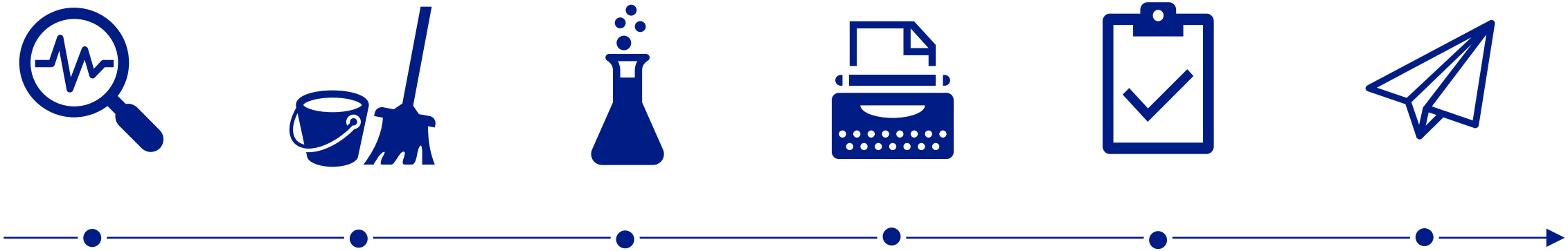
Data Governance - Roles

Build Competencies

Data Quality Guidelines

- Project launched by the Publications Office of the European Union in 2019
- Aim: analysing major quality issues and providing a set of recommendations for data providers from the EU and its Member States.
- Consisted of three parts:
  - Data analysis for identifying most common quality issues
  - Identification of data quality dimensions, indicators and metrics
  - **Recommendations for delivering high-quality data**
- For any feedback regarding the publication please contact: OP-DATA-EUROPA-EU@publications.europa.eu

Data Preparation Process



Profiling

Distilling

Enriching

Documenting

Validating

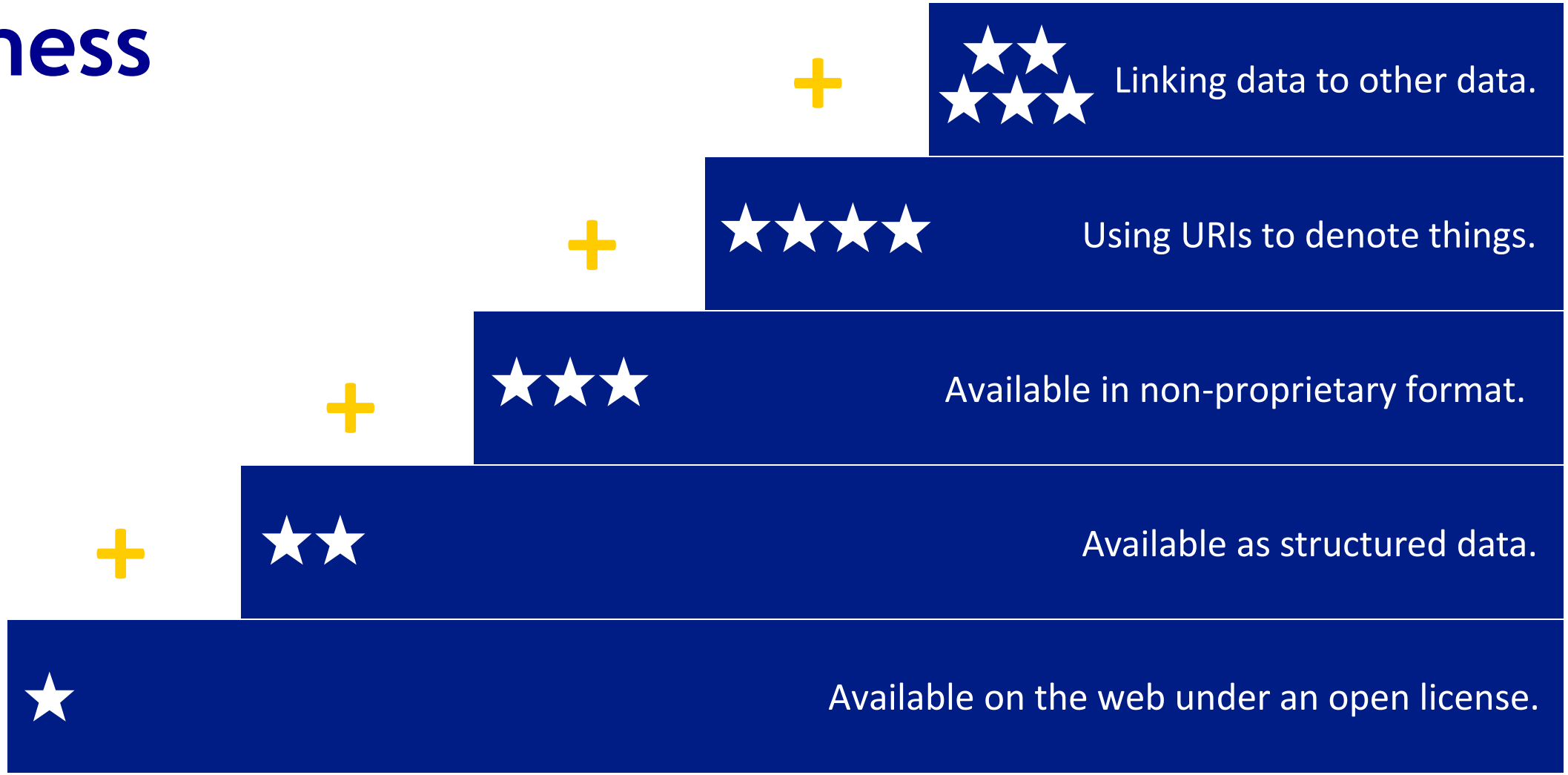
Publishing

How to improve technical openness?

- Introduce a data management in your organisation

- Choose appropriate file format

Degrees of Openness



Appropriate file formats

Machine-Readable

- CSV
- HTML*
- JSON
- ODS*
- RDF
- TXT*
- XLS*
- XLSX*
- XML

*predominantly

Non-Proprietary

- CSV
- DOCX
- HTML
- JSON
- ODS
- ODT
- PDF
- PNG
- RDF
- TXT
- XML
- XLSX

CSV, JSON, XML, RDF

PDF, DOCX, ODT, PNG, GIF, JPG/JPEG, TIFF, DOC, XLS

Format	# Datasets on data.europa.eu		1010 1010	Achievable stars
CSV	237 645	yes	yes	★★★★
JSON	99 538	yes	yes	★★★★
HTML	76 997	yes	predominantly	★
Excel XLSX	56 560	yes	predominantly	★★★★
Excel XLS	48 002	no	predominantly	★★
PDF	41 145	yes	no	★
TXT	31 645	yes	predominantly	★
XML	25 034	yes	yes	★★★★
TIFF	14 292	no	no	★
PNG	4 652	yes	no	★
RDF	4 400	yes	yes	★★★★★

Status: 5th October 2022
Geospatial and statistical formats not listed

How to convert Excel to CSV

Structure of a CSV-file

	A	B	C
1	City	Population size	
2	Berlin	3,669,491	
3	London	8,908,081	
4	Paris	2,187,526	
5			

Tabular data opened in Excel

```
City,Population size
Berlin,3669491
London,8900801
Paris,2187526
```

Tabular data in CSV – open text-based format.

