European Commission

# Using synonyms to better data discoverability

*Application to INSPIRE Spatial Objects*

Olijslagers, M.,

Vandenbroucke, D.

Hernández Quirós,L. (Editor)

2022

ELISE

Enabling digital government through geospatial & location intelligence

Joint Research Centre

**Contact information**
Name: Lorena Hernández Quirós
Address: Via E. Fermi 2749, TP263, I-21027 Ispra (VA), Italy
Email: lorena.hernandez@ec.europa.eu
Tel.: +39 003278 6653

Name: Francesco Pignatelli
Address: Via E. Fermi 2749, TP263, I-21027 Ispra (VA), Italy
Email: francesco.pignatelli@ec.europa.eu
Tel.: +39 0332 786319

**EU Science Hub**
https://ec.europa.eu/jrc

JRC128528

PDF ISBN 978-92-76-48660-2 doi:10.2760/08796

Luxembourg: Publications Office of the European Union, 2022



How to cite this report: Olijslagers, M; Vandenbroucke, D; *Using synonyms to better data discoverability*, Hernandez Quiros, L. editor(s), Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-48660-2, doi:10.2760/08796, JRC128528

## Contents

## Acknowledgements

The production of this report would not have been possible without the precious input and feedback of all those stakeholders and experts who were contacted all through the study. The authors would like to express their gratitude to all of them.

***Authors***

OLIJSLAGERS, Marc (KU Leuven University)

VANDENBROUCKE, Danny (KU Leuven University)

HERNÁNDEZ Quirós, Lorena (European Commission, Joint Research Centre)

***Reviewers***

VREČAR, Simon (External Consultant)

***Initiators of the study***

TOMAS, Robert

SMITH, Robin

## Abstract

This report summarises the findings of a study conducted by KU Leuven and the European Commission's Joint Research Centre on leveraging synonyms to improve the discovery and retrieval of data resources.

The study and proposed methodology focus on improving semantic interoperability in the geospatial domain. It develops a methodology to harvest synonyms from existing sources and provides alternative ways for Spatial Data Infrastructures (SDI) to benefit from synonyms as alternative labels to expose spatial data.

A methodology has been developed and tested in the study on three use cases or areas: noise, agriculture, and water. The resulting synonyms data sets and the developed tool "*Synonym finder*"[1] available for (re-)use on Joinup complement this report.

**Keywords**: Synonyms, vocabularies, interoperability, semantic matching, geospatial data, usability

[1] https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-finder

## Executive summary

Online sources like product catalogues, online shops and large websites offer search mechanisms for website visitors to find the products or information of their interest. The way a visitor searches and, more concretely, the terms he uses to retrieve content influence the returned results. The returned hits can be poor if the terminology used by the visitor does match that of the web source. That is frequently the case of the web sources using domain-specific language unfamiliar to the visitor. Finding the right resources is even more challenging due to the growing number of online resources and the explosion of big data in general. On average, every human created at least 1.7 MB of data per second in 2020. By 2025, 463 Exabytes of data will be generated each day by people[2]. This amount of resources requires strong data management. Artificial Intelligence (AI) is already used to create useful information from this huge amount of data. Semantic technologies like Linked Data can complement the latter and improve existing AI algorithms to create new opportunities, adding 'Semantic' to the meaning of terms and the relation between them.

These challenges are of interest to the ISA[2] action European Location Interoperability Solutions for e-Government (ELISE Action), which aims to promote location data and technologies and location interoperability, as an enabler for fostering the digital government transformation. In this context, KU Leuven and the European Commission's Joint Research Centre have produced the report titled "*Using synonyms to better data discoverability*". The report summarises the findings of leveraging synonyms to improve data resource discovery and retrieval, focusing on improving semantic interoperability across domains. Semantic interoperability refers to the ability to exchange data between parties while ensuring the data is correctly understood[3]. The proposed methodology has been tested on INSPIRE geospatial resources as a way for Spatial Data Infrastructures (SDIs) to provide alternative ways to expose spatial data.

Synonyms can be defined as words having the same or nearly the same meaning as another[4]. Synonyms exist in natural languages, like' factory' and 'plant'. In technical language, domain-specific terms can also be used; for example, a factory might be called 'production facility' within environmental legislation. The overall idea is that bringing together technical synonyms from different domains can improve interoperability between those domains while providing natural language synonyms can help the general public retrieve more and better resources.

SDIs provide catalogue services for retrieving geospatial data. The INSPIRE Geoportal, the central component of the INSPIRE infrastructure regulated under the Directive 2007/2/EC, is the web entry to the European SDIs and many geospatial data related to the environment. The data objects in the INSPIRE Geoportal are digital representations of real-world objects (e.g. a building, a protected site). They are organised in datasets for different thematic domains and legislations. Although the content looks well organised, it remains difficult for users not familiar with these classifications to find the wished information. For example, datasets including "farm" information are to be found in the thematic domain "*Buildings*" or instead "*Agricultural facilities*"? A separate tool, the "INSPIRE Catalogue of Objects", was developed to tackle this challenge partly. However, the terminology used and indexed by the system's search engine remains strongly linked to the technical language, making its usability and reuse potential weak to general users. Adding synonyms as alternative labels to the objects in the catalogue can help solve this problem.

---

2 https://techjury.net/blog/how-much-data-is-created-every-day/#gref

3 https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers#3.5

4 https://www.dictionary.com/browse/synonym

Taking the latter into account, this study is relevant for data managers (especially vocabulary managers), and users of online resources can also easier link their data to existing resources. It provides a methodology that answers three main questions:

1) Where can synonyms be found?
2) How can they be harvested efficiently?
3) How can the harvested information be applied?

The methodology has been tested in three different use cases related to agriculture, water, and noise. These case studies were selected because they represent thematic domains that go beyond the boundaries of the INSPIRE themes classification. For example, "noise" relates to transport and industry as sources of noise, but also to natural protected areas and health care, as domains impacted by noise.

The study has allowed identifying several potential sources to answer the first question, on 'where' to find synonyms. Besides the information already available in INSPIRE, the study looks into additional technical sources, including online vocabularies and ontologies such as *AGROVOC*, or *EUROVOC*, the multidisciplinary thesaurus covering the activities of the European Union. Furthermore, generic crowdsourced thesauri like *DBpedia* and *Wikidata* can provide more natural language synonyms. This activity has also produced a *Synonyms finder* tool to facilitate retrieval and cross-source browsing. The tool is available as EUPL for anyone interested to reuse it. The *Synonyms finder* tool makes use of the machine to machine services provided by the different selected input sources. Two methods to detect synonyms are exploited by the tool: *Lexical matching* looks at the words themselves, while *semantic matching* exploits the knowledge information available in the different sources, especially the semantic relations defined within and between the different sources.

The proposed methodology provide good results, delivering synonyms and/or semantic information for 63% of the input terms. It is worth emphasising that the methodology is fully applicable to any data (also non-geospatial data). Despite the initial good results, there is room for further enhancements in the methodology and tool. Some recommendations can be drawn from the tests performed:

- Best results are obtained if the data sources and the input data are semantically structured and exploitable.
- For the methodology itself and the tool, several enhancements could include; a (flexible) integration of additional vocabularies, implementation of multilingualism, fine-tuning search criteria.
- Better disclosure of the semantic results is needed to consolidate the interoperability gain provided by the methodology.

There is still a huge potential to further develop the tools into powerful instruments that support the goals of ELISE Action and *Knowledge Transfer* activities in general. Providing synonyms and other semantic information makes the knowledge more understandable, and it allows linking and integrating knowledge, increasing efficiency and opening new possibilities. The semantic operability created in that way allows breaking the vertical data silos that prohibit efficiency gain and innovative possibilities. In all places where data is shared or combined, semantic information is imperative. This work can facilitate the connection between the different upcoming European data spaces regarding data sharing in the EU easing the integration of data from public bodies, businesses and citizens.

# 1 Introduction

## 1.1 ELISE Action

Location-related information underpins an increasingly high proportion of European and national governmental policies, digital services and applications used by public administrations, companies and citizens.

The ELISE Action[5] is a package of legal, policy, organisational, semantic and technical interoperability solutions to facilitate more efficient and effective cross-border or cross-sector digital public services and processes involving location information and the insights gained from that information (location intelligence).

This Action supports Better Regulation and Digital Single Market Strategy goals, including specific actions of the e-Government Action Plan and the European Interoperability Framework (EIF). They are reinforced by the Tallinn Declaration vision and the Communications on Building the data economy and Artificial Intelligence for Europe.

Furthermore, the ELISE Action builds on the principles[6] of the INSPIRE Directive, which establishes an infrastructure for environmental spatial information in Europe. ELISE continues the work of two former ISA [7]actions: the European Union Location Framework (EULF)[8], which developed and promoted a best practice policy and guidance framework, underpinned by INSPIRE, with pilots in different countries and thematic domains, and A Reusable INSPIRE Reference Platform (ARe3NA)[9], which facilitated INSPIRE implementation in the Member States through the development of a structured implementation approach and body of reusable interoperability solutions.

ELISE continues the former work by fostering the adoption of best practice location interoperability solutions across the EU and supporting the digital transformation of public services. All the interoperability actions: EULF, Are3NA, and ELISE build further upon traditional Spatial Data Infrastructures such as INSPIRE to allow the development of location-enabled eGovernment Services and the integration of Location Intelligence in Digital Government to support our Digital Economy and Society (see **Figure 1**).

5 https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/about
6 https://inspire.ec.europa.eu/inspire-principles/9#:~:text=INSPIRE%20is%20based%20on%20a,with%20many%20users%20and%20applications
7 https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/glossary/term/isa
8 https://joinup.ec.europa.eu/collection/european-union-location-framework-eulf/about
9 https://joinup.ec.europa.eu/collection/are3na/about

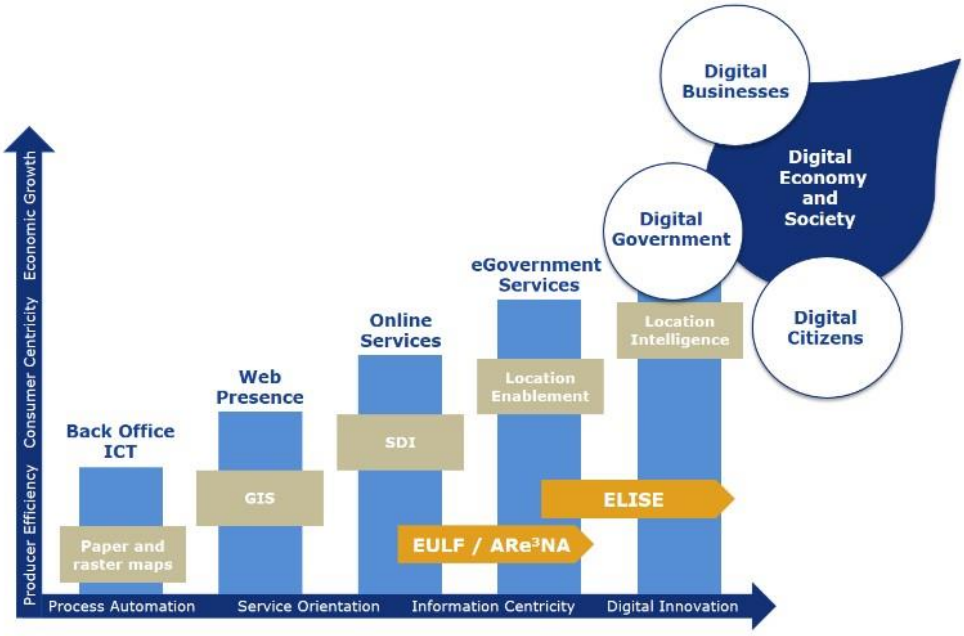



**Figure 1:** The evolution from GIS over SDI to Location Intelligence in a digital society and how ELISE fits in it (ISA², 2020)

ELISE[10] aims to break down barriers and promote a coherent and consistent approach to the sharing and reuse of location information across sectors and borders, in the context of the digital transformation of public services by:

- Supporting different policy initiatives on European and national levels,
- Providing reusable interoperable cross-border and cross-sector frameworks and solutions for public administrations, businesses and citizens,
- Discovering how emerging trends and technologies enable more effective use of location data for policy and digital public service building,
- Geo-Knowledge Base to inform and train stakeholders and promote good practices and innovations in location data.

This is being done through four types of activities and outputs:

- development of frameworks and solutions;
- conducting studies;
- developing pilots and applications, and
- providing a Geo Knowledge Base Service.

The ELISE Geo Knowledge Base Service fosters the reusability of solutions in the context of location interoperability[11]. Moreover, the underlying approach explores how knowledge about (location) interoperability can be represented and shared with different stakeholders, including the pilot/application methodologies. The ELISE Knowledge Transfer (KT) and capacity building activities

10 https://ec.europa.eu/isa2/actions/elise_en#:~:text=ELISE%20aims%20to%20break%20down,on%20European%20and%20national%20level

11 https://joinup.ec.europa.eu/node/704085

have been set up as part of the Geo Knowledge Base Service. The synonyms activity enclosed in this study is part of these efforts.

## 1.2 INSPIRE

The INSPIRE Directive (Directive 2007/2/EC[12]) aims to create a European Union Spatial data infrastructure (SDI) for EU environmental policies and policies or activities that may impact the environment. INSPIRE aims to enable the sharing of environmental spatial information among public sector organisations, facilitate public access to spatial information across Europe and assist in policy-making across boundaries and sectors.

INSPIRE is based on the infrastructures for spatial information established and operated by the Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications.

The Directive itself came into force on 15 May 2007, and its implementation takes place in various stages, with full implementation required by 2021. INSPIRE has been made operational by the EU Member States and is accessible through multiple channels such as the 'INSPIRE Geoportal' and the 'Find your scope' tool[13].

## 1.3 Scope and objectives of the "*synonyms*" activity

Users of geospatial data make use of SDI's for different purposes:

- to search datasets,
- find out about their characteristics by reading and interpreting their metadata,
- find out whether geospatial web services exist for accessing the data,
- eventually visualising the data in a web mapping viewer,
- download the data for further use.

Usually, a geoportal offers this type of function and operation. In most cases, geoportals provide different ways for searching geospatial data: by browsing, typing in keywords, selecting a resource type (data set, service), geographic extent, or by selecting a combination of criteria. However, users can encounter difficulties finding what they are looking for if they do not use the exact names of the target resources and/or are unfamiliar with the datasets' pre-defined keywords or thematic categories.

The latter also happens in the INSPIRE geoportal[14] , which proposes two ways for users to search for datasets through the 'Priority Data Sets Viewer' or the 'INSPIRE Thematic Viewer[15]. However, in both cases, the geoportal design assumes some prior knowledge, e.g. the INSPIRE data themes, environmental domains or existing legislation. There is a similar problem when using the sibling tool known as 'Find Your Scope', a tool designed to support INSPIRE implementers understanding under which INSPIRE theme and corresponding rules its datasets falls into.

---

12 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)

13 Other tools exist as well such as the INSPIRE Registry, the INSPIRE Validator, an INSPIRE Training package, etc.

14 https://inspire-geoportal.ec.europa.eu/

15 All the tools mentioned are explained and illustrated in more detail in the next sections.

To overcome these semantic silos, an approach would be to use alternative terms, i.e. synonyms, hypernyms and hyponyms, which might facilitate the discovery and reuse of the data. Better discovery of geospatial data would facilitate the exploitation and use of INSPIRE resources and other (European) SDI's.

The **objective** of this work on synonyms is **to find technical solutions to identify these synonyms**, and by extension, hypernyms and hyponyms, and how they **could be integrated into the INSPIRE toolset** to improve the search capabilities.

The starting point for the work are the *INSPIRE Objects*. However, this activity aims not to build a complete list of synonyms for all 338 spatial data objects currently available in the *Catalogue of INSPIRE objects* but rather to **elaborate and test a methodology** to collect synonyms applicable to current data objects feasible for future use. Ideally, this methodology would be automated as much as possible. To test whether this is possible and fine-tune the procedure, it will be applied to the data objects of three different use cases, i.e., in the application domains of noise, agriculture and water. For all three cases, the geospatial objects used are spread over several INSPIRE themes and, therefore, difficult to find in the INSPIRE toolset.

## 1.4 Structure of the document

This document contains seven sections:

**Section 1** introduces the ELISE Action and the ELISE Knowledge Transfer Activities in particular and outlines the scope and objectives of the work on synonyms.

**Section 2** provides more background on the challenges of finding and using INSPIRE resources. It also introduces the major INSPIRE tools for doing so: the INSPIRE geoportal and *Find Your Scope* tool and explains how searching could be improved through the use of synonyms.

**Section 3** zooms in on the methodology applied, i.e., the different mechanisms and techniques to find synonyms. This section answers where synonyms can be found and how they can be harvested.

**Section 4** describes the output files resulting from the proposed methodology and gives an overview of the results obtained in the 3 test use cases.

**Section 5** analyses the results obtained.

**Section 6** explains how the results can be used to improve data discoverability. It also proposes how the methodology to find synonyms can be further developed. Finally, it formulates recommendations for a better semantic alignment of data.

**Section 7** provides the final conclusions of this study

# 2 Background and starting point

In this section, we start setting the scene by exposing the challenge of searching and finding relevant resources. Whether the user needs to look for a product in an online shop, a book in library catalogues or a dataset in data portals, retrieving good results is a common challenge that acts as motivation and starting point for the current study.

After setting the scene, the terminology used on search results is explained. This effect is not different for the catalogue of a spatial data infrastructure like INSPIRE. INSPIRE provides several data discovery tools, i.e. the INSPIRE geoportal and the *Find Your Scope tool*, briefly explained. Finally, the idea of using synonyms for improving discoverability is introduced and why the INSPIRE Catalogue of Objects is chosen as a starting point.

## 2.1 The challenge of finding something

The challenge of finding something – a product, a piece of information … – in a catalogue is part of our daily lives. We all know the product catalogues of large companies such as Amazon[16] or IKEA[17] , which offer search mechanisms for customers to find the product(s). Because customers have different backgrounds and interests, they will use different ways to do their searches. The search results are fully influenced by the words and terms they know and are used to typing, having only a very rough idea of what they want versus an exact wish (list). For example, a person might look on the IKEA website for a wardrobe[18] offered in various formats, materials and colours. When writing this report, 432 products were found **Figure 2**: Search results for 'wardrobe' (IKEA, 2021).

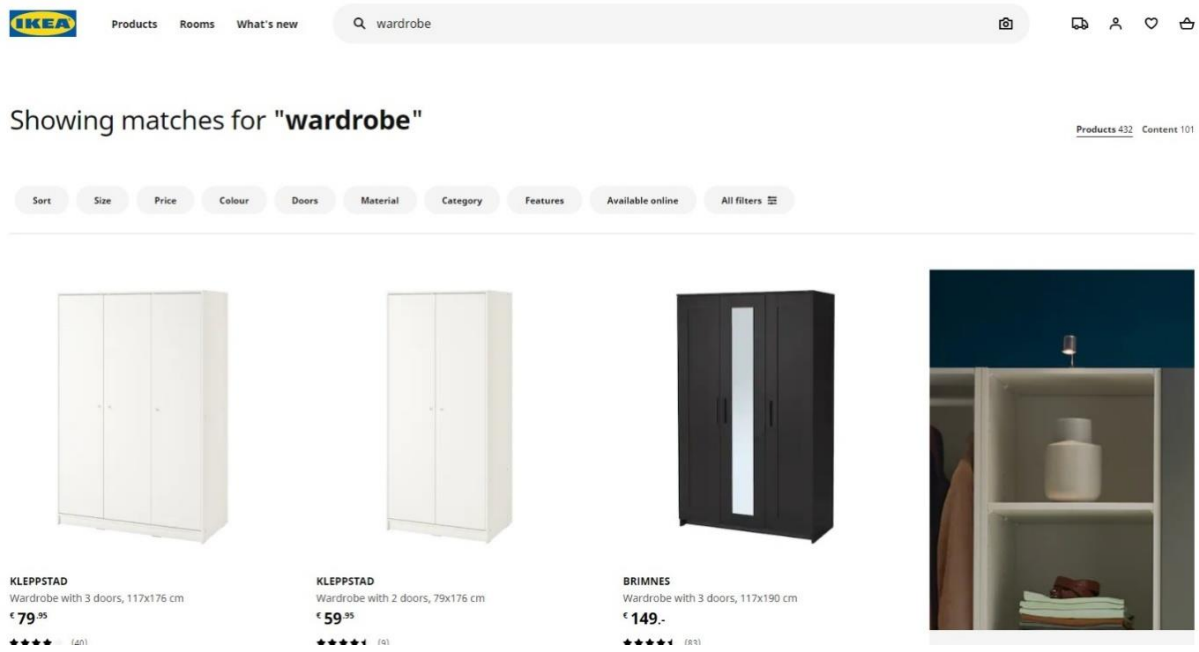



**Figure 2**: Search results for 'wardrobe' (IKEA, 2021)

[16] https://www.amazon.com/
[17] https://www.ikea.com/
[18] A wardrobe or 'armoire' is a standing closet used for storing clothes (Wikipedia, 2021).

