The European Single Access Point:

Harvesting guidelines for National Single Information Points

Speakers

Gianfranco Cecconi
Capgemini Invent

**Maria Rosaria
Coduti**
European Commission

Juan Carvajal B.
Fraunhofer FOKUS

Torben Jastrow
Fraunhofer FOKUS

Agenda

- **Opening and welcome**

- **Introduction to the Data Governance Act: European Single Access Point**

- **Overall implementation plan for the ESAP on data.europa.eu**

- **Harvesting guidelines: technical requirements**

- **Next steps: Roadmap and objectives**

- **Q&A regarding harvesting guidelines**

- **Final questions and closing of the webinar**

Introduction to the Data Governance Act:
European Single Access Point

Data Governance Act

▶ Requires Member States to establish **National Single Information Points** (NSIPs).

▶ The NSIP shall receive requests for the re-use of data held by public sector bodies protected on grounds of:
- a) commercial confidentiality;
- b) statistical confidentiality;
- c) third parties' intellectual property;
- d) protection of personal data.

▶ It defines a common framework for the access of such data, which shall be non-discriminatory, proportionate, and allow competition.

▶ Overall objective is to increase sharing and re-use of non-open public sector data.

DGA - ESAP

- Requires the Commission to establish a European Single Access Point (ESAP).

- The ESAP should offer a searchable register of data available in the National Single Information Points.

- It also should offer further information on how to request data via the National Single Information Points.
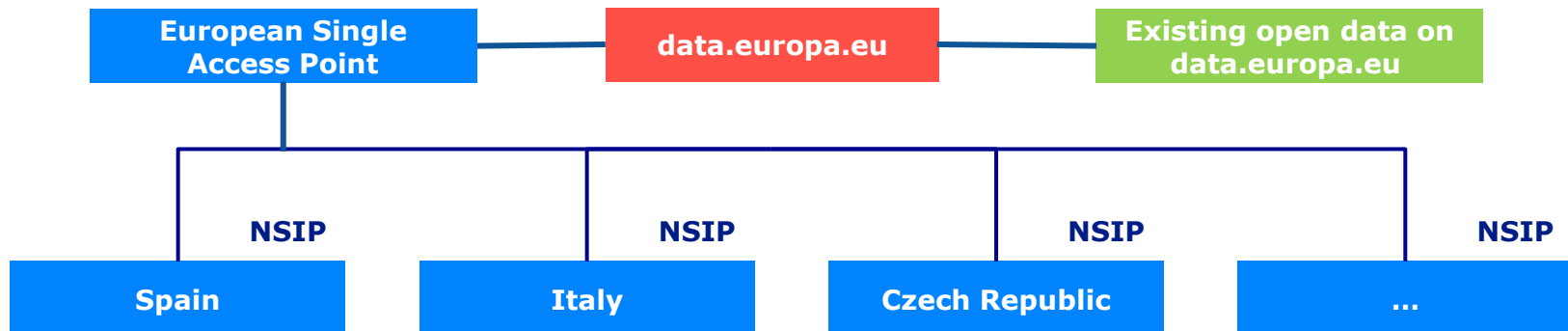
Overall implementation plan for the ESAP on data.europa.eu

European Single Access Point

> The **European Single Access Point** (ESAP) will be integrated into data.europa.eu.

> The ESAP will build upon the **existing harvesting infrastructure** and data architecture.

> Member States have experience with the existing **APIs and workflows**.

> Metadata is **automatically translated** and discoverable via data.europa.eu's **multilingual search**.

ESAP - Structure

> All data included in the ESAP will be are arranged in a **dedicated catalogue** structure and is searchable via filters as well as existing search features on data.europa.eu.

> On data.europa.eu, users will be able to search and find metadata relating to
> - **open data** (from open data portals, already available) and
> - **non-open data** (via the European Single Access Point).

European Single Access Point	data.europa.eu	Existing open data on data.europa.eu

NSIP	NSIP	NSIP	NSIP
Spain	Italy	Czech Republic	...

