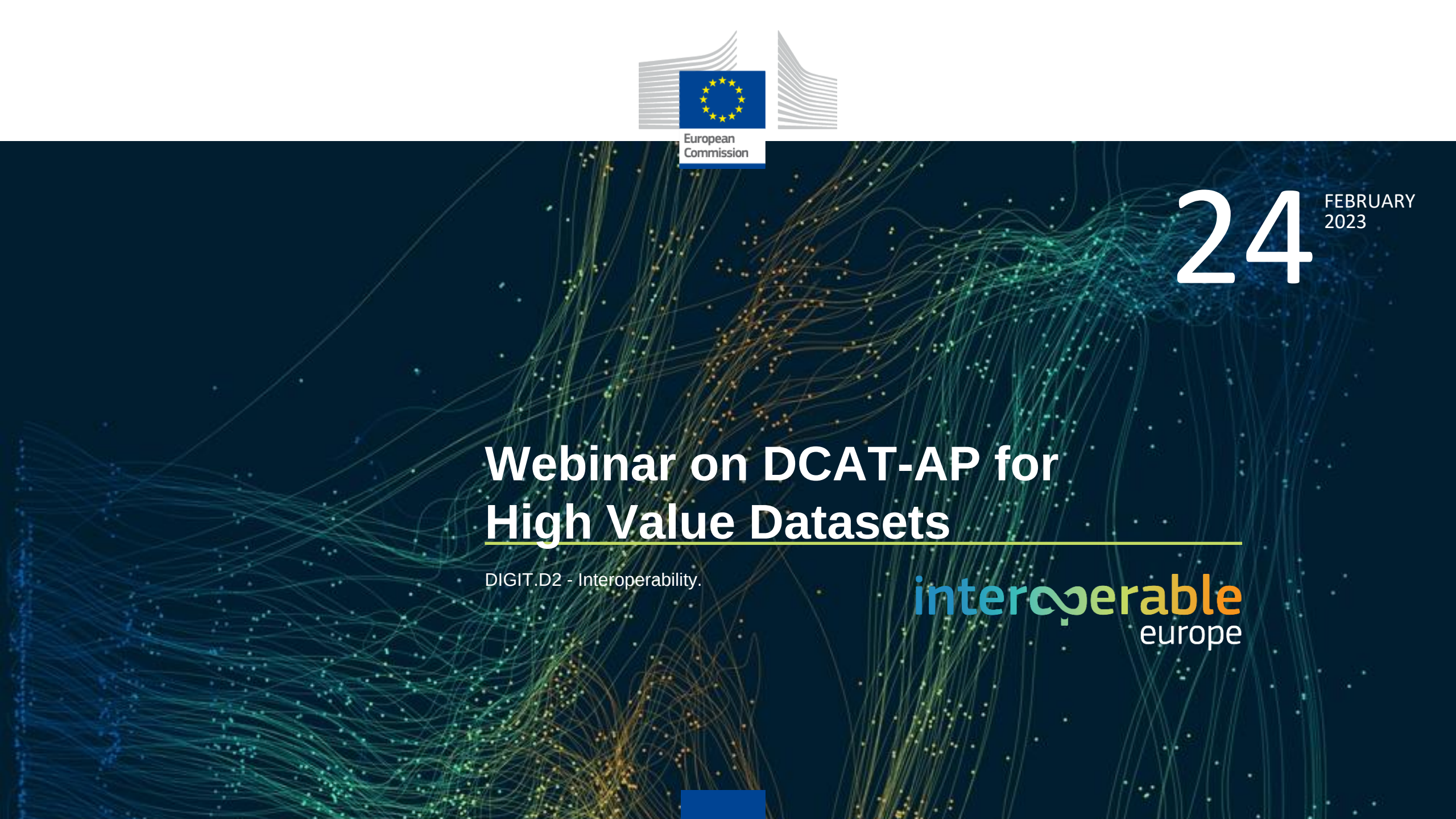

European Commission
24
FEBRUARY
2023
Webinar on DCAT-AP for High Value Datasets
DIGIT.D2 - Interoperability.
interoperable europe

Objective of this webinar

- Creating a common metadata descriptions for High Value Datasets in DCAT-AP in a collaborative way

Agenda

1. **Welcome**
2. **Context of DCAT-AP**
3. **Assessment of HVD from a metadata perspective**
4. **Short break**
5. **Proposal building DCAT-AP for HVD**
6. **Wrap up and next steps**

Workshop practicalities

Please mute your microphones

You can also share your questions for the Q&A session via the chat

The workshop will be recorded, and will be shared publicly.

Introduce yourself in the chat:

- Name
- Affiliation
- Contact email address

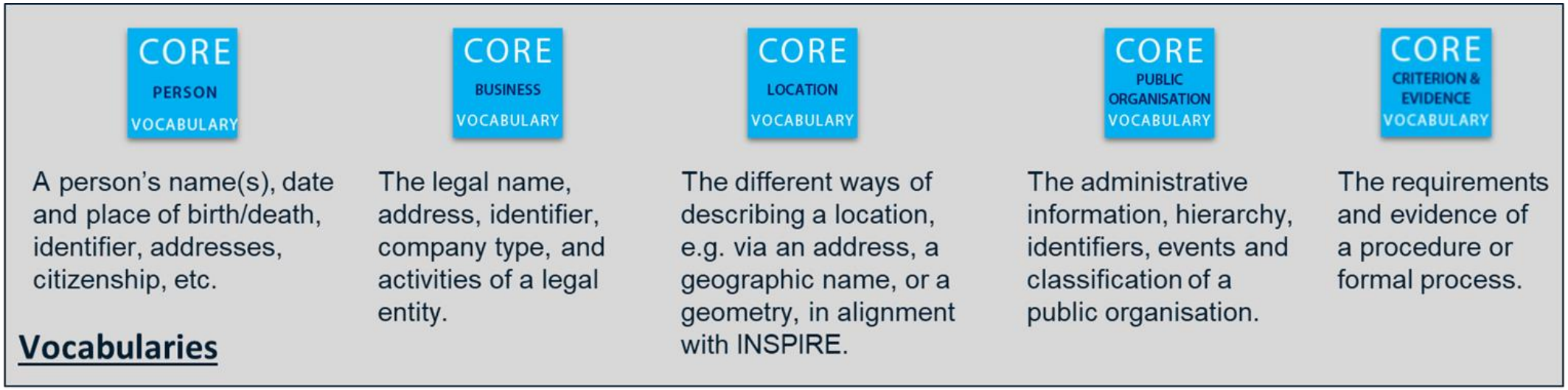
Context of the SEMIC assets

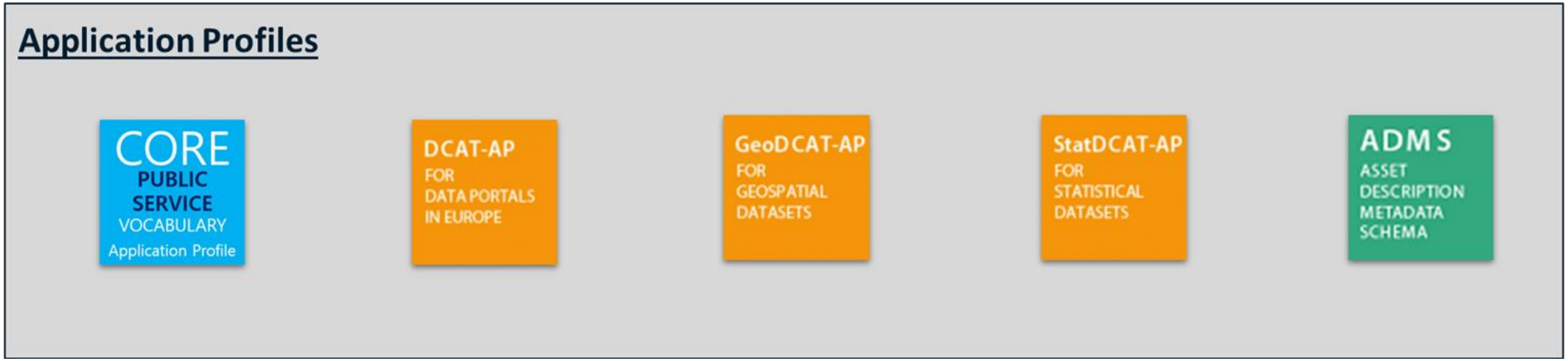
Introduction to SEMIC

The objectives of the SEMIC action is to promote Semantic Interoperability amongst the EU Member States by:

- Promoting, share and reuse of semantic assets, experience and tools and facilitating agreements in key areas.
- Identifying opportunities for alignment on semantic definitions, metadata and reference data sources with special focus on identification and definitions of Core Concepts / Vocabularies.
- Raising awareness on the importance of data and metadata management.