However, the person might be looking for a specific type of wardrobe, not a standalone one, but one that can be built into a house wall, often using spare spaces otherwise lost. These are called *closets*[19] , of which 30 were found in the IKEA catalogue. In that sense, we can say that different words or terms can give different search results.

In storing cloths, the meaning of closet and wardrobe are very close to each other. Suppose that IKEA customer uses to word closet while looking for a wardrobe. In that case, he/she will be disappointed only to get 30 search results.

This example shows that the challenge of finding the right data is equally dependent on the words or terms used in the search process.

This dependency is also the case in the context of SDI's and INSPIRE, which is used as a case study in this study. INSPIRE currently contains more than 180.000 resources (data sets, data sets series, services). Users might use the geoportal and follow different paths to search for data sets, and it is not always evident that users find what they are looking for. While the geoportal can search for data sets, the Find Your Scope tool might help find particular (spatial) objects. But also here, different words or terms used might lead to different results. The user is not necessarily aware of the themes used by INSPIRE or the precise naming of an object (type). Synonyms might help solve that problem by integrating them in the INSPIRE catalogues (or in dedicated synonyms vocabularies) so that the tools can use them when users are performing their search operations.

## 2.2 What are synonyms, hypernyms and hyponyms?

Although the original scope of this study only mentions *synonyms*, other relations between words, especially the hypernym-hyponym relationships, are also relevant.

WordNet ([20]), one of the resources used in this study, gives the following definitions:

- **Synonyms**: Two words that can be interchanged in a context are said to be synonymous relative to that concept;
- **Hypernym**: a word that is more generic than a given word;
- **Hyponym**: a word that is more specific than a given word.

Hypernyms and hyponyms are on the two sides of an "is a type of" relation. A hyponym is a type of hypernym. A car (hyponym) is a type of vehicle (hypernym). ). The following image illustrates the hierarchical structure created by hyponym-hypernym relations.

[19] A closet is an enclosed space, with a door, used for storage, particularly that of clothes. "Fitted closet" are built into the walls of the house so that they take up no apparent space in the room. Closets are often built under stairs, thereby using awkward space that would otherwise go unused (Wikipedia, 2021).

[20] Princeton University "About WordNet." WordNet. Princeton University. 2010.

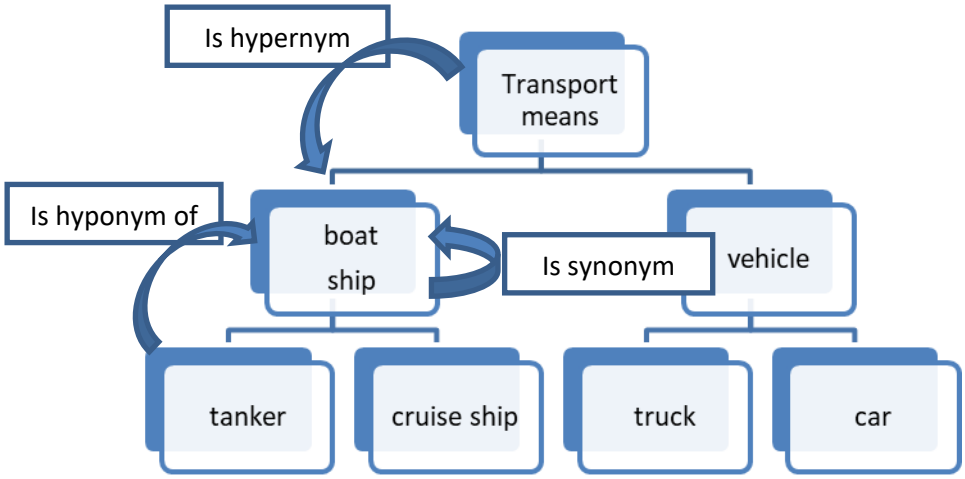



**Figure 3**: Synonyms, hyponyms and hypernyms

The term "*synonym*" will be mainly used in the rest of the document, although the three relation types can be implicated.

## 2.3 The INSPIRE geoportal and *Find Your Scope* tool

The INSPIRE Geoportal is the central access point to the geospatial data provided by the EU Member States and several EFTA countries under the INSPIRE Directive[21]. It is the entry point for discovering and accessing datasets considered priority datasets used for environmental reporting and the 34 different INSPIRE data themes. **Figure 3** shows what this looks like.

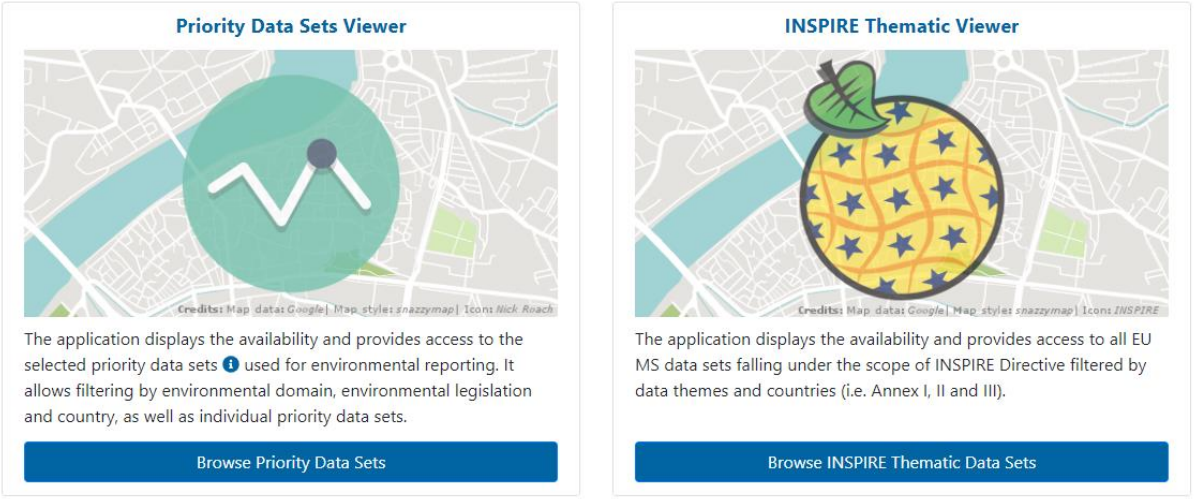



**Figure 3**: The homepage of the INSPIRE Geoportal focusses on priority datasets and INSPIRE Themes

The *priority datasets viewer* provides a rearrangement of data based on the country, legislation or environmental domain they belong to (**Figure 4**). The tool allows browsing per country, per environmental domain or environmental legislation. The INSPIRE Thematic Viewer also allows to

[21] https://inspire-geoportal.ec.europa.eu/

search per country or per INSPIRE data theme. Other dataset compartmentalisations of datasets are in development.

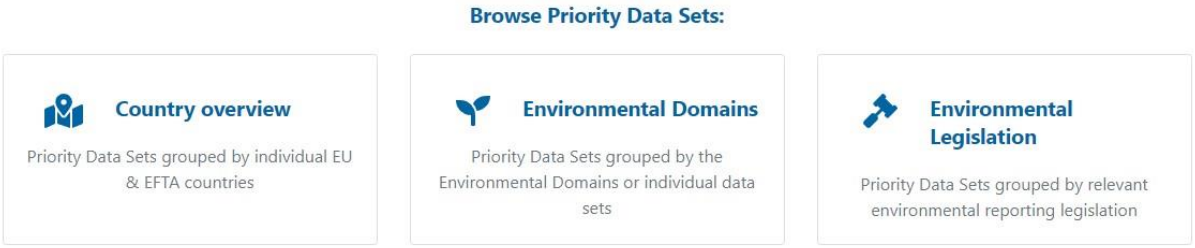



**Figure 4**: Three ways to browse Priority Data Sets in the INSPIRE Geoportal

Another way to browse the INSPIRE geoportal is through the *INSPIRE thematic Viewer*. It provides a Country Overview (**Figure 5**), and data can be explored by going through the INSPIRE Data Themes. INSPIRE is organising its geospatial data in 34 themes defined in the annexes of the INSPIRE Directive. **Figure 6** shows the themes of Annex I of the Directive. Still, it remains challenging to find the data for people not familiar with the provided dataset arrangements or people only interested in one specific data type.

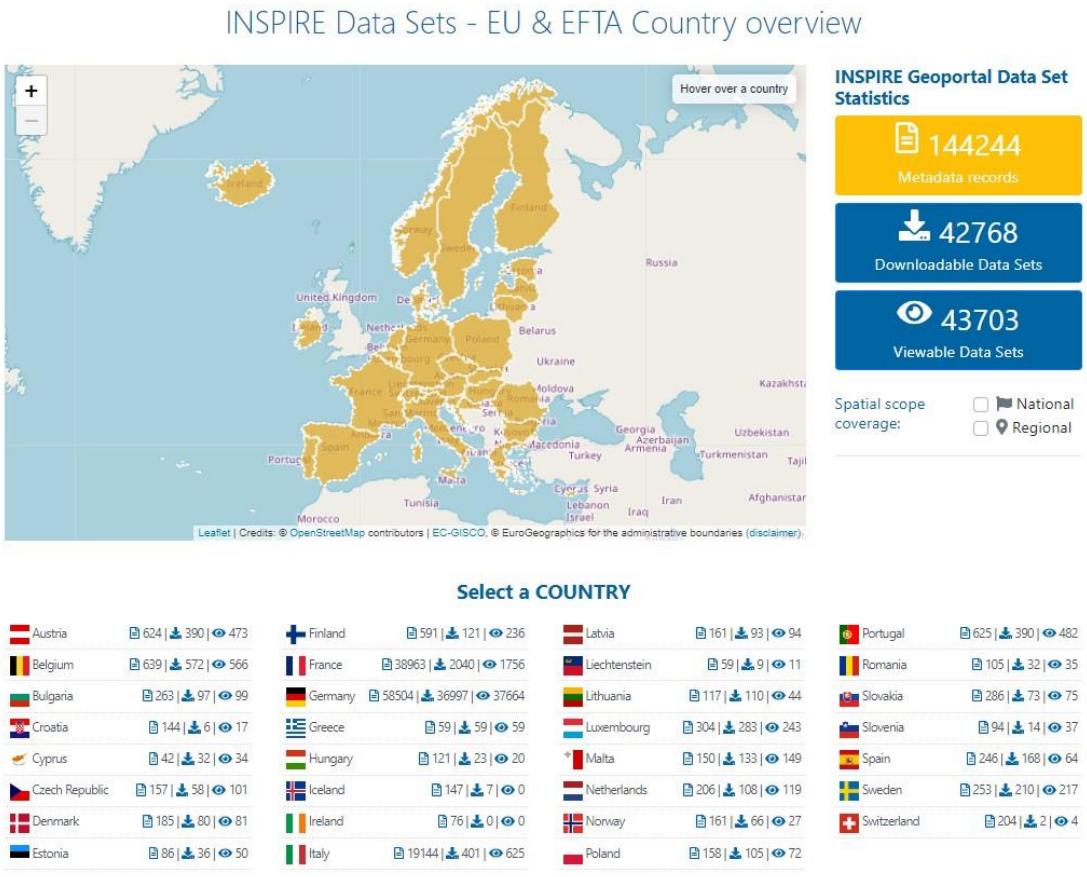



**Figure 5**: INSPIRE Data Sets organised per EU and EFTA country

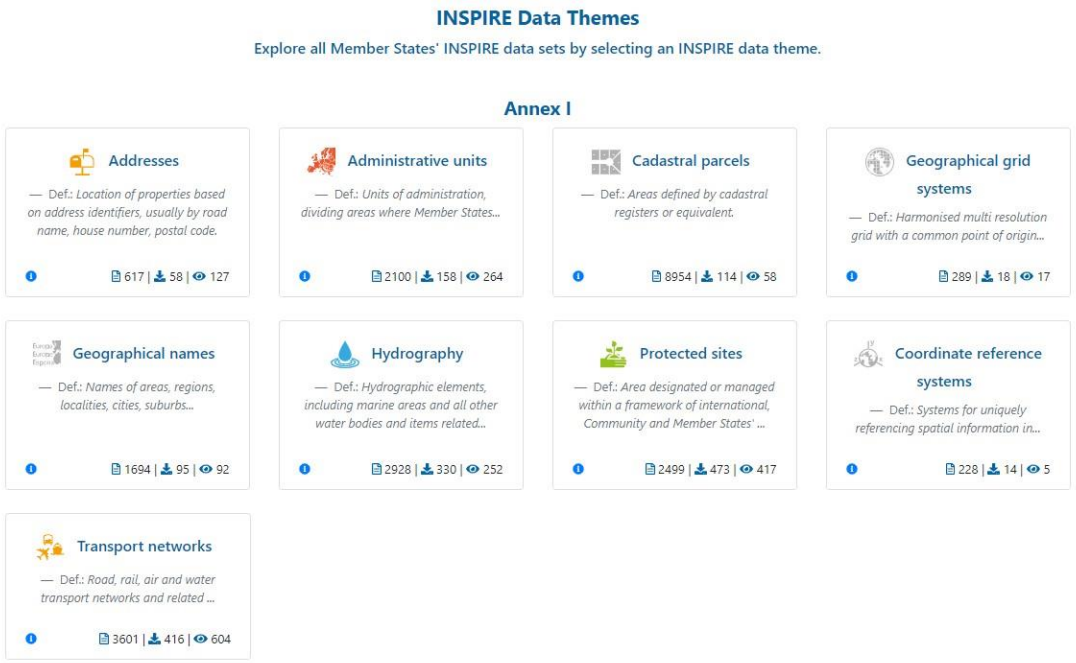



**Figure 6**: Exploring Member States' data sets through one of the 34 themes

For users not working at the dataset level but instead working with geospatial objects and object types, the general functionality of the INSPIRE geoportal is less valuable. The *Find Your Scope tool* was developed to help implementers in an early stage to understand under which INSPIRE data theme(s) their objects of interest fall and to find more tailored information on implementation guidelines. The Find your scope tool allows the user to assess the usability of an object based on its description and additional information about the object's attributes. But Find Your Scope does not directly access the individual datasets containing the data itself.

The *Find Your Scope tool* contains a **catalogue of INSPIRE objects**, an **Interactive Workflow tool** and a **Direct Search** (see **Figure 7**).

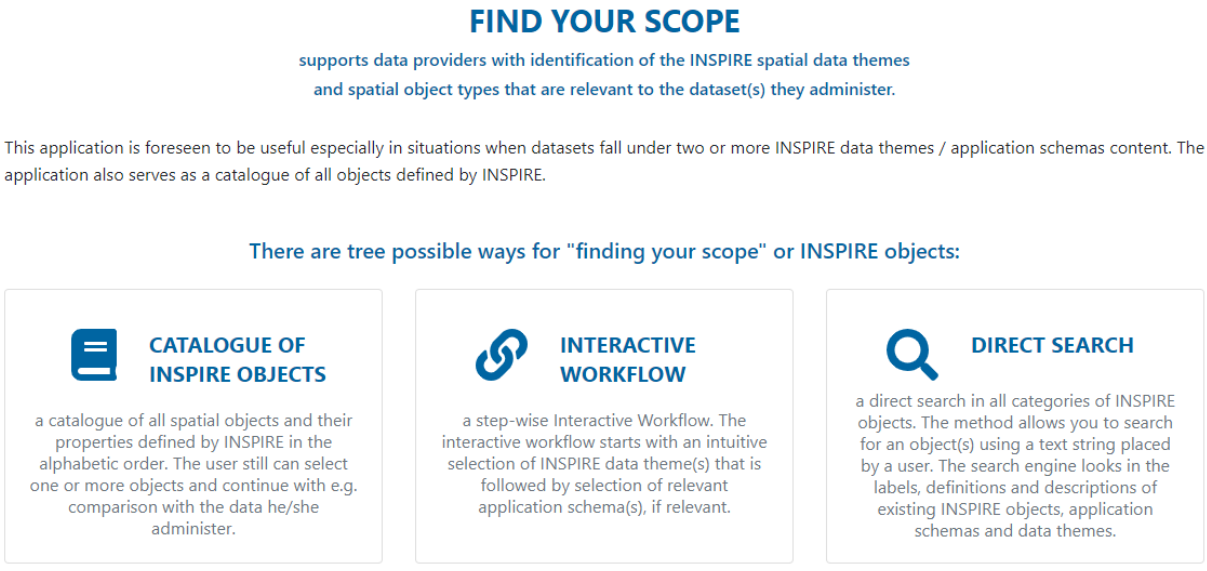



**Figure 7**: The main page of the Find your scope tool

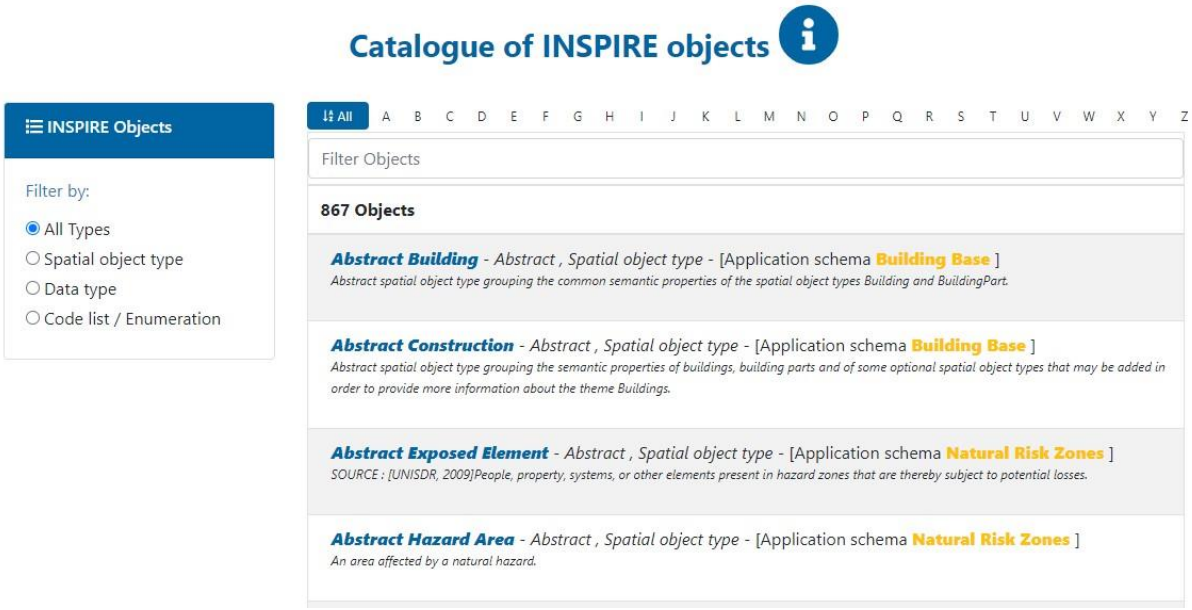



**Figure 8**: The Catalogue of INSPIRE objects tool

The '**Catalogue of INSPIRE Objects**' tool (see **Figure 8**) allows users to search for ordered objects alphabetically. The catalogue can also be filtered by showing only spatial object types, data types or code lists/enumerations. Once an object or objects are found, they can also be selected (add to favourites). They could then serve, e.g. comparison with the content of data providers databases utilising several output options (PDF/DOCX, Matching table).

