Report on Data Homogenisation for High-value Datasets

Contents

This study has been prepared as part of data.europa.eu. Data.europa.eu is an initiative of the European Commission. The Publications Office of the European Union is responsible for contract management of data.europa.eu.

For more information about this paper, please contact:

Disclaimer

Executive summary

Following the European Union's open data policies, Directive 2019/1024 (EU), also known as the open data directive, defines the concept of high-value datasets (HVDs) as datasets whose reuse is associated with significant benefits for society and the economy that public sector bodies will have to make available under very reuse friendly conditions. On 21 December 2022, the EU adopted an Implementing regulation lying down a list of specific high-value datasets in each of the six main categories identified in the directive: geospatial, Earth observation and environment, meteorological, statistics, company and company ownership, and mobility. The Regulation must be implemented by June 2024. Starting in February 2025, Member States will have to report to the Commission every two years on specific high-value datasets available at the Member State level, including links to the licensing conditions and the APIs. Where available, Member States should also report on their guidance documents, data protection impact assessments, and possible temporary exemptions from the no-charging rule. Effectively used interoperable metadata will play a crucial role in the findability of the datasets by re-users and for easy reporting by Member States.

The identification of these specific HVDs is a relevant step towards improving cross-border interoperability and dataset reuse. However, the high level of heterogeneity in the data formats, structures and levels of granularity in which these datasets are available, the lack of good-quality metadata about datasets (including metadata that allows one to determine whether a dataset can be considered a HVD or not) and the lack of interoperability of published datasets means that companies, entrepreneurs, researchers, data journalists, policymakers and citizens will still find it difficult to make use of these important resources, especially when using data from several countries.

In this study, we propose a methodological approach that facilitates the identification and homogenisation of HVDs. We implement this approach using selected examples of HVDs from three distinct categories, showcasing its adaptability for different types of HVDs.

Following desk research and an interactive workshop held in May 2023, featuring representatives from European countries, a set of challenges and corresponding recommendations has emerged. These insights provide a roadmap for homogenising datasets across European countries and enhancing interoperability. The key findings are as follows:

- There is a need to raise awareness about the concept and significance of HVDs at the administrative level and among all stakeholders involved in the ecosystem.
- In certain domains, such as geospatial, a lack of interoperability and homogenisation exists between metadata standards, such as the EU's infrastructure for spatial information (Inspire) and the Data Catalogue Vocabulary Application Profile (DCAT-AP). This should be addressed.
- Additional efforts are required to encourage developing and utilising controlled vocabularies and ontologies for data publication. This should involve leveraging existing ones available in EU vocabularies and reinforcing other related initiatives.

- In alignment with the open data directive, application programming interfaces (APIs) must be made available and recognised licences should be used as part of the HVD metadata.
- Further detailed guidelines should be established for the comprehensive description of High-Value Datasets, building upon the existing activities associated to Data Catalogue Vocabulary Application Profile (DCAT-AP).

1. Introduction and motivation

Directive (EU) 2019/1024 of the European Parliament and of the Council on open data and the reuse of public sector information was published in June 2019, establishing a framework for the reuse of public sector information, such as geographical, statistical or legal information held by public sector bodies, and publicly funded research data. It is in this directive that the idea was introduced of establishing conditions to facilitate reuse of certain datasets associated with important socioeconomic benefits that have a particularly high value for the economy and society.

On 21 December 2022, the European Commission published Commission Implementing Regulation (EU) 2023/138 laying down a list of specific HVDs and the arrangements for their publication and reuse. Within this implementing regulation, the Commission defines a series of conditions that must be fulfilled by the Member States in the publication of these datasets to harmonise their reuse. It also identifies the concrete datasets in the six categories of HVDs identified in the directive (geospatial, Earth observation and environment, meteorological, statistics, companies and company ownership, and mobility) and the conditions that must be fulfilled when making them available. In addition, it indicates that by February 2025 and every two years after that all Member States must provide the Commission with a report listing the specific datasets that meet the conditions set out in the implementing regulation, including their metadata, which must follow an open standard, applicable licences, links to the APIs ensuring access to the high-value datasets. Where available, Member States should also report on their guidance documents, data protection impact assessments and on possible temporary exemptions from the no-charging rule. In today's increasingly connected world, the availability and accessibility of HVDs is of vital importance in effective decision-making, research and innovation. However, the heterogeneity of data formats, structures and semantics across borders poses significant challenges for data interoperability and the efficient use of these valuable resources. This is where the process of homogenisation takes a key role. Homogenisation in this context would be the set of processes and actions whose ultimate goal is to standardise and harmonise datasets in order to enable greater interoperability between datasets and between applications using them. The homogenisation of HVDs in this report is seen as a crucial step towards realising a good practice, such as 'write your application code once and run it everywhere', in a cross-border context among Member States.

To address the need for data homogenisation, various initiatives and organisations have been actively working towards the development of data models, APIs and (meta)data standards (e.g., DCAT-AP ([1])) that promote interoperability between datasets. One prominent initiative is the INSPIRE directive, which seeks to establish a European spatial data infrastructure by harmonising and standardising geospatial information for the purposes of EU environmental policies. Inspire addresses 34 spatial data themes needed for environmental applications and provides a common framework for interoperability, facilitating data sharing and integration across national borders. In a more reduced scope, another remarkable data homogenisation initiative is the smart data models initiative from the Fiware Foundation, which aims to standardise and

---

[1]() https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe

harmonise data models for smart cities, the internet of things (IoT) and industrial applications. It provides a common framework for structuring and sharing data, promoting interoperability and simplifying development. The models cover various domains, enabling easy integration and reusability of data. Another important initiative to mention is the [open cities project](#), which advocates for open and standardised city data, enabling cities to share and exchange data in a uniform manner.

A number of important studies on HVDs have been conducted in recent years. For example, the report *High-value Datasets: Understanding the data provider perspective* (Publications Office et al., 2020) aims to understand the perspective of data providers in publishing HVDs. The impact assessment study on the list of HVDs (European Commission et al., 2020), from the European Commission Directorate-General for Communications Networks, Content and Technology, analyses the possible impacts that interventions can have at the national level. These reports have served as inspiration for the work that is presented in this document, especially in terms of the methodology that has been followed for the selection and analysis of HVDs. However, none of them address aspects related to the challenges and opportunities that a higher homogenisation of data can bring, as we do in this report.

Traditionally, two approaches have been pursued for homogenisation within the broader context of data governance. The first approach involves working with ontologies and reference data (commonly represented using thesauri and authority tables). Ontologies provide a structured representation of knowledge, enabling semantic interoperability by establishing common concepts and relationships. Additionally, the use of SKOS (simple knowledge organisation system) thesauri, or authority tables also contributes to this effort by ensuring consistent and controlled vocabularies. A second complimentary approach focuses on developing common APIs for data access and utilisation. By defining standardised APIs, data providers can expose their HVD to external parties, facilitating integration and utilisation. This approach aligns with the open data directive, which emphasises the publication of HVDs as APIs to promote openness and accessibility.

Given that the publication of the implementing regulation is still very recent, and that Member States are still working on the identification and creation of inventories for HVDs, within this report we only mention clearly identified examples of HVDs from some Member States. By exploring the initiatives mentioned above, along with the management approaches adopted by different Member States, we can gain valuable insight into the challenges and opportunities associated with HVD homogenisation. With this report, we aim to contribute to the advancement of data interoperability and foster the effective utilisation of HVDs in cross-border contexts. To do this, first we talk about the methodology that we propose to facilitate the identification and homogenisation of HVDs, then we apply this methodology to some sample HVDs, and finally we mention a series of recommendations that facilitate homogenisation.

2. A method for high-value dataset identification and homogenisation

In this section we describe the method that we have used for the identification and the proposal for the homogenisation of HVDs published in the open data portals of Member States, in the open data portals of some of their associated governmental agencies and in the official portal for European data (data.europa.eu). This method consists of a crucial first step, which is the identification of HVDs in these data portals, a second step which consists of finding or developing common data models, ontologies, controlled vocabularies and/or APIs for achieving such homogenisation in the future, and a third step where such data models, ontologies, controlled vocabularies and/or APIs are applied.

2.1. Step 1. Dataset identification

The identification of HVDs among the long list of datasets commonly published by Member States (MSs) in the data.europa.eu portal or in their own open data portals (or in the data portals of their governmental agencies, e.g., statistical offices and geographical institutes) is a challenging and error-prone task. There are some initial agreements on how to annotate or provide metadata (e.g., by extending metadata models such as the DCAT) for those datasets, so that they are easier to identify. In the context of the HVD implementing regulation, the Semantic Interoperability Centre Europe (SEMIC) team, in collaboration with the Directorate-General for Communications Networks, Content and Technology, has prepared usage guidelines on how to use the DCAT-AP for HVDs. This document provides guidelines on how to use the DCAT-AP for a dataset that is subject to the requirements imposed by the implementing regulation. However, these agreements are not yet implemented.