Current SEMIC assets





Objectives of DCAT-AP

- Supporting the discovery of/access to (open) data in a cross-border and cross-domain environment, by harvesting data from distributed portals.

- In the form of an application profile of W3C DCAT, by:
  - Expressing constraints and usages on DCAT properties and classes, and
  - Including additional properties and usages of controlled vocabularies

  In such a way that the metadata descriptions are maximally harmonised across Europe, and provide a reliable source for the European Data portal.

- Extensions exist to serve different communities better: BregDCAT-AP, GeoDCAT-AP, StatDCAT-AP.

Expected outcome of today's activity

A document that expresses how to apply DCAT-AP in order to satisfy the metadata requirements expressed in the Implementing Regulation for High Value Datasets.

Note: not foreseen that this becomes an extension, such as GeoDCAT-AP, but more an expression of use.

Governance

This activity is supported by

- DG CNECT for endorsement and alignment with policy implementation
- DIGIT/SEMIC for editorial and community facilitating services

This document will become a part for the DCAT-AP ecosystem

- SEMIC will take care that this document will be up to date with the latest DCAT-AP evolution.
- **Note:** *as the objective is to impose specific usage rules on the existing DCAT-AP there is a long term stability expected***.**

Community

Initiated as a joined activity of the

- PSI expert group (DG CNECT)
- DCAT-AP community (SEMIC)

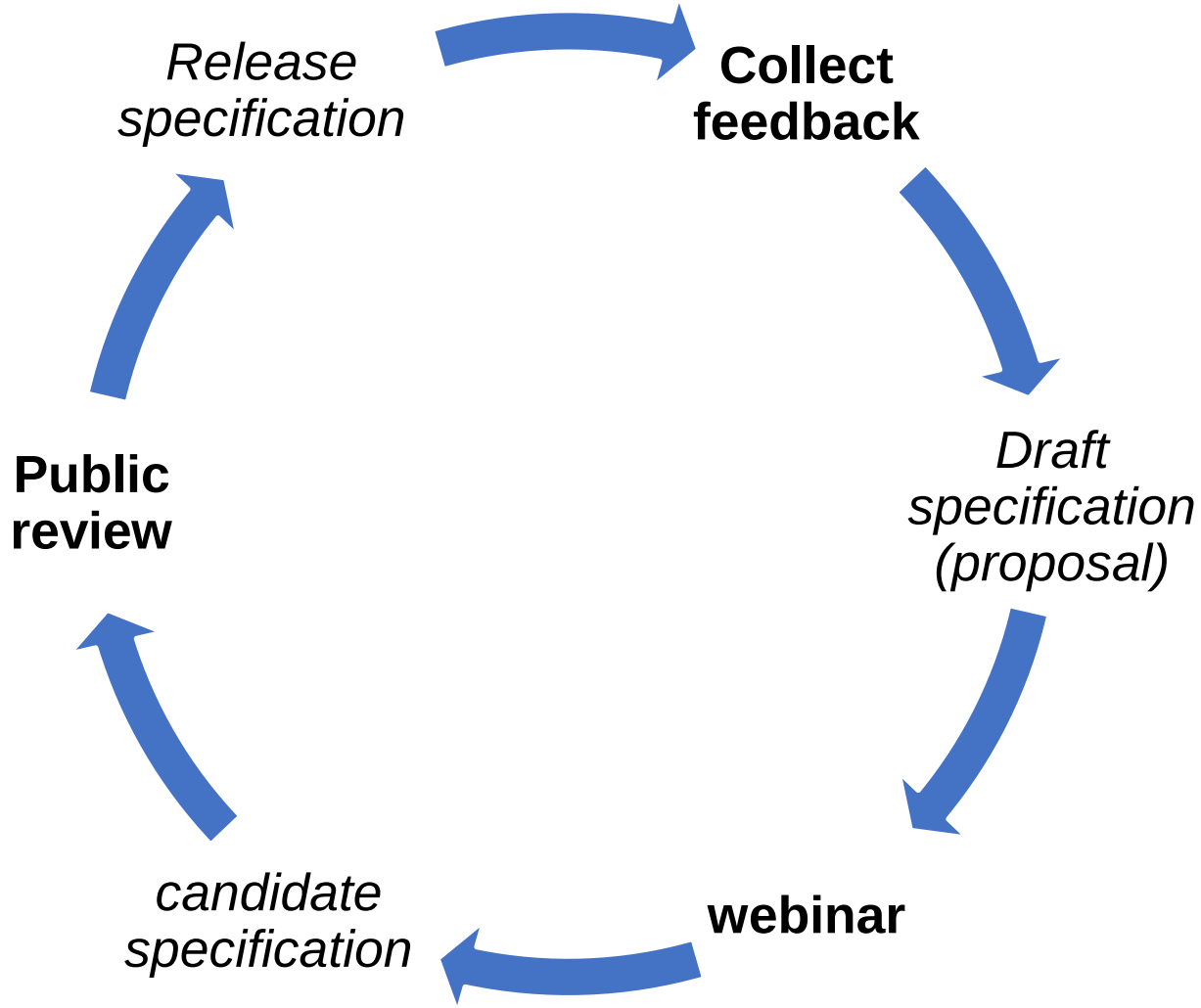
Feedback via

- GitHub
- Regular webinars

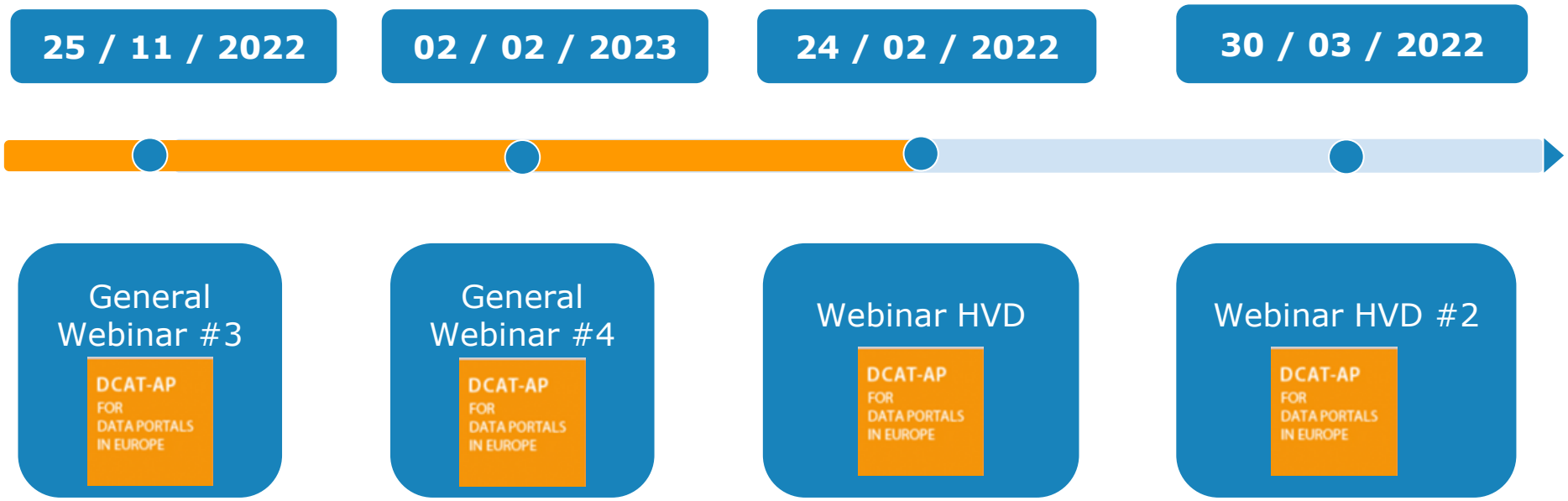
DCAT-AP data specification consists of

- Human readable text
- Support for implementers:
  - JSON-LD context
  - SHACL shapes for validation