The '**Interactive Workflow**' (see **Figure 9**) starts with an intuitive selection of INSPIRE data theme(s) that is followed by the selection of relevant application schema(s), if relevant. Both selections are made based on the definition of themes and application schemas. The following step is about concrete spatial objects selection based on their definitions and, if needed, based on the interactive UML diagram. Then the workflow shows a preliminary list of all selected objects, including their properties (attributes) and all associated objects. This list could be refined by changing the selections based on the additional resources, definitions, detailed comparisons etc. Once the final list of objects is complete, the user has two options to save and print the final result. The PDF/DOCX shows all selected INSPIRE Objects and their properties a list of associated objects, including their properties (INSPIRE, 2020).

**Figure 9**: The Interactive Workflow tool

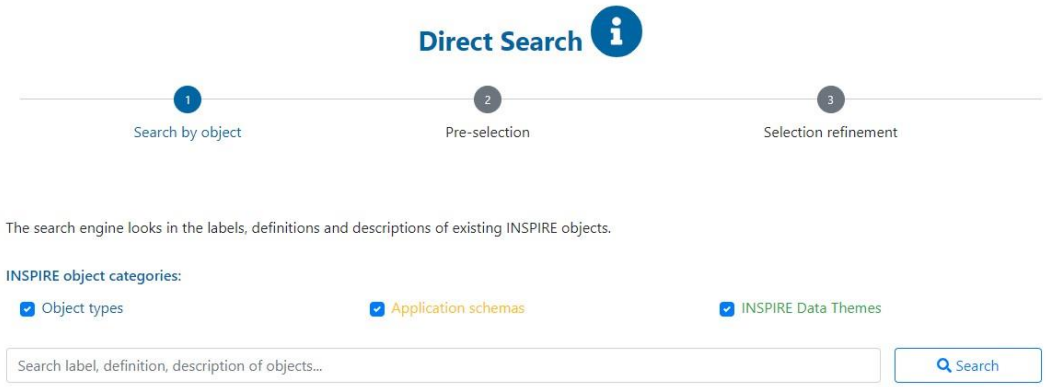



**Figure 10:** The Direct Search tool

The '**Direct Search**' tool allows you to search for an object(s) using a text string placed by a user (see **Figure 10**). The search engine looks in the labels, definitions and descriptions of existing INSPIRE objects, application schemas and data themes. The most relevant objects from the list can be added to a "favourite" objects collection, and this step can be repeated. Thus, more objects can be added to the "favourite" list. The workflow shows a preliminary list of all selected objects, including their properties (attributes) and associated objects. This list could be refined by changing the selections based on the additional resources, definitions, detailed comparisons etc. Once the final list of objects is complete, the user has two options to save and print the final result. The PDF/DOCX shows all selected INSPIRE Objects and their properties + a list of associated objects, including their properties.

How the 'Catalogue of INSPIRE Objects' and the 'Direct Search' tool are used in the context of this study will be explained in Section 3.

## 2.4 Improving search facilities through the use of synonyms

The Catalogue of INSPIRE Objects eliminates the need to select an INSPIRE theme before reaching the spatial object types it contains. All INSPIRE objects are directly accessible in the catalogue. This accessibility increases the discoverability of the data objects for users not familiar with the 34 INSPIRE themes. It also facilitates the use of data in cross-domain use cases. The catalogue provides a filter functionality to select data types based on search keywords. Keywords are compared to the label and definition of the available data types. However, the labels of the objects are often domain-specific related to the INSPIRE theme from where they originate. Users active in a different domain might not be aware of the jargon used in the INSPIRE catalogue. A solution to this would be the use of "synonyms".

*Synonyms* might be jargon originating from different domains or alternative words in natural language. *Natural language synonyms* will make the object catalogue more accessible to the general public.

Next will be shown how and where to find synonyms and exploit them.

## 2.5 The catalogue of INSPIRE objects as the starting point

The Catalogue of INSPIRE Objects provides the starting point for this study. The goal is to examine methods to identify and provide alternative names for the objects in this catalogue. However, some preparations are needed to facilitate the search for synonyms.

The catalogue contains some object types resulting from the INSPIRE modelling process (i.e., developing the data specifications for each of the 34 INSPIRE data themes). They are not related to real-world objects but rather abstract objects[22]. These abstract objects are excluded from the study because these artefacts are not linked to any real-world domain or use case. For example, the *AbstractBuilding* Type in the *Buildings Base application schema* groups the common properties of Building and Building Part but is not instantiable. The catalogue also tells us about whether they are abstract or not for all spatial object types in the Catalogue of Objects. Filtering them out is trivial (see **Figure 11**).

Secondly, some object names only identify the object correctly when considering it inside the schema it was defined. These names have to be better specified before a search for synonyms starts. For example, the "Agricultural and Aquaculture Facilities Model" application schema contains a spatial object type "*Site*". The term "*Site*" is a much too generic concept to be identified as an object related to agriculture outside the schema. Before any (automated) effort to find synonyms can be successful, the name must be changed. In this example, the name "agricultural or aquaculture site" could be an option to contextualise a bit, or taking the next point into account, one could prefer to create two names, "agricultural site" and "aquaculture site".

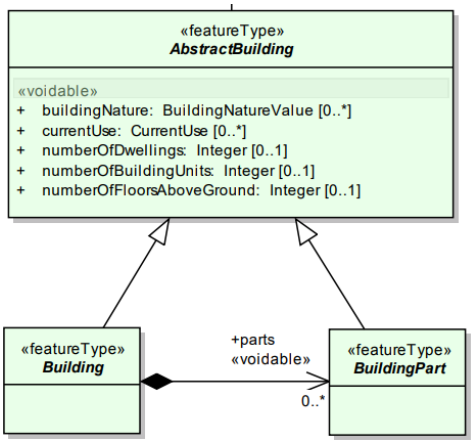



**Figure 11:** Abstract Building type in the Building Base application schema[23]

A third adaptation is related to object types that can be considered "collective types", which are containers for different subtypes. For example, "*Governmental Service*", the code list '*service type value*' gives better information about the real-life objects present in this object type (e.g. *school, hospital*, *fire station*). Therefore, it is suggested to search for synonyms for the different code list values. This problem is already recognised in the INSPIRE data specifications. For INSPIRE View Services, separate layers are requested for spatial object types whose objects can be further

[22] In geospatial data modelling distinction is made between abstract and real world objects (see glossary for definitions)
[23] INSPIRE_dataspecification_bu_v3.0.pdf

classified using a code list-valued attribute. The INSPIRE layer register ([24]) contains the list of layers thus provided.

The previous point also indicates that only looking for synonyms is not enough. "*School*" and "*governmental service*" are not synonyms. "*School*" is "*a type of*" a governmental service. Linguistically this relation is expressed by stating that "*school*" is a hyponym of "*governmental service*", and "*governmental service*" is a hypernym of "*school*". In general, real-life objects are grouped in INSPIRE object types according to the INSPIRE theme's logic. This grouping might be less relevant in other domains. Therefore, it is suggested not only to look for synonyms but also for hyponyms and hypernyms.

Finally, some minor corrections might be needed concerning spelling. As an example, statistical tessellation is misspelt as 'tesselation'.

[24] https://inspire.ec.europa.eu/layer

# 3 Methodology

Three questions must be answered to implement the use of synonyms:

1. Where can synonyms be found?
2. How can they be harvested efficiently?
3. How can the harvested information be applied?

The methodology implemented in this section answers the first two of these questions. Synonyms resources are identified, and methods to efficiently access them are implemented. It is possible to manually assign synonyms for each of the 338 spatial object types in the Catalogue of INSPIRE Objects. In some cases, this is still a valid approach. However, as stated in Section 1.3, this study aims to elaborate a methodology to generate a list of synonyms in the most automated way possible. **The proposed approach explores the possibility to maximise the use of existing data sources that already group or connect (geospatial) terms**. The related terms can be harvested automatically after relating the INSPIRE object labels to corresponding terms in the existing data sources. The proposed methodology to harvest synonyms is presented in the following schema

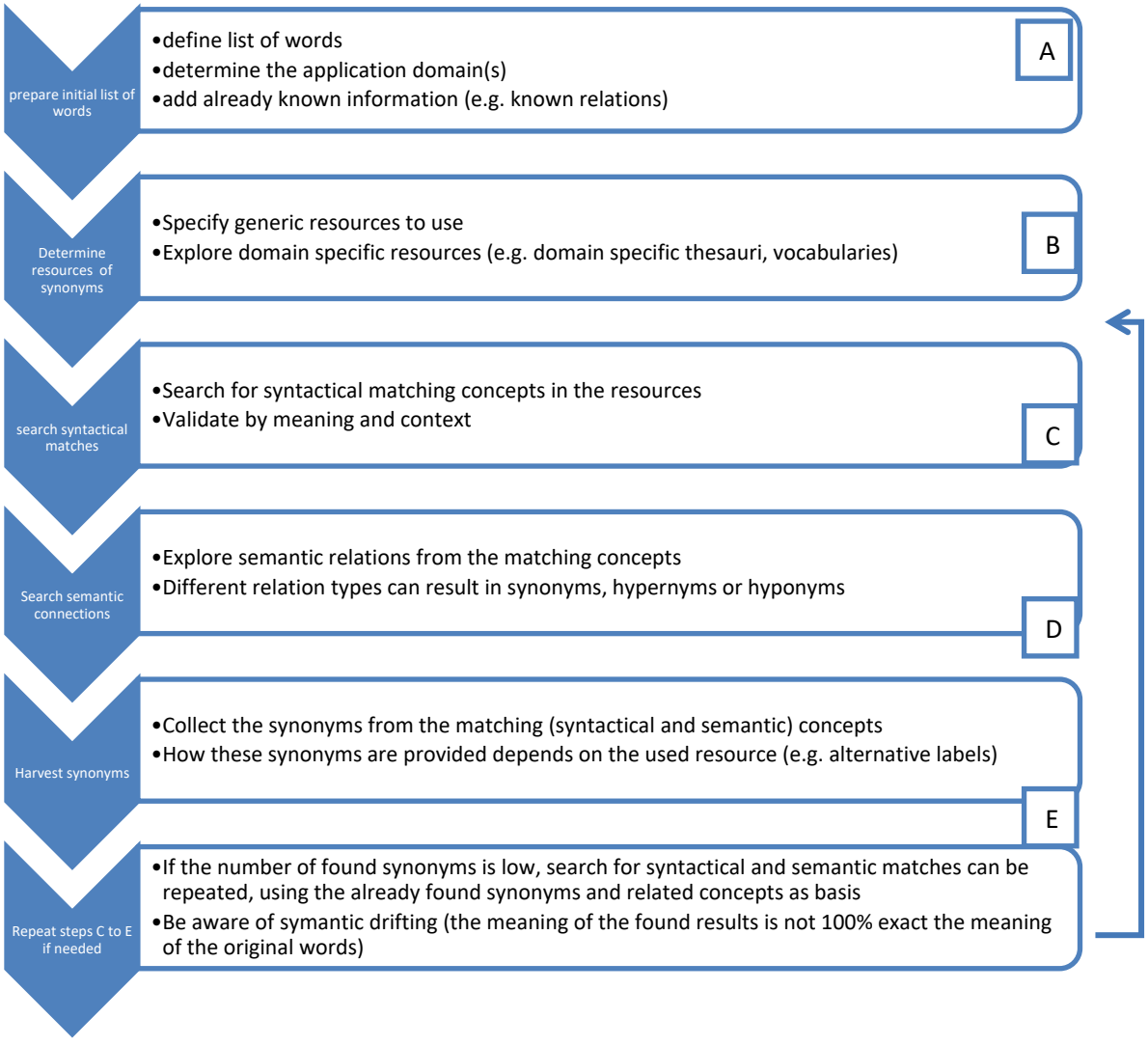



**Figure 12**: Schema for the proposed methodology

The first two steps are preparatory. Step A prepares the initial list of words for which synonyms are requested. In step B, possible resources of synonyms are determined, answering the question where synonyms can be found.

In steps C to E, the resources are searched for synonyms for the initial list of words. If needed, these steps can be repeated iteratively: the results from a previous search is used as input for a new cycle. These two steps answer the question *of how synonyms can be harvested efficiently*.

## 3.1 Preparing the initial list of words: INSPIRE

Step A in the schema is preparing the initial list of words. This step aims to harvest all information already known and prepare this to be used in the procedure.

In this study, INSPIRE is used to demonstrate the synonyms methodology. In particular, the catalogue of objects provides the initial list of words. The catalogue objects related to 3 domains are selected for this test: noise, water and agriculture. These three application domains will be used to determine specific data sources in the next step. But before that, it is helpful to analyse the information already known about the initial list of words. This information is found in the INSPIRE registry

The INSPIRE registry[25] contains four registers directly related to INSPIRE data in themes, application schemas, objects and layers.

The INSPIRE theme register[26] contains the 34 themes defined by INSPIRE. Because the goal is to make the spatial object types discoverable outside their original domain, the theme register is considered less important in the scope of this study.

The INSPIRE application schemas model, the data for one or more applications within an INSPIRE theme.

The INSPIRE Feature Concept Dictionary[27] (IFCD) contains terms and definitions required for specifying thematic spatial object types. These terms correspond to the spatial object types in the INSPIRE Catalogue of Objects. The IFCD is used as the starting point of this study.

The INSPIRE layer register[28] is essential in this study for the spatial object types defined as collective terms. For most of these collective terms, the layers in the register correspond to different (sub) types present in the collection.

IFCD contains the objects from the Catalogue of INSPIRE objects and can be used as base vocabulary to create links with other resources. The INSPIRE theme register provides hypernyms for the concepts in IFCD. The INSPIRE layer register is an important source of hyponyms for the given concepts. Indeed, layers are defined for spatial object types whose objects can be classified further using a code list-valued attribute. So each layer has "a type of" relation with the object type. In this study, the INSPIRE layer register is not used directly. Instead, the code list related to the layer is used.

The different code lists provide more detailed information on their related features. The relation between the code list and its features can have different meanings: code list values can be

25 https://inspire.ec.europa.eu/registry
26 https://inspire.ec.europa.eu/theme
27 https://inspire.ec.europa.eu/featureconcept
28 https://inspire.ec.europa.eu/layer

hyponyms for the feature label, but they can also indicate a more general relationship or property value.

The search functionality of the INSPIRE Object Catalogue only considers a limited part of the information (e.g. de object label, the definition). It does, e.g. not consider the layers related to the data objects. The INSPIRE documentation contains even more information than that, as can easily be illustrated. The Find Your Scope tool also has a Direct Search functionality, which uses a much broader part of the object description to search in. **Figure 13** and **Figure 14** show the effect of this. Looking for 'noise' in the Object Catalogue returns no results (**Figure 13**). Using Direct Search, five object types are returned (**Figure 14**). No spatial objects are found for noise in the catalogue of INSPIRE objects, while five are found when using the Direct Search functionality.

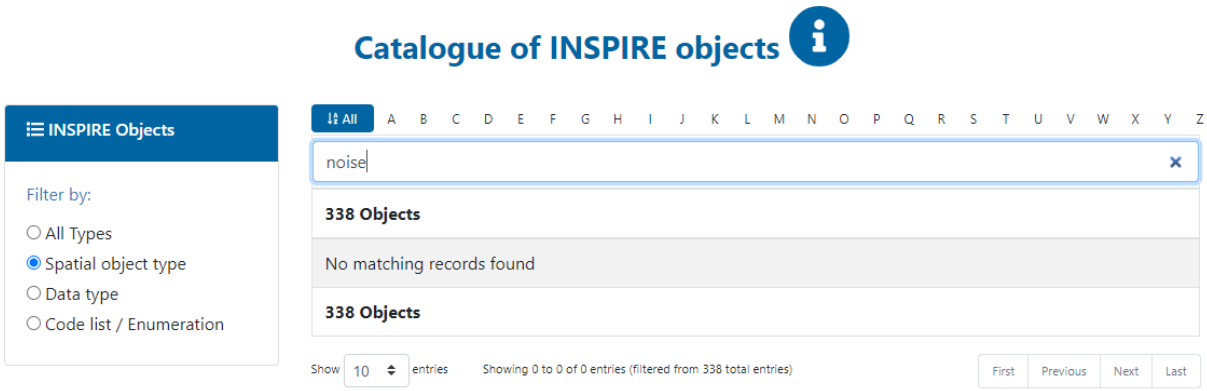



**Figure 13**: Search result for "noise" in the Catalogue of INSPIRE objects

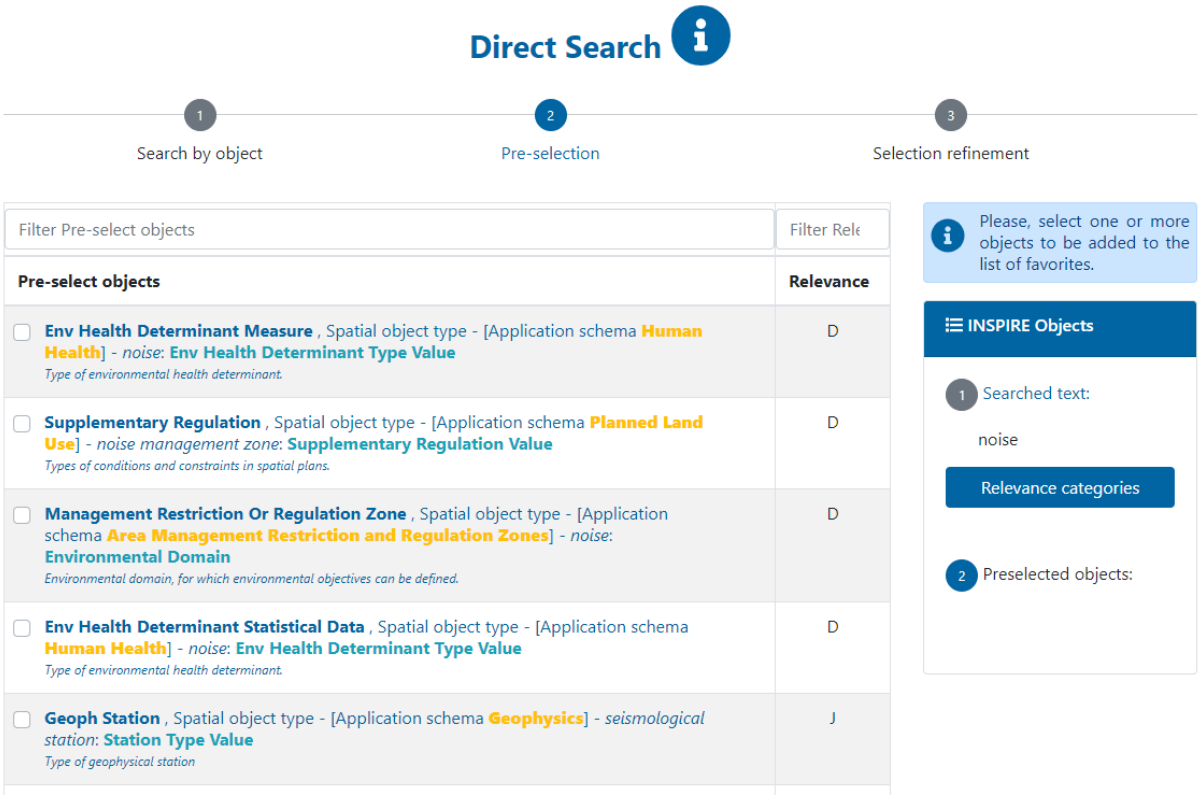



**Figure 14:** Search result for "noise" in Direct Search

This study will focus on the code lists related to the relevant object attributes, including synonyms, hypernyms, and hyponyms related to available INSPIRE documentation. These code lists can provide hyponyms and terms of type 'related to' in addition to those provided by the INSPIRE layers. In the example above, noise is, among others, one of the *Environmental Health Determinant Types* for the measures and statistical data.