During an interactive workshop in May 2023, several European countries (ECs) representatives indicated that it would be favourable to have an official methodology indicating how to label these datasets. If European countries provide this additional piece of metadata, then Sparql queries may be evaluated on the data.europa.eu Sparql endpoint to identify them more easily. For example, some of them suggested creating specific inventories for HVDs, as is already being done in some European countries. Another option that some European countries are exploring (e.g., Poland) is to add a label, filter or keyword to HVDs published in their national open data portals (see Annex 1). Although this is for now a tool that makes it easier to locate which datasets are considered as HVDs by the European countries, it should also be taken into account that it may favour heterogeneity, since there is still no standard that indicates how these datasets should be named/labelled. The only aspect mentioned in in Article 3 of the implementing regulation is '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'.

In our report, we have decided not to wait for the stabilisation and wide application of such a set of recommendations on metadata items for the task of dataset discovery, and instead work with the current status quo of metadata provision by Member States and data.europa.eu. We have proposed and applied a protocol for the identification of datasets that can be partially automated. As a side result, this may show

the difficulties that people who reuse data may have in understanding which HVDs are available in these data portals, with the current status quo.

We start with the information contained in Annex I to Commission Implementing Regulation (EU) 2023/138, which lists a number of datasets according to the classification of the aforementioned six categories of HVD. In the 'Datasets in scope' section, the document provides a definition of the datasets that would fall into each category, along with a list of attributes associated with each dataset. For example, for datasets in the companies and company ownership category, the information that can be found in this annex is indicated in Table 1.

Table 1. List of attributes provided by the regulation for the identification of HVDs of the category of companies and company ownership.

Datasets	Basic company information	Company documents and accounts
		The descriptions of the datasets and their key attributes shall be understood in accordance with Articles 4 and 5, Articles 9 to 18, Article 19, point (a), Article 24, Article 28, Article 29, point (a), Articles 31, 35, 36, 39, 40 and 43, and Article 48, point (c) of Directive 2013/34/EU of the European Parliament and of the Council and Articles 4 to 6 of Directive 2004/109/EC of the European Parliament and of the Council.
	● Name of the company (full version; alternative names when applicable); ● company status (such as when it is closed, struck off the register, wound up, dissolved (as well as the date of these events), economically active or inactive as defined in national law); ● registration date; ● registered office address; ● legal form; ● registration number; ● Member State where the company is registered; ● activity/activities that are the object of the company, such as the NACE code.	Accounting documents, which include: — financial statements (including the list of participating interests, subsidiary undertakings and associated undertakings, their registered office address and proportion of capital held), audit reports, — non-financial statements, management reports and other statements or reports, — annual financial reports.

Taking this information into account, we have proposed a protocol for dataset identification, which we have tested with datasets in the companies and company ownership, statistics (datasets on population and

national accounts) and mobility categories. The testing of this protocol has been part of the field work done in preparation for the aforementioned interactive workshop.

The first step in the protocol consists of searching for HVDs in the data.europa.eu portal by **using a list of keywords that are based on the attributes** included in the aforementioned Annex I (e.g., those in Table 1). These keywords have been used both in English and in the official language of the Member State, so as to maximise recall. Our search results showed that this keyword-based approach generated too much noise in the search results, resulting in difficulties in locating the corresponding HVDs, if they were published.

Taking this into account, we supplemented this type of search with a **manual search for these datasets in other data portals** (not data.europa.eu, nor the national data portals that are harvested by data.europa.eu), considering the fact that these are more specific to the type of data that we are looking for. The data portals used to identify HVDs of the categories mentioned above are listed below:

- **Companies and company ownership category**
  - o **The European e-Justice Portal.** The European e-Justice Portal was used since it provides information on how each Member State publishes company records, and company information for each Member State. Within this information, it is often indicated in which data portal or page this data is published and under what conditions and restrictions.
- **Statistics category**
  - o **National statistical institutes and offices**. Eurostat publishes and updates a list of national statistical institutes and other national authorities which provide data portals and catalogues containing statistical datasets. Within these portals, it is common to find datasets in themes that match the datasets described in the statistics category in the implementing regulation.

  - o **Eurostat.** The statistical office of the European Union publishes a database with official statistical data at the level of each Member State and at the level of Europe in general. The data published by Eurostat complies with many of the requirements mentioned in the regulation. Each dataset has associated metadata in the Euro SDMX (statistical data and metadata exchange) metadata structure (European Commission et al., 2021). In addition, Eurostat provides a series of resources to ensure the quality of the published data.
- **Mobility category**
  - o **National geospatial data portals.** Many of the Member States have specific portals dedicated to the publication of geospatial data and Earth observation and environmental data. To find these data portals, data.europa.eu itself provides a catalogue of data portals for each Member State, where these geospatial portals are easily found. It is quite common for MSs to also publish their mobility data in these kinds of portals.
  - o **INSPIRE Geoportal.** As indicated in the regulation, the HVDs of the Earth observation and environment, geospatial and mobility categories include the datasets that are within the scope of Inspire data themes. Several of the Member States publish their high-value geospatial, mobility and environmental data following these Inspire standards and procedures. Therefore, in order to promote data harmonisation, these datasets are located on the INSPIRE Geoportal. Within this portal, we can search for these datasets by category in the section 'Inspire Thematic Viewer'.

This manual approach gave better results in identifying HVDs. However, we still encountered many difficulties and obstacles in identifying which of the datasets published by the MSs can be considered as HVDs.

As a summary of our proposed method to identify HVDs in existing data portals, we can provide the following set of points.

1) Obtain keywords from the definitions and attributes of HVDs that appear in Annex I to the implementing regulation, and from the corresponding underlying legal acts for those datasets, and use those keywords (in English and in the local language of each Member State) to search for datasets in the data.europa.eu portal.
2) Search for datasets in specialised sectoral data portals that are normally maintained by governmental agencies. Some examples of relevant data portals are national statistical data portals, Eurostat, national geospatial data portals, the Inspire Geoportal and the European e-Justice Portal.
3) Identify whether there are advanced search functionalities that allow adding special filters for HVDs, including their categories or not.

Of course, the results of all these steps may always be refined further with the participation of European countries representatives, as was the case in the aforementioned interactive workshop in May 2023, which enabled the correction of existing HVDs and identification of new ones from the original list. This is the process that was followed in previous reports that have been already referenced, for example, in the study *Impact assessment study on the list of high value datasets to be made available by the Member States under the PSI directive* (European Commission, 2020), where a series of interviews were conducted with the parties interested in using these datasets.

The following sections present examples of searches of HVD following the method just explained for those categories.

A) **Companies and company ownership**

Regarding this category, the annex establishes that 'The companies and company ownership thematic category includes datasets containing basic company information and company documents and accounts at individual company level and with the key attributes listed in the table below (see Table 1). As we have mentioned before, we based our search of the HVDs on three key elements.

- **List of keywords**. For this search, we have focused on HVDs about a company's basic information. Keeping in mind the list of attributes provided, Table 2 presents the list of keywords employed for the search.
- **Legal basis.** Key elements that allows a user to identify HVDs and classify a dataset as one are the guidelines and legal basis mentioned in the annex. Regarding this category, Directive 2013/34/EU of the European Parliament and of the Council and the Directive 2004/109/EC of the European Parliament and of the Council are mentioned. However, these are mentioned for the datasets that address company documents and accounts, so these are not applicable for the datasets we looked for.
- **Data portals.** As mentioned above, an essential part of the search is knowing where this type of data can be found in data portals. In this case, besides data.europa.eu, the HVDs of company
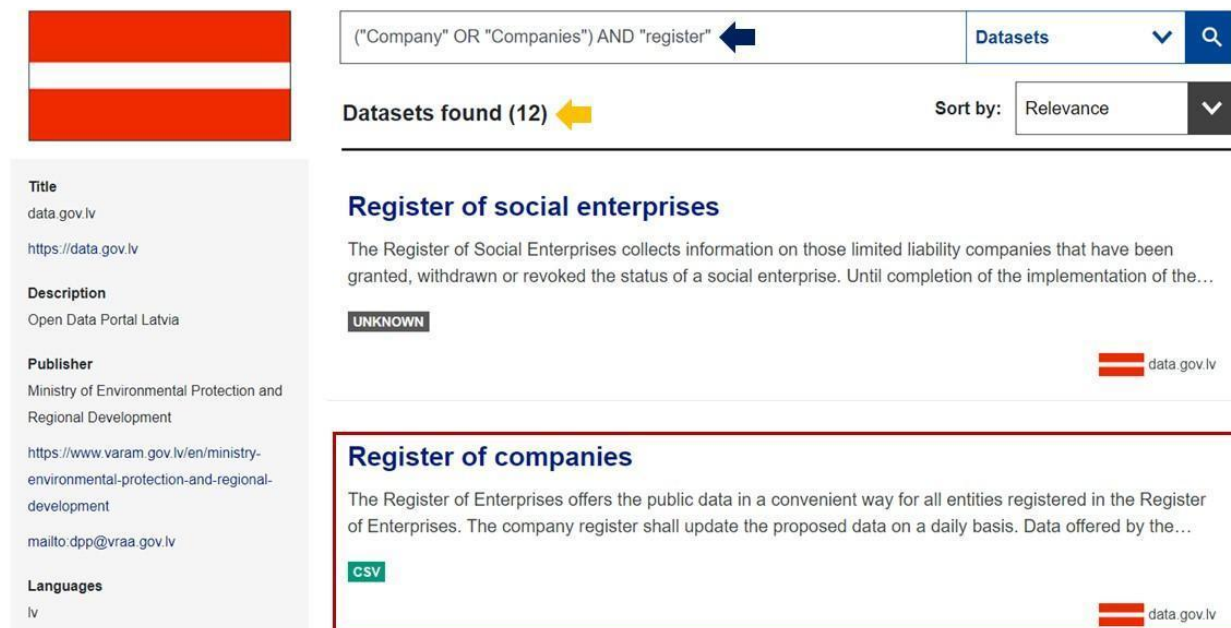
information have been found in data portals that contain company records, and for this the European e-Justice Portal is a very useful resource.