Consensus building



Webinars


25 / 11 / 2022
02 / 02 / 2023
24 / 02 / 2022
30 / 03 / 2022
General Webinar #3
DCAT-AP FOR DATA PORTALS IN EUROPE
General Webinar #4
DCAT-AP FOR DATA PORTALS IN EUROPE
Webinar HVD
DCAT-AP FOR DATA PORTALS IN EUROPE
Webinar HVD #2
DCAT-AP FOR DATA PORTALS IN EUROPE

Ongoing activities

Issue handling

DCAT 3.0

High Value Datasets

Long term planning

- Further addressing issues raised by the community
- Supporting data spaces
- Supporting other legislative requirements

Participation (today) in this activity

- Assess the proposals and share your perspective

- How?
  - Each proposal will be introduced
  - Followed by an open discussion
  - If consensus reached, proposal is adopted.
  - If no consensus, or discussion takes too long
    - Take it off-line to github
    - Develop for the next webinar a new proposal.

- More feedback, questions, suggestions, ...
  - Meeting minutes and recording will be shared
  - File them on GitHub (or reach out via any of the channels) for open discussion with the community,

HVD metadata requirement assessment

Legislative background

Open Data Directive

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019L1024

Implementing Regulation for High Value Datasets

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R0138

General requirements (article 3)

1. *Public sector bodies holding high-value datasets listed in the Annex shall ensure that the datasets described or referenced in the Annex are made available in machine-readable formats via **APIs** corresponding to the reasonable needs of re-users. Where indicated in the Annex, the datasets shall also be made available as **a bulk download**.*
2. *Those public sector bodies referred to in paragraph 1 shall set out and publish the **terms of use** of the API and **the quality of service** criteria on its performance, capacity and availability. The terms of use shall be available **in a human-readable and machine-readable format**. Both the terms of use and the quality of service criteria shall be compatible with the arrangements for the re-use of high-value datasets laid down in accordance with Article 4.*
3. *API terms of use shall be accompanied **by API documentation in a Union or internationally recognised open, human-readable and machine-readable format**.*

General requirements (article 3)

4. *Public sector bodies referred to in paragraph 1 shall designate **a point of contact** for questions and issues related to the API with a view to ensure the availability and maintenance of the API and ultimately the smooth and effective publication of the high-value datasets.*
5. *Public sector bodies holding high-value datasets listed in the Annex shall ensure that the datasets are **denoted as high-value datasets** in their metadata description*

Requirements Assessment Annex

Publish datasets with

- ‘open legal conditions’, i.e. CC-BY 4.0
- With a bulk download
- With an API
- According to the best practices and agreements imposed by the relevant legislation

Requirements Assessment Annex

Publish the identified datasets with

- ‘Open legal conditions’, i.e. CC-BY 4.0
- With a bulk download
- With an API
- According to the best practices and agreements imposed by the relevant legislation
- **Identified by a circumscription of the content**

Example **identification**

The HVD for Buildings are

- The building datasets that are in scope of the INSPIRE data theme building (Directive 2007/2/EC)
- With the granularities up to the scale of 1:5000
- And must provide the following content:
- Unique identifier;
  - Geometry (footprint of the building);
  - Number of floors;
  - Type of use.

Requirements Assessment Annex

Publish the identified datasets with

- ‘Open legal conditions’, i.e. CC-BY 4.0
- With a bulk download
- With an API
- According to the best practices and agreements imposed by the relevant legislation
- Identified by a circumscription of the content
- **With documentation about the structure and semantics of the data**

  The datasets shall be described in a complete and publicly available online documentation describing at least the data structure and semantics (e.g. Annex 3.2.c)

Reporting requirements (article 5.3)

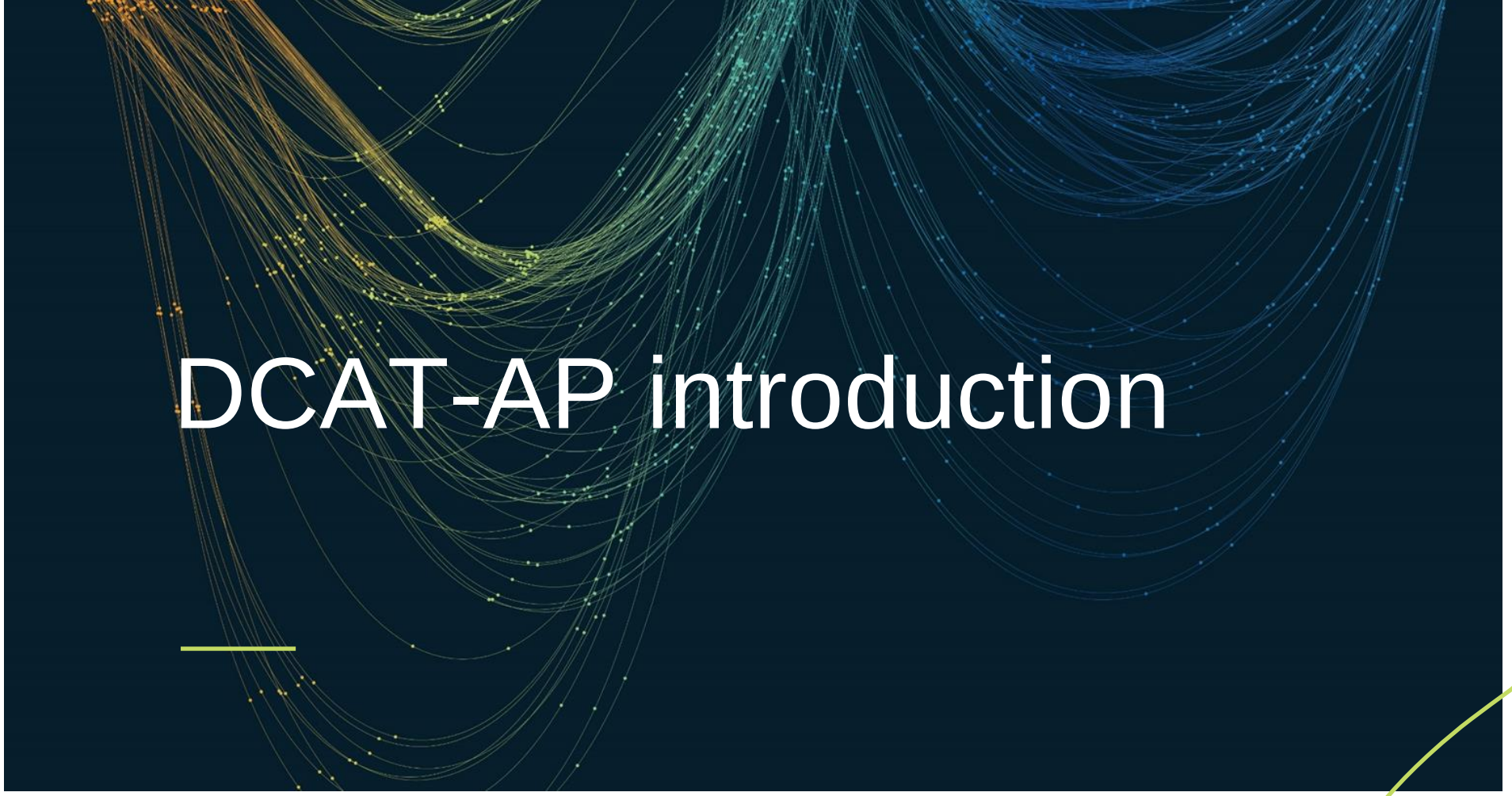
1. A **list of specific datasets** at Member State level (and, where relevant, subnational level) corresponding to the description of each high-value dataset in the Annex to this Regulation and with **online reference to metadata** that follow existing standards, such as a single register or open data catalogue;
2. A **Persistent link** **to the licensing conditions** applicable to the re-use of high-value datasets listed in the Annex to this Regulation, **per dataset** referred to in point a);
3. A **Persistent link** to **the APIs** ensuring access to the high-value datasets listed in the Annex to this Regulation, per dataset referred to in point a);