## 3.2 Different sources to identify synonyms information for INSPIRE

Where can synonyms be found? After gathering the initial list of words, the next step, B, is to identify possible sources for synonyms. These **data sources must be linkable to the original terms** (INSPIRE objects) to automate the synonyms identification. The type of link indicates whether it concerns a synonym, hyponym or hypernym, or it might be another relation altogether. Four source types were identified to provide links between the original term and related alternatives. The source types were identified during meetings with JRC and based on previous work in the field of the semantic web. In particular, the eENVplus project examined several sources related to the environmental domain.

An additional consideration for the selection of resources is their accessibility. **Only recourses that are open and accessible through a machine-readable interface are retained**. Most sources also have a human-readable web interface. This allows exploring the content of the resources before integrating them into the study.

Two resource types are not linked to INSPIRE and can be used to find synonyms in general:

- Domain-specific vocabularies, thesauri, ontologies
- Generic synonym sources

Two additional resources are specific for INSPIRE:

- Log files from Find Your Scope
- Transformation schemas from the user community

Manually adding known synonyms can be considered a valuable fifth source. This manual intervention might be needed if an automated procedure doesn't give initial results. Any final methodology should provide this option to add synonyms manually.

### 3.2.1 Using vocabularies and ontologies

Recommendations 11 and 12 in the INSPIRE Metadata Implementing Rules[29] promote the use of controlled vocabularies when selecting keywords for an INSPIRE dataset. Logically, one would also use this approach when describing the object types present in those datasets. This approach is also applicable outside the context of INSPIRE.

**Vocabularies, thesauri and ontologies are essential because they define a common, shareable and reusable language within a domain**. As such, they promote the interoperability of information. In general, a vocabulary has a relatively flat structure. If a structure is present, it mainly provides a few hierarchical levels. A thesaurus adds information about synonyms (and sometimes antonyms, words with the opposite meaning). An ontology is used for more complex structures, allowing more relations between the different concepts in the collection.

[29] http://inspire.ec.europa.eu/file/1557/download?token=UaQBcRvQ

Ontologies expose information about synonyms, hypernyms and hyponyms in different ways. They often provide alternative labels for a concept. These can be considered synonyms. Besides that, a hierarchical structure provides broader and narrower concepts (hypernyms and hyponyms). Relations between concepts from different ontologies can refer to synonyms (exact match, close match relation), hypernyms (broad match) or hyponyms (narrow match).

Ontology alignment is the process of mapping concepts between the ontologies. When an ontology is aligned with other ontologies, links between concepts in both ontologies are created. When aligning ontologies from different domains, the links provide cross-domain synonym (and hyponym/hypernym) information. This cross-domain information is precisely what is needed in the scope of this study.

There are numerous linked open ontologies and thesauri available. It is not possible to cover them all in the scope of this study. In general, the best results can be expected if the application domain of the selected thesauri is somehow related to the start list of keywords. This study focuses on GEMET and AGROVOC because they are closely related to the application domains chosen for this study. GEMET covers the environmental domain, AGROVOC covers food and agriculture. As an additional advantage, both resources are, to a certain degree, aligned to an important number of other ontologies. GEMET and AGROVOC are both open; they both have a human-friendly web interface and are machine-accessible through web services. Both are published as linked data, and both ontologies are also available in downloadable format (in RDF). It is also important to note that both ontologies are multilingual, which opens other opportunities outside the scope of this study.

**GEMET** is the General Multilingual Environmental Thesaurus and has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA), Copenhagen[30]. The development of GEMET aimed to define a common general language, a core of general terminology for the environmental field. GEMET contains almost 5300 concepts. These are arranged hierarchically in 3 supergroups and 30 groups. Besides that, 40 themes are defined. Each concept can be assigned to as many themes as necessary. It is also important to note that GEMET is often referred to in INSPIRE dataset metadata. **Figure 15** shows the web interface for GEMET for 'road network'.

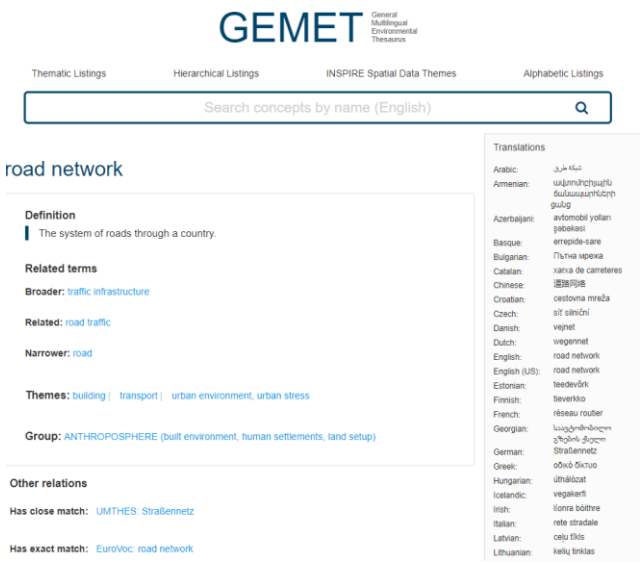



**Figure 15**: The GEMET web interface

30 https://www.eionet.europa.eu/gemet/en/about/

The GEMET interface also provides access to the INSPIRE Spatial Data Themes. Where available, relations between GEMET and the Spatial Data Themes are shown through the interface as shown in **Figure 16**.

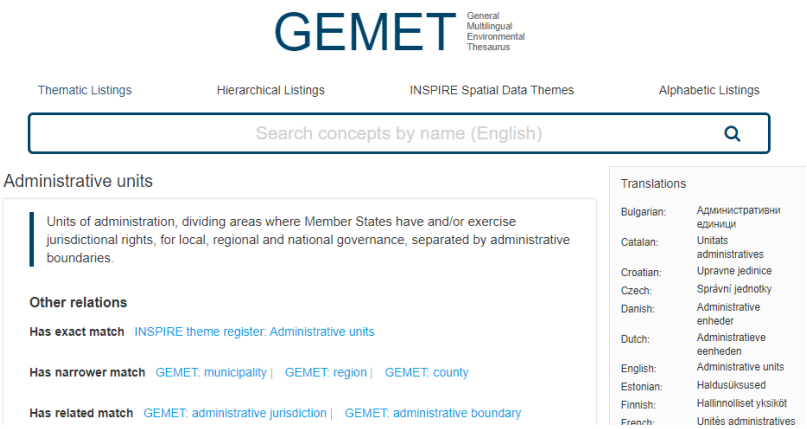



**Figure 16**: INSPIRE Theme Administrative units with matches in the GEMET web interface

**AGROVOC** is an open controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO). It is published by FAO and edited by a community of experts[31]. AGROVOC concepts are grouped in 25 subject areas and are available in up to 29 languages. The ontology contains more than 37700 terms and 10340 alternate terms in English. Those alternate terms are synonym candidates. AGROVOC also contains information on the alignment with 16 other open ontologies, partly related to agriculture. GEMET is one of those datasets. **Figure 17** shows the AGROVOC web interface[32] with the concept' land cover'

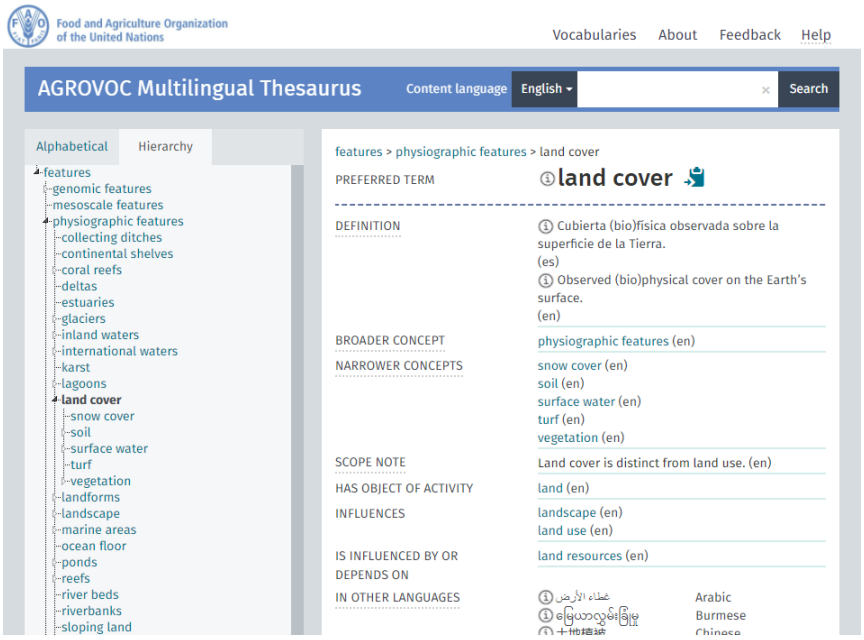



**Figure 17:** The AGROVOC web interface

31 http://aims.fao.org/standards/agrovoc/concept-scheme
32 http://aims.fao.org/standards/agrovoc/functionalities/search

As mentioned, the alignment between different thesauri or ontologies is important and creates cross-domain information. The LusTRE framework contains such reason why it was included in the study. LusTRE results from the eENVplus project ([33]) on eEnvironmental services for advanced applications in INSPIRE. LusTRE is a Linked Thesaurus Framework for Environment. The Framework aims to provide shared standard and scientific terms for a common understanding of environmental data among the different communities operating in the various fields of the environment[34]. LusTRE comprises several ontologies in the environmental domain and focuses on matching concepts in these different ontologies. AGROVOC and GEMET are both integrated into LusTRE. It is important that LusTRE also integrates the INSPIRE theme register and the INSPIRE feature concept dictionary.

The advantage of LusTRE is that it provides unified access to additional ontologies. Therefore all ontologies available in Lustre can be included in the test. Lustre contains the ontologies and also provides additional ontology alignment information in the form of inter-ontology concept mappings. The most important additional ontologies (besides GEMET and AGROVOC) available through Lustre are Eurovoc and EARTh. The complete list of sources accessible through Lustre is available in the vocabularies list of Lustre[35].

**Eurovoc**[36] is the EU's multilingual and multidisciplinary thesaurus, containing keywords in 21 domains and 127 sub-domains. These keywords are used to describe the content of documents in EUR-Lex.

**EARTh** is the Environmental Applications Reference Thesaurus. It has been compiled and is maintained by the CNR-IIA-EKOLab to facilitate the indexing, retrieval, harmonising and integration of human- and machine-readable environmental information from disparate sources across the cultural and linguistic barriers[37]. EARTh is bilingual, English and Italian.

An overview of the mappings between resources in LusTRE is shown in **Figure 18**. Links to the INSPIRE Feature Concept register are few (the figure only shows resources with >200 links). These could have been very valuable for this study. On the other hand, there are already several links to DBpedia. These create links between the domain-specific and technical resources and the more generic content of Wikipedia.

Mappings between Thesauri: Number of links for skos linksets (filter > 200)

| | EARTh | ThIST | AGROVOC | GEMET | EuroVoc | DBpedia | UMTHES |
|---|---|---|---|---|---|---|---|
| EARTh | | 1140 | 1425 | 4328 | 1346 | | |
| | | 1140 | 1425 | 4328 | 1346 | 1862 | 2970 |
| ThIST | 1140 | | 1695 | 792 | 792 | 733 | 921 |
| | 1140 | | 1741 | 835 | 835 | 797 | 948 |
| AGROVOC | 1425 | 1695 | | 1175 | 1269 | | |
| | 1445 | 1741 | | 1181 | 1269 | 11,014 | |
| GEMET | 4328 | 792 | 1175 | | 1683 | | |
| | 4328 | 835 | 1175 | | 1683 | 2035 | 3482 |
| EuroVoc | 1346 | 733 | 1269 | 1683 | | | |
| | 1346 | 797 | 1269 | 1683 | | | |

For each pair of thesauri, the first and second rows indicate the number of *skos:exactMatch* and *skos:closeMatch*, respectively

[33] http://www.eenvplus.eu/
[34] http://linkeddata.ge.imati.cnr.it/
[35] http://linkeddata.ge.imati.cnr.it/terminologies_new.jsp
[36] https://eur-lex.europa.eu/browse/eurovoc.html
[37] https://old.datahub.io/dataset/environmental-applications-reference-thesaurus

**Figure 18**: Mappings between resources in LusTRE[38]

Although there are not many direct mappings identified towards INSPIRE, it is clear that the chosen thesauri cover the range of INSPIRE themes well. This coverage was one of the criteria for vocabulary selection in eENVplus. The following figure shows this coverage.

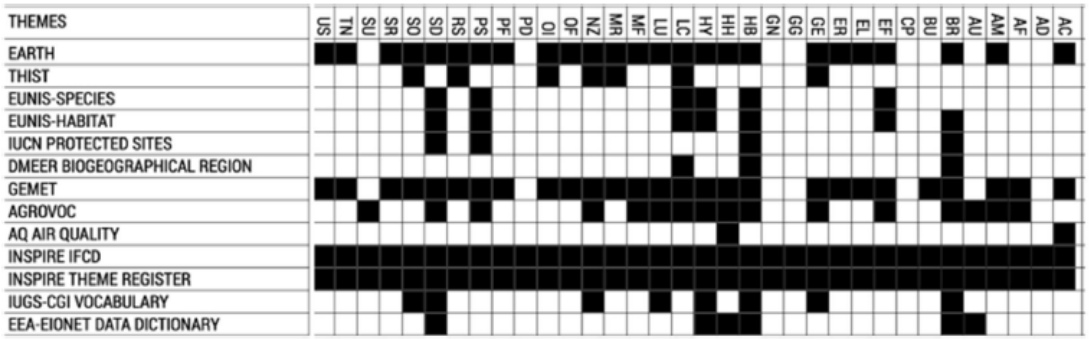



**Figure 19**: Coverage of INSPIRE themes by the resources integrated into LusTRE [39]

It can be concluded that the content of LusTRE is very well fit for this study.

LusTRE provides a common interface to all the integrated resources in a graphical user interface[40], through a web API and directly through a SPARQL endpoint. The example shows that the INSPIRE Feature Concept register is integrated. However, there is no mapping available to the other resources in LusTRE.

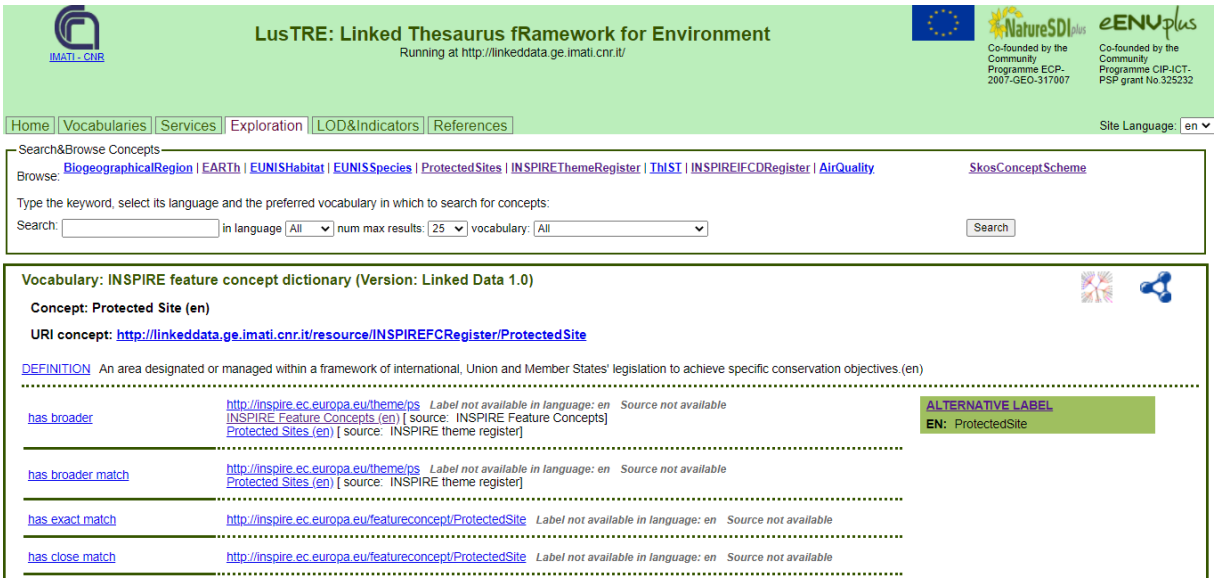



**Figure 20**: INSPIRE Protected Site (INSPIRE Feature Concept Register) in the LusTRE web interface

Finally, besides the resources represented in LusTRE, the study also briefly looks at the OSM map feature list[41]. This is a simple hierarchical structure vocabulary, mostly 2 or 3 levels deep. The OSM map feature list will not be considered a source that can automatically extract new synonyms.

38 https://doi.org/10.1007/s12145-018-0344-8
39 https://doi.org/10.1007/s12145-018-0344-8
40 http://linkeddata.ge.imati.cnr.it/exploration.jsp
41 https://wiki.openstreetmap.org/wiki/Map_Features

Instead, it will be examined if the terms used in the OSM feature list end up in the list of synonyms, hypernyms and hyponyms collected through the other methods explained.

### 3.2.2 Use of generic synonyms thesauri (NLP, WordNet ...)

Connecting an INSPIRE object type to existing vocabularies and ontologies creates added value. But these resources often use formal language. Natural Language Processing (NPL) thesauri can close the gap to a more common language. *Thesaurus.com*, *Wiktionary.org*, *Merriam-webster.com* and *WordNet* are some common thesaurus examples.

#### *3.2.2.1 WordNet*

In this study, WordNet is chosen because it is free and easily accessible, and several mappings to WordNet are already present in some of the ontologies used. But the main reason to use it is its structure using *synsets*.

WordNet[42] groups words into "*synsets*", sets of synonyms matching a certain definition. A word with more than one meaning is part of one *synset*. Relations connect synsets, not separate words. As a result, once the correct *synset* is selected, one knows that all other terms present in that *synset* are valid synonym candidates. No extra check on these terms is needed. **Figure 21** provides a visual presentation of the synsets related to the term "school". Red dots are synsets as noun; green indicates synsets as verb. This visualisation is created in WordVis[43].