ESAP Standard – DCAT-AP

- The Data Governance Act does not specify a technical standard for the implementation of NSIPs or the ESAP-

- The **accepted specification** for harvesting and storing (open) data portals in Europe is the Data Catalog Vocabulary - Aplication Profile (**DCAT-AP**).

- Because the ESAP builds on existing infrastructure it also **taps into existing expertise** on DCAT-AP (both in EU institutions and MS governments).

- The ESAP will be implemented based on a minor adaptation of **the DCAT-AP specification**, mapping the required data to DCAT-AP properties.

DCAT-AP https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

ESAP Catalogues

- NSIP metadata on datasets and distributions are automatically grouped into **country-specific NSIP catalogues** in the ESAP.

- For each country **only one NSIP catalogue** exists, e.g. "Italy National Single Information Point".

- All NSIP catalogues are sorted into a **meta-catalogue** labelled "European Single Access Point".

- This data structure allows users to retrieve metadata either across the **entire range** of the ESAP (i.e. discovering data across all NSIPs) or to limit their search to only one NSIP.

Representation of NSIP metadata in ESAP
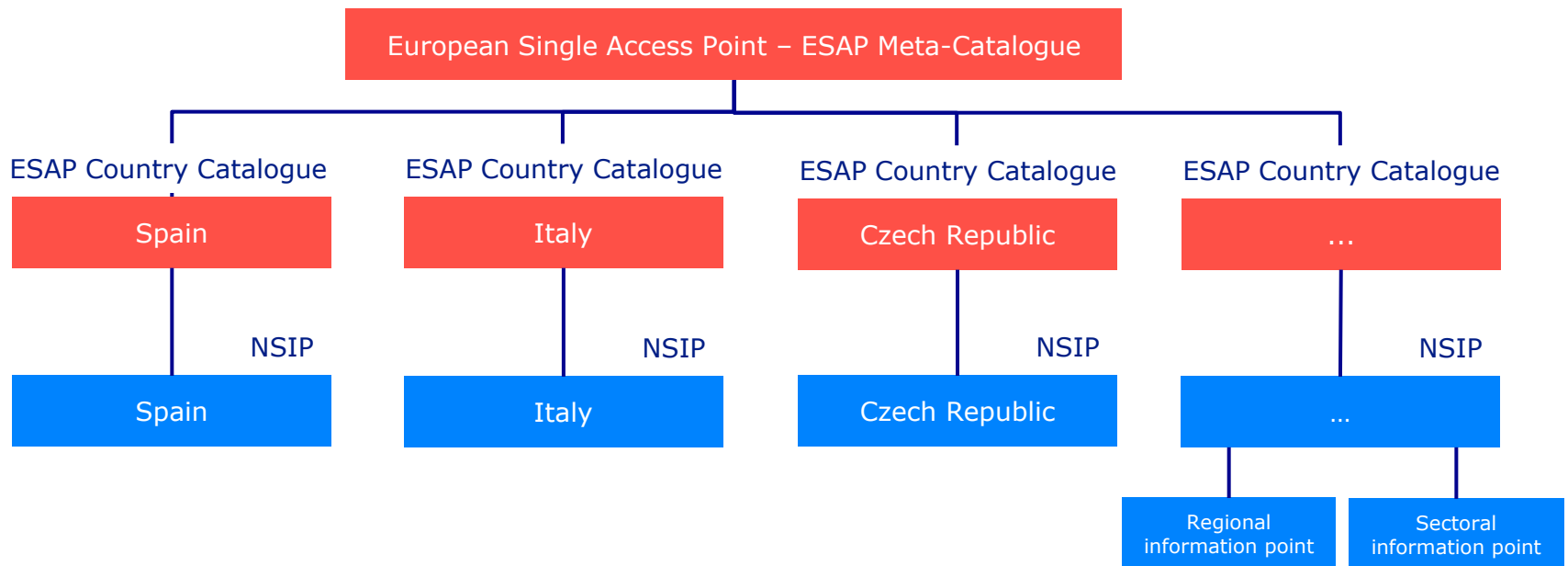
Harvesting guidelines: technical requirements

Endpoint

Members States National Single Information Points (NSIP) must offer

➤ A dedicated endpoint exclusively for the harvesting of NSIP data **or**

➤ A filter mechanism in the endpoint that allows the retrieval of NSIP data only

Flexibility

This approach gives some flexibility to member states in the implementation of their NSIPs. But correct harvesting relies on a solution to differentiate NSIP-data from open data.