Reporting requirements (article 5.3)

4. Where available, guidance documents issued by the Member State on publishing and reusing their high-value datasets;
5. Where available, the existence of data protection impact assessments carried out in accordance with Article 35 of Regulation (EU) 2016/679;
6. The number of public sector bodies exempted in accordance with Article 14(5) of Directive (EU) 2019/1024.

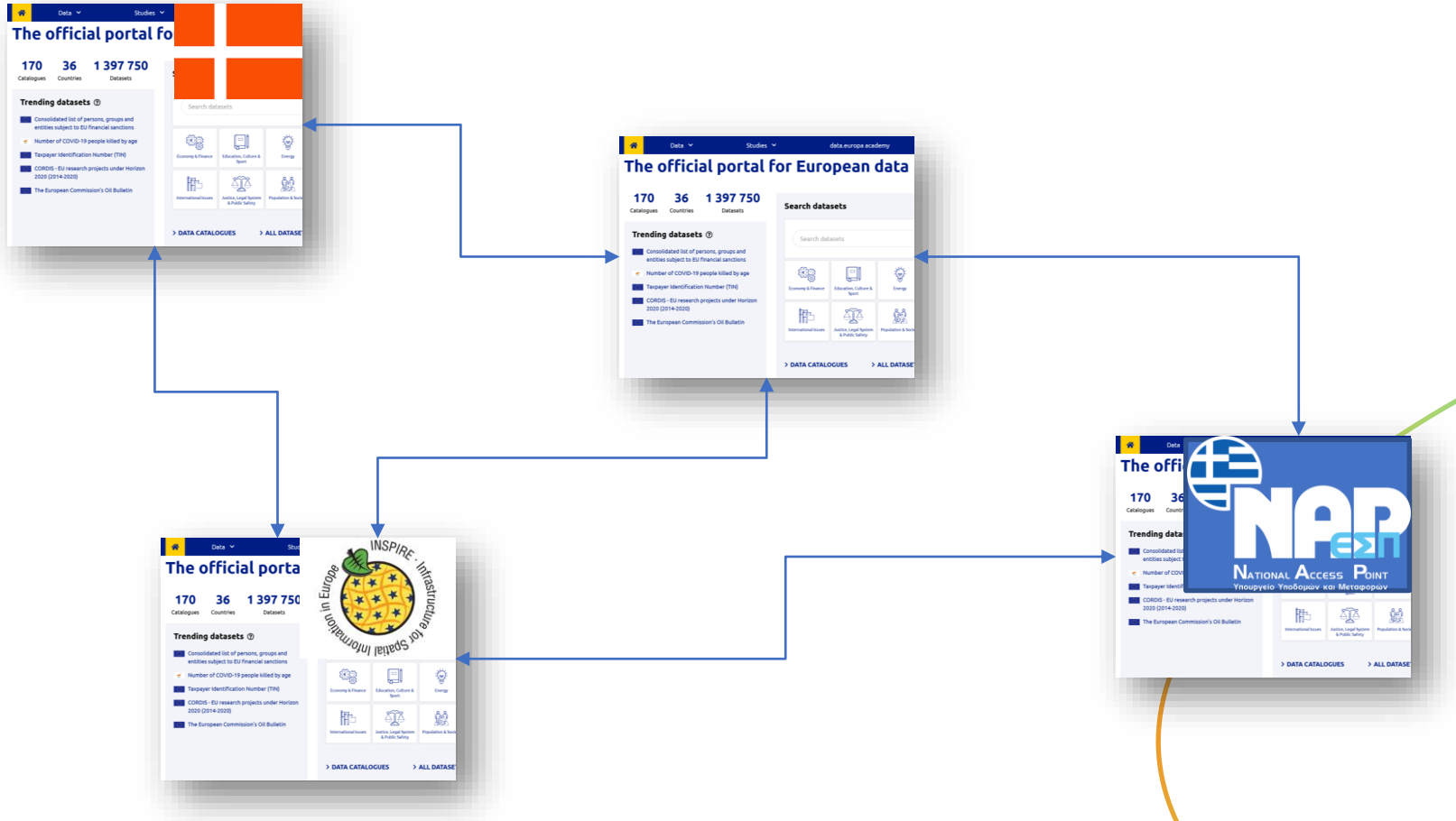
Challenges

- How can MS adapt current metadata descriptions to satisfy the mentioned requirements?
- How can MS fulfil the reporting requirements?
- How can citizens/businesses see the HVD status (outer world perspective)?