Table 2. List of attributes provided by the regulation for the identification of HVDs of the companies and company ownership category versus the keywords which were deduced from them and employed in the search.

Attributes from the annex	Search keywords
- Name of the company (full version; alternative names when applicable); - company status (such as when it is closed, struck off the register, wound up, dissolved (as well as the date of these events), economically active or inactive as defined in national law); - registration date; - registered office address; - legal form; - registration number; - Member State where the company is registered; - activity/activities that are the object of the company, such as the NACE code.	('Company' OR 'Companies') AND 'register'
	('Company' OR 'Companies') AND 'registration'
	('Business' OR 'Businesses') AND 'registration'
	('Business' OR 'Businesses') AND 'register'
	('Company' OR 'Companies') AND 'information'
	('Business' OR 'Businesses') AND 'information'

Testing the keywords step in the open data catalogue of Latvia on data.europa.eu, we can see that with the keyword used it returns a reasonable number of results (Figure 1), and that also among the first results we can find the possible HVD. On the other hand, when using the keywords in the Estonian data portal, the datasets that interest us do not appear. Therefore, we proceed to identify where this dataset may be published using the European e-Justice Portal. In it we can see that it is located in the e-Business Register on the website of the Centre of Registers and Information Systems of Estonia (https://ariregister.rik.ee/eng), which has an open data section allowing access to the dataset (https://avaandmed.ariregister.rik.ee/en/downloading-open-data).

Figure 1. Results from the Latvian open data catalogue in the data.europa.eu portal after employing keywords.

B) Statistics

This category covers a great diversity of datasets, which is why we wanted to focus on the population HVDs. For these datasets, the annex establishes the legal acts on which they are based in order to define the variables to be considered (see Figure 2). Taking this information and the variables into account, we define the list of keywords for the search, shown in Table 3. As with the HVD from the companies and company ownership category, some Member States with the usage of these keywords identified the potential HVDs (see Figure 3). However, the search with keywords introduced a lot of noise and it was not feasible to find HVDs. We therefore proceeded to the second step, which consisted of searching manually in the national statistical portals. For example, Statistics Norway (https://www.ssb.no/en) allows users to search, visualise and download the statistical datasets by category.

A

Population, Fertility, Mortality	Regulation (EU) No 1260/2013 of the European Parliament and of the Council ([13])
	Annex II to Commission Implementing Regulation (EU) No 205/2014 ([14])
	Regulation (EC) No 862/2007 of the European Parliament and of the Council ([15])
	Commission Regulation (EU) No 351/2010 ([16])
	Regulation (EU) No 1260/2013
	Implementing Regulation (EU) No 205/2014
	Regulation (EU) No 2020/851 of the European Parliament and of the Council ([17])
Population (see Table 3 below for variables in scope)	Regulation (EU) No 1260/2013
	Implementing Regulation (EU) No 205/2014
	Regulation (EC) No 862/2007
	Regulation (EU) No 351/2010

Specification for the yearly population high-value dataset

B

Note: The breakdowns simultaneously indicated in a column should be offered for the crosstabulation of all variables involved.

Legend for breakdowns: ● = mandatory, ● = required for Member States that meet conditions set out in the relevant Regulation; ○ = voluntary

Breakdowns	Key variables	Population on 1 January					Median age	Old age dependency ratio — Proportion of people 65+ to people 20-64
Sex		●	●	●	●	●	●	
Age		●	●	●	●	●		
Education attainment level	ISCED 2011 ([2])					○		
Marital status					○			
Citizenship			●					
Country of birth				●				
Human development index	HDI is a regrouping of country of birth and of country of citizenship				●	●		
Region	NUTS 3	●					●	●

Figure 2. A) Legal acts laying down the key variables of the population datasets and their breakdowns. **B)** The variables and their breakdowns for HVDs of population.

Table 3. List of keywords employed for searching HVDs on population

Search keywords
'population'
'population' AND ('sex' AND 'age')
'population' AND ('sex' AND 'age' AND 'region')
'population' AND ('sex' AND 'age' AND 'country of birth')

Figure 3. Results from the Luxembourg open data portal in the data.europa.eu portal after employing keywords.

C) **Mobility:** In this category the annex establishes that it includes *datasets within the scope of the Inspire data theme 'Transport networks' as set out in Annex I to Directive 2007/2/EC, at all levels of generalisation available up to the scale of 1:5 000 covering the entire Member State when combined. If datasets are not available at the scale of 1:5 000, but are available at higher spatial resolution(s), they shall be provided at the available spatial resolution. The datasets include as key attributes national identification code, geographical position as well as links with cross-border networks, where available.* In addition, *for those Member States to which Directive 2005/44/EC of the European Parliament and of the Council applies, this category also includes the datasets in Figure 4; the datasets are understood as described in Directive 2005/44/EC.*

For this category we focused on datasets within the scope of the Inspire data theme 'Transport networks'. To identify a list of keywords to search for these datasets in the data.europa.eu portal we search in the Inspire documentation about the data theme (https://Inspire.ec.europa.eu/theme/tn). The list of keywords employed are presented in Table 4. Using these keywords in the data.europa.eu portal allowed us to find HVDs, for example in the open data portal of the Netherlands (see Figure 5). However, as in the other categories, the use of keywords is sometimes insufficient to find the datasets. In those cases, we proceeded to perform a manual search of the datasets in the Inspire Geoportal. This portal allows us to search by data theme category and by Member State.

Inland waterways datasets	Data type
— Fairway characteristics — Long-time obstructions in the fairway and reliability — Rates of waterway infrastructure charges — Other physical limitations on waterways — Regular lock and bridge operating times — Location and characteristics of ports and transhipment sites — List of navigation aids and traffic signs — Navigation rules and recommendations	Static data
— Water depths contours in the navigation channel — Temporary obstructions in the fairway — Present and future water levels at gauges — State of the rivers, canals, locks and bridges — Restrictions caused by flood and ice — Short term changes of lock and bridge operating times — Short term changes of aids to navigation	Dynamic data
— Waterway axis with kilometres indication — Links to the external xml-files with operation times of restricting structures — Location of ports and transhipment sites — Reference data for water level gauges relevant to navigation — Bank of waterway at mean water level — Shoreline construction — Contours of locks and dams — Boundaries of the fairway/navigation channel — Isolated dangers in the fairway/navigation channel under and above water — Official aids-to-navigation (e.g. buoys, beacons, lights, notice marks)	Inland electronic and navigational charts (Inland ENC according to the Inland ECDIS Standard)

Figure 4. List of HVDs that are included in the mobility category for those Member States that apply Directive 2005/44/EC of the European Parliament and of the Council.

Table 4. List of keywords employed for searching HVDs on mobility

Keywords
'Transport networks' OR ('Transport networks' AND 'Inspire')
'Road transport networks' OR ('Roads networks' AND 'Inspire')
'Rail transport networks' OR ('Rail transport networks' AND 'Inspire')
'Air transport networks' OR ('Air transport network' AND 'Inspire')

Figure 5. Results from the open data portal of the Netherlands in the data.europa.eu portal after employing keywords.

2.2. Step 2. Identification or development of common data models, controlled vocabularies, ontologies and APIs to promote interoperability

As we have already emphasised in the previous step with the use of the DCAT-AP for the metadata used to describe datasets, there is also an opportunity for the homogenisation of the content of datasets, so as to promote more dataset reuse and increase interoperability across Member States. This can be done using controlled vocabularies, ontologies and common data models, along with standardised APIs, which provide guidelines about the structure of the datasets and about the content limitations for some of the dataset attributes.

The implementing regulation focuses on describing the main characteristics of HVDs and some of the recommended attributes. The descriptions of the exact data structures to use or how the data should be entered in each of the identified datasets is out of the scope, but they are important for interoperability and homogenisation.

In this section we describe some useful resources and resource centres where common data models, controlled vocabularies, ontologies and APIs can be found, for the six different dataset categories identified in the directive. The lists that are provided in the following sections are not meant to be exhaustive, since there may be other relevant resources and resource centres (general-purpose or specialised) that may be consulted to locate these types of resources.

As regards general-purpose resources of importance across all domains, we can point to the EU Vocabularies (2) website, which contains a large quantity of EU reference data that allows for homogenisation. All EU Vocabularies lists are multilingual and available in multiple formats, including RDF, and many alignments also exist between the thesauri.