Harvesting frequency

- The harvester accesses the NSIP endpoints weekly, depending on the total size of data provided.

- The harvester is configured individually for each NSIP, exact **harvesting schedules depend on the amount of data** to be harvested.

- Only updated data is effectively updated in the triplestore.

Access & retrieval

▶ Endpoints that require **authentication, must provide credentials** to access the data.

▶ NSIPs must provide **one dedicated endpoint** or **filtered endpoint** to harvest the data (see slide 14, Endpoint).

▶ The preferred format and protocol are DCAT-AP via the Open Archives Initiative Protocol for Metadata Harvesting **OAI-PMH.**

▶ Harvested datasets should have a **consistent unique id** to avoid duplication.

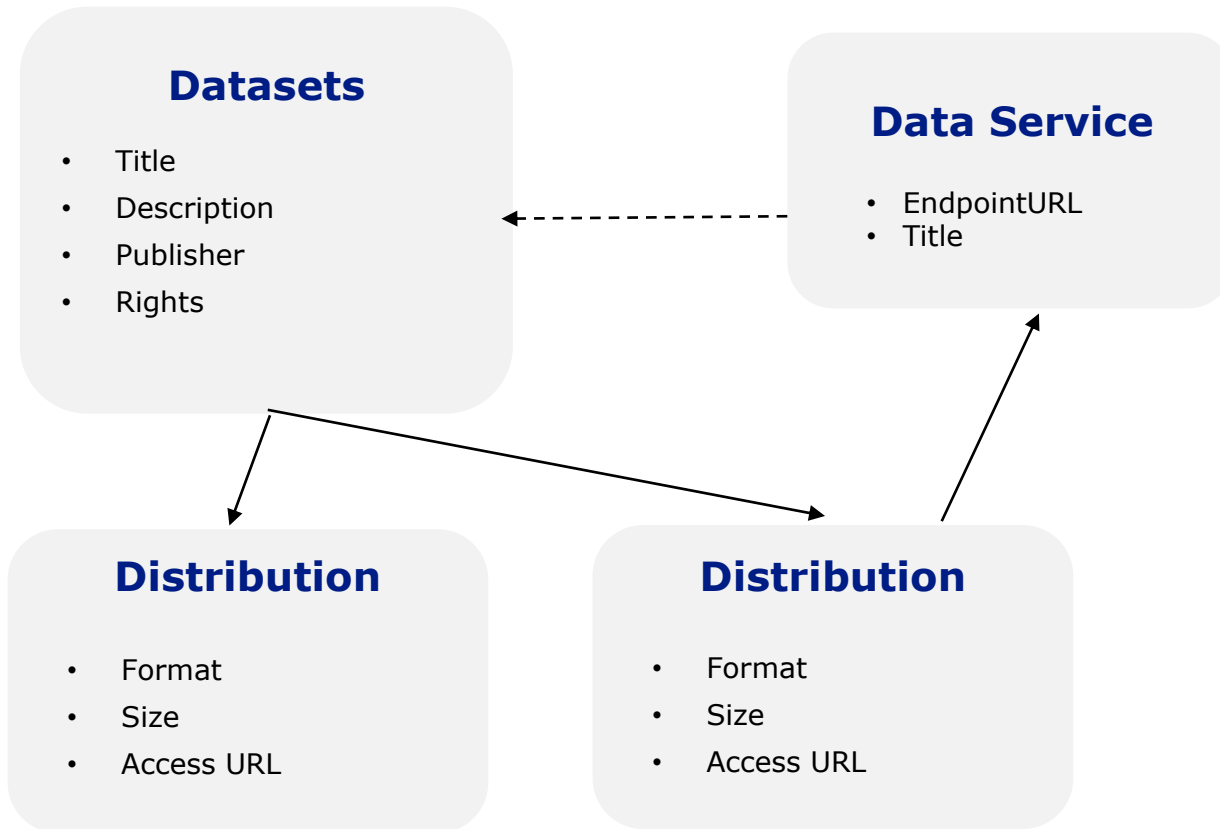| Hint | DCAT-AP via OAI-PMH is the preferred way of data harvesting! |