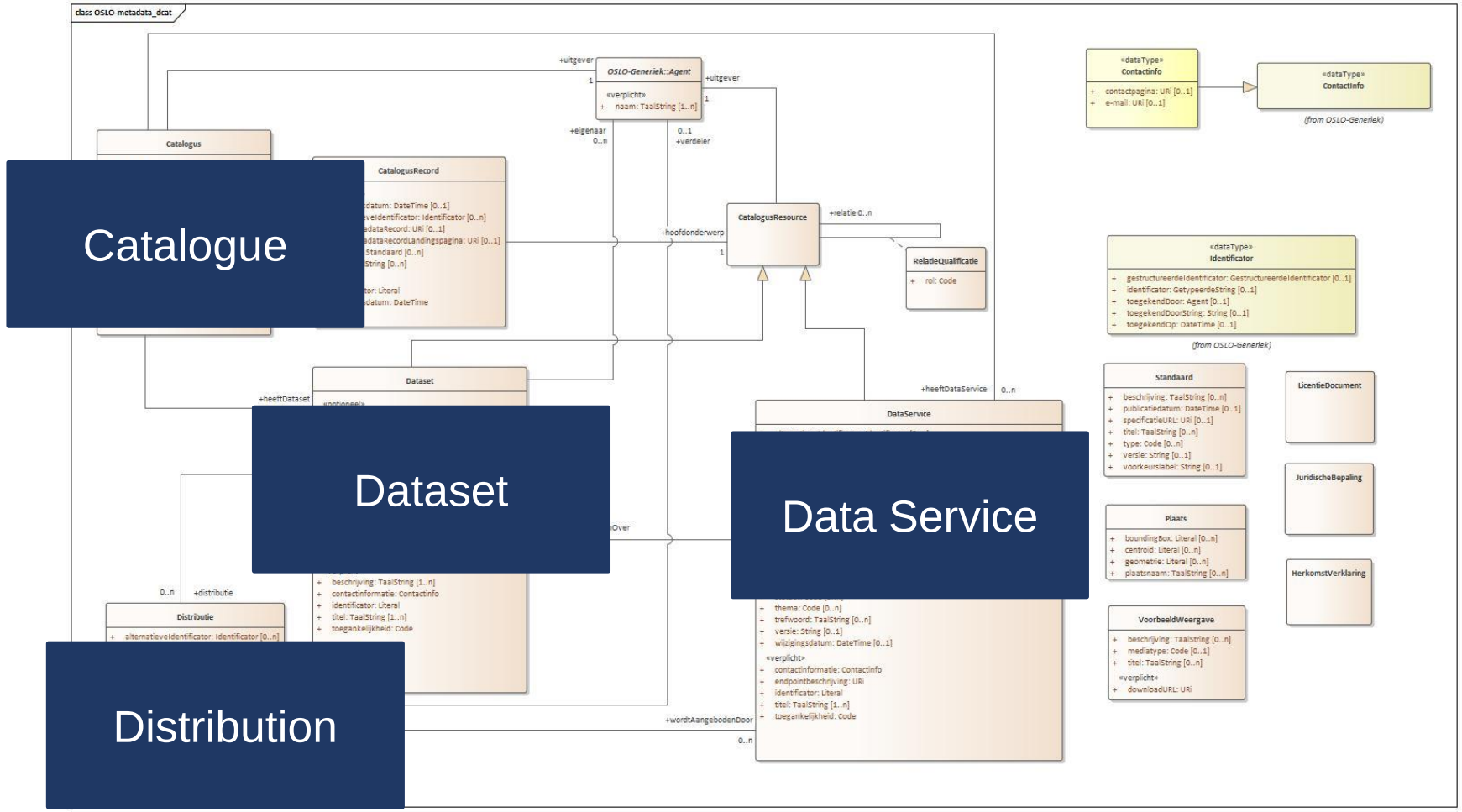
DCAT-AP introduction

DCAT-AP Use Case



DCAT-AP overview

https://github.com/SEMICeu/DCAT-AP/



Short break (5’)

Common agreements

Challenges

C1. How to know which dataset is within the scope of HVD directive?

C2. Reference to metadata descriptions

C3. Legal information

C4. Bulk download

C5. API

C6. Point of Contact

C7. Adhere to specific information requirements

C8. How to know which MS that is taking up the HVD responsibility

GitHub #251

C1. Part of HVD scope

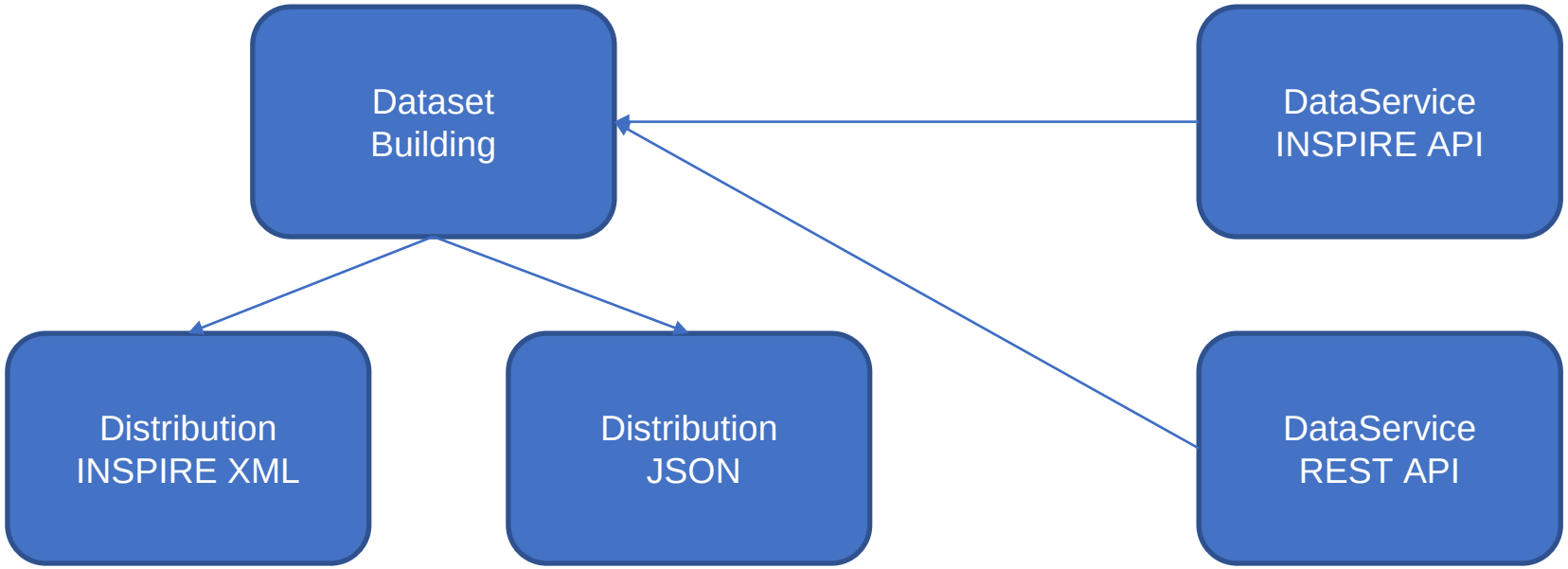
The HVD defines 6 thematic data categories:

geospatial; earth observation and environment; meteorological; statistics; companies and company ownership; mobility

Proposal:

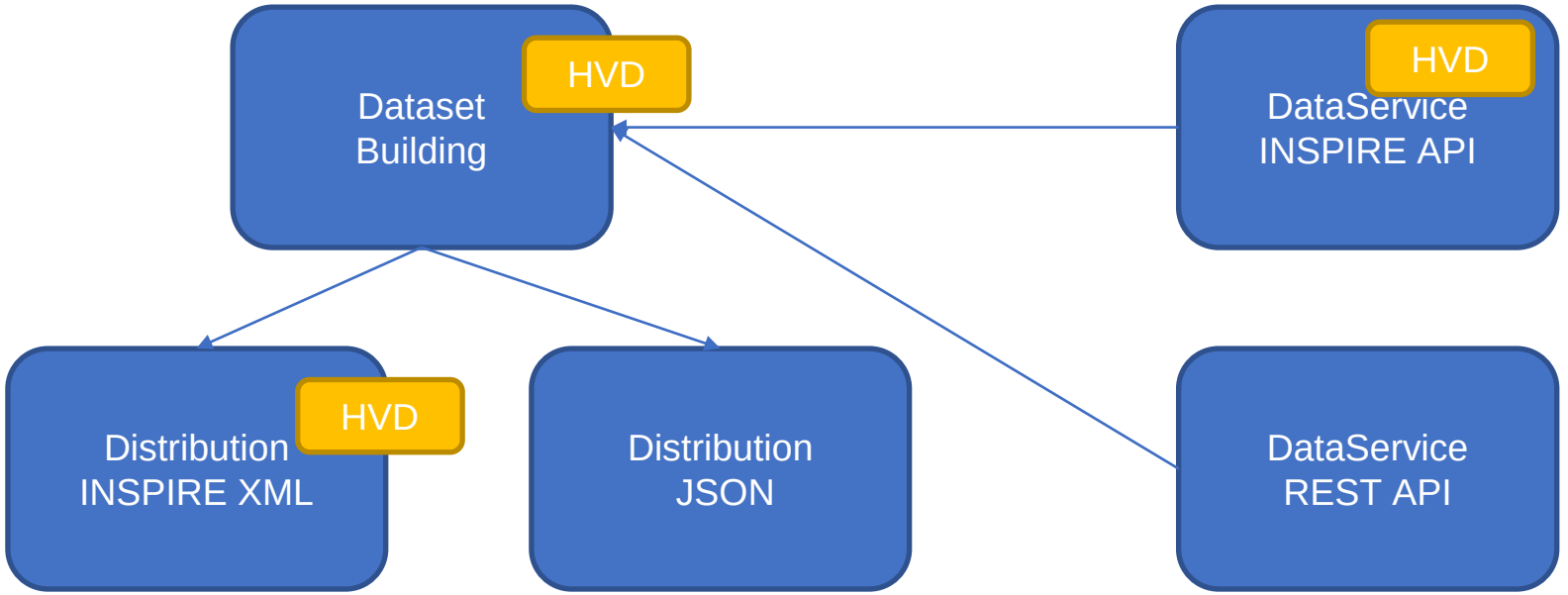
- Create new property **m8g:hvdCategory** defining **the HDV category to which this resource belongs.**
- The codelist will be created and maintained by the Publications Office.
- A resource may belong to more than one data category

C1. Part of HVD scope: example



In DCAT-AP are Dataset, Distribution, DataService and Dataset Series **independent entities**, with their own life cycle.