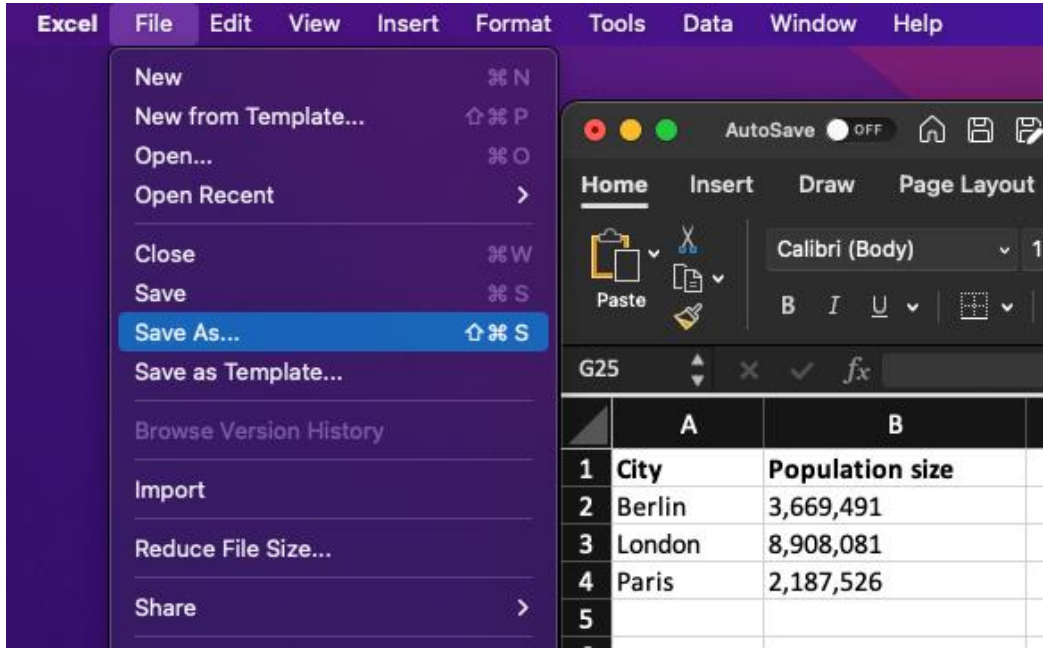
	A	B	C	D
1	Jahr	Standort	Gesamtkosten in Euro	Landeszuwendung in Euro
2	2017	Detmold - Turnleistungszentrum Detmold am Grabbe-Gymnasium	65.987,17	39.724,00
3	2017	Dortmund - Helmut-Koernig-Halle	75.021,03	30.008,00
4	2017	Dortmund - Helmut-Koernig-Halle	23.600,00	5.721,00
5	2017	Hennef - Sportschule	240.860,50	102.767,00
6	2017	Hennef - Sportschule	137.036,76	50.251,00
7	2017	Hennef - Sportschule	48.100,92	33.671,00
8	2017	Winterberg - Bob- und Rodelbahn	14.200,00	4.260,00
9	2017	Winterberg - Bob- und Rodelbahn	373.765,75	112.130,00
10	2017	Winterberg - Bob- und Rodelbahn	34.005,00	10.202,00
11	2017	Winterberg - Schanzenanlage	89.273,35	40.425,00
12	2018	Bielefeld - NRW-Sportschule Helmholtz-Gymnasium	553.398,71	4.442.719,00
13	2018	Bochum - LA Halle Hollandstrasse	362.407,95	173.689,00
14	2018	Bonn - Baseballstadion	1.124.726,64	539.869,00
15	2018	Bonn - Fechtzentrum	38.007,79	7.602,00
16	2018	Bonn - Fechtzentrum	130.375,48	26.075,00
17	2018	Dortmund - Helmut-Koernig-Halle	35.350,50	14.140,00
18	2018	Dortmund - Helmut-Koernig-Halle	23.600,00	5.721,00
19	2018	Dortmund - Ruderleistungszentrum	85.356,38	34.143,00
20	2018	Winterberg - Bob- und Rodelbahn	119.863,92	35.959,00
21	2018	Winterberg - Bob- und Rodelbahn	55.886,77	16.766,00
22	2018	Winterberg - Schanzenanlage	5.121,40	2.438,00
23	2019	Bergisch-Gladbach - Rheinische Landesturnschule	15.683,92	10.979,00
24	2019	Bochum - LA Halle Hollandstrasse	26.733,60	14.315,00
25	2019	Bonn - Fechtzentrum	10.200,00	2.040,00
26	2019	Bonn - Fechtzentrum	74.935,00	14.987,00
27	2019	Dortmund - Helmut-Koernig-Halle	3.699.118,09	1.664.603,00
28	2019	Dortmund - Eissportzentrum	61.146,54	24.459,00
29	2019	Essen - Schwimmzentrum Essen-RÄÄuettenscheid	858.444,69	429.222,00
30	2019	Hennef - Sportschule	174.506,83	63.992,00
31	2019	Hennef - Sportschule	233.861,58	163.703,00
32	2019	Krefeld - CHTC 1890 e.V.	112.549,00	67.529,00
33	2019	Leverkusen - NRW-Sportschule Landrat-Lucas-Gymnasium	7.769.052,15	6.215.424,00
34	2019	Neuss - HTC SW Neuss	120.760,00	50.[illegible]9,00
35	2019	Winterberg - Bob- und Rodelbahn	72.750,00	[illegible]750,00
36	2019	Winterberg - Bob- und Rodelbahn	38.505,00	11.552,00

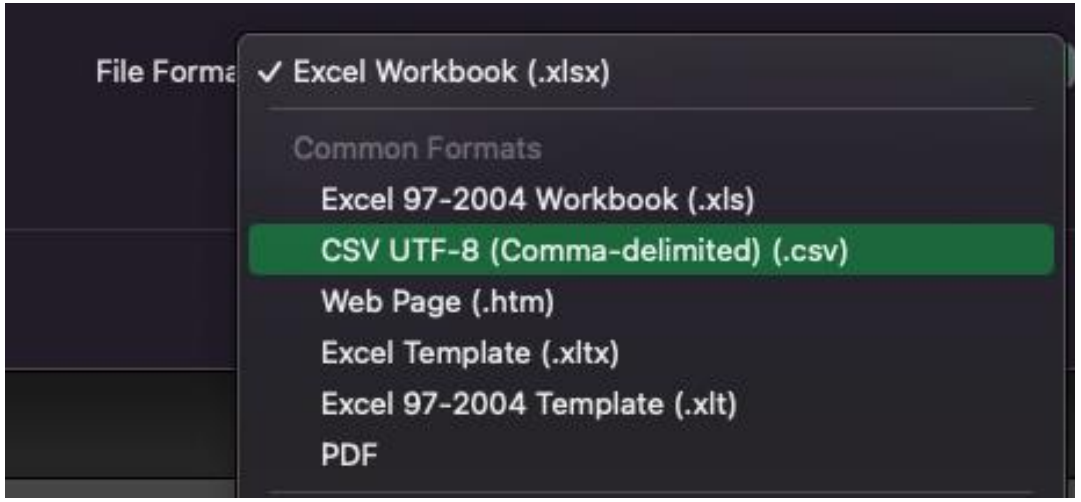
Förderung herausragender Sport +

	A	B	C
1	Förderung herausragender Sportstätten		
2			
3	Standort	Gesamtkosten in Euro	Landeszuwendung in Euro
4	Jahr 2017		
5	Detmold - Turnleistungszentrum Detmold am Grabbe-Gymnasium	65.987,17	39.724,00
6	Dortmund - Helmut-Koernig-Halle	75.021,03	30.008,00
7	Dortmund - Helmut-Koernig-Halle	23.600,00	5.721,00
8	Hennef - Sportschule	240.860,50	102.767,00
9	Hennef - Sportschule	137.036,76	50.251,00
10	Hennef - Sportschule	48.100,92	33.671,00
11	Winterberg - Bob- und Rodelbahn	14.200,00	4.260,00
12	Winterberg - Bob- und Rodelbahn	373.765,75	112.130,00
13	Winterberg - Bob- und Rodelbahn	34.005,00	10.202,00
14	Winterberg - Schanzenanlage	89.273,35	40.425,00
15	Jahr 2017 Gesamt (ab Juli)	1.101.850,48	429.159,00
16			
17	Jahr 2018		
18	Bielefeld - NRW-Sportschule Helmholtz-Gymnasium	553.398,71	4.442.719,00
19	Bochum - LA Halle Hollandstrasse	362.407,95	173.689,00
20	Bonn - Baseballstadion	1.124.726,64	539.869,00
21	Bonn - Fechtzentrum	38.007,79	7.602,00
22	Bonn - Fechtzentrum	130.375,48	26.075,00
23	Dortmund - Helmut-Koernig-Halle	35.350,50	14.140,00
24	Dortmund - Helmut-Koernig-Halle	23.600,00	5.721,00
25	Dortmund - Ruderleistungszentrum	85.356,38	34.143,00
26	Winterberg - Bob- und Rodelbahn	119.863,92	35.959,00
27	Winterberg - Bob- und Rodelbahn	55.886,77	16.766,00
28	Winterberg - Schanzenanlage	5.121,40	2.438,00
29	Jahr 2018 Gesamt	25.069.238,16	17.550.661,00
30			
31	Jahr 2019		
32	Bergisch-Gladbach - Rheinische Landesturnschule	15.683,92	10.979,00
33	Bochum - LA Halle Hollandstrasse	26.733,60	14.315,00
34	Bonn - Fechtzentrum	10.200,00	2.040,00
35	Bonn - Fechtzentrum	74.935,00	14.987,00
36	Dortmund - Helmut-Koernig-Halle	3.699.118,09	1.664.603,00
37	Dortmund - Eissportzentrum	61.146,54	24.459,00
38	Essen - Schwimmzentrum Essen-RÄÄuettenscheid	858.444,69	429.222,00
39	Hennef - Sportschule	174.506,83	63.992,00
40	Hennef - Sportschule	233.861,58	163.703,00
41	Krefeld - CHTC 1890 e.V.	112.549,00	67.529,00
42	Leverkusen - NRW-Sportschule Landrat-Lucas-Gymnasium	7.769.052,15	6[illegible].424,00
43	Neuss - HTC SW Neuss	120.760,00	5[illegible]
44	Winterberg - Bob- und Rodelbahn	72.750,00	21.7[illegible]
45	Winterberg - Bob- und Rodelbahn	38.505,00	11[illegible]