Required metadata

> The metadata should meet the requirements of **both the DGA and DCAT-AP**.

> The DGA requires relevant information describing the available data including at least: **format**, **size**, **conditions for its re-use,** and **access procedure**.

> The DCAT-AP specification lists the following mandatory properties for datasets and distributions: **title**, **Description**, **Publisher**, **Access URL**.

Background

Because the DGA requires certain fields, other mandatory properties from the DCAT-AP specification come into play. The requirements for the metadata come both from the DGA and the DCAT-AP specification.

Structure

Datasets

- Title
- Description
- Publisher
- Rights

Data Service

- EndpointURL
- Title

Distribution

- Format
- Size
- Access URL

Distribution

- Format
- Size
- Access URL

Metadata - Datasets

Property	URI	Range	Usage note
Title (M)	dct:title *	rdfs:Literal	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name.
Description (M)	dct:description *	rdfs:Literal	This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description.
Publisher (M)	dct:publisher	foaf:Agent	This property refers to an entity (organisation) responsible for making the Dataset available.
Conditions for re-use (Rights) (M)	dct:rights	dct: RightsStatement	This property refers to a statement that specifies rights associated with the Dataset

Metadata - Distributions

Property	URI	Range	Usage note
Format (M)	dct:format	dct:MediaTypeOrExtent	This property refers to the file format of the Distribution.
Size (M)	dcat:byteSize	rdfs:Literal	The size in bytes can be approximated (as a decimal) if the precise size is not known.
Access procedure (O)	dcat:accessURL*	rdfs:Resource	A URL of a Website that contains information on how to request the data.

Metadata - Data Service

Only if data is available via an endpoint that is accessible **without prior registration**:

> We recommend using the **DataService** class to state information on endpointURL and title, ideally also indicating which Dataset this relates to via servesDataset property

Interfaces

The harvester can read DCAT-AP compliant metadata delivered via

⬤ a) a dump file containing the RDF/XML representation of the datasets,

⬤ b) a DCAT-AP data from a SPARQL endpoint, or

⬤ c) OAI-PMH.

Best practice!

OAI-PMH

Open Archives Initiative Protocol for Metadata Harvesting is a protocol is **specifically designed to gather metadata.**

Datasets can be returned in various metadata formats by an OAI-PMH endpoint, but DCAT-AP is preferred.

CKAN

> The CKAN version **must support the action API**. The legacy APIs of CKAN are not supported.

> **Only the 'package_search' API endpoint** is needed to harvest the metadata.

> The endpoint should be able to **query and paginate** the metadata.

> It must accept query parameters and **return dictionaries** with the datasets.

Legacy

Categorization

AGRI	Agriculture, fisheries, forestry and food
ECON	Economy and finance
EDUC	Education, culture and sport
ENER	Energy
ENVI	Environment
GOVE	Government and public sector
HEAL	Health
INTR	International issues
JUST	Justice, legal system and public safety
REGI	Regions and cities
SOCI	Population and society
TECH	Science and technology
TRAN	Transport

https://op.europa.eu/en/web/eu-vocabularies/dataset/-
/resource?uri=http://publications.europa.eu/resource/dataset/data-theme

Getting Started

To get your NSIP harvested you should take two steps:

➤ **1st**: Check that your NSIP is suitable for harvesting (Checklist).