2.2.1. Resources and resource centres with common data models, controlled vocabularies, ontologies and APIs for geospatial data

In the geospatial category, datasets that fall within the scope of the **Inspire** categories, such as administrative units, geographical names, addresses, buildings and cadastral parcels, are identified as HVDs. This is why the Inspire data model can be considered as a key data model that allows for the harmonisation of data among Member States. Inspire provides a series of resources that promote and facilitate data harmonisation. We will highlight the following.

- **Inspire data specifications**. Inspire defines data specifications that establish data models, schemas and encoding rules for the different thematic areas in order to ensure consistency and interoperability across diverse spatial datasets. The technical guides where the data model for each of the categories defined by Inspire is explained in detail is available at https://Inspire.ec.europa.eu/data-specifications/2892.
- **Inspire network services**. Inspire defines a set of common interfaces for web services that enable the discovery, viewing, download and transformation of spatial data. The technical guidance documents for the implementation of these services can be found at: https://Inspire.ec.europa.eu/Technical-Guidelines2/Network-Services/41.
- **Inspire technical guidelines for metadata**. The regulation establishes that those datasets that fall within the scope of Inspire must contain at least the metadata elements defined in the Commission Regulation (EC) No 1205/2008. Inspire provides technical guidelines (2) to facilitate and guide the creation of this metadata. There is also a mapping between Inspire metadata and the Geo-DCAT-AP (3), so as to make datasets discoverable on national cross-domain data portals.

In addition to these resources, Inspire provides other useful resources, such as tutorials and good practice guides, together with a large community involved in promoting the harmonisation of geospatial data in Europe. Furthermore, the Inspire Geoportal must be mentioned, since it serves as a central access point for discovering and accessing the geospatial data and services of Member States. It allows users to search, visualise and download data and (meta)data based on themes, keywords and geographic locations, thereby promoting data discovery and exchange.

From the point of view of the SEMIC Support Centre, the **Core Location Vocabulary** data model also falls under this HVD category, and can be also helpful in promoting greater homogenisation. This is a simplified, reusable and extensible data model that captures the fundamental characteristics of a location, represented as an address or geographic name or with geometry.

Finally, another very important resource for datasets in this category would be the **General Multilingual Environmental Thesaurus (GEMET)**, a controlled vocabulary specialising in environmental information. In

2() https://op.europa.eu/en/web/eu-vocabularies; https://Inspire.ec.europa.eu/Technical-Guidelines2/Metadata/6541

3() https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/geodcat-application-profile-data-portals-europe

fact, GEMET has a section of concepts belonging to Inspire spatial data themes. It is highly recommended to use this type of thesaurus, for example, when linking keywords to datasets.

2.2.2. Resources and resource centres with common data models, controlled vocabularies, ontologies and APIs for earth observation and environmental data

As with geospatial data, many of the HVDs that fall under this category fall within the scope of the Inspire categories. That is why the use of the aforementioned Inspire resources may be considered essential to promote the harmonisation of these datasets. As with geospatial HVDs, the use of the GEMET thesaurus is highly recommended, as it may improve terminological consistency and facilitate interoperability.

2.2.3. Resources and resource centres with common data models, controlled vocabularies, ontologies and APIs for meteorological data

Within this category there are several types of datasets: those with measurements made by weather stations, validated observations (climate data), weather alerts, radar data and numerical weather prediction models.

In Inspire there is a data specification on atmospheric conditions and meteorological geographical features, which is in fact recommended by the regulation itself, to represent and document these meteorological datasets.

Another relevant resource that can be used for data homogenisation is the **Semantic Sensor Network ontology** (Haller et al., 2017), a W3C recommendation that allows us to describe sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, the observed properties and the actuators.

In terms of controlled vocabularies or other types of ontologies, we may also mention the potential use of the **quantity, unit, dimension and type (QUDT) ontology** to provide units of measurement that can be used for the values obtained in these observations.

2.2.4. Resources and resource centres with common data models, controlled vocabularies, ontologies and APIs for statistical data

When it comes to promoting the harmonisation of statistical datasets, **Eurostat** has created and developed a series of highly important resources over the years. Among these resources, we find the following.

- A list of statistical classifications available as linked open data in XKOS (extended knowledge organisation system) (Cyganiak et al., 2019), the SKOS extension for modelling statistical classifications. These classifications are a structured set of well-described categories, which are normally represented in a hierarchy with numeric or alphabetic codes assigned to them. These classifications are divided into families characterised by domains and subdomains. Some of these domains and subdomains are:
  o business statistics: Classification of Business Functions;
  o economic statistics: statistical classification of economic activities in the European Community (NACE), standard goods classification for transport statistics (NST), LUCAS (Land-Use / Cover Area Frame Survey), classification of environmental protection activities and expenditure (CEPA), etc.;

        o   regional and geospatial statistical information: NUTS (nomenclature of territorial units for statistics) and SR (statistical regions).
- [Standard code lists](), which are predefined, organise sets of items that present statistical concepts. These are used in the dimensions of the datasets to define unique concepts with unique codes.

The statistical data and metadata exchange ([SDMX]()) is an international standard (ISO standard) for the exchange and sharing of statistical data and metadata. It provides a framework and a set of rules for structuring, describing and transmitting statistical information in a standardised format. It consists of technical standards (the SDMX information model), guidelines (content-oriented guidelines), IT architecture and tools, and a series of tutorials to help the users. Another important resource would be the **RDF Data Cube Vocabulary** (Cyganiak et al., 2014), an ontology that allows us to describe multidimensional data (such as statistics). This vocabulary builds upon the core of the SDMX 2.0 information model.

2.2.5. Resources and resource centres with common data models, controlled vocabularies, ontologies and APIs for companies and company ownership data

Mentioned by the regulation itself, the **Core Business Vocabulary** developed by SEMIC is an indispensable resource. It consists of a simplified, reusable and extensible data model that captures the fundamental features of a legal entity (their legal name, activity and address, for example). It includes a set of reusable and well-defined classes, properties and relationships that describe business entities and their attributes. It is a general vocabulary, not limited to a specific industry or domain.

Another important element mentioned by the regulation itself is the use of the **NACE code** to represent the companies' activity/activities. In October 2022 the European Commission published the updated version, the statistical classification of economic activities [NACE Revision 2]() so that the different categories of classification of activities in the EU are interpreted in a uniform way in all Member States.

Additionally, the regulation emphasises that all MSs are encouraged to go beyond the minimum requirements established in it. For this reason, we also want to mention other resources that favour the interoperability of these datasets, such as **W3C's Organization ontology** (Reynolds and Epimorphics Ltd, 2016). This resource is aimed at supporting linked data publishing of organisational information across a number of domains by adding the classification of organisations and roles, along with extensions to support neighbouring information such as organisational activities.

Finally, another important resource is the **Global Legal Entity Identifier Foundation**, a centralised database that contains information about legal entities participating in global financial markets. It assigns to each entity a unique Legal Entity Identifier (LEI) code which serves as a globally recognised identifier.

2.2.6. Resources and resource centres with common data models, controlled vocabularies, ontologies and APIs for mobility data

As with the geospatial, earth observation and environment categories, within the HVDs of the mobility category are those that match the scope of the Inspire 'Transport networks' [data theme](). In addition, it is also advisable to use GEMET if possible when defining the keywords for the metadata by example.

Another resource for HVDs in this category is the **NST (2007) taxonomy**. This is a classification system used to categorise goods transported via four main modes of transport: road, rail, inland waterways and sea. This classification takes into account the economic activity associated with the origin of the goods.

Finally, we would like to mention the '**Transport service**' table of authorities (also known as a named authority list) provided by the EU Vocabularies section of the Publications Office. This table provides a list of codes for the different types of transport services.

2.3. Step 3. Application of the common data models, controlled vocabularies, ontologies and APIs to existing datasets

The obvious final step in this process is the actual harmonisation of the data that will be published in national or specialised data portals, as well as in the data.europa.eu portal, or that are going to be made available via APIs. The common data models, controlled vocabularies, ontologies and/or APIs that have been selected in the previous step can hence be used to structure the datasets or to determine the content to be used for some of the values applicable to some attributes (e.g., by using controlled lists of terms from the controlled vocabularies).

The way in which these transformations can be applied will differ depending on the final foreseen output, and the type of common data model, controlled vocabulary, ontology or API that is selected for this purpose. It may fall into any of these major categories (the list is non-exhaustive, but sufficiently representative of common practices).

- The transformation of original tabular data sources (comma-separated values (CSV), Excel, relational databases) from the data providers into other tabular data sources that follow the structure provided in the common data models.
- The transformation of the original tabular data sources into tree-based representations (e.g., JavaScript Object Notation (JSON)) according to the selected common data models or APIs.

- The transformation of the original data sources (tabular or non-tabular) into RDF-based representations according to the ontologies and controlled vocabularies that are selected, following good practices such as the ones described by **W3C for government linked data**.

3. Evaluating homogeneity and heterogeneity in high-value datasets: an analysis of key criteria

In order to identify first-hand what are the main challenges faced by Member States in applying the implementing regulation, we have applied the method described in Section 2 to a set of datasets from three of the identified categories (companies and company ownership, statistics and mobility).