Förderung herausragender Sport | Sportstätten ab 2015 | Schlüsselinformationen

Converting Excel-file to CSV





Choose in “Save-as” Dialog “CSV UTF-8”

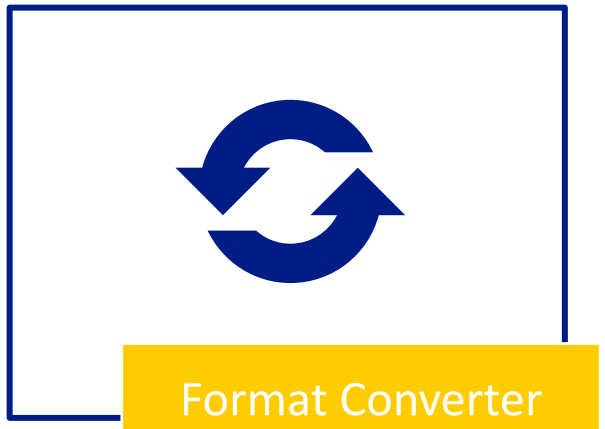
How to improve technical openness?

- Introduce a data management in your organisation

- Choose appropriate file format

- Make use of tooling

Make use of tooling

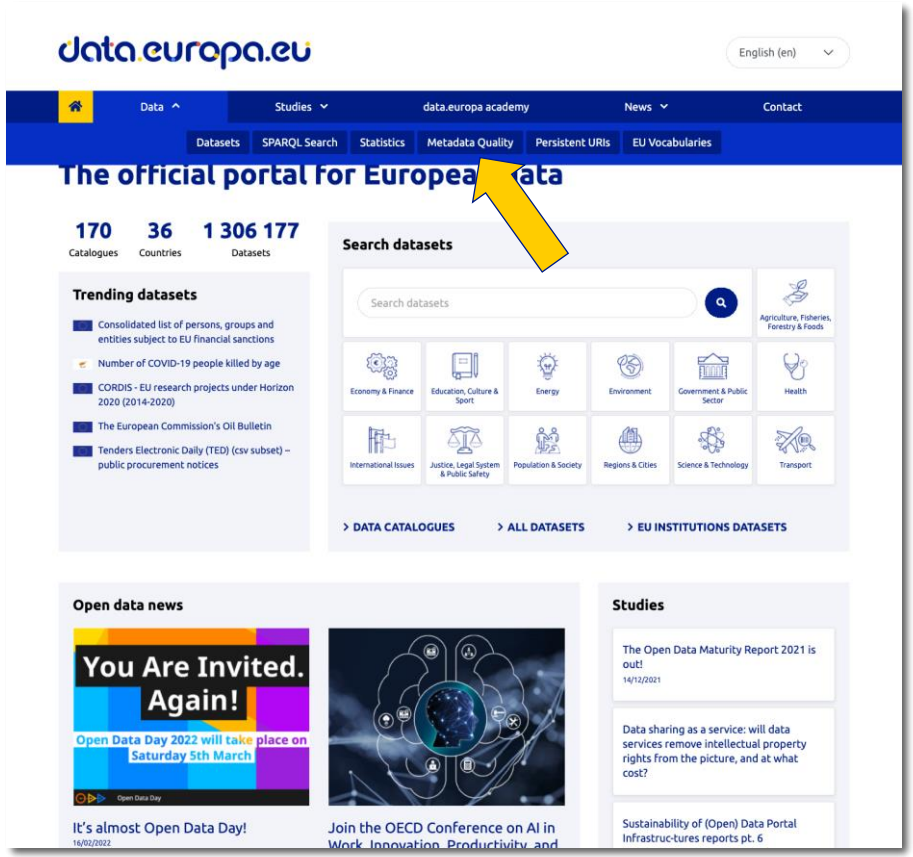


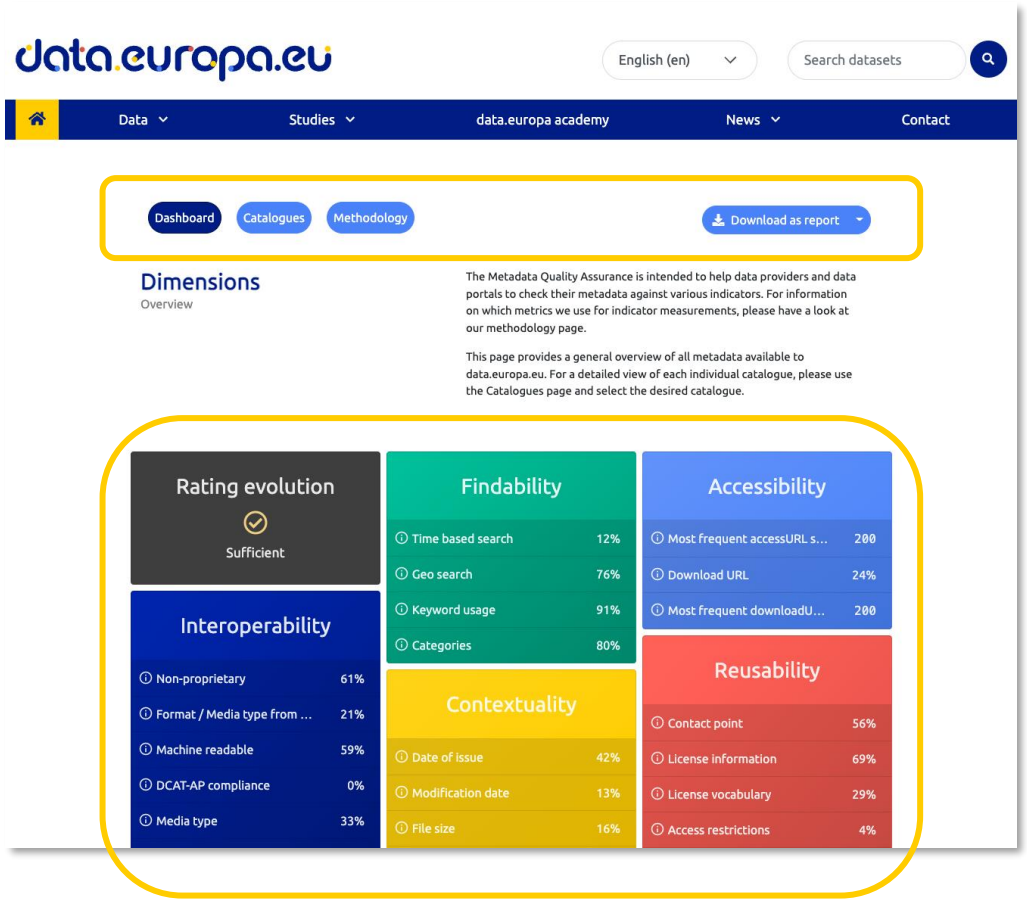
Format Converter

Cleansing tools

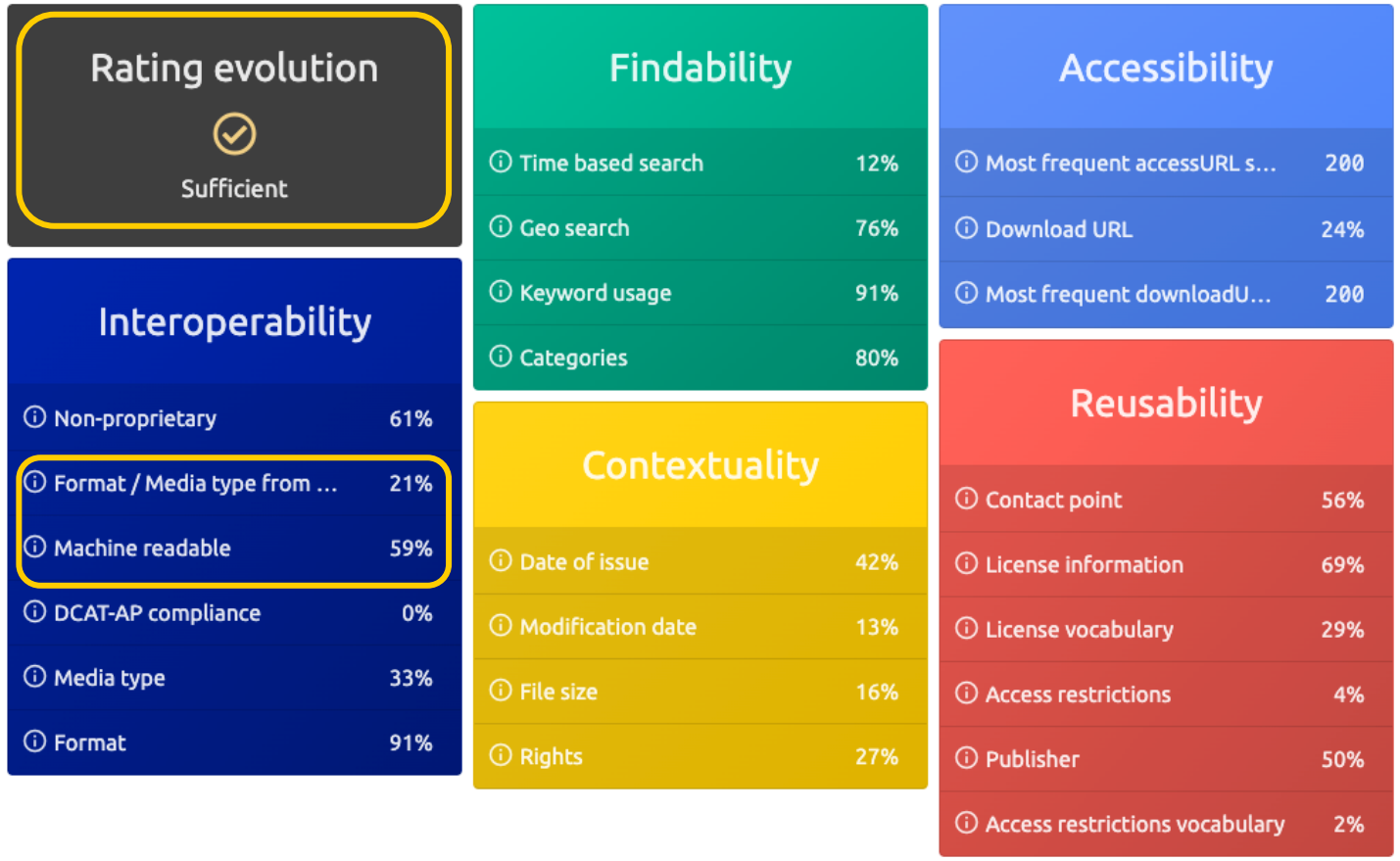
Quality Checking tools