C1. Part of HVD scope: example



Apply categorisation to **all** entities: Dataset, Distribution, DataService and Dataset Series
Usage Note: Related entities should have the same categorisation.

GitHub #252

C2. Identifiers for each HVD entity in the DCAT-AP metadata

Reporting requirement 5.3.a: with online reference to metadata

Online

- Dereferenceable: URI → URL (thus, no uuid)
- Guidelines on identifiers for DCAT-AP
- Best practices for URIs:
  - Organization agnostic
  - System agnostic

GitHub #253

C3. License information in DCAT-AP

Principle: ***indicate legal information at the most precise level in the metadata description***

- Distribution and DataService, not at the Dataset or Dataset Series level
- Because the legal conditions for one representation can be different than for the other.

Usage Note

- Use the codelist from the Publications Office, if no MS specific codelist is provided.
- http://publications.europa.eu/resource/authority/licence/CC_BY_4_0 → Only machine readable data

C3. Licence information

For HVD the licence information shall be given by

- Property dct:licence
- With a URI value (persistent link)
  - URI should be dereferenceable:
    - machine readable (provide RDF representation)
    - Human readable text

To indicate relationship with CC-BY 4.0

Proposal: add properties to express a relationship with another license.

- Proposed properties: rdfs:seeAlso and owl:sameAs
- + restrict target list to the codelist of the Publications Office

C3. Licence information

Current state of affairs at the DEU (early Feb 2023):

700 000 (approx.) distinct licences of:

- 500 000 (approx.) are blanks (no URI)
  - 170 000 (approx.) have dct:identifier, having text values like "cc-by4“
- 180000 are dereferenceable URIs per distribution in CZ
  - Alignment with CC-BY by dereferencing
- 1000s are well-known URIs for BE
  - Alignment with CC-BY by DEU sparql endpoint querying
- + 500 candidates
  - Alignment with CC-BY by manual inspection

GitHub #254

C4. Bulk download

- Bulk download is a functionality that allows to get the whole dataset as a copy on your local computer
- In DCAT-AP terminology: a Distribution
- Proposal: A HVD bulk download is denoted as a Distribution for a HVD dataset.

C4. Bulk download

- Assumption: The download URL is not always public, it might be only given on request.

- Use the current DCAT-AP practice:
  - Property *access URL* is mandatory
  - Property *download URL* is optional

GitHub #255

C5. API representation

Terminology:

- API = *a set of functions, procedures, definitions and protocols for machine-to-machine communication and the seamless exchange of data (source HVD)*
- Data Service = *A collection of operations that provides access to one or more datasets or data processing functions. (source DCAT-AP)*

Proposal: a HVD API is DataService

C5. API representation

DataService has properties:

- Endpoint URL = *The root location or primary endpoint of the service (a Web-resolvable IRI)*.
  - i.e. the physical endpoint to be used
- Endpoint description = *A description of the services available via the end-points, including their operations, parameters etc.*
  - i.e. the documentation in the form of OpenAPI, OGC GetCapabilities,...

C5. Reference

Reporting requirement 5.3.c: persistent link to APIs

DCAT-AP recommends persistency of the DataService URI, but does not impose any requirements on the Endpoint URL

Proposal:

It is recommended to perform efforts to maintain persistency for both.

C5. API

The HVD mentions:

- Terms of use
- Quality of service

C5. API: Terms of use

Proposal:

Documenting the *terms of use* is considered as the same activity as documenting the *license*.

C5. API: Quality of Service

- DCAT-AP does **not** have a specific property for quality of service

- Quality of service covers a broad spectrum of topics

Proposal:

Add a usage note on the generic documenting property **dcat:landingPage** that is should contain information or a reference to information about the quality of service.

GitHub #256

C6. Point of contact

Public sector bodies shall designate **a point of contact** for questions and issues related **to the API** with a view to ensure the availability and maintenance of the API and ultimately the smooth and effective publication of the high-value datasets

Proposal:

A contact point is **mandatory** for *HVD DataServices* and **recommended** for *HVD Datasets* either in the form of *an (persistent) email-address* or a *webform address* (service desk)

GitHub #257

C7. Adhere to specific data requirements

(annex 3.2): the datasets shall be described in a complete and publicly available online documentation describing at least the data structure and semantics

Proposed approaches:

Option A

Implicit

Option B

Explicit with state of the art information

C7. Adhere to specific data requirements

Option A

Implicit

- The label that is has been identified as HVD is sufficient to know that it follows the directive
- No technical check can be performed if this is satisfied or not (self declaration)

Option B

Explicit with state of the art information

C7. Adhere to specific data requirements

Option A

Implicit

Option B

Explicit with state of the art information

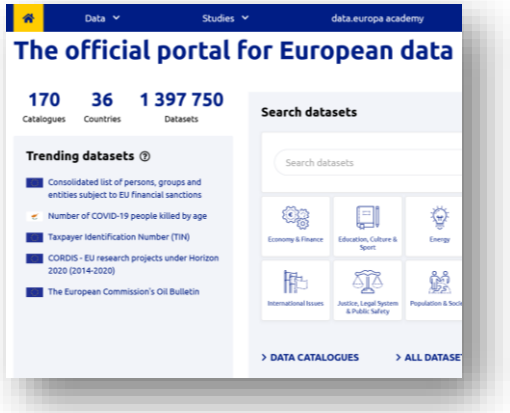
- (from DCAT-AP) Use dct:conformsTo to express relationship with a **public** document (standard) that describes the internals of the resource.
- This ensures that the information about the content is made publicly accessible for reusers
- Technical check can be performed by experts when using the provided information to assess if that matches the HVD requirements.

Reporting queries

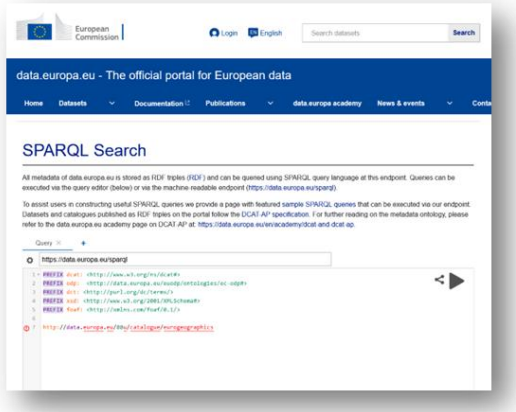
Share your reporting needs on GitHub

Reporting

All metadata is harvested by data.europa.eu.