First, a keyword-based search was performed in the data.europa.eu portal, as proposed in our method. The results obtained were not as successful as originally expected, since there were many datasets that were wrongly retrieved (false positives), that is, they could not be considered to fall into any of the categories of HVD. After this step, a manual search was conducted in more specific data portals for each category, as discussed in our methodological section. For example, the HVDs of the statistics category were found mostly in the portals of national statistical institutes, while the HVDs belonging to the mobility category were located mostly in the Inspire Geoportal.

This method for HVD search and identification is far from being the most effective one since, on the one hand, even if translations into English are provided in the data.europa.eu portal, some content may be better searched for in local languages, which adds some additional challenges. On the other hand, another of the great disadvantages of this method is that it prevents it from being easily replicated to facilitate future searches for these HVDs. There is a remarkable lack of automation in the search for these HVDs, which points to the great need to standardise the labelling process. Faced with this situation, as mentioned above, the most optimal solution would be to use a standard tagging element that could be applicable for the publication of HVDs across all MSs. In this case, the most recommendable would be the use of DCAT-AP high-value datasets to create the metadata associated with each dataset, since it allows a dataset to be labelled as an HVD in a standard way and also allows us to indicate to which category it belongs. However, this adaptation of the DCAT-AP to HVDs is still very recent, which is why, for the most part, the MSs have not had the opportunity to implement it.

That is why, at this point, another essential element for the identification of these HVDs was the interaction with data providers from different European countries in the dedicated workshop. In this event the representatives of different European countries indicated what the situation of each participating country was regarding the implementation of HVDs and what the main challenges they faced were. Among the results collected (see Annex 1), a series of European countries were identified that have already created a kind of filter or label in their open data portals that allows for faster identification of HVDs (Annex 1, 1. Ongoing activities, group 1.C). Another element that facilitates the search is that European countries identify who are the agencies and entities in charge of the publication and use of the HVDs.

As a final result, we were able to identify 8 HVDs from 8 European countries in the companies and company ownership category, 17 HVDs from 19 European countries in the statistics category and 13 HVDs from 13 European countries in the mobility category. For these datasets, listed in the annex, a comparison exercise was conducted during the workshop where the following dimensions were analysed.

1. **Documentation and accessibility.** Here we have analysed the metadata and how it is provided by MSs when publishing datasets. We have analysed whether:
    - they provide a clear and basic description of the data and provide publicly available information describing the structure and semantics of the data;
    - they indicate what the licence of the data is;
    - they publish easily accessible metadata;
    - they provide datasets via APIs and bulk download options;
    - they indicate their update frequency.

2. *Technical aspects.* Here we analysed:
    - granularity, which refers to the level of detail of the dataset;
    - format(s) in which the dataset is published, particularly whether it is in a machine-readable format;
    - the attributes of the data;
    - the use of internationally recognised controlled vocabularies and ontologies.

3. *Metadata.* In the datasets where we have found metadata, we have checked:
    - the format of the metadata;
    - whether it is easily identifiable, accessible and downloadable;
    - the information that they provide within the metadata.

The following results from this comparison exercise can be highlighted.

3.1. Comparison of HVDs belonging to the companies and company ownership category

This category may be considered as the most conflictive within those identified as HVDs, since in many of the MSs a payment rate is established to access these datasets, or they do not directly allow access to the information about the companies at the dataset level, only allowing the search of information of individual companies. That is why the identification and accessibility of these datasets is much more complicated. In Annex 2, Table 1 lists these datasets, providing information on their location in data portals and a review of them based on the attributes and functionalities established by the regulation.

Regarding **documentation and accessibility**. A description section has been provided in all datasets, however in some of them it consisted of the title of the dataset. We located only in two datasets that the attributes of the data were mentioned in the description, and only in one of these were the semantics of the attributes explained. All datasets had a licence associated with them, some of them have a Creative Commons Attribution 4.0 or CC BY 0.0 licence and others have a national domain licence. Most of them have been published on data.europa.eu, therefore most of them have associated metadata in the DCAT-AP format. The only dataset that was published in the national portal has a web-based metadata format. All datasets are available for bulk download. Regarding the use of APIs, only one of them did not identify the availability of an API to access them. Of the rest, the most used API is the CKAN API. In order to use the API of one of the datasets, it is necessary to register. Regarding the update frequency, the vast majority indicate it within their metadata. However, the frequency varies between the MSs, between daily, weekly, biweekly and monthly.

Regarding *technical aspects*. In terms of granularity, all MSs provide information at individual companies level. All datasets were available for download in CSV format. Only one dataset was also accessible in other

formats such as JSON, GeoJSON and shapefile. Although the CSV format is machine-readable, it would be advisable for MSs to publish datasets in various formats whenever possible. Regarding the attributes of the datasets, it is difficult to compare the attributes of the datasets due to the great heterogeneity observed in naming them. Some MSs use abbreviations in the attributes, complicating the comparison, and this is even more difficult when the attributes are in the language of the MS. Compared to the attributes describing the regulation, these are the most frequently missing attributes among the datasets: status of the enterprise, MS where the company is registered and activity/activities object of the enterprise, such as NACE code. Most of the MSs were not identified as using controlled vocabularies in the datasets. The only thing that was identified was the use of national codes to classify companies and, in the case of one dataset, the use of the European Unique Identifier (EUID).

Regarding **the metadata**. As mentioned above, all the datasets that have been located on data.europa.eu have associated metadata in the DCAT-AP format. As for the dataset in the national portal, the metadata is visible on the dataset web page but there is no option to download the metadata directly from the dataset. With respect to the content of the metadata, since most of it is in the DCAT-AP format, they are generally quite complete, although it should be noted that a few datasets were missing the information on the point of contact and on what licence and conditions of use each dataset has.

While there are increasing efforts by the MSs to implement the regulation concerning HVDs of this category, it is worth noting that there is still a long way to go in terms of harmonising these datasets. The main challenge lies in the accessibility of this data. As mentioned, several MSs still charge fees for access or restrict access to individual company-level data. This is largely due to the sensitive nature of the data, which, if not published in a controlled and documented manner, may conflict with the general data protection regulation. Therefore, the mandatory publication of such data requires a clear indication of not only the licence but also the conditions of use. In reference to the structure of the data and semantics, as mentioned before, we found that there is great heterogeneity among MSs when choosing which attributes to represent for the data, and also in the way of naming these attributes. To address these discrepancies, it is highly recommended to adhere to the attribute table and information provided in the regulation and employ controlled vocabularies and code lists, as mentioned in the methodology section. Furthermore, it is crucial to have detailed documentation of the data attributes and structure, in addition to the necessity of standardised metadata in formats such as the DCAT-AP, which has been scarcely identified in the analysed datasets. Many datasets use country-specific codes and terminologies that are challenging to comprehend without proper reference.

From this analysis, several key points emerge. European initiatives for dataset harmonisation are essential to facilitate the implementation of the regulation by the Member States. In this regard, we find two distinct scenarios within the HVD categories. On one hand, there are categories such as geospatial, mobility, and Earth observation and environment, where initiatives like Inspire have made significant progress in their harmonisation efforts. On the other hand, in the case of the companies and company ownership category, while some resources like the Core Business Vocabulary are available, a considerable number of resources are still required, not only to promote the reuse of these datasets but also in the initial step of their publication. During the interactive workshop organised with data providers, the main challenge highlighted by the participants was the lack of resources and funding available to the Member States for implementing

the regulation. This is why major European institutions are promoting the development and dissemination of resources that facilitate the publication of these datasets. For instance, the work done to develop the DCAT-AP extension for HVDs will be a crucial resource for ECs to identify and group their HVDs so that they can be added to the list in the report that must be completed and provided to the European Commission by 2024.

One area where European countries face challenges is the establishment of APIs for accessing and downloading data. To incorporate these resources into their data portals, strong coordination is required at the internal level by public administrations. At the European level, multiple initiatives are underway to promote API implementation, such as the publication of *An Application Programming Interfaces (APIs) Framework for Digital Government* (Boyd et al., 2020) by the European Commission's Joint Research Centre. This framework proposes a standard for adopting APIs in government environments. Additionally, in April 2023, the Digital Economy Unit of the European Commission's Joint Research Centre, in collaboration with APIdays, conducted a webinar discussing the opportunities and challenges associated with APIs for implementing the open data directive in the public sector and how to collaborate effectively with the private sector. The webinar also addressed the implementation of the API strategy for HVDs. During the webinar, the need for standardised API regulation at the European Union level was emphasised to facilitate the implementation of the regulation by European countries.

3.2. Comparison of HVDs belonging to the statistics category

In the regulation, up to 22 types of datasets that can fall into this category are identified and described, indicating the types of attributes that they must contain. This report focuses on a single type of dataset: population. As previously mentioned, a number of these datasets have been located on the portals of national statistical offices and institutes. In many cases these datasets have been published as tables that can be downloaded in various formats, CSV being the most commonly used. Some of these datasets, after being found in the statistical portals, were easier to locate later on data.europa.eu. In Annex 2, Table 2 shows the list of these datasets, providing information on their location in data portals and a review of them based on the attributes and functionalities established by the regulation. After analysis and comparison, the following conclusions stand out.