➤ **2nd**: issue a harvesting request via the data.europa.eu contact form.

Checklist

Requirement	Value
Does your NSIP provide the metadata listed under in section 4 of the documentation?	See section 4 in the guidelines. Only metadata can be harvested, not the data itself.
Which metadata standard is supported by your NSIP?	DCAT-AP or CKAN (mapped to DCAT-AP)
Which representation of the metadata is used?	XML, JSON, or any RDF representation
Which API is used to retrieve the data?	OAI-PMH; RDF dump file; CKAN; SPARQL endpoint
Is authentication required for you to access your API?	yes/no
Does your data include complete vocabulary for categorization, or other fields that use a defined vocabulary (for example update frequency)?	Free text. Please provide vocabularies with translations, if available.
What is the Uniform resource locator (URL) to your NSIPs interface / endpoint?	Free text (URL)

Checklist

Requirement	Value
How often can/should the site be harvested?	e.g. daily/weekly/monthly/etc.
Are there any times when the site should not be harvested (e.g. scheduled maintenance)?	Free text
Who is the **publisher** of the National Single Information Point (Name and email address)?	Free text
Does your data use standard date/time formats?	ISO8601
How can NSIP data be filtered from other (i.e. open) data that is part your infrastructure?	Free text. This is essential to correctly distinguish data provided under the DGA from other data.