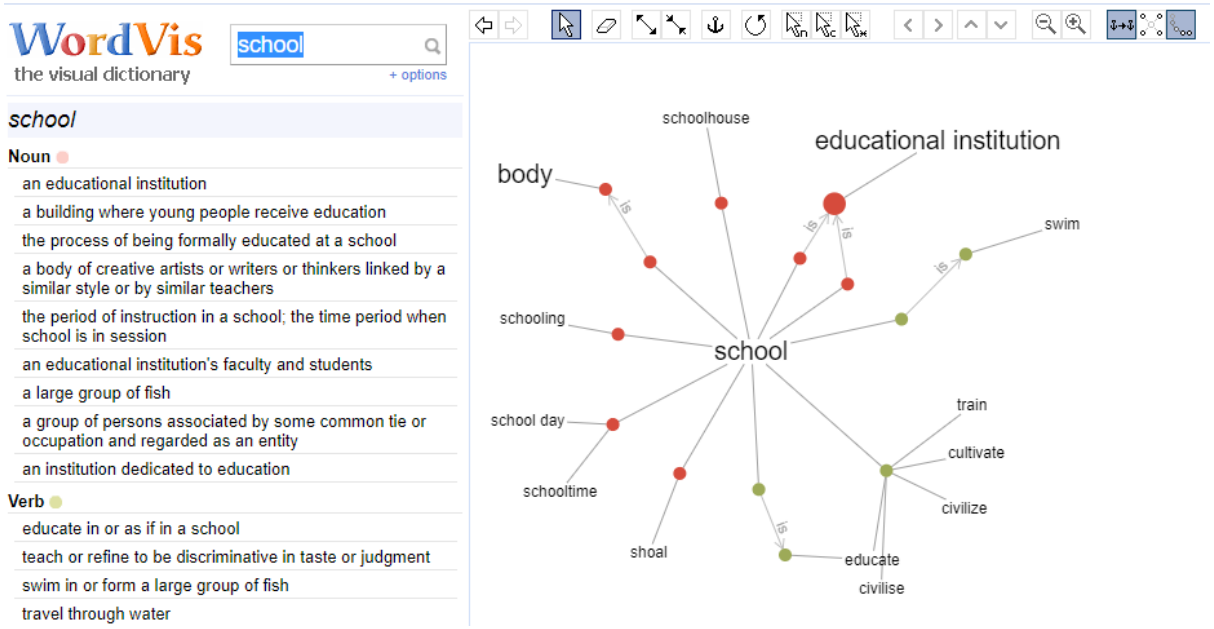



**Figure 21:** Visualisation of the term "school" in WordNet

WordNet has some known shortcomings. The coverage of compound words is not great. Besides that, domain-specific language is not extensively covered. Therefore also, some wiki sources are considered. These crowdsourced datasets contain a broad spectrum of information, and labels can range from 1 letter to complete sentences. In this study, DBPedia and Wikidata are integrated.

---

42 https://wordnet.princeton.edu/
43 http://wordvis.com/

### *3.2.2.2 DBPedia*

DBpedia is a cross-domain ontology, which has been manually created based on the most commonly used info boxes within Wikipedia. DBpedia contains more than 4.200.000 instances[44]. While most collections contain mainly concepts describing object classes, DBpedia also contains many class instances. For example, more than 700.000 places are present in DBpedia. Other vocabularies mostly only contain the concepts' country', 'city', etc. DBpedia has separate concepts for 'Paris', 'New York', 'Belgium', 'Germany', etc.

DBpedia doesn't explicitly define synonyms. Instead, the redirect information can be used. This information indicates Wikipedia terms that all redirect to the same Wikipedia page. For example (**Figure 22**), "*street network*" and "*road network*" both land on the page "*https://en.wikipedia.org/wiki/Street_network*". As shown in Figure 22 for the input term "INSPIRE", each original term has a redirect page. Redirects are used for synonyms and abbreviations, different spelling forms (also incorrect ones), etc. As a result, DBpedia terms always need review before accepting them as a synonym.


**Figure 22**: Wikipedia: Street network redirected from search term Road network

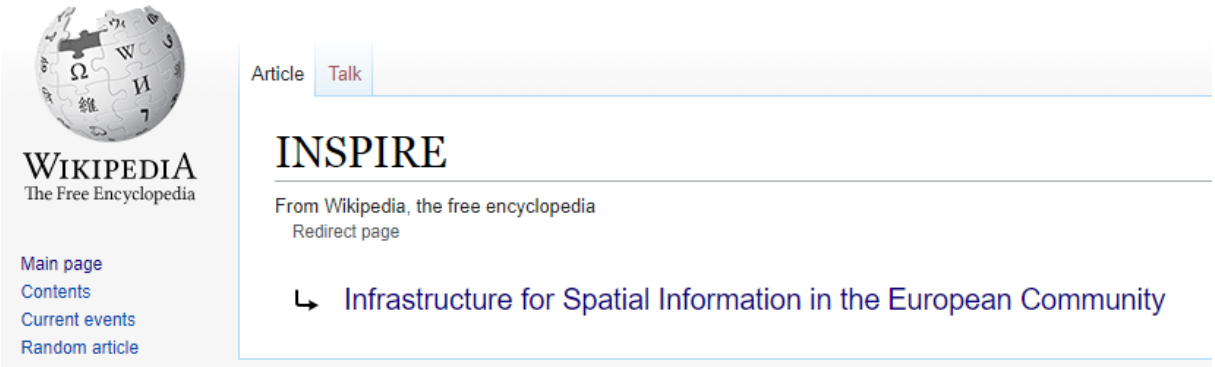
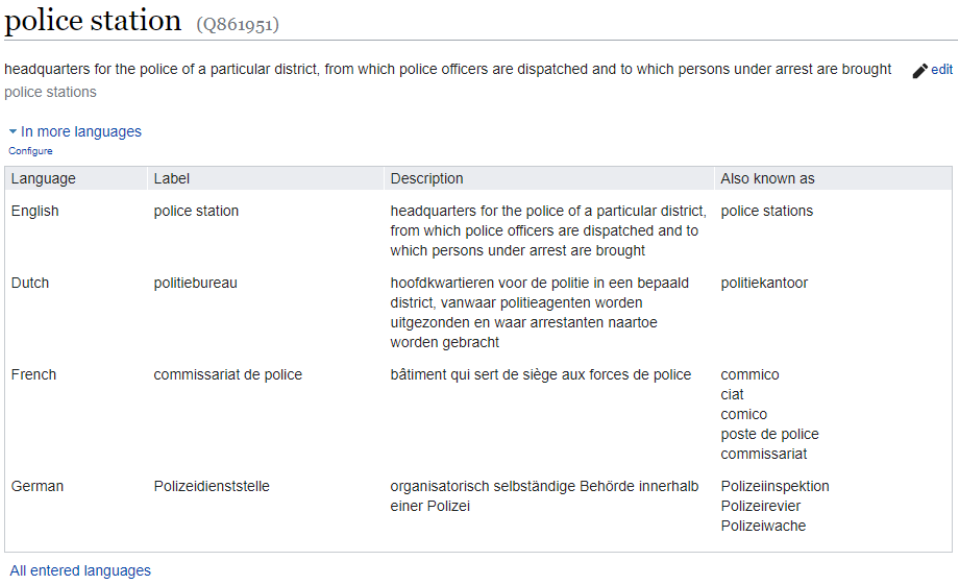



**Figure 23**: Wikipedia: the redirection page for INSPIRE.

[44] https://wiki.dbpedia.org/services-resources/ontology

### *3.2.2.3 Wikidata*

Finally, Wikidata is examined as well. Wikidata is a free and open knowledge base that acts as central storage for the structured data of projects like Wikipedia, Wiktionary, Wikisource etc.[45] Wikidata is more structured than DBpedia, making it easier to reuse the data. Concerning this study, it is important that Wikidata also includes the identifiers for the same concept in different resources. This renders Wikidata into a hub, connecting many dictionaries, thesauri and other data sources. As such, Wikidata can also form a link to OpenStreetMap because it contains the property 'OpenStreetMap tag or key', which is present for almost 3000 Wikidata items. The example in **Figure 24** gives part of the content for the concept '*police station*':



police station (Q861951)

headquarters for the police of a particular district, from which police officers are dispatched and to which persons under arrest are brought edit

police stations

In more languages

Configure

| Language | Label | Description | Also known as |
|---|---|---|---|
| English | police station | headquarters for the police of a particular district, from which police officers are dispatched and to which persons under arrest are brought | police stations |
| Dutch | politiebureau | hoofdkwartieren voor de politie in een bepaald district, vanwaar politieagenten worden uitgezonden en waar arrestanten naartoe worden gebracht | politiekantoor |
| French | commissariat de police | bâtiment qui sert de siège aux forces de police | commico<br>ciat<br>comico<br>poste de police<br>commissariat |
| German | Polizeidienststelle | organisatorisch selbständige Behörde innerhalb einer Polizei | Polizeiinspektion<br>Polizeirevier<br>Polizeiwache |

All entered languages

**Figure 24:** Wikidata: concept police station

Additional information is available for the same concept, such as superclasses, OpenStreetMap tags, and identifiers in other data sources. **Figure 25** indicates two superclasses, the OpenStreetMap tag and the equivalent class in schema.org. Besides this, identifiers for other sources like Freebase, GeoNames, Nomenclatura for Museum Cataloging and several others are available.

[45] https://www.wikidata.org/wiki/Wikidata:Main_Page

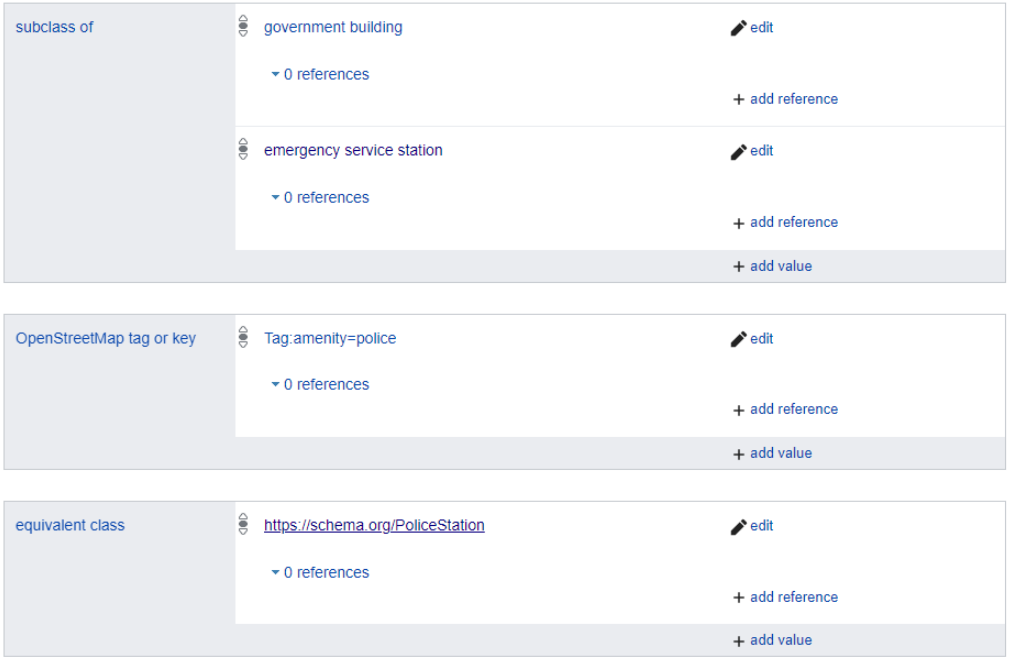



**Figure 25:** Wikidata: additional statements for "police station"

### 3.2.3 INSPIRE specific resources

The last two resources are specific to the INSPIRE framework and the INSPIRE Geoportal.

#### *3.2.3.1 Log files from Find Your Scope*

The web analytics tool on the Find Your Scope server logs all the search terms requested through the webpage. If these search terms can be linked to the object type selected by the user, this connection between the search term and spatial object might be preserved to propose an alternative term for the dataset.

The log files of a search function might link the search term(s) to the search results. Although there is a link between the search terms and the result, it is difficult to deduct information about the type of relation between search terms and resulting search results in an automated way.

#### *3.2.3.2 Using transformation schema from the user community*

In general, INSPIRE data providers manage their geospatial information in their data model and format that does not directly correspond with the INSPIRE data model. As a result, the data provider must apply a data transformation to create INSPIRE compliant data sets. This process is generally known as Extract, Transform and Load (ETL). An ETL process reads data from a source and writes it to a destination where it is presented in another format. The data representation needs to be changed to destination format between reading and writing. The ETL process is guided by a project file containing information about the mapping, renaming and transforming data files or tables, attributes and their values etc.

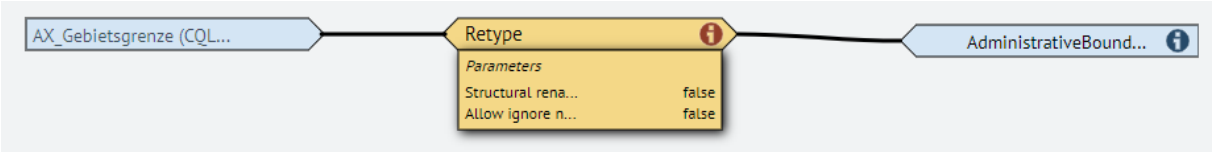


**Figure 26**: Visualisation of a mapping in Hale tool

The goal is to examine whether the information stored in the ETL project can link the INSPIRE objects to the original data and identify the original data name as a synonym for the related INSPIRE object type.

## 3.3 Search for matches and harvest synonyms: a practical approach

With those different synonym resources defined, the next question is: *How can these synonyms be harvested efficiently*?

After identifying and selecting resources to use, the next steps in the methodology are to explore these resources and harvest synonyms. In the schema, in figure 11, these steps are separated in

- Step C: search lexical matches
- Step D: search semantic matches
- Step E: harvest synonyms

The different proposed resources should not be considered as standalone solutions. The heterogeneity of thematic domains, ontologies, and entered search terms asks for a combined approach. The different methods can be combined in parallel or in an iterative way. Parallel means that each approach is applied on the same input. Iterative means that the output (concept or term) from one method is used as input for another method.

The start of the procedure is the prepared list of input terms. In this study, spatial object types are identified as concepts in the INSPIRE Concept Feature register and labelled as *literal* starting term.

The different sources can provide the following information related to a concept or term as input, as shown in **Table 1**.

**Table 1:** Overview of the output for the different information sources

| **Information source** | Concepts | Synonym, hypernym, hyponym terms |
|---|---|---|
| **Original Spatial data type** | 1 concept in the IFC register | 1 Label for that concept |
| **INSPIRE documentation** | | Hyponyms by using related layers and/or code lists |
| **Vocabularies/Ontologies** | Term > concept:<br>Matching/searching concepts by lexical matching of labels<br>Concept > concept:<br>Synonyms by traversing 'exactMatch' or 'closeMatch' connections<br>Hypernyms by traversing 'broader' connections<br>Hyponyms by traversing 'narrower' connections | Synonym, hypernym, hyponym or related terms using the preferred, alternative and hidden labels from the concept itself and/or from the found concepts |

| | | |
|---|---|---|
| | Other related concepts by 'related' connections | |
| **WordNet** | -synset(s) containing the original term | Synonyms, hyponyms or hypernyms according to the synsets found from an input term |
| **DBpedia** | - | Candidate synonyms from the redirection information |
| **Wikidata** | Matching Wikidata concept<br>The link to related concepts in other data sources | Synonyms "also known as"<br>"Subclass of" indicates hypernyms |
| **Logfiles from Find in Scope** | - | Related terms if search term can be linked to the objects retained in the end |
| **Transformation information** | | Candidate synonym, hypernym or hyponym terms by looking at the names of the data sources that are transformed to a spatial object type |

Each applied method results in 0 or more additional concepts and/or terms related to the original. After each applied method, other methods can be applied to the original concept/label or the results from a previous step. In the latter case, consecutive processing should be applied with care to avoid 'meaning drifting': minor differences between input and output might become significant after a few iterations.

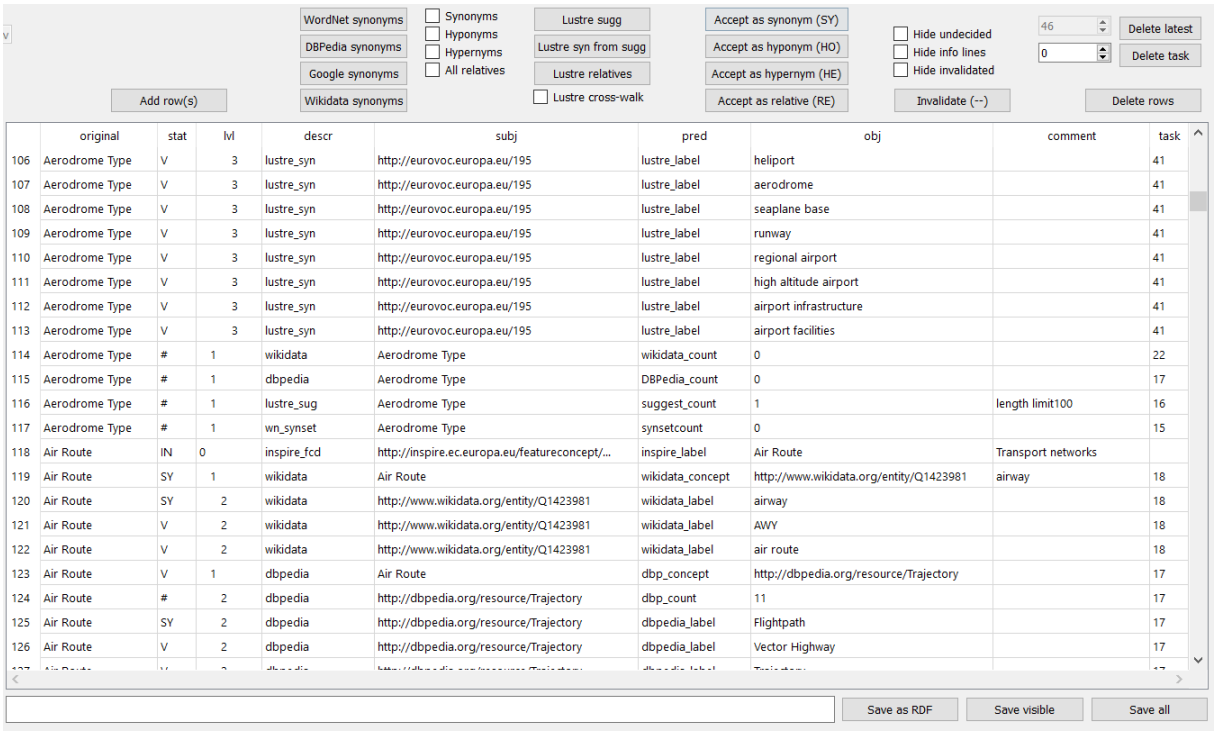
The combination of different resources, different ways of linking (lexical or semantic) and the option to work iteratively result in a difficult process to manage. The danger of meaning drift stresses the importance to keep track of the different steps taken in the process. To tackle this complexity, a *Synonyms finder* tool is developed. Its functionality focusses on:

- providing easy access to the different resources
- providing information on all steps taken
- allowing the operator to make informed decisions on accepting or rejecting results
- allowing manual intervention on each process step
- selecting approved results
- linking directly to the web interfaces of the different resources
- loading input and saving output in CSV format
- saving results in RDF format

Selection-approved results (and saving them) can be considered the synonyms' harvesting.

**As part of the output from this study, the *Synonyms finder* is available on Joinup**[46]. The tool itself can be downloaded[47], and a Quick guide[48] is also available to get started with the tool.

As illustrated in **Figure 27**, the *Synonyms finder* provides a tabular overview showing the steps and related results.




| | original | stat | lvl | descr | subj | pred | obj | comment | task |
|---|---|---|---|---|---|---|---|---|---|
| 106 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | heliport | | 41 |
| 107 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | aerodrome | | 41 |
| 108 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | seaplane base | | 41 |
| 109 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | runway | | 41 |
| 110 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | regional airport | | 41 |
| 111 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | high altitude airport | | 41 |
| 112 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | airport infrastructure | | 41 |
| 113 | Aerodrome Type | V | 3 | lustre_syn | http://eurovoc.europa.eu/195 | lustre_label | airport facilities | | 41 |
| 114 | Aerodrome Type | # | 1 | wikidata | Aerodrome Type | wikidata_count | 0 | | 22 |
| 115 | Aerodrome Type | # | 1 | dbpedia | Aerodrome Type | DBPedia_count | 0 | | 17 |
| 116 | Aerodrome Type | # | 1 | lustre_sug | Aerodrome Type | suggest_count | 1 | length limit100 | 16 |
| 117 | Aerodrome Type | # | 1 | wn_synset | Aerodrome Type | synsetcount | 0 | | 15 |
| 118 | Air Route | IN | 0 | inspire_fcd | http://inspire.ec.europa.eu/featureconcept/... | inspire_label | Air Route | Transport networks | |
| 119 | Air Route | SY | 1 | wikidata | Air Route | wikidata_concept | http://www.wikidata.org/entity/Q1423981 | airway | 18 |
| 120 | Air Route | SY | 2 | wikidata | http://www.wikidata.org/entity/Q1423981 | wikidata_label | airway | | 18 |
| 121 | Air Route | V | 2 | wikidata | http://www.wikidata.org/entity/Q1423981 | wikidata_label | AWY | | 18 |
| 122 | Air Route | V | 2 | wikidata | http://www.wikidata.org/entity/Q1423981 | wikidata_label | air route | | 18 |
| 123 | Air Route | V | 1 | dbpedia | Air Route | dbp_concept | http://dbpedia.org/resource/Trajectory | | 17 |
| 124 | Air Route | # | 2 | dbpedia | http://dbpedia.org/resource/Trajectory | dbp_count | 11 | | 17 |
| 125 | Air Route | SY | 2 | dbpedia | http://dbpedia.org/resource/Trajectory | dbpedia_label | Flightpath | | 17 |
| 126 | Air Route | V | 2 | dbpedia | http://dbpedia.org/resource/Trajectory | dbpedia_label | Vector Highway | | 17 |

**Figure 27:** Processing table in the interface

To test the validity of the different approaches, they are executed using the available web interfaces or manually looking for values in the available sources. These user interfaces provide a good overview of the available information, showing whether the different sources can provide synonyms, hyponyms or hypernyms. These interfaces are not well fit for machine to machine communication. Therefore the data sources are accessed and searched automatically through services to automate the process. The selected sources offer different methods for this. These different automated search methods are integrated into the *Synonyms finder*.