Regarding ***documentation and accessibility***. In this category we can find several cases within the MSs. Many of these provide in their statistical portals exhaustive documentation on how the data has been collected, what are the quality processes to which they are submitted, a description of the statistical concepts together with a minimum of metadata elements, all in one document. Others provide only a document describing the established methodology regarding data collection and maintenance without providing a specific metadata file for the dataset. In the latter case, it is also where the licence and the conditions of use of the datasets are often not clearly indicated. All the datasets are accessible by bulk download. However, there have been several cases in which no API has been identified that allows access to the data, or cases in which there is an API but it does not have adequate documentation that allow its use.

Regarding ***technical aspects***. When it comes to the granularity and the attributes provided, we found some heterogeneity between the datasets. In general, regarding population datasets, there is a tendency to publish population data according to sex and age, but with respect to publishing data with other attributes such as citizenship, human development index or region, there is a greater heterogeneity. This is not a big

problem at the level of homogenisation since the situation that usually occurs is that they publish different datasets to indicate the different attributes, but it is advisable to follow the attribute specifications indicated in the regulation. All the datasets found are available at least in CSV or Excel (XLSX) (machine-readable formats) but in most MSs there are also download options in other formats like JSON or Extensible Markup Language (XML).

Regarding *metadata*. It is in this section where we find greater heterogeneity. Much of the metadata that was associated with the datasets is in a web-based format or in a PDF that usually does not include enough information, only a description of the content of the dataset. The datasets that could be found in the data.europa.eu portal did have some metadata available in the DCAT-AP version.

There is a notable trend towards greater homogeneity in the publication of statistical data. Eurostat's resources, such as its guidelines, standards and codes, have played a significant role in promoting interoperability. However, there are still areas that need attention, such as the need for MSs to publish standardised linked data metadata, preferably in the DCAT-AP format, to facilitate better identification of these datasets as HVDs. Another aspect to address is the implementation of standardised APIs with associated comprehensive documentation. Overall, while progress has been made, there is room for improvement in ensuring consistent data documentation, accessibility and utilisation across the Member States.

3.3. Comparison of HVDs belonging to the mobility category

Within this category, this report focuses on locating datasets that belong to the 'Transport networks' area. Most of the HVDs in this category have been found using the Inspire Geoportal, using the 'Transport networks' data theme. In Annex 2, Table 3 shows the list of these datasets, providing information on their location in data portals and a review of them based on the attributes and functionalities established by the regulation. The publication of the datasets following the data specification established by Inspire has a series of advantages that promotes homogenisation.

Regarding *documentation and accessibility*. As previously mentioned, Inspire has detailed and extensive documentation on the data models to follow for publication on its portal. They describe both the structure and the semantics that these datasets must follow. On the other hand, thanks to its network services, it thus complies with the requirements of the regulation that indicate that the datasets must be available through download services based on Directive 2007/2/EC. In addition, the publication of these datasets on Inspire ensures that a set of metadata is published in XML format.

Regarding *technical aspects*. Inspire's network services allow data to be accessed in a machine-readable format such as Geography Markup Language (GML). This is the format that has been identified as mostly used among MSs for this type of dataset. In addition, publishing on Inspire encourages that datasets across different MSs contain the same attributes.

Regarding *metadata*. As already indicated, in the Inspire Geoportal, each dataset has the metadata available for download in XML format, which allows quick and easy access to it. In addition, Inspire promotes compliance with the requirements established for the creation and maintenance of metadata for spatial datasets indicated in Regulation (EC) No 1205/2008.

Despite the fact that the Inspire initiative proves to be an indispensable and highly useful resource to promote the harmonisation of these HVDs, we also found some elements that need to be polished. For example, it is not mandatory on Inspire to establish the licence and the conditions of use of the datasets. In fact, in several of the HVDs that were located, the licence was not indicated. It is essential that the MSs add it since one of the requirements is the publication of these datasets under the conditions of the Creative Commons BY 4.0 licence or any equivalent or less restrictive open licence. In addition, another element that is not mandatory to indicate and that we have identified as missing in multiple HVDs is the frequency with which the data is updated. Although the regulation only indicates as a requirement that the datasets must be available in their most up-to-date version, it is highly recommended to also indicate the frequency of updates.

4. Summary of findings and recommendations

Thanks to the analysis conducted with sample datasets from three HVD categories and the information collected during the workshop with European countries held in May 2023, we can obtain the following key findings about the identification and homogenisation of HVDs across Member States.

There is still a lot of **uncertainty and confusion surrounding the identification of HVDs and the application of the implementing regulation across European countries**. One of the main challenges that the workshop participants mentioned is the need to raise awareness and disseminate the concept and importance of HVDs, not only at the administrative level, but also at the level of all the possible stakeholders involved. As mentioned above, in HVD categories where there are large data harmonisation initiatives involved, such as Inspire on geospatial data or Eurostat on statistical HVDs, there is a greater tendency for European countries to publish data in a more interoperable and harmonised manner. On the contrary, in categories where the lack of standards is notable, such as companies and company ownership, there is still a long way to go to implement the regulation.

However, it has been also detected that even in categories where standards such as Inspire are used, there are some elements that still need to be investigated. For example, the lack of obligation on the part of Inspire for European countries to indicate the licence and conditions of use of the datasets. In addition, during the workshop several participants wanted to highlight the **lack of standardisation between the Inspire ISO standard and the DCAT-AP**, even though the mapping between Inspire metadata and the Geo-DCAT-AP is highlighted as a good practice. Differences have been reported regarding the key features of the DCAT and Inspire, which makes the mapping process more difficult. For geospatial data categories, Inspire is essential when it comes to harmonising, however, the [DCAT-AP](#) is an essential tool when it comes to declaring metadata, and its extension to HVDs is going to be key when it comes to identifying them. That is why a process is necessary at the level of institutions/organisations where an effort is made to try to harmonise these two models of interoperability, in order to put an end to this incompatibility.

Another key element is the **lack of use of controlled vocabularies and ontologies for the publication of data**. In the analysed datasets, hardly any use or mention of use of these resources, which are essential in ensuring the interoperability of the datasets, has been found. Given this situation, the action of the European institutions in charge of promoting interoperability is necessary, focusing on greater development and dissemination of this type of resources, to facilitate European countries' task of implementing them. For example, the EU Vocabularies website is a key reference in this respect, as are all the efforts made in the context of SEMIC. On the part of the ECs, it is also necessary to promote among their institutions and data providers a greater willingness to use these resources.

There is still much progress to be made by the European institutions and by ECs at the individual level in taking measures to ensure the implementation of the regulation on the publication and reuse of HVDs. With this report, however, we propose a series of recommendations aimed at promoting a greater harmonisation and homogenisation in the publication and use of HVDs, at the data level and at the metadata level.

A) **Data homogenisation recommendations**

To promote data homogenisation, as mentioned above, it is essential to use common data models, controlled vocabularies, ontologies and APIs in order to avoid ambiguities in the data published by ECs. In the methodology section (Step 2) we mentioned some useful resources and resource centres that can be used for each of the six categories. Among the resources to be highlighted, we find the following.

- For **geospatial, mobility, meteorological and Earth observation and environment data**, the data specifications are provided by Inspire.
- For the **statistics** category, the use of the classifications and data provided by Eurostat is strongly recommended.
- For the **companies and company ownership** category, the SEMIC Core Business Vocabulary is an essential resource. Within this category, the use of code lists that promote the disambiguation of terms, such as the NACE Revision 2 classification or the LEI code, is also recommended.

At a more general level applicable to all datasets, a series of controlled vocabularies that allow for a consistent description of the data are available on the portal of the Publications Office of the European Union. Within these we find authority tables, thesauri and taxonomies.

To facilitate the reuse of datasets, it is recommended to use APIs and bulk downloads to access data. While bulk downloads are commonly available on data portals, the implementation of APIs is still an area that needs improvement. Therefore, it is crucial for institutions and organisations at the national level to prioritise the establishment of API resources. Additionally, it is important to address the lack of documentation regarding the usage of available APIs for accessing each dataset. APIs should also be associated with persistent links to ensure long-term accessibility.

Another essential requirement is the use of open and internationally recognised licences. The regulation specifies that datasets should be licensed under Creative Commons BY 4.0 or any equivalent open licence with less restrictive terms. Furthermore, it is necessary to describe the chosen licence with a persistent Internationalised Resource Identifier (IRI) in the metadata. The Publications Office provides a named authority list licence, which includes a collection of licences used within the EU. This resource offers authority codes and labels in most official EU languages.

B) Metadata homogenisation recommendations

As established in the regulation, a common requirement for all datasets is the creation of metadata that fully describes the dataset and is available on the web in an open and widely used machine-readable format. For this, the fundamental resource to use is the DCAT-AP standard with the additional guidelines to describe HVDs. The DCAT-AP is essential for creating metadata for datasets, not only because it is an international standard but also because it enables the quick and efficient identification and retrieval of datasets. With the guidelines to describe HVDs, the DCAT-AP will become even more valuable. These guidelines allow for the categorisation of datasets as high value, facilitating their prioritisation and utilisation. In cases where the metadata originates from another metadata source, such as the Inspire model, the original identifiers can be employed for the dataset by indicating the relationship using the dct:source property. This ensures interoperability and consistency while leveraging existing identifiers for efficient dataset retrieval and integration.