Data.europa.eu offers SPARQL endpoint



Objective:
Design queries that highlight the HVD status of a MS

Reporting queries (example 1)

Find all datasets that are within scope of HVD

```
Select ?hvd where {
    ?hvd a dcat:Dataset.
    ?hvd r5r:hdvCategory ?cat.
}
```

Reporting queries (example 2)

Find all APIs that are within scope of HVD

```
Select ?hvd where {
    ?hvd a dcat:DataService.
    ?hvd r5r:hdvCategory ?cat.
}
```

Reporting queries (example 3)

Find all bulk downloads that are within scope of HVD

```
Select ?hvd where {
    ?hvd a dcat:Distribution.
    ?hvd r5r:hdvCategory ?cat.
}
```

Reporting queries (example 4)

Find all datasets that are within scope of HVD and have a bulk download and an API

```
Select ?hvd ?bulk ?api where {
    ?hvd a dcat:Dataset.
    ?hvd r5r:hdvCategory ?cat.
    ?hvd dcat:distribution ?bulk.
    ?bulk r5r:hdvCategory ?catb.
    ?api dcat:servesDataset ?hvd.
    ?api  r5r:hdvCategory ?cata.
}
```

Reporting queries (example 5)

Find all datasets that are within scope of HVD and have a bulk download and an API with their licences

```
Select ?hvd ?bulk ?bulklic ?api ?apilic where {
    ?hvd a dcat:Dataset.
    ?hvd r5r:hdvCategory ?cat.
    ?hvd dcat:distribution ?bulk.
    ?bulk r5r:hdvCategory ?catb.
    ?bulk dct:license ?bulklic
    ?api dcat:servesDataset ?hvd.
    ?api  r5r:hdvCategory ?cata.
    ?api dct:license ?apilic
}
```

Reporting queries (example 6)

Find all datasets that are within scope of HVD and have a bulk download and an API with their licences for a MS <catalogueMS>

```
Select ?hvd ?bulk ?bulklic ?api ?apilic where {
    <catalogueMS> a dcat:Catalog.
    <catalogueMS> dcat:dataset ?hvd.
    <catalogueMS> dcat:service ?api.
    ?hvd a dcat:Dataset.
    ?hvd r5r:hdvCategory ?cat.
    ?hvd dcat:distribution ?bulk.
    ?bulk r5r:hdvCategory ?catb.
    ?bulk dct:license ?bulklic
    ?api dcat:servesDataset ?hvd.
    ?api  r5r:hdvCategory ?cata.
    ?api dct:license ?apilic
}
```

How do we know <catalogueMS> ?

GitHub #258

C8. Indicating MS

- DCAT-AP does not provide the notion of MS responsibility
- MS data in data.europa.eu is the result of multiple catalogues
- Some catalogues may provide data from multiple MS
- DCAT-AP dct:spatial has a broader meaning than jurisdiction/legal responsibility, i.e. the “geographical coverage”.

C8. Indicating MS

Proposal option A

MS provide a separate DCAT-AP HVD catalogue containing only the metadata that is relevant in the context of the HVD.

- Note that in case the HVD rules on persistent identifiers this will never lead to duplicates in the aggregated data.europa.eu. The same data can be supplied via different ways.

- \+ No new property introduced

\+ Clear scope of reporting

\+ Use of persistent identifiers is mandatory

\+ MS can include information from other catalogues

- Portal system implementers may have to provide support for such separate catalogue.
- Visitors to a portal that would like to see the MS perspective need access to this catalogue. This must be provided by portal implementers.

C8. Indicating MS

Proposal option B

Introduce a property to indicate MS responsibility in the context of the HVD

+ Clear indication by a publisher
+ Probably a trivial extension for portal implementers
- High editorial effort
- In case of multiple catalogues, all portals should participate
- In case of multiple catalogues, no global overview exists unless one maintains it. Thus the reporting may become still complex and may lead to implement proposal option A.

Wrap-up and next steps

DCAT-AP for High Value Datasets

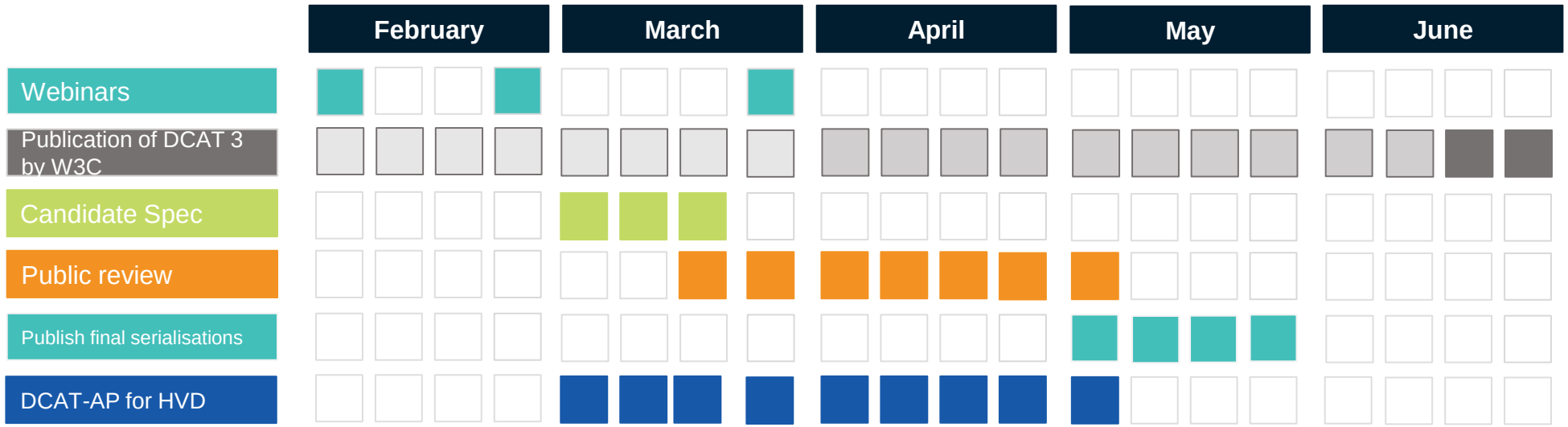
- Discussed 8 proposals on how to apply DCAT-AP in order to satisfy the metadata requirements expressed in the Implementing Regulation for High Value Datasets.
- Available for discussion on GitHub (open issues)
- Welcome any feedback