Getting Started

Use the contact form to initiate the onboarding process for your NSIP. Please also provide the following information:
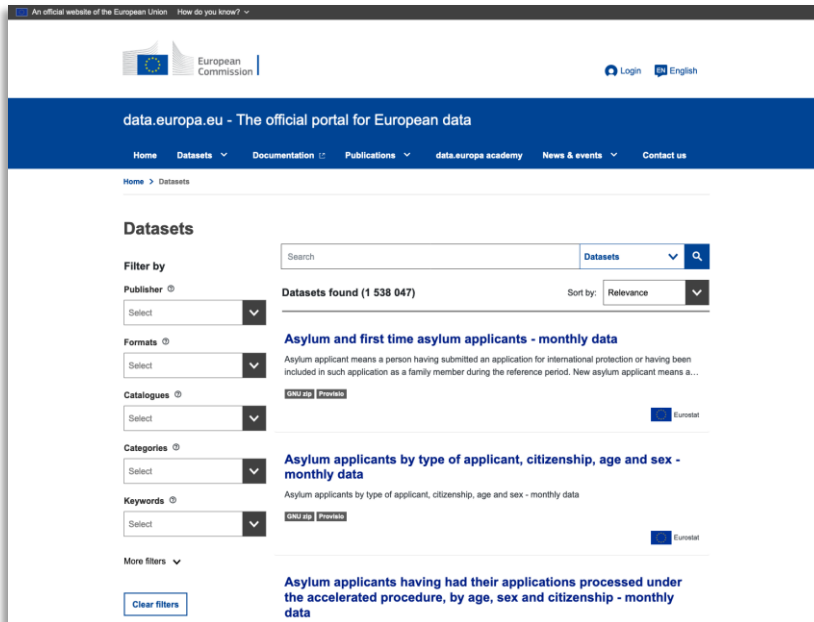
- Which country does your NSIP cover?
- Is your NSIP already being harvested by another portal?
- What is the URL to the homepage of the NSIP?
- What is the default language of the catalogues' datasets?
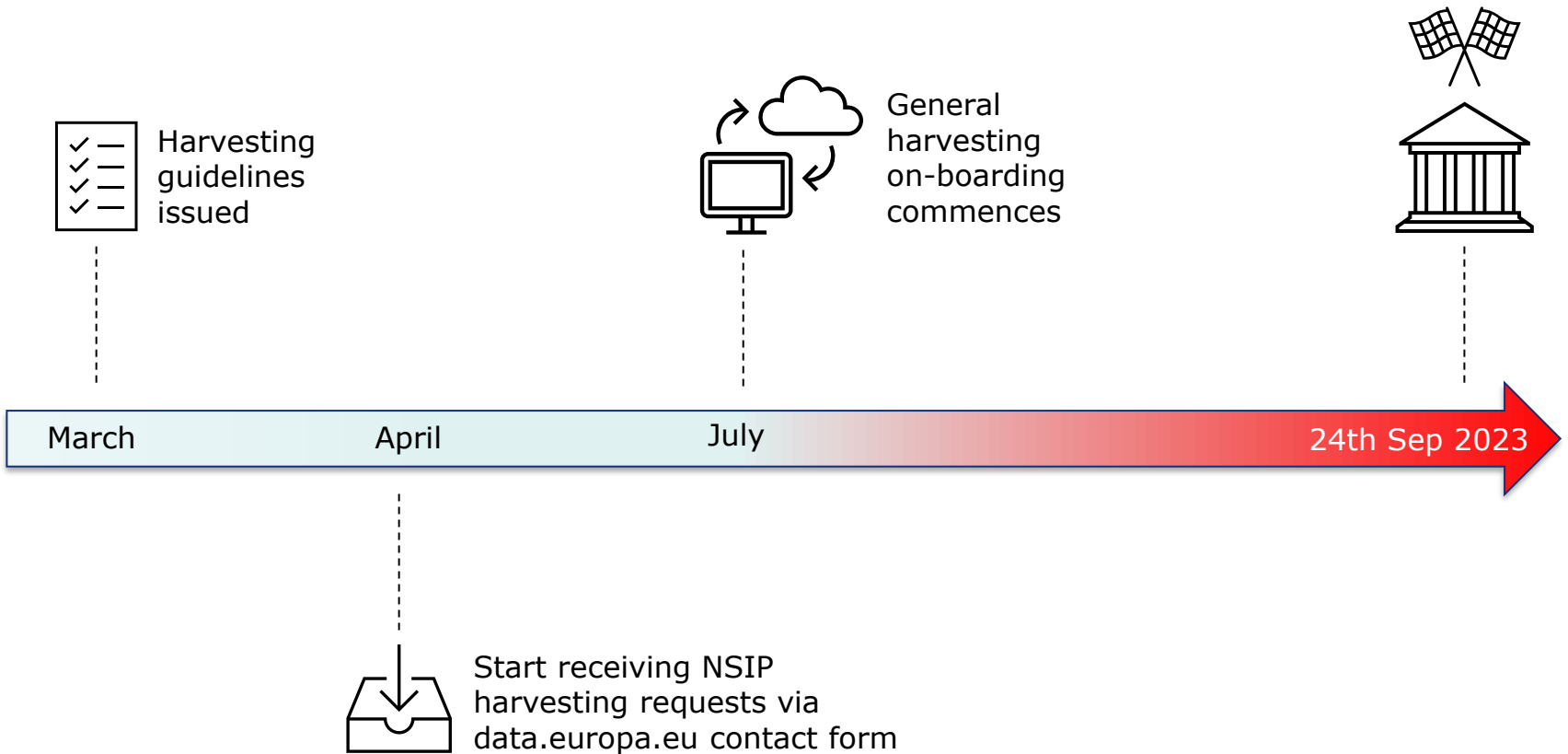
Next steps: Roadmap and objectives

Implementation objectives

- ESAP and ESAP Country Catalogues as additional facets on data.europa.eu search.

- An implementation of the ESAP that blends with the existing data.europa.eu UI.

- Users can opt to search either in the ESAP or in individual ESAP Country catalogues.

- Users can search both for catalogues and datasets.

Roadmap

Harvesting guidelines issued

General harvesting on-boarding commences

March   April   July   24th Sep 2023

Start receiving NSIP harvesting requests via data.europa.eu contact form

Q&A regarding harvesting guidelines

For further questions and feedback, please contact us at:

CNECT-G1@ec.europa.eu