References

Boyd, M., Vaccari, L., Posada, M. and Gattwinkel, D. (2020), *An Application Programming Interfaces (APIs) Framework for Digital Government*, Publications Office (https://data.europa.eu/doi/10.2760/772503).

Cyganiak, R., Gillman, D., Grim, R. T. A. M., Jacques, Y. and Thomas, W. (2019), *XKOS – RDF/RDFS representation of the cross-domain knowledge organization systems (KOS) model*, The DDI Alliance (Retrieved from https://rdf-vocabulary.ddialliance.org/xkos.html).

Cyganiak R. and Reynolds, D. (2014), 'The RDF Data Cube Vocabulary', W3C Recommendation, W3C (https://www.w3.org/TR/vocab-data-cube/, accessed 17 June 2023).

European Commission, Directorate-General for Communications Networks, Content and Technology (2020), *Impact assessment study on the list of high value datasets to be made available by the Member States under the open data directive*, Publications Office of the European Union (https://data.europa.eu/doi/10.2759/493091).

European Commission, Eurostat (2021), *European Statistical System Handbook for Quality and Metadata Reporting – 2021 re-edition*, Publications Office (https://data.europa.eu/doi/10.2785/616374).

Fragkou, P. (2023), *Usage Guidelines of DCAT-AP for High-value Datasets*, Semantic Interoperability Community of Practice (SEMIC) (retrieved from https://semiceu.github.io/DCAT-AP/releases/2.2.0-hvd).

Haller, A., Janowicz, K., Cox, S. J., Le-Phuoc, D., Taylor, K. L. and Lefrançois, M. (2017), 'Semantic Sensor Network ontology', W3C Recommendation, W3C-OGC.

Publications Office of the European Union, Huyer, E. and Blank, M. (2020), *High-value Datasets: Understanding the perspective of data providers*, Publications Office, (https://data.europa.eu/doi/10.2830/363773).

Reynolds, D. and Epimorphics Ltd, (2016), 'Organization ontology', W3C Recommendation, W3C (retrieved from http://www.w3.org/TR/vocab-org/).

European Commission, Commission Implementing Regulation (EU) 2023/138 of 21 December 2022 laying down a list of specific high-value datasets and the arrangements for their publication and re-use, OJ L 19, 20.1.2023, p. 43.

European Parliament and Council, Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an infrastructure for spatial information in the European Community (Inspire), OJ L 108, 25.4.2007, p. 1.

European Parliament and Council, Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (Text with EEA relevance), OJ L 326, 4.12.2008, p. 12*.*

European Parliament and Council, Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the reuse of public sector information, OJ L 172, 26.6.2019, p. 56.

European Parliament and Council, Commission Delegated Regulation (EU) 2023/137 of 10 October 2022 amending Regulation (EC) No 1893/2006 of the European Parliament and of the Council establishing the statistical classification of economic activities NACE Revision 2 (Text with EEA relevance), OJ L 19, 20.1.2023, p. 5.

Annex 1. Results from the workshop

During the interactive research workshop on HVDs with data providers from the European countries (ECs), the attendees were asked to attempt to answer the following questions.

- What are the ongoing activities in your country (national/regional/local) for the implementation of high-value datasets?
- Who is responsible for these activities?
- Are you experiencing challenges in these activities? If yes, what?

To answer these questions, participants were provided with the following categories in which to add their response.

- Ongoing activities
- Responsibles
- Challenges

We have clustered the responses that were received. In order to respect the general data protection regulation, the name of the participants or the European Country/ies to which they belong will not be mentioned.

1) Ongoing activities

Within this category, participants were asked to indicate which HVD-related initiatives are being implemented in their country or are planned to be implemented in the near future. In this category, we have identified six clusters based on participants' responses.

1.A. Creation of a working group / task force / committee related to the publication and management of HVDs. A small percentage of European countries representatives indicated in their response that, as an initiative, the creation of working groups that are focused on the work required to identify, process and publish HVDs had been (or will be) conducted.

1.B. Initiatives related to HVD dissemination and education. One of the most common and frequent initiatives among European countries is the organisation of workshops, webinars, documents and resources in general to disseminate the concept of HVDs and the regulation implemented by the European Union.

1.C. Add a tag or filter on national open data portals for HVD identification. Several European countries indicated a willingness to add in their national open data portals a label option in the datasets, or keyword, or filter, which would allow users to identify HVDs more easily. In fact, a small percentage of these ECs indicated that this initiative has already been implemented in their data portals.

1.D. Initiatives to establish coordination and agreements with stakeholders. Another of the most frequently mentioned initiatives was the action of the ECs in contacting data owners, providers and local authorities to promote the publication and transaction of datasets.

1.E. Mapping and creation of HVD inventories. Several ECs indicated that one of their main initiatives was the self-identification of these HVDs as indicated in the regulation, and the creation of specific inventories of these datasets.

1.F. Control and publication of quality (meta)data. Although this cluster was created on the basis of more heterogeneous responses, several Ecs launched initiatives to conduct a more thorough control of the quality of the data.

Although these are the main clusters that we have identified among the ECs in terms of initiatives, we also want to highlight initiatives that have been conducted at a more individual level, such as the creation of templates that follow the guidelines indicated in the implementing regulation or the creation of a thesaurus of attributes to promote data interoperability.

2) Responsibles

Based on the initiatives that the ECs indicated in the previous section, in this category they had to answer who were the groups/people in charge of directing and participating in them. In this category, we have identified four clusters based on participants' responses.

2.A. Ministries related to digital transformation. Several European countries indicated that the ministries related to digital transactions, such as the Ministry of e-Government and Ministry of Digital Transformation, are responsible for policies related to the HVD.

2.B. Other types of ministries. It has been seen in several European countries that the management of HVDs often falls to more specialised ministries, according to the category of HVD. Among these we find that the most mentioned are the ministries of finance, ministries of economy and ministries related to environmental policies.

2.C. Agents with roles in the open data portals or the open data unit. Some participants mentioned that the initiatives related to the publication of the HVDs were conducted by the people responsible for maintaining the national open data portals.

2.D. Other national data institutes. With regard to HVD categories such as statistical data or mobility data, in several European countries it is national institutions such as statistical institutes or federal transport agencies that are responsible for managing and publishing these data.

Other things that we can highlight on a more individual level is, as some ECs mention, that it is the data owners and providers themselves who are in charge of managing this type of initiative.

3) Challenges

Finally, the participants indicated the main challenges that were encountered when trying to implement this type of initiative and in relation to HVDs in general. Within this category is where we find a greater heterogeneity in terms of the answers provided. In this category, we have identified four clusters based on participants' responses.

3.A. Lack of resources. The most frequently mentioned challenge among ECs is the lack of resources from the administrations to conduct this type of initiative. The participants emphasise that an increase in funding and a greater education and interest in public workers acquiring the necessary skills to conduct these types of tasks would be necessary.

3.B. Need of standardisation between Inspire and the DCAT-AP. Several representatives indicated that they had identified an incompatibility between the Inspire model and the DCAT-AP vocabulary when creating the metadata for the HVD.

3.C. Denoting HVDs on the national data portal. Unlike the Member states that have already implemented this initiative, other ECs are unclear on how to indicate which sets on their data portals are of high value. Several indicated that they were waiting for the DCAT-AP extension to mark in the metadata that it is an HVD.

3.D. Difficulty in understanding the definition and the escape of the datasets in the implementing regulation. Another of the most frequently mentioned challenges among the ECs was that there are still many doubts and some confusion about how to identify and define what an HVD is despite the information provided in the EU regulation. It was mentioned that it would be quite useful to have representations and practical examples to help identify and publish them.

Annex 2. High-value datasets found from the categories companies and company ownership, statistics and mobility

Table 1. List of HVDs found within the **companies and company ownership** category. The columns are: MS to which the dataset corresponds; title of the dataset; check if the dataset is found on data.europa.eu; check if it is found on other portals; what is the licence and/or terms of use; how often is it updated; their formats; whether they are accessible through API and/or bulk download; and which of the attributes specified in the regulation do they provide.