The 'Metadata Quality Assurance' of data.europa.eu

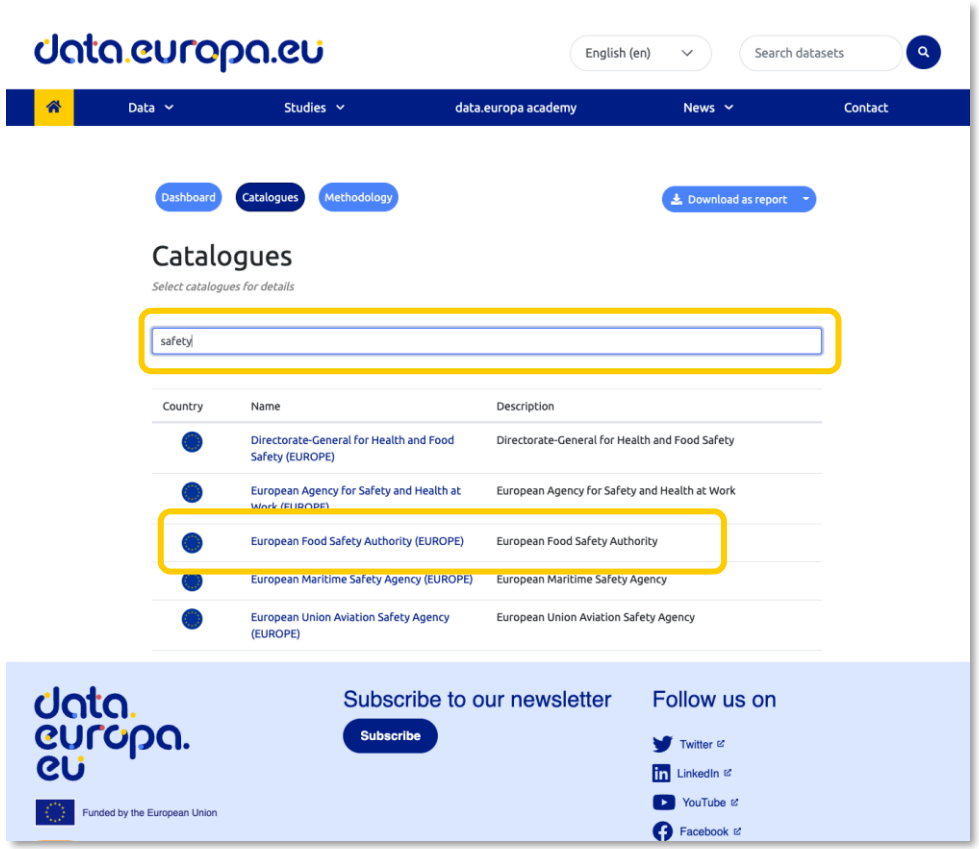


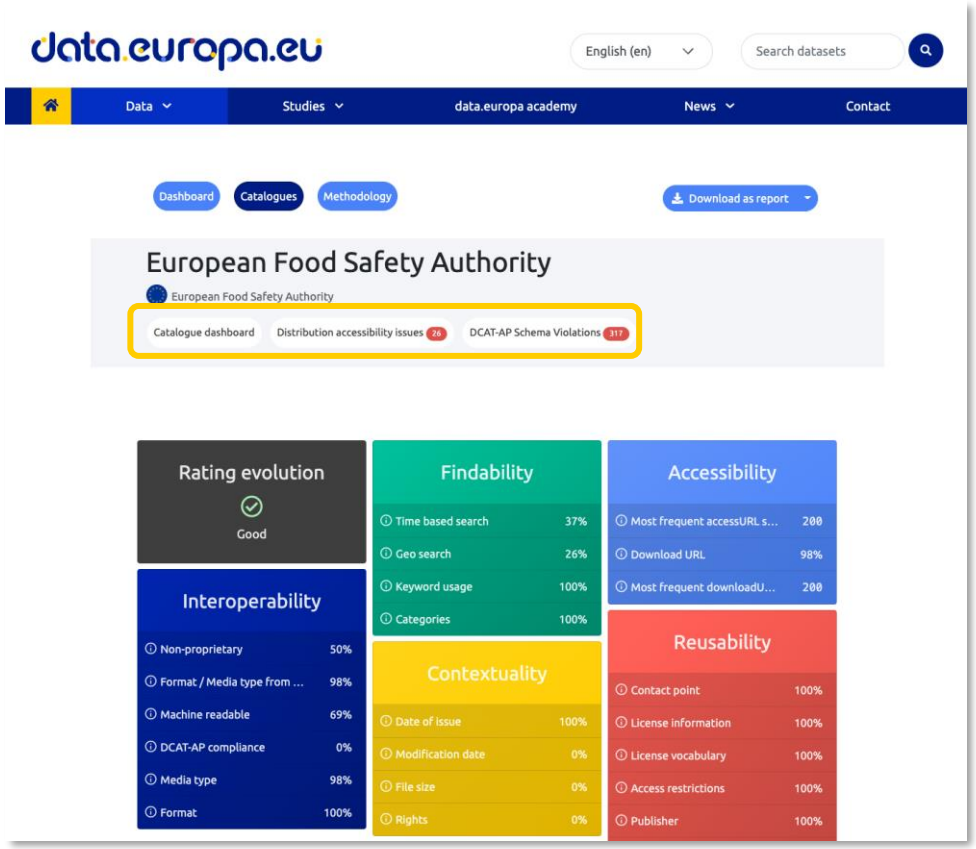


The 'Metadata Quality' section



The 'Metadata Quality' section





Questions?

Hands on: Best Practices from Data Provider

Open Data (OD) Best Practices

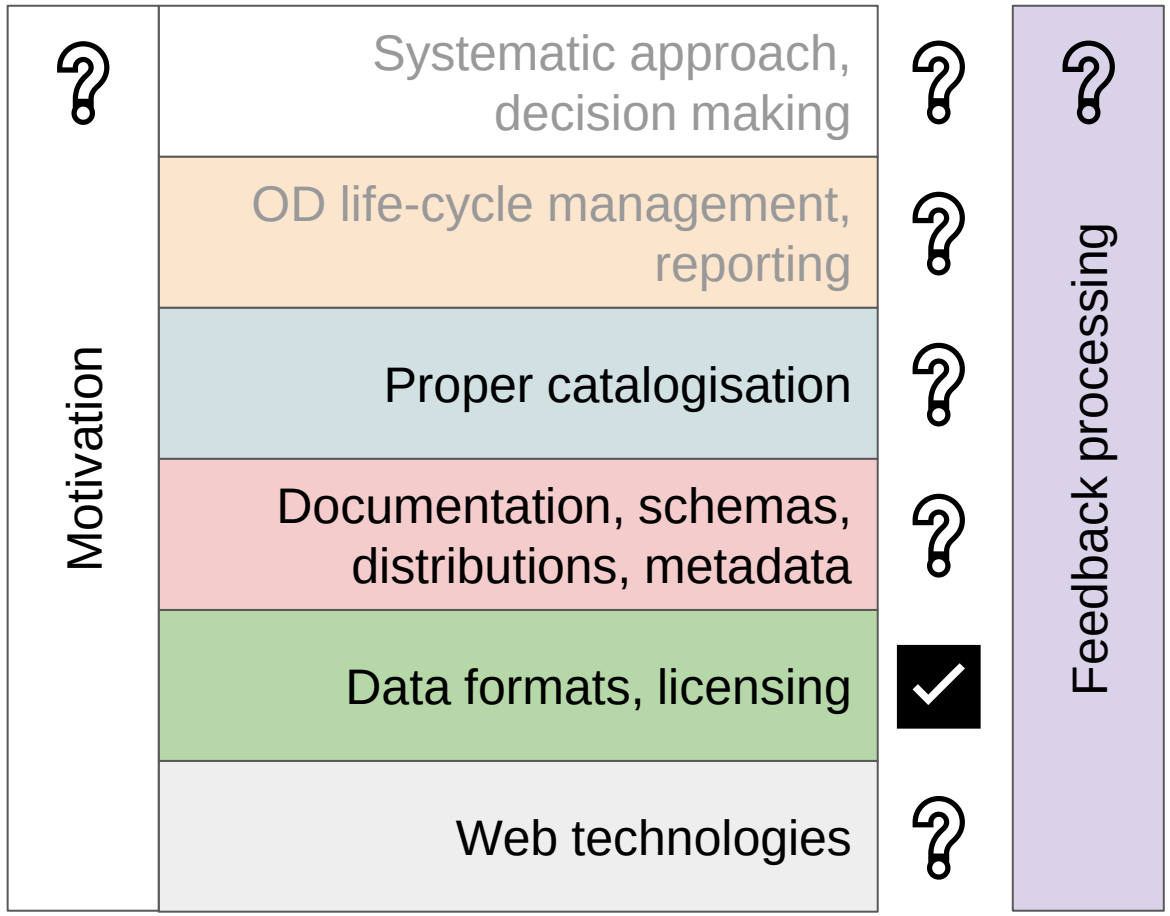
Jakub Klímek

#1: Charles University
(OD publisher, consumer, researcher)

#2: Ministry of the Interior of the Czech Republic
(OD publisher and consumer)

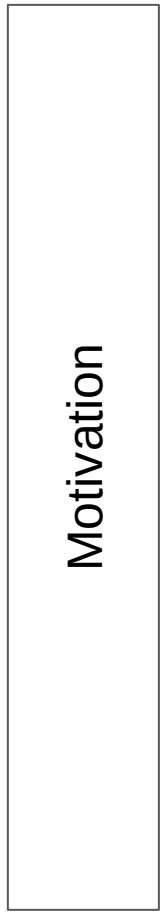
#3: Ministry of the Interior of the Czech Republic