It can be noted that step C, searching lexical matches, and step D, searching semantic matches, can be executed in reversed order. First, searching semantic matches is recommended if the input list of terms already has several connections with the search vocabulary. However, in this INSPIRE test case, such relationships are almost entirely missing. Therefore the starting point must be lexical matching to provide a first link to the vocabulary.

46 https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-finder#q5

47 https://joinup.ec.europa.eu/rdf_entity/http_e_f_fdata_ceuropa_ceu_fw21_f3e1c3ffd-9aab-4676-8946-ed182f3b3a76

48 https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-finder-get-started

# 4 Output sample data

The processing tool creates a dataset in CSV format that reflects the table displayed in the interface. The file can be reloaded for the user to continue working on it. The second dataset in RDF format contains only the 'end result': where relevant links between INSPIRE registry entities and related concepts in other resources, and preserved alternative terms for the INSPIRE concepts. For each of the three use cases, noise, agriculture and water, two CSV files and an RDF file are provided. Additionally, results are visualised online using the Flourish tool[49]. **Figure 28** shows the overview in Flourish.

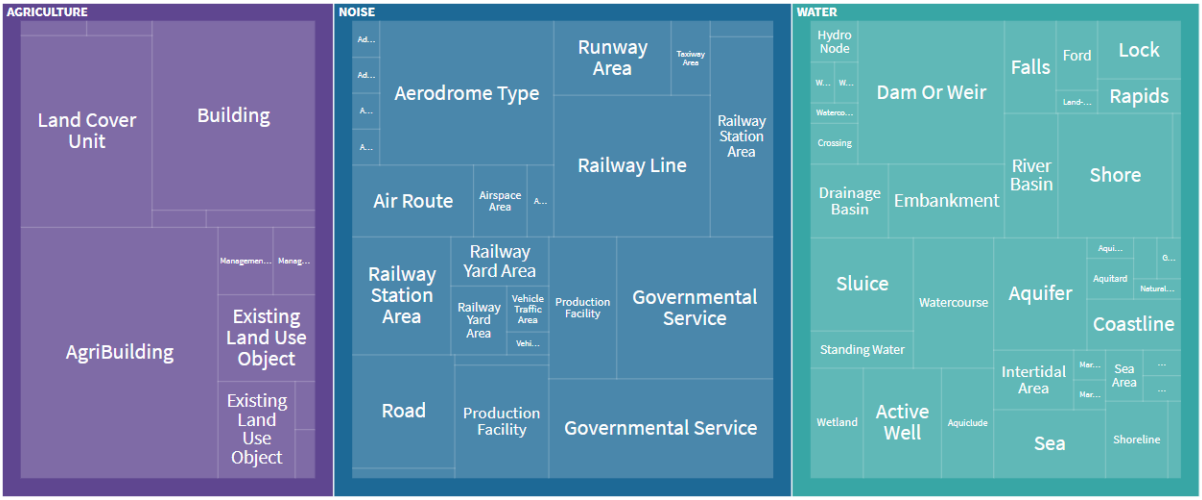



**Figure 28:** Visualisation of the results in Flourish

## 4.1 Structure of the CSV datasets

The process to create a list of synonyms takes several steps. Synonyms are collected from the different resources mentioned before. Besides that, manual interaction is sometimes needed to get the process started. The synonyms tool builds a table containing the most important information regarding the methods used in the workflow and the order in which they are used. An example could be seen in **Figure 27**. That table helps to understand the workflow used to create a result. It is important to replicate the process and understand the origin from the different resulting terms and identify possible issues.

The table can be saved in CSV format with a vertical separator '|'. The table can be re-imported in the developed tool, allowing further process results. Saving and importing intermediate results as CSV is also useful if different domain specialists assess the terms in the table.

An example of the CSV file imported in Excel is shown in **Figure 29**.

[49] https://public.flourish.studio/visualisation/6075551/

| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| **original** | **stat** | **lvl** | **descr** | **subj** | **pred** | **obj** | **comment** | **task** |
| Runway Area | IN | 0 | inspire_fcd | http://inspire.ec.europa.eu/featureconcept | inspire_label | Runway Area | Transport networks | |
| Runway Area | SY | 2 | wikidata | Runway | wikidata_concept | http://www.wikidata.org/entity/Q1 | runway | 18 |
| Runway Area | SY | 3 | wikidata | http://www.wikidata.org/entity/Q184590 | wikidata_label | landing strip | | 18 |
| Runway Area | SY | 2 | dbpedia | Runway | dbp_concept | http://dbpedia.org/resource/Runway | | 17 |
| Runway Area | SY | 3 | dbpedia | http://dbpedia.org/resource/Runway | dbpedia_label | Landing strip | | 17 |
| Runway Area | SY | 3 | dbpedia | http://dbpedia.org/resource/Runway | dbpedia_label | Airplane Landing Field | | 17 |
| Runway Area | SY | 3 | dbpedia | http://dbpedia.org/resource/Runway | dbpedia_label | Runway strip | | 17 |
| Runway Area | HO | 3 | dbpedia | http://dbpedia.org/resource/Runway | dbpedia_label | Parallel runway | | 17 |
| Runway Area | RE | 2 | lustre_sug | Runway | lustre_sug_concept | http://eurovoc.europa.eu/195 | runway | 16 |
| Taxiway Area | IN | 0 | inspire_fcd | http://inspire.ec.europa.eu/featureconcept | inspire_label | Taxiway Area | Transport networks | |
| Taxiway Area | SY | 2 | dbpedia | Taxiway | dbp_concept | http://dbpedia.org/resource/Taxiway | | 17 |
| Taxiway Area | SY | 3 | wn_synonym | taxiway.n.01 | wn_label | taxi_strip | | 15 |
| Railway Line | IN | 0 | inspire_fcd | http://inspire.ec.europa.eu/featureconcept | inspire_label | Railway Line | Transport networks | |
| Railway Line | SY | 1 | wikidata | Railway Line | wikidata_concept | http://www.wikidata.org/entity/Q7 | railway line | 18 |
| Railway Line | SY | 2 | wikidata | http://www.wikidata.org/entity/Q728937 | wikidata_label | rail line | | 18 |
| Railway Line | SY | 2 | wikidata | http://www.wikidata.org/entity/Q728937 | wikidata_label | railroad line | | 18 |
| Railway Line | A | 2 | dbpedia | http://dbpedia.org/resource/Glossary_of_r | dbpedia_label | Rail line | | 17 |
| Railway Line | SY | 1 | lustre_sug | Railway Line | lustre_sug_concept | http://linkeddata.ge.imati.cnr.it/res | Railway Line | 16 |
| Railway Line | SY | 1 | lustre_sug | Railway Line | lustre_sug_concept | http://linkeddata.ge.imati.cnr.it/res | railway line | 16 |
| Railway Line | SY | 1 | lustre_sug | Railway Line | lustre_sug_concept | http://eurovoc.europa.eu/3430 | railway line | 16 |
| Railway Line | HE | 2 | lustre_syn | http://eurovoc.europa.eu/3430 | lustre_label | rail network | | 16 |
| Railway Line | SY | 2 | lustre_syn | http://eurovoc.europa.eu/3430 | lustre_label | railway track | | 16 |
| Railway Line | SY | 2 | wn_synonym | railway.n.01 | wn_label | railway | | 15 |
| Railway Line | SY | 2 | wn_synonym | railway.n.01 | wn_label | railroad | | 15 |
| Railway Line | HE | 2 | wn_synonym | railway.n.01 | wn_label | railway_system | | 15 |
| Railway Line | SY | 1 | wn_synset | Railway Line | wn_synset | line.n.14 | | 15 |
| Railway Line | SY | 2 | wn_synonym | line.n.14 | wn_label | rail_line | | 15 |

**Figure 29**: Resulting CSV format opened in Excel (part of noise use case)

When a processing step is executed on a particular start row, the tool will add new rows to the table. Each new row builds on the start row and adds other data resulting from the process. The content of the different fields in each row is explained in **Table 2**

**Table 2:** Structure of the CSV datasets

| | | |
|---|---|---|
| **Original:** | The value in this column is copied from the start row. It usually contains the original INSPIRE concept label. This allows directly to see to which concept the row is related. | |
| **Status (stat):** | This indicates if a line is a comment ('#') or a possible valid process output ('V'). The user can invalidate rows ('--') or select them as results to harvest from the process. For harvesting, the user can indicate the relation between the output and the original input term: | |
| | IN | INPUT: this row is input for the process |
| | SYX | Exact synonym: This row contains an exact match (stronger than SY) |
| | SY | Synonym: This row contains a synonym or close matching concepts |
| | HE | Hypernym: this row contains a hypernym |
| | HO | Homonym: this row contains a homonym |
| | RE | Relation: The term in this row is related to the input, but the relation is not a synonym, hypernym or homonym |
| **Level (lvl):** | The level row simulates a tree-view like behaviour. The original input gets level 0. If a process is run on a row, the results get one level higher than the start row. Within a process, sub-processes can create additional levels. For example, | |

| | |
|---|---|
| | if related concepts are added, the concept itself gets start-level +1. The labels for that concept get start-level +2. This indicates that the labels are harvested from the concept. The indication of the level is also important once the user starts to iterate processing steps. |
| **Description (descr):** | The description row indicates the process used to get to this result. |
| **Subject (subj):** | The starting point for this row. In general, this is the input of the processing step leading to this row. |
| **Predicate (pred):** | Indicates the relation between object and subject. These relations are formulated in a form indicating the processing step. Most of them can be translated to the standard RDFs relations (exact match, narrower (match), broader (match), preferred label, alternative label). |
| **Object (obj):** | This is the result of the processing step, and it has the relation indicated in the predicate to the subject. In general, this is an additional related concept (represented by its URI) or a label as a new candidate synonym. |
| **Comment:** | The comment line contains information that might be useful for the operator. Some processing steps automatically add a comment. The operator can also add comments manually. |
| **Task:** | Each time the operator pushes an execution button, all resulting rows get the same task number. This is useful to replicate a workflow, but it can also be used to remove/undo the processing step (e.g. remove the results if the process was started in the wrong way). |

The rows' subject-predicate – object' contain the main results of each processing step. When a process starts, the object from the start row becomes the subject of the result. The predicate and new object are results from the process.

The CSV files'*_final_all.csv'* contains the complete information as shown in the *Synonyms finder* interface. The file *'*_final_select.csv'* only contains the rows accepted by the user. More information on the tool and its CSV structure can be found on Joinup[50]. The CSV results [51]for the three use cases are also available on Joinup.

## 4.2 Structure of the RDF datasets

CSV is a well-known and easily accessible data format. But there are better ways to present semantic information. The Resource Description Framework (RDF) is a standard model for data interchange on the Web. RDF especially supports linking structures and is, therefore, the ideal choice to share the results of this study. The Simple Knowledge Organisation System (SKOS) is a common data model for sharing and linking data sources. SKOS provides, among other things, a standard definition for relations between concepts. The combination of RDF and SKOS provides a format known to most knowledge systems. Although RDF is mainly developed for a machine to machine communication, it is also human-readable, as shown in **Figure 29**.

[50] https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-finder

[51] https://joinup.ec.europa.eu/rdf_entity/http_e_f_fdata_ceuropa_ceu_fw21_fe436aa2e-b7dc-47bc-b44e-045df8d6c7c6

```
<rdf:Description rdf:about="http://inspire.ec.europa.eu/featureconcept/LandWaterBoundary">
</rdf:Description>
<rdf:Description rdf:about="http://inspire.ec.europa.eu/featureconcept/Lock">
   <skos:closeMatch rdf:resource="http://www.wikidata.org/entity/Q105731"/>
   <skos:altLabel xml:lang="en">sluice</skos:altLabel>
   <skos:hiddenLabel xml:lang="en">canal lock</skos:hiddenLabel>
   <skos:closeMatch rdf:resource="http://linkeddata.ge.imati.cnr.it/resource/EARTh/43010"/>
</rdf:Description>
<rdf:Description rdf:about="http://inspire.ec.europa.eu/featureconcept/Rapids">
   <skos:closeMatch rdf:resource="http://www.wikidata.org/entity/Q695793"/>
   <skos:closeMatch rdf:resource="http://linkeddata.ge.imati.cnr.it/resource/EARTh/113620"/>
</rdf:Description>
<rdf:Description rdf:about="http://inspire.ec.europa.eu/featureconcept/RiverBasin">
   <skos:closeMatch rdf:resource="http://linkeddata.ge.imati.cnr.it/resource/EARTh/13720"/>
   <skos:closeMatch rdf:resource="http://aims.fao.org/aos/agrovoc/c_8334"/>
   <skos:closeMatch rdf:resource="http://linkeddata.ge.imati.cnr.it/resource/EARTh/29370"/>
   <skos:altLabel xml:lang="en">fluvial basin</skos:altLabel>
   <skos:altLabel xml:lang="en">watershed</skos:altLabel>
</rdf:Description>
<rdf:Description rdf:about="http://inspire.ec.europa.eu/featureconcept/Shore">
   <skos:closeMatch rdf:resource="http://www.wikidata.org/entity/Q468756"/>
   <skos:altLabel xml:lang="en">bank</skos:altLabel>
   <skos:hiddenLabel xml:lang="en">riverbank</skos:hiddenLabel>
   <skos:hiddenLabel xml:lang="en">stream-bank</skos:hiddenLabel>
   <skos:hiddenLabel xml:lang="en">coast</skos:hiddenLabel>
   <skos:hiddenLabel xml:lang="en">beach</skos:hiddenLabel>
   <skos:closeMatch rdf:resource="http://linkeddata.ge.imati.cnr.it/resource/EARTh/131950"/>
   <skos:altLabel xml:lang="en">shoreline</skos:altLabel>
   <skos:altLabel xml:lang="en">coastline</skos:altLabel>
   <skos:closeMatch rdf:resource="http://aims.fao.org/aos/agrovoc/c_1700"/>
   <skos:hiddenLabel xml:lang="en">Oceanfront</skos:hiddenLabel>
   <skos:hiddenLabel xml:lang="en">Beachline</skos:hiddenLabel>
   <skos:hiddenLabel xml:lang="en">Beachfront</skos:hiddenLabel>
   <skos:hiddenLabel xml:lang="en">Sea shore</skos:hiddenLabel>
</rdf:Description>
```

**Figure 28:** RDF file format example, part of the output for the water use case.

Providing RDF results with SKOS makes them directly usable in combination with the original data sources. The translation of the CSV results to RDF is as follows:

**Table 3:** RDF tags used for different status values

| Status value | RDF tag for concepts | RDF tag for labels |
|---|---|---|
| Exact Synonym (SYX} | skos:exactMatch | skos:altLabel |
| Synonym (SY) | skos:closeMatch | skos:altLabel |
| Hyponym (HO) | skos:narrowMatch | skos:hiddenLabel |
| Hypernym (HE) | skos:broadMatch | skos:hiddenLabel |
| Relation (RE) | skos:relatedMatch | Skos:hiddenLabel |

The RDF file contains an *rdf:Description* element for each entry in the original list of terms. The subject-predicate –object information of the CSV file is added to these elements, but only for those entries accepted by the operator (this matches the content of the *_final_selected.csv' file). If an original input term doesn't preserve the information, the *rdf:Description* element in the file is empty.

**The synonyms tool provides two types of information: relations to concepts** in other resources **and alternative terms** harvested from other resources. If the object of an accepted line in the tool is a concept in another vocabulary, the result is a semantic relation between the INSPIRE concept and that other concept. The relations skos:exactMatch or skos:closeMatch are used for synonyms. Hypernyms are linked with the relation skos:*broadMatch* and hyponyms have a

skos*narrowMatch* connection. If the relationship is not fully clear, a more generic skos:*relatedMatch* can also be used.

On the other hand, if the object of a selected line in the table is a text label, this results in an RDF statement providing an alternative label for the original INSPIRE concept. For exact or close matches the predicate *skos:altLabel* (alternative label) is used. Hyponyms and hypernyms cannot be considered alternative terms for the original concept, and therefore these terms are provided as *skos:hiddenLabel*. A hidden label is used by machine-to-machine communication and will be used by the search engine, but it will not be visible in any graphical user interface.

It must be noted that the output's relation is always defined towards the original input term. Suppose the output is a synonym of a hyponym of the initial input. In that case, it will be indicated as a narrower (hyponym) in the RDF file, not as a synonym.

The World Wide Web Consortium (W3C) online RDF Validation service [52] is used to validate the RDF output created for the three use cases. The results are available on Joinup[53].

## 4.3 Test datasets for the selected use cases

The methodology and *Synonyms finder* tool have been tested on three test datasets related to agriculture, water and noise. Data for these domains are spread over several INSPIRE themes. The input list of terms was selected from the INSPIRE Concept Dictionary. The INSPIRE Label for each concept is used in the input list for each concept. Additionally, code values of relevant code lists are added to specify the objects types further. These code values are especially useful for INSPIRE Concepts labelled by a collective term. Sometimes additional editing is done before starting the search. A clear example is the concept "DamOrWeir", of which the label is manually split into two sub-concepts, "Dam" and "Weir"

**Figure 30** gives a graphical overview of the terms in the three use cases, visualised in Flourish[54]. The Flourish visualisation can be used to easily browse through the results for the use cases. The visualisation is created using the CSV output file of the *Synonyms finder*.

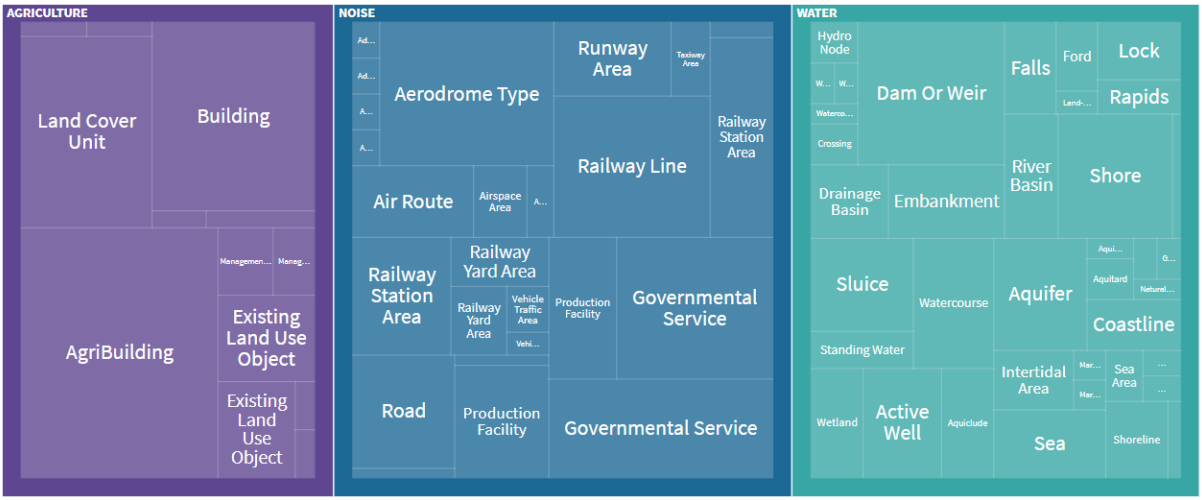



**Figure 30:** Overview of the selected terms for the three use cases, visualised in Flourish

[53] https://joinup.ec.europa.eu/rdf_entity/http_e_f_fdata_ceuropa_ceu_fw21_f9a46d378-a9b0-4171-beef-ac0b6ab4f3c8

[54] https://public.flourish.studio/visualisation/6075551/

**Table 4** gives an overview in numbers of the results. The next paragraphs describe the results by use case, showing the information retained by the user.