Member State	Title HVD	data.europa.eu	Other portals	Licence/Copyright
Estonia	Public free data of the commercial register	Not found	E-Business Register Open Data	Creative Commons Attribution-ShareAlike 4.0 International
Finland	Company register	Yes	Finland's open data catalogue	Creative Commons Attribution 4.0
Latvia	Register of companies	Yes	Latvian open data portal	Creative Commons — CC0 1.0 Universal
Netherlands	KvK HR Open Data Set	Yes	Open data van de Overheid	Creative Commons — CC0 1.0 Universal
Norway	Central Coordinating Register for Legal Entities	Yes	The Brønnøysund Register Centre	Norwegian Licence for Open Government Data (NLOD)
Romania	Companies registered with the Trade Register until April 7, 2023	Yes	Romania's Open Data Portal	Creative Commons Attribution 4.0
France	Companies registered in 2023	Yes	Open platform for French public data	Open Licence
Poland	Przedsiębiorcy telekomunikacyjni	Yes	Poland' open Data Portal	Creative Commons Attribution 4.0 International
Slovenia	Business Register of Slovenia	Yes	Slovenia's open data portal	Creative Commons Attribution 4.0 International

Member State	Update frequency	Format	APIs / bulk download	Key attributes from the Regulation
Estonia	Daily	XML, CSV, JSON	Yes	Name of the company, company status, registered office address, registration number.
Finland	Monthly	CSV	API not found	Name of the company, registration date, registered office address
Latvia	Daily	CSV	Yes	Name of the company, company status, Registered office address
Netherlands	Weekly	CSV	Yes	Registration date, registered office address, legal form, activity/activities that are object of the company
Norway	Daily	CSV	Yes	Name of the company, registration date, registered office address, registration number
Romania	Not found	CSV	Yes	Name of the company, registration number
France	Not found	CSV, JSON, GeoJSON, Shapefile	Yes	Name of the company, company status, Registration date, registered office address, Registration number
Poland	Monthly	CSV	Yes	Name of the Company, legal form, registration date, registered office address, registration number and Member State where the company is registered
Slovenia	Bi-weekly	CSV	Yes	Name of the company, registration number, legal form, registered office address, Member State where the company is registered

Table 2. List of HVDs found within the **statistics** category. The columns are: MS to which the dataset corresponds; title of the dataset; check if the dataset is found on data.europa.eu; check if it is found on other portals; what is the licence and/or terms of use; their formats; whether they are provided with metadata and documentation that explain their structure and semantics; whether they are accessible through API and/or bulk download; whether they provide the population on 1 January and with what breaks; whether they provide the median age and with what breaks; and whether if they provide the old age dependency ratio (proportion of people aged 65+ to people aged 20–64) and with what breaks.

Member State	Title HVD	data.europa.eu	Other portals	Licence/Copyright	Format
Croatia	Statistics in series: Population	Not found	Croatia Bureau of Statistics	Not found	PX, XML, CSV, XLSX
Denmark	Population at the first day of the quarter by region, marital status, sex, age and time	Not found	StatBank Denmark	Creative Commons, CC 4.0 BY	XLSX, CSV, PX, DBF, SAS, TXT
Estonia	RV0212: Population at the beginning of the year and mean annual population by sex and age.	Not found	Statistics Estonia	Creative Commons Attribution 4.0 International	PX, CSV, TXT, XLSX
Finland	11rf -- Population according to age (1-year) and sex by area and the regional division of each statistical reference year, 2003-2022	Not found	Statistics Finland's free-of-charge statistical databases	Creative Commons Attribution 4.0 InternationalYou are moving to another service.	PX, XML, CSV, XLSX, TXT, JSON
Ireland	PEA04: Estimated population (persons in April)	Yes	Central Statistics Office's PxStat Open Data Platform	Creative Commons Attribution 4.0 International	CSV, JSON-stat, PX, XLSX
Latvia	Population by sex, country of birth and age group at the beginning of year 2008 - 2023	Not found	Official statistics portal of Latvia	Licence and copyright notice - All statistics (tables, maps, infographics and other materials) published on the official statistics portal can be used for non-commercial and commercial purposes, indicating the reference to the data source (responsible authority) and the Official Statistics Portal.	PX, XML, CSV, XLSX, TXT, JSON

Luxembourg	Population by age and sex on 1st January	Yes	Statistics portal of Luxembourg	Creative Commons Zero Public Domain Transfer	CSV, XLSX
Netherlands	Population; key figures	Yes	CBS Open Data StatLine	Creative Commons, CC 4.0 BY	CSV, JSON
Norway	05196: Population, by sex, age and citizenship 1977-2023	Not found	Statistics Norway	Commons Attribution 4.0 International (CC BY 4.0)	PX, XML, CSV, XLSX, TXT, JSON
Serbia	Estimates of population by age and sex (beginning, middle and end of year)	Not found	Statistical Office of the Republic of Serbia	Copyright:https://www.stat.gov.rs/en-us/copyright/	XLSX, JSON, CSV, XML, SDMX
Slovakia	Population Age status of the Slovak Republic by sex and age [om2024rs]	Yes	DATAcube	Creative Commons, CC 4.0 BY	XLSX, PDF
Sweden	Population by region, marital status, age and sex. Year 1968-2022	Not found	Statistical Database from the Official Statistics of Sweden	Creative Commons Attribution 4.0 International (CC-BY 4.0)	PX, XML, CSV, XLSX, TXT, JSON
Bulgaria	Population by statistical regions, age, place of residence and sex	Yes	National Statistical Institute of Bulgaria	Licence of the National Statistical Institute	XLS
Germany	Population: Germany, reference date, age years, nationality, gender, marital status	Yes	Statistisches Bundesamt	Data licence Germany	CSV
Lithuania	Resident population at the beginning of the year	Not found	Official statistics portal of Lithuania	Not found	XLSX, CSV, XML, TXT
Slovenia	Population by gender and age, municipalities and settlements, Slovenia, annually	Yes	Statistical office of Slovenia	Copyright	PX, CSV, XLSX, JSTAT
Romania	POP111A - Population on 1 January by age group, age and sex	Not found	Statistics database of Romania	License for Open Government	XLSX, CSV

Member State	Metadata and documentation describing dataset	APIs / bulk download	Population on 1 January	Median Age	Old age dependency ratio (proportion of people aged 65+ to people aged 20-64)
Croatia	Not found	API not found	(*)By sex, counties (NUTS3) and age groups	No	No
Denmark	Yes	Yes	By sex, age and country of birth	No	No
Estonia	Yes	Yes	By sex and age	By sex and age	No
Finland	Yes	Yes	By sex, age, area and regional division	No	No
Ireland	Only metadata	Yes	(*)By sex, counties (NUTS3) and age groups	No	No
Latvia	Not found	Yes	By sex, age group and country of birth	No	No
Luxembourg	Not found	Yes	By sex and age groups	No	No
Netherlands	Yes	Yes	By sex, marital status, age groups, country of birth	No	No
Norway	Not found	Yes	By sex, citizenship and age	No	No
Serbia	Yes	API not found	By sex and age	No	No
Slovakia	Yes	API not found	By age and sex	No	No
Sweden	Not found	Yes	By region, marital status, age and sex	No	No
Bulgaria	Yes	API not found	By region, age and sex	No	No
Germany	Yes	Yes	By age, nationality, sex and marital status	No	No
Lithuania	Not found	Yes	By age, sex, place of residence	No	No
Slovenia	Yes	Yes	By sex, age and municipalities	No	No
Romania	Yes	API not found	By age group and sex	No	No

(*) Instead of 1 January it is 31 December.

Table 3. List of HVDs found within the **mobility** category. The columns are: MS to which the dataset corresponds; title of the dataset; check if the dataset is found on data.europa.eu; check if it is found on other portals; what is the licence and/or terms of use; their formats; whether they provided metadata according to Regulation (EC) No 1205/2008; and whether they are accessible through API and/or bulk download.

Member State	Title HVD	data.europa.eu	Other portals
Croatia	Transport networks	Not found	Inspire Geoportal
Denmark	DK Inspire Road transport network	Not found	Inspire Geoportal
Finland	FTIA Inspire Transport Networks Theme Dataset	Yes	Open Data Finland // Inspire Geoportal
Ireland	National Public Transport Access Node	Not found	Inspire Geoportal
Luxembourg	Inspire - Annex I Theme Transport Networks - Roads - RoadNode	Yes	Inspire GeoPortal
Poland	Transport networks	Yes	Inspire Geoportal
Romania	Road Transport Network	Not found	Inspire Geoportal
Slovakia	Transport networks	Yes	Open data portal Slovakia
Bulgaria	Road Transport Network	Not found	Inspire Geoportal
Lithuania	Annex I. Inspire Dataset for Transport networks Theme	Not found	Inspire Geoportal
Italia	Road Area	Yes	Inspire Geoportal
Spain	Transport Networks of Spain	Yes	Inspire Geoportal
France	Transport network French territory (LoD1 et LoD2)	Not found	Inspire Geoportal

Member State	Licence/Copyright	Format	Metadata according to Regulation (EC) No 1205/2008	APIs / bulk download
Croatia	Public access limited according to Article 13(1)(e) of the Inspire directive	GML	Yes	Download via Inspire
Denmark	Terms of use	GML	Yes	Download via Inspire
Finland	Creative Commons Attribution 4.0 International	N / A	Yes	WFS for download and WMS for visualise
Ireland	Creative Commons Attribution 4.0 International (CC BY 4.0)	N / A	Yes	Not download services available

Luxembourg	Creative Commons 1.0 Universal (CC0 1.0) Public Domain Dedication	GML	Yes	Download via Inspire and data.europa.eu
Poland	No conditions to access and use	N / A	Yes	Download via Inspire and data.europa.eu
Romania	No conditions to access and use	N / A	Yes	Not download services available
Slovakia	Creative Commons Attribution	GML, geopackage, esri gdb	No	Download via data.europa.eu
Bulgaria	No conditions apply to access and use	N / A	Yes	Not download services available
Lithuania	Intellectual property rights	GML	Yes	Download via Inspire
Italia	CC0 1.0 Universal (CC0 1.0)	N / A	Yes	Not download services available
Spain	Terms of use and Creative Commons Attribution 4.0 International	GML	Yes	Download via Inspire and via data.europa.eu
France	No public access restriction according to Inspire	N / A	Yes	Not download services available