https://github.com/SEMICeu/DCAT-AP/issues

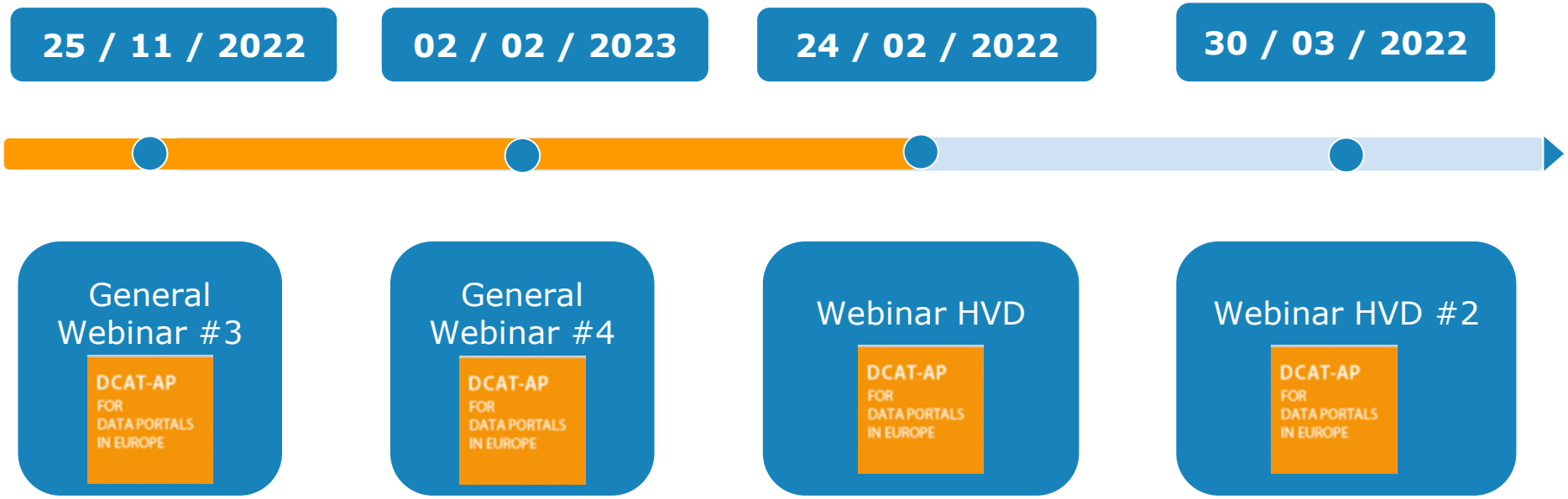
New release DCAT-AP

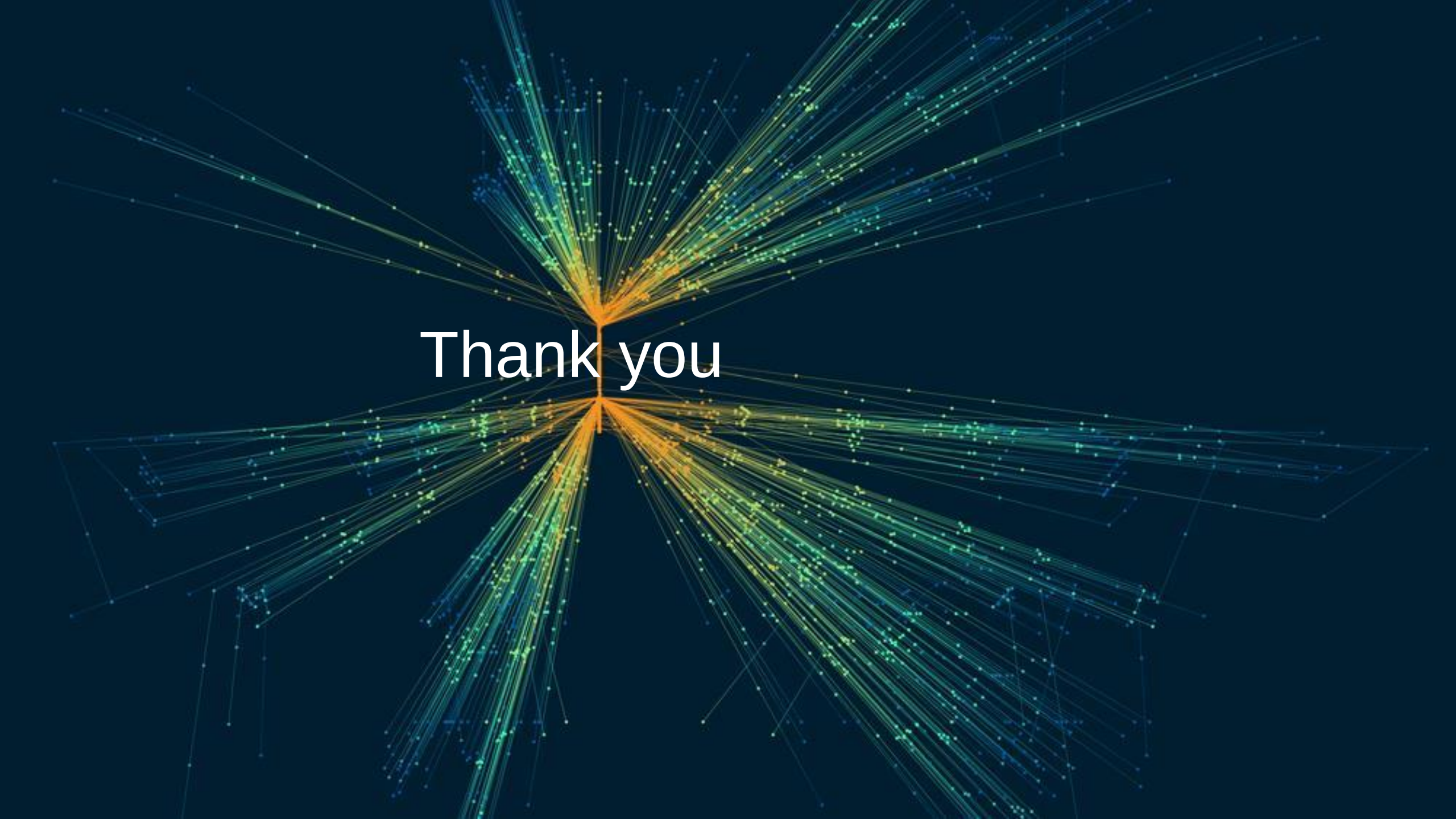
- Bug fixes
- Addressing issues according to agreed resolution from webinar November 2022
- HTML representation
- Alignment with W3C DCAT 3.0
- High Value Datasets

Timeline release



Next webinar




Thank you

interoperable europe

innovation ∞ govtech ∞ community

Stay in touch

(@InteroperableEU) / Twitter

Interoperable Europe - YouTube

Interoperable Europe | LinkedIn

DIGIT-INTEROPERABILITY@ec.europa.eu

https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe

Short break (5’)

Proposal summary

Additional profile statements: HVD Dataset

property	uri	range	card	definition	Usage note
HVD category	**m8g:hvdCategory**	**Concept**	**1..n**	**The HDV category to which this dataset belongs**	**The codelist <HVDCategory> must be used**
contact point	dcat:contactPoint	**HDV Kind**	**0..n (rec)**	contact information that can be used for sending comments about the Dataset	
conforms to	dct:conformsTo	Standard	**1..n**	refers to an implementing rule or other specification.	**The provided information should enable to the verification whether the detailed information requirements by the HVD is satisfied.**
Dataset distribution	dcat:distribution	**HVD BulkDownload Distribution**	**1..n**	links the Dataset to an available Distribution	**At least one of the associated distributions must satisfy the Bulk Download requirement of HVD.**

Additional profile statements: HVD Bulk download Distribution

property	uri	range	card	definition	Usage note
HVD category	**m8g:hvdCategory**	**Concept**	**1..n**	**The HDV category to which this dataset belongs**	**The codelist <HVDCategory> must be used**
conforms to	dct:conformsTo	Standard	**1..n**	refers to an implementing rule or other specification.	**The provided information should enable to the verification whether the detailed information requirements by the HVD is satisfied.**
Access URL	Dcat:accessURL	Resource	1..n	a URL that gives access to a Distribution of the Dataset. The resource at the access URL may contain information about how to get the Dataset.	
Download URL	Dcat:downloadURL	Resource	0..n	a URL that is a direct link to a downloadable file in a given format.	
Licence	Dct:license	**HVD Licence Document**	**1..1**	the licence under which the Distribution is made available.	**The licence must be a dereferenceable URI, provided as machine and human readable description.**

Additional profile statements: HVD Licence Document

property	uri	range	card	definition	Usage note
Same as	owl:sameAs	URI	0..n	An alternative formulation for the licence with the same legal conditions	From a legal perspective both licences are the same.
For more information	rdfs:seeAlso	URI	0..n	For a more information about the legal statements.	The provided information gives more insight in the legal conditions.

Additional profile statements: HVD Contact Point

property	uri	range	card	definition	Usage note
Contact page	Foaf:page	URI	0..1	A webpage that either allows to make contact (i.e. a webform) or the information contains how to get into contact.	
email	Vcard:hasEmail	URI	0..1	A email address via which one can make contact.	

Additional profile statements: HVD Data Service

property	uri	Range	Card	definition	Usage note
HVD category	**m8g:hvdCategory**	**Concept**	**1..n**	**The HDV category to which this dataset belongs**	**The codelist <HVDCategory> must be used**
conforms to	dct:conformsTo	Standard	**0..n (rec)**	refers to an implementing rule or other specification.	**The provided information should enable to the verification whether the detailed information requirements by the HVD is satisfied.**
Endpoint description	dcat:enpointDescription	Resource	**1..n**	a description of the services available via the end-points, including their operations, parameters etc.The property gives specific details of the actual endpoint instances, while dct:conformsTo is used to indicate the general standard or specification that the endpoints implement.	
Endpoint URL	Dcat:endpointURL	Resource	1..n	The root location or primary endpoint of the service (an IRI).	
Licence	Dct:license	**HVD Licence Document**	**1..1**	the licence under which the Distribution is made available.	**The licence must be a dereferenceable URI, provided as machine and human readable description.**