**Table 4:** Overview of validated results

| | Agriculture | Water | Noise |
|---|---|---|---|
| **Number of input terms** | 11 | 38 | 19 |
| **Output data (not validated)** | >1500 | >850 | >2000 |
| **Output data validated & selected** | Alternative labels: 4<br>Hidden labels: 132<br>Semantic relations: 49 | Alternative labels: 68<br>Hidden labels: 62<br>Semantic relations:71 | Alternative labels: 50<br>Hidden labels: 52<br>Semantic relations: 67 |
| **Input terms without results** | 5 | 14 | 6 |
| **% of input terms with results** | 55% | 63% | 68% |

### 4.3.1 Agriculture

For agriculture, the list of terms is loosely based on use case B1, 'Safe Plant and Animal Production' in the INSPIRE Data Specifications for Agricultural Facilities. These results in 11 INSPIRE concepts. However, those INSPIRE concepts often only partially relate to agriculture. Therefore code list values are used to focus more on agriculture-related objects. For example, for Buildings, the code list for Building Nature Value is analysed. The values greenhouse, shed, silo, and storage tank are added as a subtype for building. For land cover, INSPIRE does not provide its own code list, and instead, it refers to the standard Corine Land Cover code list. Again, only the values related to agriculture are preserved.

Processing the list in the synonyms tool provided four alternative labels and 132 hidden labels, with hidden labels mainly pointing to sub-concepts of the input term. The process also provided 49 semantic relations to the different resources.

No additional information from the selected resources is retained for five terms in the start list.

As an example, **Figure 31** shows alternative terms found for the INSPIRE concept '*Building*'. Because buildings are not restricted to the agricultural domain, only specific types of buildings are retained, resulting in a list of mainly hyponyms.


**Figure 31**: alternative terms for the INSPIRE concept Building (in the agricultural domain)

### 4.3.2 Water

A broad selection of INSPIRE concepts is selected for water, mainly from the INSPIRE data specifications for Hydrography, Geology and Sea regions. This results in a list of 38 input terms.

The list processing in the synonyms tool provided 68 alternative labels and 62 hidden labels, with hidden labels mainly pointing to sub-concepts of the input term. The process also provided 71 semantic relations to the different resources.

For 14 terms in the start list, no additional information from the selected resources is retained.

The example in **Figure 32** shows the alternative labels and concepts found for the INSPIRE concept '*Watercourse*'.

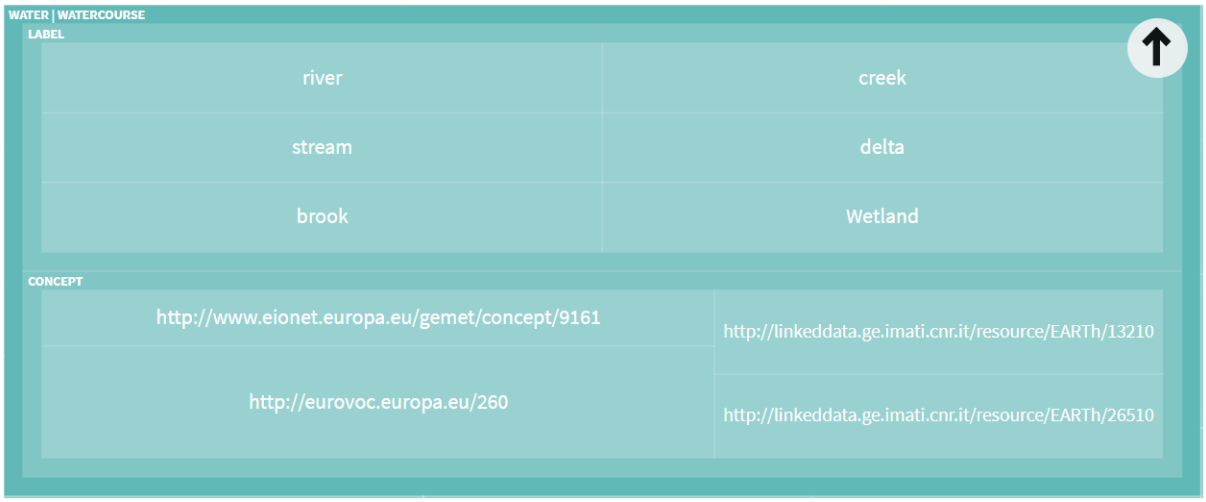



**Figure 32:** Visualisation of the results for 'Watercourse'

### 4.3.3 Noise

The terms in the start list for the noise dataset is selected based on the ongoing work for Environmental Noise Directive (END) reporting guidelines. Terms are related to noise sources and facilities affected by noise, e.g. hospitals. Nineteen concepts are selected.

The processing of the list in the synonyms tool provided 50 alternative labels and 52 hidden labels, with hidden labels mainly pointing to sub-concepts of the input term. The process also provided 67 semantic relations to the different resources.

No additional information from the selected resources is retained for six terms in the start list.

**Figure 33** shows the results for the concept '*Railway line*'.

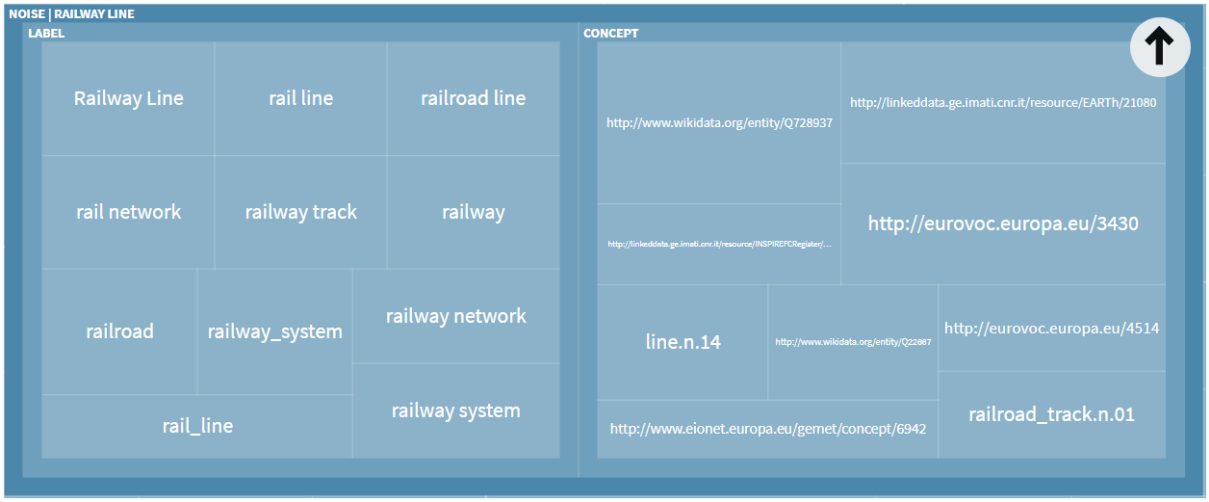



**Figure 33:** results for INSPIRE concept 'Railway line'

# 5 Analysis of the results

This study aims to provide a reusable methodology to find synonyms, not limited to the scope of INSPIRE. Therefore, the resulting datasets and the different steps in the process are analysed.

The first section shows the additional information provided by the original source of the list of keywords. The INSPIRE use case shows how the INSPIRE registry provides additional information, creating a better start point for the process.

The following paragraphs document the feasibility of the selected vocabularies and thesauri. Both the domain-specific and the generic resources are evaluated, followed by the specific INSPIRE related resources are evaluated after that. Although the final approach uses the synonyms tool, the data sources are illustrated using their web interfaces because they show all information provided by the resource. The synonyms tool provides easy access to these different web interfaces.

## 5.1 The initial list of words: INSPIRE and its documentation

Layers and code lists related to a spatial object type are identified as possible sources for related terms, mainly of the type 'narrower' or 'related'. A few examples immediately show the validity of this approach.

'Governmental services' is a collective object which not really indicates the type of real-world objects it represents. But its layers, built on the *serviceTypeValue* code list, return terms like *Fire station*, *Barrack*, *Hospital service*; **Figure 34** only shows part of the terms in the *Service Type Value* code list

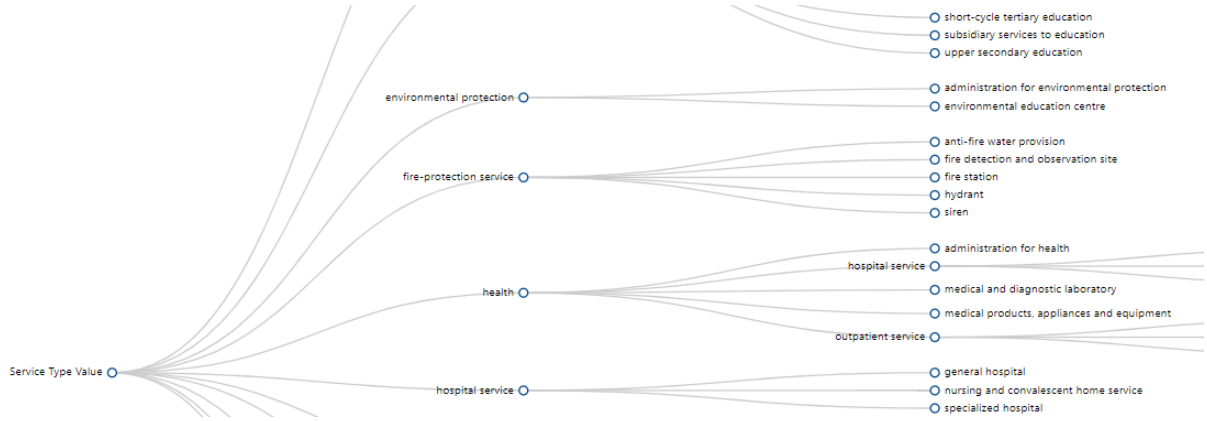



**Figure 34:** Part of the Dendogram graph for the Service Type Value Code list[55]

Another example was illustrated before when showing the results in Find your Scope of the search using *'noise'* as a search term (see Section 3.1). The narrow search applied in the Catalogue of objects doesn't provide results. The more extensive search method used in Direct Search gives five results, indicating that '*noise*' is one of the 'ENV Health Determinant Type Value' values and the 'Environmental domain' code lists. The second list is too broad, but the first list is directly related to this spatial object type. Its values (**Figure 35**) can be added as 'related' terms.

[55] https://inspire-reqadmin.jrc.ec.europa.eu/dataspecification/ScopeObjectDetail.action?objectDetailId=10210

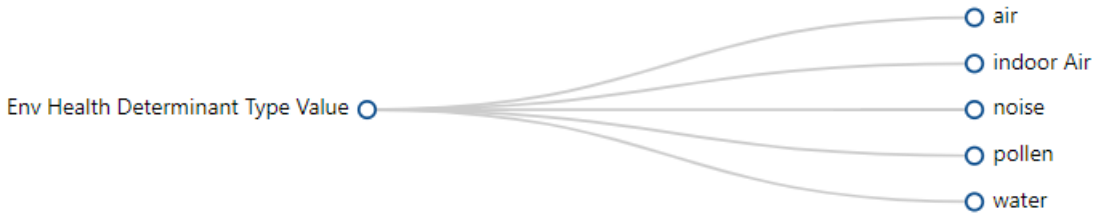



**Figure 35:** Dendogram graph for the Env Health Determinant Type Value code list[56]

This method directly provides a list of additional terms, mostly 'hyponym' (a Fire station is a type of Governmental service) or 'related' (environmental health statistical data are related to air, noise, pollen…) type. But not all code lists used in the definition of a spatial object type provide this added value. Which code lists are retained and what relation (hyponym or related) should be used cannot be decided automatically.

The importance of this additional information is clearly shown in the agriculture test case, where only four direct synonyms are retained. But 132 hidden labels, pointing mainly to sub-concepts found not starting from the original concepts but the code list values related to them.

Conclusion: In the case of INSPIRE, it is important to analyse INSPIRE registry information related to the original list of words. It provides a valuable source for hyponyms and related terms. Selecting the code lists to use needs human intervention.

| In general, it is recommended to use all information provided by the original source before searching for other resources. |
|---|

## 5.2 Selected sources to identify synonyms

### 5.2.1 Using vocabularies and ontologies

As mentioned in Section 3.2.1, the alignment of ontologies provides information on the relation between concepts in both ontologies. In short, the approach used is to align IFCD and relevant parts of the layer register (through code lists used for the layer) with the selected ontologies and vocabularies. Matches are created between the INSPIRE concepts and concepts in other ontologies. Subsequently, these matched concepts are interrogated for synonyms (alternative labels, exact matches), hypernyms (broader match) and hyponyms (narrower match). In the LusTRE platform and many other ontologies, the alignment between different ontologies is already (partially) accomplished. This alignment allows to efficiently harvest the synonyms from the linked concepts in other ontologies.

It must be noted that the final exercise doesn't include hypernyms. On the INSPIRE side, hypernyms for a spatial object type are the dataset or the INSPIRE theme. The labels used for these two are mostly compound and collective terms, and they provide very sparse results. Moreover, the INSPIRE use case explicitly wants to enhance the discoverability of object types *outside* their INSPIRE theme. Finally, the INSPIRE registries are not organised in a directly exploitable semantic structure, which implies mostly manual work to integrate it in the exercise. All these points lead to the decision not explicitly to search for hypernyms.

[56] https://inspire-reqadmin.jrc.ec.europa.eu/dataspecification/ScopeObjectDetail.action?objectDetailId=10503

**Figure 36** shows the result for "*Protected Site*". Through the link to the EARTh Thesaurus, "*Protected area*" is identified as a synonym, and through narrower relations, sites related to different protection types are available. These can serve as hyponyms (protected landscape, world heritage site…).

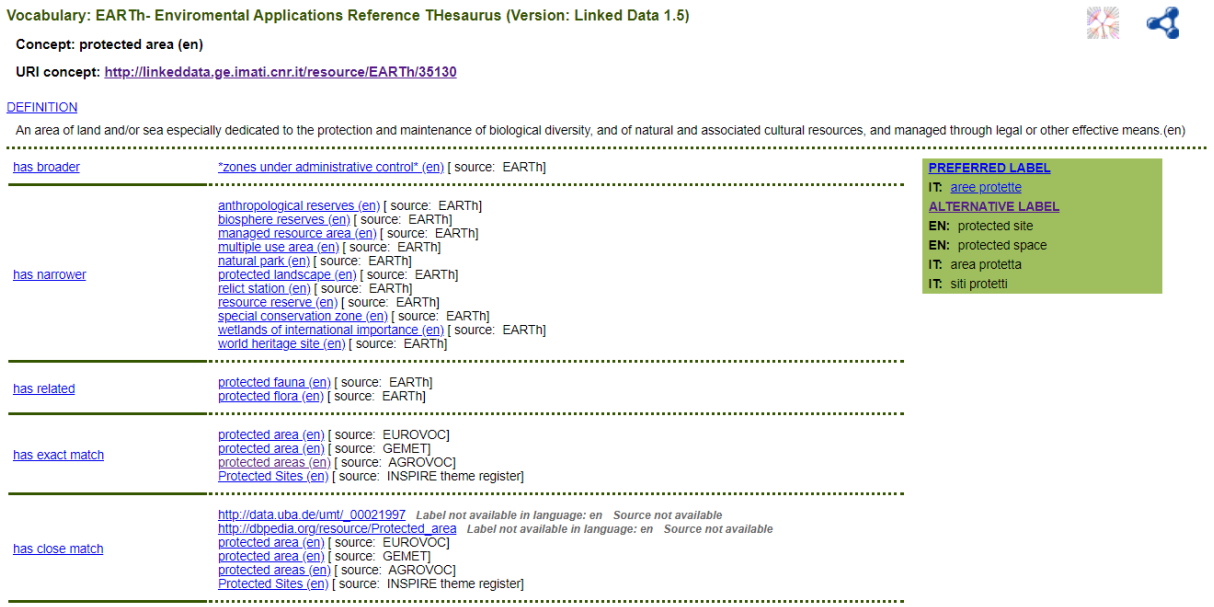



**Figure 36:** LusTRE web interface showing results for Protected Site (through its direct match with Protected Area in the EARTh thesaurus)[57]

Additional synonyms or hyponyms might be available from the exactly matched concepts in other ontologies. The link to the exact match in Eurovoc provides the results shown in **Figure 37**, showing an additional list of possible hyponyms (indicated with 'has narrower'. This example shows that the traversing of ontologies, following exact matches, can provide extra information, especially if the aligned ontologies cover different domains.

[57] http://linkeddata.ge.imati.cnr.it/resource/page/EARTh/35130?language=en

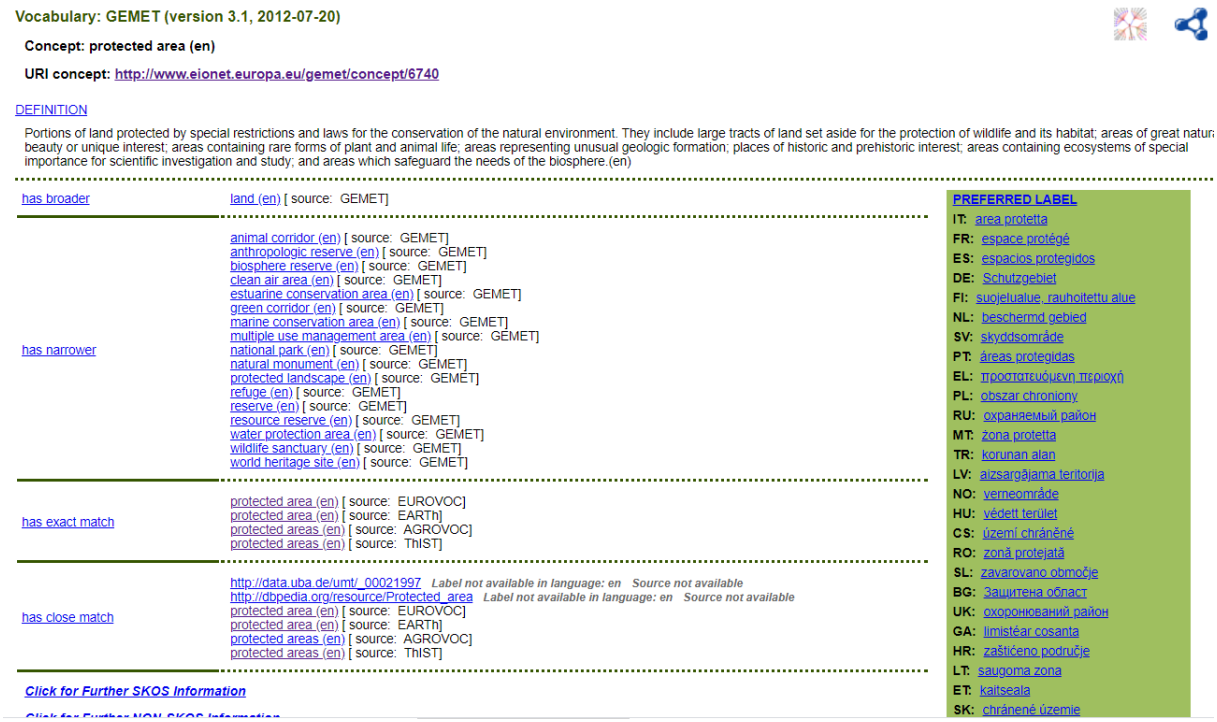



**Figure 37**: LusTRE web interface showing results for Protected Site (through its direct match with Protected Area in the GEMET[58]

It shows that the approach of discovering synonyms using the link to ontologies provides valuable results. It can be automated through the services provided by LusTRE or by direct connection with the SPARQL endpoint of LusTRE or, where available, the individual ontologies and thesauri.