(team of National OD Coordinator)

Open data - way more than just a format and a license



Lessons learned

motivation



OD by public administration is primarily a service to others

For financial profit of others, not publisher's

For popular apps of others, not publisher's

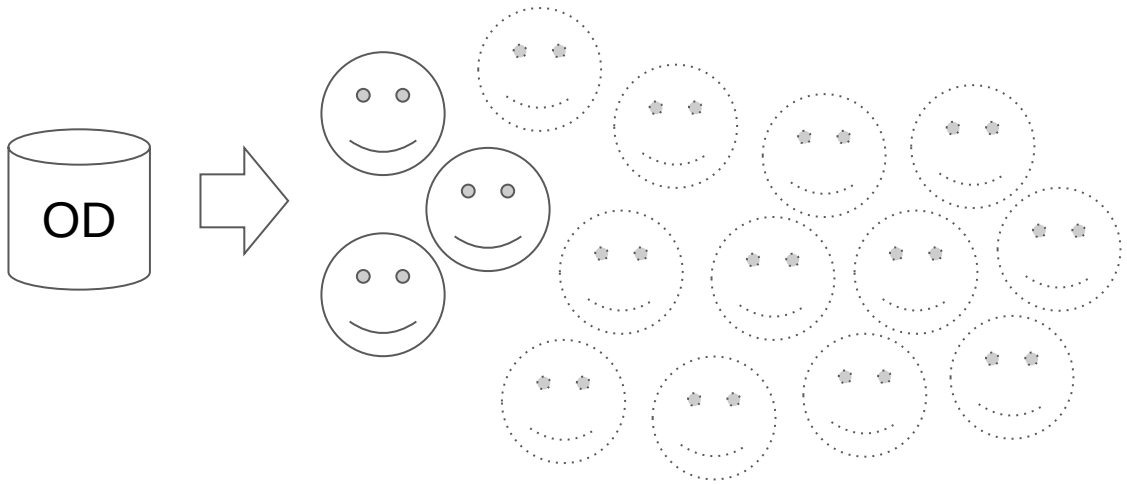
Exceptions, e.g.

- dogfooding
- PR
- better SLAs

Publisher's goal: to enable **others** to use the data in as many use cases as possible, as easily as possible

Usage often unknown to the publisher, e.g.

- internal usage in enterprises
- schools and universities
- OD re-use

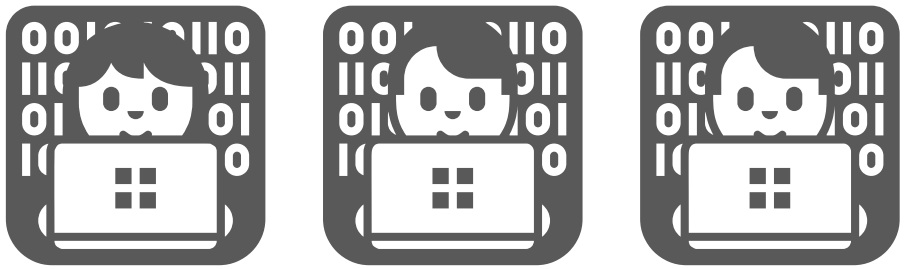


Open data is not for general public

Open data is for technically-savvy users

- analysts
- journalists
- computer scientists
- application developers
- technical students

Users who can work with data in various data formats and do something with the data.



For non-technical consumers

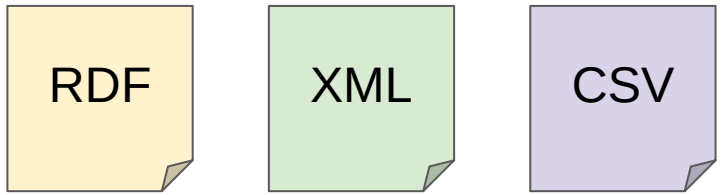
- products based on open data
  - Apps
  - News articles
  - Plot charts
  - Maps
  - …
- prepared by technically-savvy users

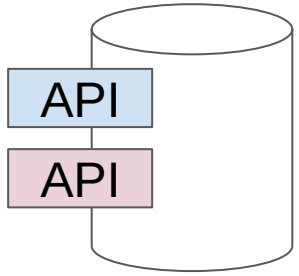
Lessons learned

OD distribution design

Documentation, schemas, distributions, metadata

OD Distribution design - Bulk download and/or APIs?





enables **most** use cases

- anyone can download the whole dataset
- do whatever they want with the data

inefficient for **some** use cases

- e.g. “get info about 1 item”
- frequently updated data

usually tailored to **some** use cases

- e.g. “get info about 1 item”
- “list items with label starting with A“

inefficient, possibly even prohibiting **some other** use cases, e.g.

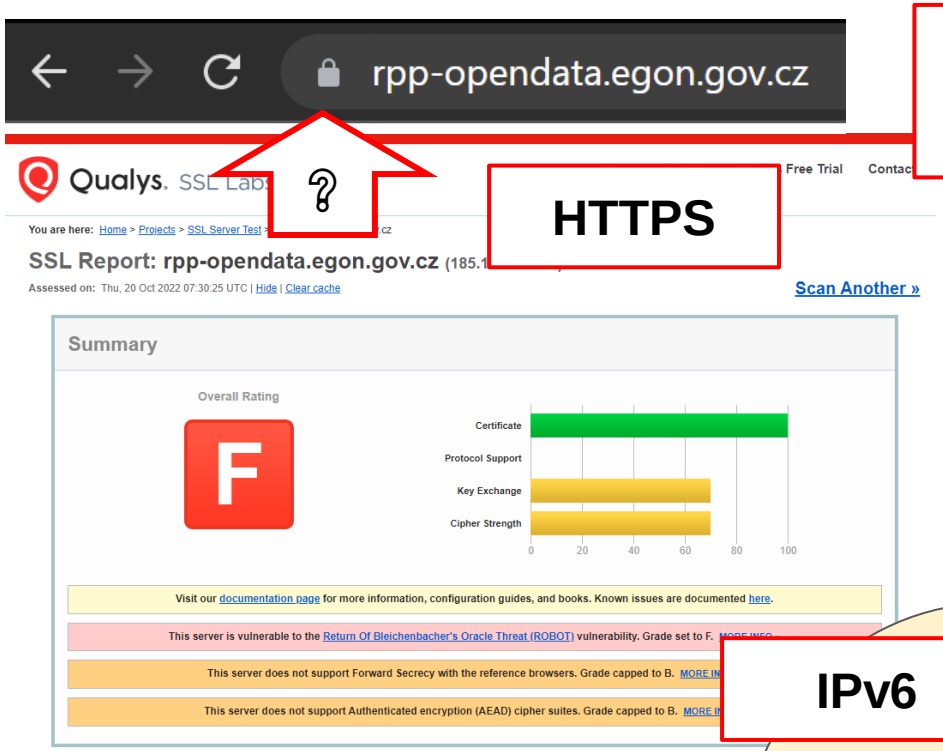
- download the whole dataset
- maintain a fresh copy of a dataset

Key OD use case: Maintain a fresh copy of the data => bulk download

Lessons learned
web technologies

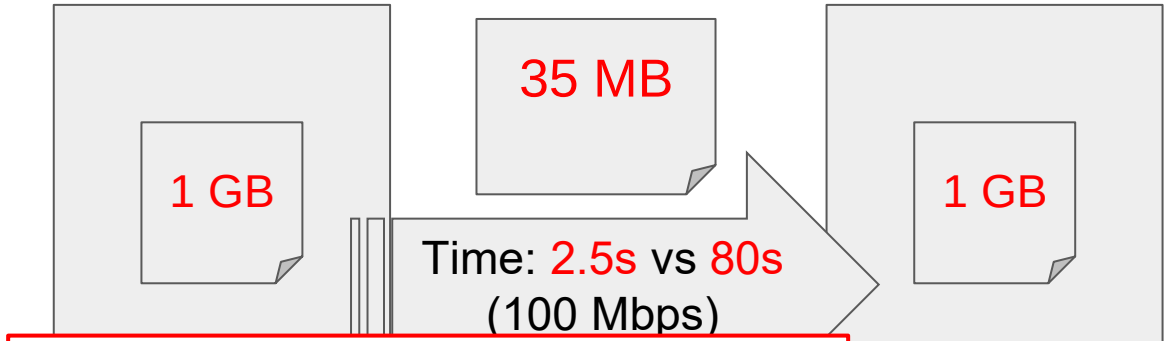
Web technologies

Lessons learned - web technologies

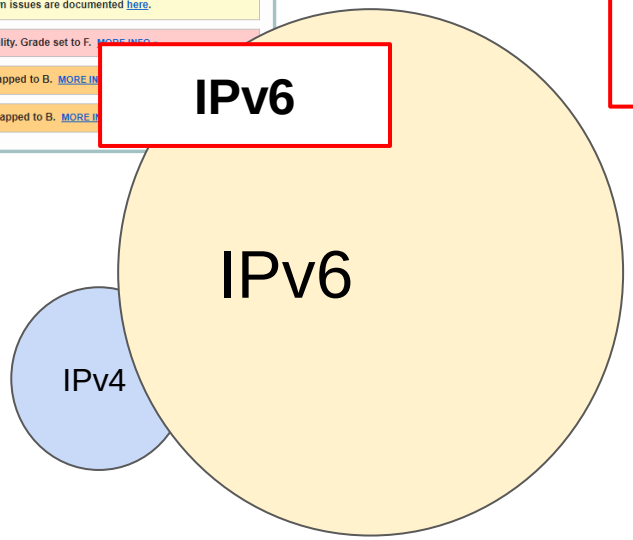


HTTPS

Cross-Origin Resource Sharing (CORS)



HTTP Compression

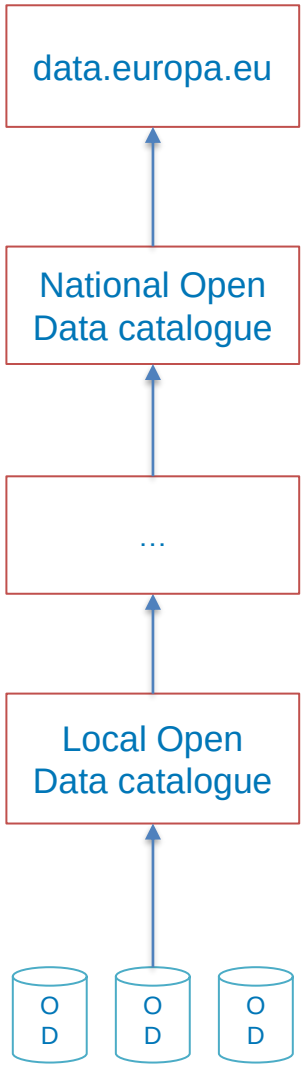


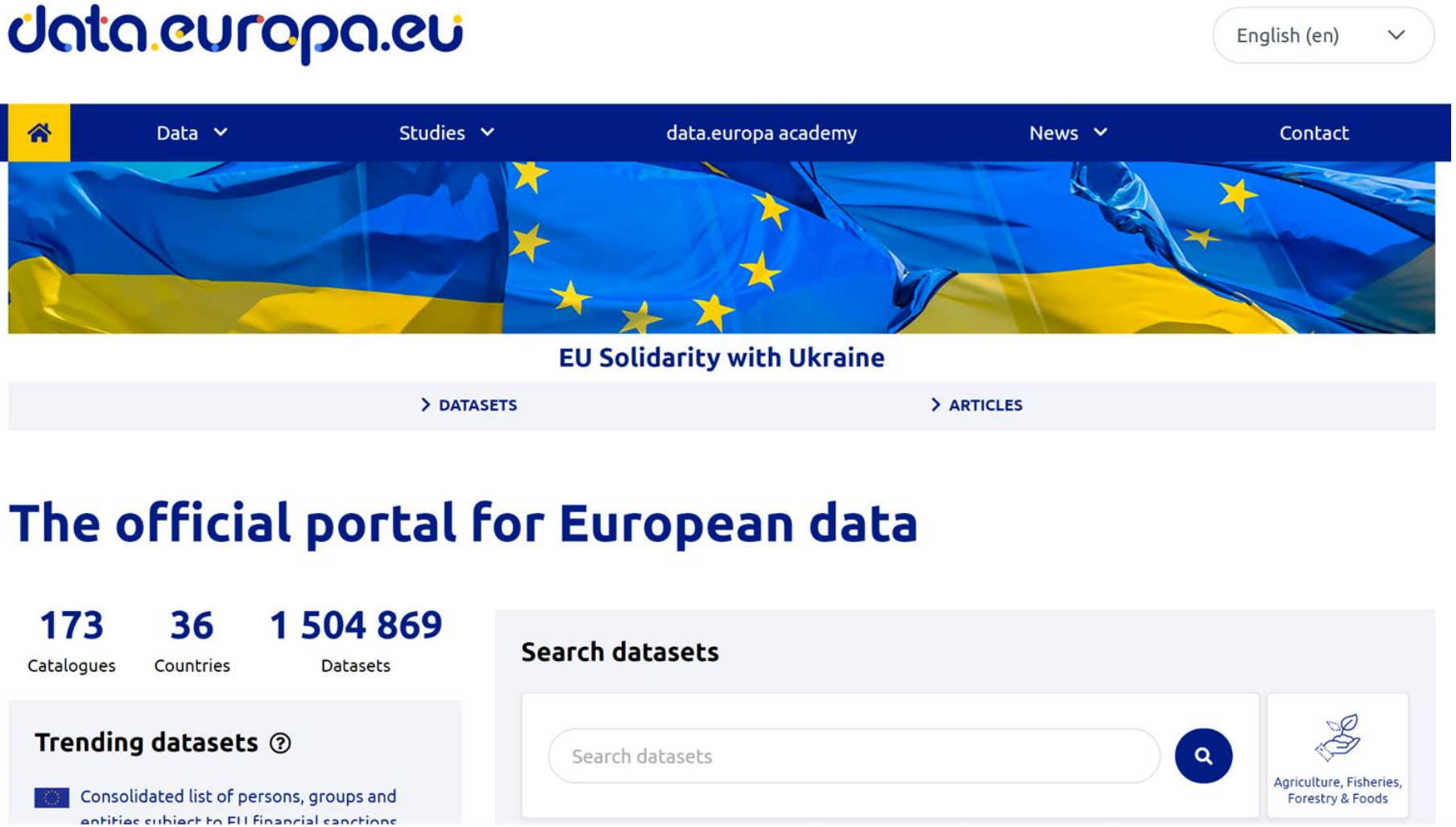
HTTP Content-Type header
e.g. application/trig

Lessons learned
catalogisation

Proper catalogisation

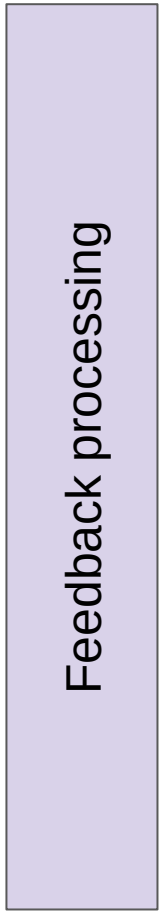
Data catalogue API - DCAT-AP





Lessons learned

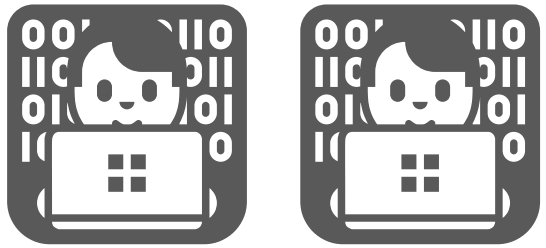
feedback processing



2 kinds of open data consumers

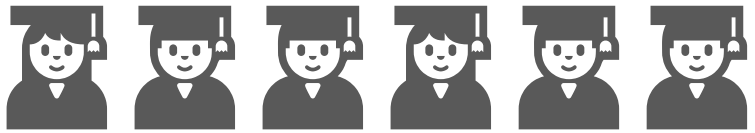
Power user

- knows the publisher
- knows what data a publisher has
- already has a use case for the data
- knows the context of the data
- can deal with may data quality problems



Novice, or potential user

- does not know the publisher exists
- does not know what data the publisher has
- does not know the data
- does not have a use case for the data
  - nor motivation to jump through hoops for it
- does not know the context of the data



Feedback processing - some of it is just lazy consumers

Lazy consumers asking for something they can do themselves

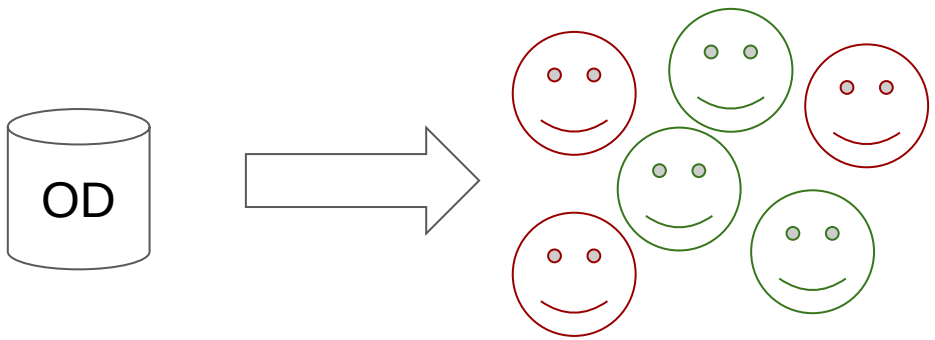
- “Give me CSV in addition to a SPARQL endpoint”
- “Give me Excel in addition to XML”
- “Give me a subset of the data”

It is OK to say “you are able to do this yourself”

Real obstacles on publisher’s side

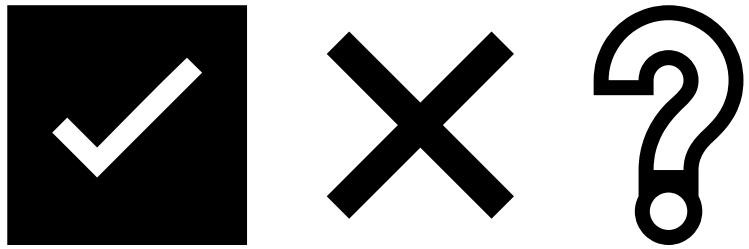
- something missing in data
- insufficient update frequency
- inability to keep a fresh copy of a dataset
- low data quality

Blocking a real use case of publisher’s data



Lessons learned

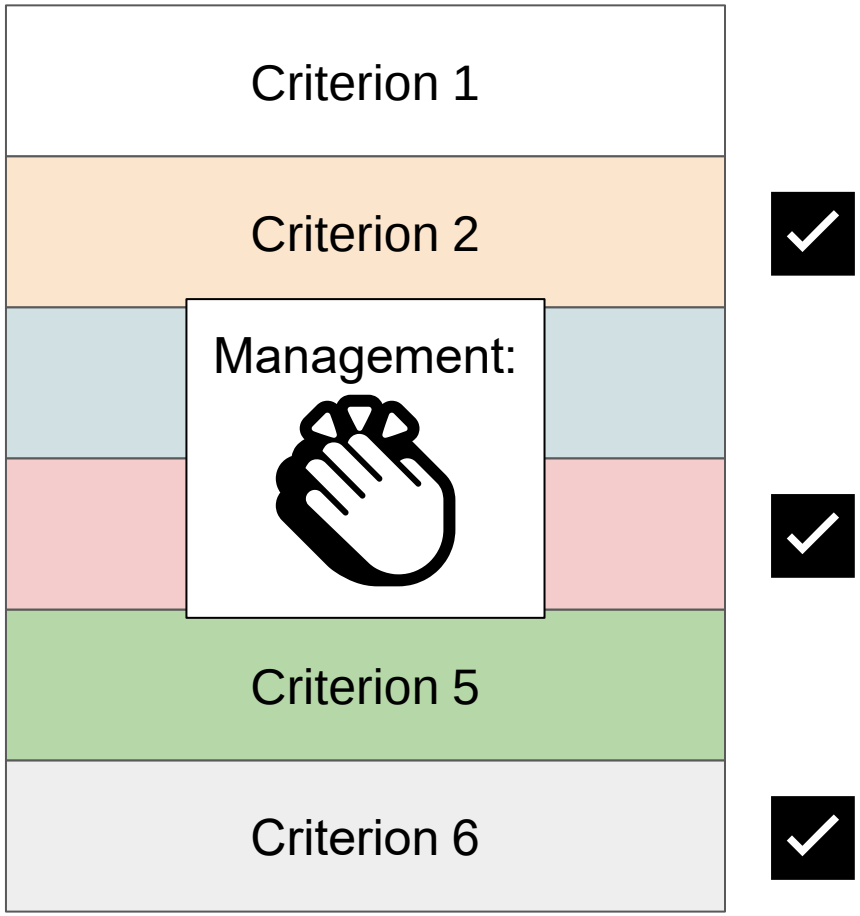
evaluation



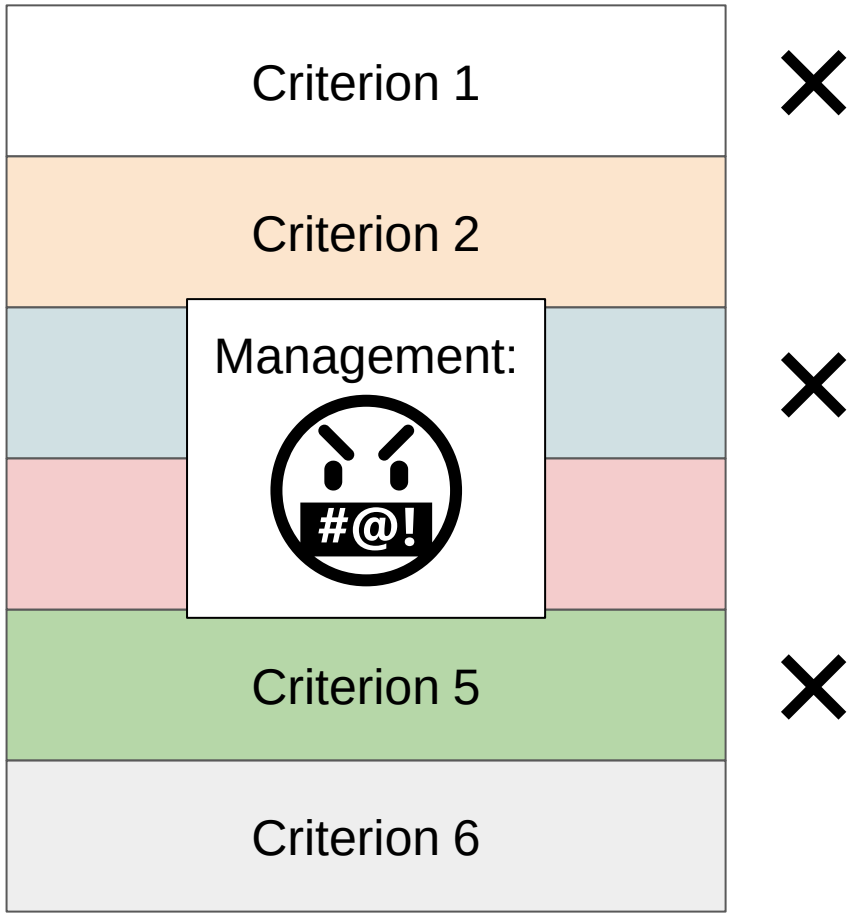
Evaluation - Best vs. Bad practice in OD

Specific to Czechia?

Ministry of ABC



Ministry of ABC



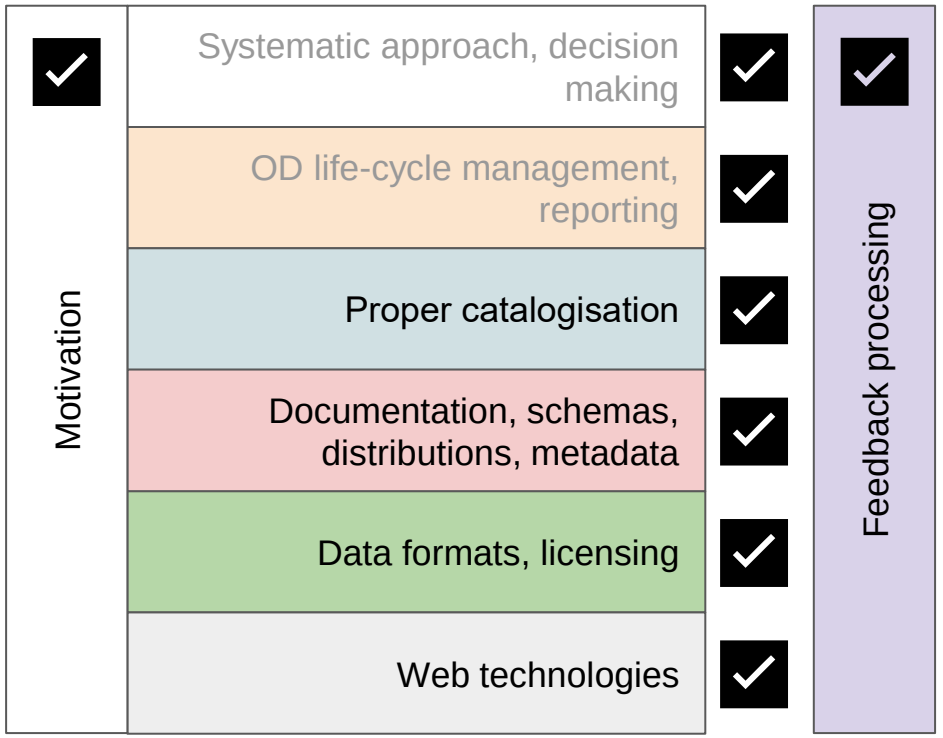
Open Data (OD) Best Practices

Jakub Klímek

Charles University

Ministry of the Interior of the Czech Republic

jakub@jakubklimek.com
@jakub_klimek



Questions?

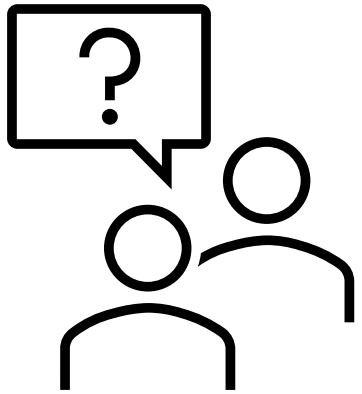


Photo by Matt Jones on Unsplash

One word or line on your experience today

on sli.do

Future webinars

Please provide us your feedback!

Stay updated!

Sign up for the newsletter: data.europa.eu/newsletter

Follow us on social media:

- EU_opendata
- Publications Office of the European Union
- data.europa.eu

data.europa.eu The official portal for European data