Not only are the retained labels important, but the semantic link between INSPIRE and concepts in other vocabularies also has even more value for further development. Over the three use cases, 19 links to Agrovoc, 25 to Eurovoc and 21 to GEMET are retained in the test.

### 5.2.2 Using WordNet, DbPedia redirects, Wikidata

#### *5.2.2.1 Wordnet*

If the input word is known in WordNet, it will return the *synsets* containing the given name. Using the definitions of the synsets, it is then to be decided which synset(s) is (are) related to the INSPIRE concept. The other words present in the selected synsets are added as a synonym to the object type. The hypernyms and hyponyms are extracted and added to the object type for those synsets.

**Figure 38Figure 38:** Web interface of WordNet showing the results for 'railway' provides the results from Wordnet for input 'railway'. It shows 1 of the 2 synsets returned, related to the railway as a physical object or an organisation. It also shows several hyponyms (e.g. Underground, funicular)

[58] http://linkeddata.ge.imati.cnr.it/resource/page/gemet/concept/6740?language=en


**Figure 38:** Web interface of WordNet showing the results for 'railway'[59]

The results contain the new term 'railroad'. The current Catalogue of objects does not return a result for 'railroad' because most documentation uses 'railway'.

In another example, WordNet gives only very limited results for *'contour line'* (only the word '*contour*'). *'Contour line'* is a compound word, and it is more technical. Both compound words and technical terms are underrepresented in WordNet.

The synonyms tool only applies a lexical comparison of words, and therefore it is up to the user to select the correct synsets. This user selection can be made efficiently because of the systematic organisation of words in synsets.

[59] http://wordnetweb.princeton.edu/perl/webwn?s=railway&sub=Search+WordNet&o2=1&o0=1&o8=1&o1=1&o7=1&o5=1&o9=&o6=1&o3=1&o4=1&h=0

### 5.2.2.2 DBpedia

DBpedia contains a much broader set of terms, and more search attempts return a result. There is no graphical web interface to get this result, and it is acquired using a SPARQL request, as shown in **Figure 39**.


**Figure 39**: The DBpedia SPARQL query and (part of) the resulting terms from Wikipedia redirects

Contrary to WordNet, in DBpedia, there is little organisation in the results. Additionally, the alternative terms provided by DBpedia are often incorrect English to allow people to find a term even if a spelling error was made. As a result, manual filtering of results is always needed.

DBpedia provides more synonym candidates, but manual intervention is always needed to filter the output. Due to the unstructured organisation of the information retained by DBpedia, only 14 relations to DBpedia objects are retained in the tests.

### 5.2.2.3 Wikidata

Wikidata is a structured system that also supports multilingualism. One of the caveats of Wikidata is that it contains both classes and instances covering a vast field. Wikidata, for example, also contains information about individuals or movies. The following figure shows the search term “police station” results. It returns 1721 results, of which most are instances, meaning descriptions of specific police stations. It also contains the information for a television series named ‘*Police Station*’.

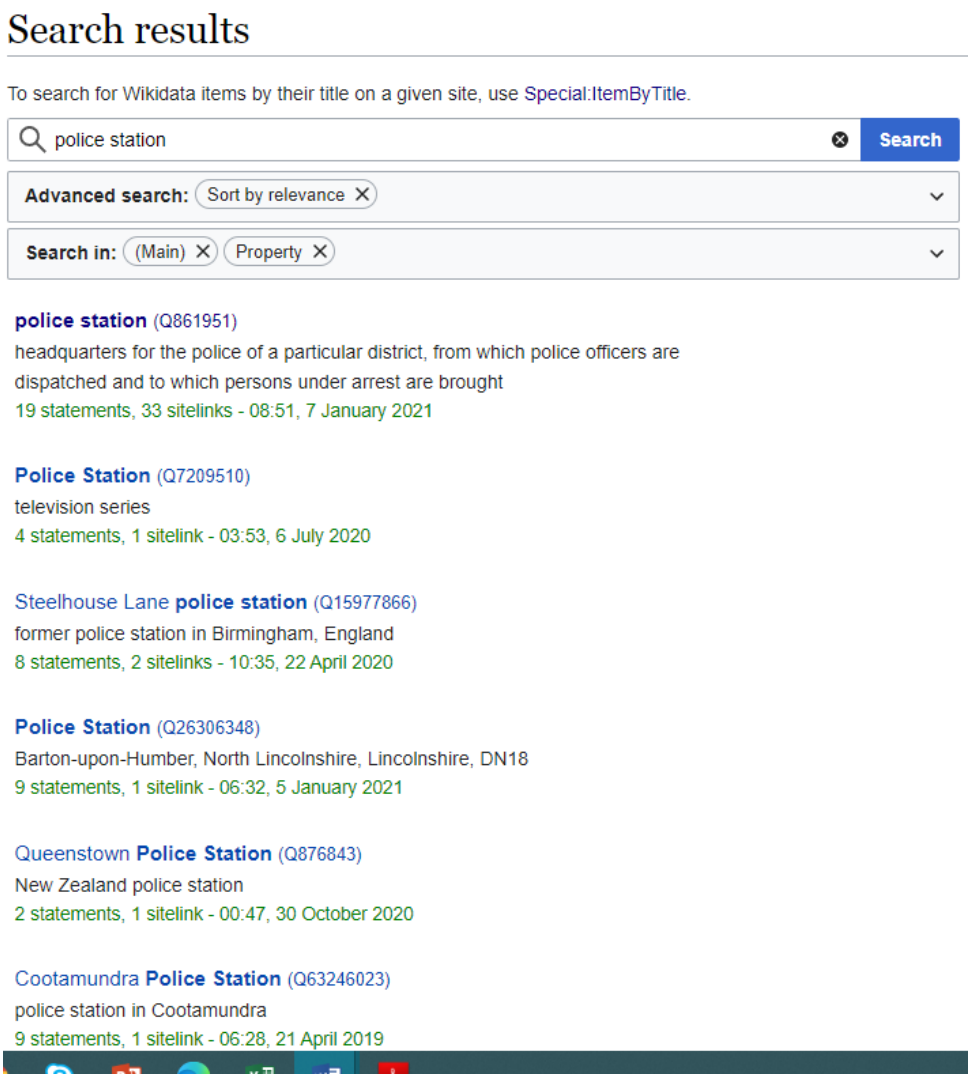



**Figure 40:** Wikidata: search results for 'police station'

For the synonyms task, results describing instances are not helpful. The Wikidata data model defines detailed property type and instance type definitions. These can be used in the SPARQL queries to filter the query results by excluding people or movies. Unfortunately, the data model does not directly define the distinction between class and instance. The query used in the synonyms tool only returns results with an attribute 'is subclass of' to make the distinction.

Wikidata can be considered a valuable source to link to. Over the 3 test cases, more than 60 connections to Wikidata were retained, which is already a significant number of links.

On top of this, Wikidata concepts often contain several relations to other sources. In this geospatial test case, concepts can have an OpenStreetMap key or tag, and the Wikidata connection allows linking INSPIRE concepts to OpenStreetMap object types. Other examples of provided links are schema.org, GeoNames feature codes, Library of Congress, Encyclopedia Britannica Online, etc. In this test, these additional links are not further explored.

### 5.2.3 INSPIRE specific sources

#### *5.2.3.1 Log files from Find your scope*

The usability of the *Find Your Scope* log files is examined starting from a list of search terms containing 500 search terms that were entered at least two times. The list doesn't contain any link to the spatial objects selected after the search. The only information about how the user behaves after entering the search term is the average time on page—the Search exits (indicating when a user does not interact with the resulting page). There is no way to indicate why a user stays on the page or does/does not interact with it.

**Table 5:** The first 15 search terms from the Find Your Scope log

| Label | Searches | Page views | Total time spent by visitors (in sec) | Exits | Search Results pages | Avg. time on page | Bounce Rate | % Search Exits | Metadata: segment |
|---|---|---|---|---|---|---|---|---|---|
| **Protected Site** | 84 | 263 | 15750 | 16 | 3.1 | 0:01:00 | 0% | 19% | siteSearchKeyword==Protected+Site |
| **Watercourse** | 53 | 128 | 10731 | 14 | 2.4 | 0:01:24 | 0% | 26% | siteSearchKeyword==Watercourse |
| **ISO 19103** | 44 | 45 | 455 | 44 | 1 | 0:00:10 | 0% | 100% | siteSearchKeyword==ISO+19103 |
| **noise** | 44 | 97 | 3506 | 17 | 2.2 | 0:00:36 | 0% | 39% | siteSearchKeyword==noise |
| **soil** | 41 | 77 | 3778 | 7 | 1.9 | 0:00:49 | 0% | 17% | siteSearchKeyword==soil |
| **soil body** | 41 | 43 | 388 | 7 | 1 | 0:00:09 | 0% | 17% | siteSearchKeyword==soil+body |
| **Address** | 38 | 71 | 6184 | 14 | 1.9 | 0:01:27 | 0% | 37% | siteSearchKeyword==Address |
| **Building** | 38 | 87 | 2958 | 12 | 2.3 | 0:00:34 | 0% | 32% | siteSearchKeyword==Building |
| **geology** | 38 | 67 | 2781 | 2 | 1.8 | 0:00:42 | 0% | 5% | siteSearchKeyword==geology |
| **watercourse** | 38 | 81 | 6035 | 14 | 2.1 | 0:01:15 | 0% | 37% | siteSearchKeyword==watercourse |
| **protected sites** | 34 | 70 | 2155 | 2 | 2.1 | 0:00:31 | 0% | 6% | siteSearchKeyword==protected+sites |
| **station** | 34 | 39 | 334 | 29 | 1.1 | 0:00:09 | 0% | 85% | siteSearchKeyword==station |
| **flood** | 32 | 57 | 3098 | 11 | 1.8 | 0:00:54 | 0% | 34% | siteSearchKeyword==flood |
| **road** | 28 | 51 | 5186 | 5 | 1.8 | 0:01:42 | 0% | 18% | siteSearchKeyword==road |
| **Cadastral Parcel** | 26 | 54 | 1987 | 11 | 2.1 | 0:00:37 | 0% | 42% | siteSearchKeyword==Cadastral+Parcel |

This lack of information makes it impossible to deduce synonyms from this list. However, the list is still helpful as it gives insight into what terms the users are entering.

One remark is that the search terms don't always refer to a spatial object but more to a domain or phenomenon. Examples are noise, waste, water, fish, dredging, climate, energy… These terms

cannot be linked to a spatial object using synonyms or hyponyms and hypernyms. This problem might be solved by considering the more generic SKOS relation type “is related”.

It can be concluded that with the information provided by the current logging system, this approach does not provide new terms that can be automatically related to spatial object types.

#### *5.2.3.2 Using transformation schema from the user community*

Many project files from the HALE ETL software from Wetransform[60] were examined to test the use of ETL transformation schemas as a source for synonyms. These files contain different mappings between the original dataset and the resulting INSPIRE compliant data. The name of the original dataset (filename, database table name) is extracted and evaluated as a possible synonym for the INSPIRE data type. The following table gives some results, showing source and target names.

**Table 6:** Hale: examples of data mappings

| | |
|---|---|
| River | Watercourse |
| bodenschutzwald | ManagementRestrictionOrRegulationZone |
| Communes | AdministrativeUnit |
| Transportation | RoadLink |
| Districts | AdministrativeUnit |
| Villages_Councils | AdministrativeUnit |
| Settlements | AdministrativeUnit |
| ROADS | RoadLink |
| Hauskoordinaten | Address |
| AX_Gebietsgrenze | AdministrativeBoundary |
| AX_Kondominium | Condominium |
| AX_Flurstueck | CadastralParcel |
| AX_Gebaeude | Building |
| AX_Turm | Building |
| AX_Bahnverkehr | ExistingLandUseObject |
| AX_Bergbaubetrieb | ExistingLandUseObject |
| AX_FlaecheBesondererFunktionalerPraegung | ExistingLandUseObject |
| AX_FlaecheGemischterNutzung | ExistingLandUseObject |
| AX_Fliessgewaesser | ExistingLandUseObject |
| AX_Flugverkehr | ExistingLandUseObject |
| AX_Friedhof | ExistingLandUseObject |
| AX_Gehoelz | ExistingLandUseObject |
| AX_Hafenbecken | ExistingLandUseObject |

[60] https://www.wetransform.to/products/haleconnect/

| | |
|---|---|
| kmmarkeringen | MarkerPost |
| vaarwegvakken | WaterwayLink |
| River | Watercourse |
| ritagliato_nuovo | Building |
| wegvakken | FunctionalRoadClass |
| hectopunten | MarkerPost |
| rijstr_wv | NumberOfLanes |
| vlakken | RoadArea |
| wegvakken | RoadLink |

Looking at these results, it seems that terms used in the original, local datasets can give valuable information but are not fit for automated processing. This was discussed with *Wetransform*, the developers of the HALE software. Their experience with many public data transformation projects confirms that using different languages, prefixes, and suffixes makes it difficult to automate HALE data mappings. Because of this, these mappings are finally not considered a data source in the automated approach. But it might be a source to select terms manually if other more automated methods fail.

## 5.3 Integrated search and human interaction

The alignment of concepts or terms must be based on the concept's meaning on both sides of the alignment. Even if the terms match exactly, alignment is not correct if the meaning (context) is different. An incorrect alignment will result in wrong synonyms. In WordNet, the synsets group words by meaning. to reach valid results, the correct synsets must be selected. It also applies to the ontology approach, where a lexical match of the concept's label does not guarantee a 100% semantic match. For DBpedia and Wikidata, the number of returned results is often large, and the DBpedia results do not contain much structure. These points indicate that it is possible to provide synonym candidates, but human interaction is still needed to make a selection.

Furthermore, the user must also decide when to include hyponyms or hypernyms and run iterative searches using the output from one step as input for the next. This decision can, for example, be based on the results already acquired at a particular moment. It is recommended not to go too far because the output can become chaotic. Combining different heterogeneous resources with different interfaces hampers the cross-resource search for synonyms, especially if one wants to work iteratively.

An automated approach must facilitate flexible access and search of the different data sources and allow human interaction between different steps.

**This flexibility is what the *Synonyms finder* provides as a tool.**

## 5.4 Semantic links open the door (but are sparse for INSPIRE concepts)

Semantic links connect concepts within an ontology or thesaurus, or they align between different resources by linking concepts between the resources. These **semantic links take meaning and context into account**. Therefore, using these links can provide more reliable results and reduce

the human interaction needed in the process. In LusTRE, these semantic links are implemented in the search for synonyms service.

Unfortunately, the number of existing direct semantic relations between the INSPIRE registry and other sources is very low. Although LusTRE is an integrated system, no semantic links exist between the IFCD and the other data sources. On the Eurovoc website, an alignment file between Eurovoc and INSPIRE is available. However, this file only contains six links between Eurovoc and an INSPIRE Theme and eight links between Eurovoc and an INSPIRE concept (on a total of 260 INSPIRE concepts). These 14 links are all exact matches.

This number of links is far too low to find synonyms by semantic relations alone. As a result, the process mainly relies on lexical matching of labels. Once a lexical match is found in another vocabulary, additional information can be harvested starting from the semantic relations in that other vocabulary.

This additional information is not limited to the resources the user starts from. For example, Wikidata contains numerous links to other thesauri like the Library of Congress or the UK Parliament thesaurus. In a geospatial context, Wikidata provides links to OpenStreetMap and GeoNames, and these links to other geospatial sources might be beneficial in the context of INSPIRE.

## 5.5 Lexical matching is not straightforward

The sources used to collect synonyms have different origins. Some are built from a technical perspective, others from a legal perspective or just natural language. Less formal, collaborative sources like DBpedia and Wikidata are also used. Still, the number of fully automatic matches is quite small. If we look at results related to natural language resources, only 18% of INSPIRE spatial object labels are present in Wordnet, and for DBpedia, this is almost 26%. These low percentages might explain the difference between common language and the more technical language used in INSPIRE.

However, it is remarkable that also in LusTRE, there are only 20% direct matches (of course, the matches to the INSPIRE FCD itself, also present in LusTRE, are not counted here). INSPIRE contains datasets related to European environmental legislation. The Eurovoc vocabulary is part of the LusTRE dataset and provides keywords directly related to European Legislation. GEMET and EARTh, also present in LusTRE, provide terminology for the environment. Therefore, one would expect that the number of matches would be larger, but it isn't, although queries towards the LusTRE system combine all these sources.

**This low number of direct matches can indicate that INSPIRE uses a very specific language**, but the analysis is also valid for other input lists.

The use of collective concepts, compound labels, and generic (geospatial) words negatively impacts the results.

### 5.5.1 Collective concepts can be complicated

A collective concept is a concept that is defined to group several sub-concepts. For some concepts, this is trivial. For example, the concept 'Dam Or Weir' can be split into the two sub-concepts, 'Dam' and 'Weir'. Another clear example is 'Oil, Gas And Chemicals Pipe', which can be split into three concepts.

Other concepts provide a generic name, grouping several subconcepts with a mutual characteristic. A good example is '*Governmental Service*'. This is a valid expression that might exist in other data sources. But the term might be too generic, and users will more often search for specific services like schools, hospitals or police stations. Adding these specific sub-concepts will produce extra links to other data sources and provide terms closer to what the user is looking for.

On the other hand, when concepts are grouped based on domain-specific criteria, it will be harder to find synonyms outside this specific context. In both cases, too generic grouping and grouping based on too specific criteria, the use of sub-concepts opens more options to find synonyms and links to other sources.

In the specific case of INSPIRE, information of sub-concepts can often be derived from code lists, as is explained before.

### 5.5.2 Compound words and generic (geospatial) words

Input terms labelled by compound words often return limited results. Most of those compound words are not present in the consulted data sources. In the INSPIRE use case, this is specifically true when one of the words indicates the spatial nature of the object. Node, area, link are such words. These words are needed to explain the nature of the spatial object in the INSPIRE context, but they have no added value for linking the concept to other resources. Therefore these generic geospatial words are probably best left out of the search. 'Aerodrome area' and 'Aerodrome node' are examples of this. As such, these terms are not present in other resources. Removing the word 'node' or 'area' results in several synonyms and hyponyms and connections to concepts in other vocabularies. Of course, the word can be added again for a final interpretation.

Alternatively, a search strategy can search for all individual parts of the compound word. However, in testing this approach, it is found that this most often results in too many returned results. Therefore this option was not retained in the *Synonyms finder*, and the implemented methodology searches for the complete concept label. In specific cases, generic words are removed manually to allow better search results.

### 5.5.3 Formulation of the concept label

Technically speaking, it has to be noted that different resources use different rules for writing the labels for their concepts. Words capitalisation and the use of single or plural word forms influence results if the queries used by the different services do not tackle this. Queries in SPARQL are by default case sensitive. If exact matching is used in the service, plural and single word forms are interchangeable.

The different rules for writing concept labels can be illustrated with some examples:

The INSPIRE feature concept dictionary use single forms with the capitalisation of each word, e.g. '*Protected Site*';

The INSPIRE theme register uses plural forms with a capitalisation of only the first word, e.g. '*Protected sites*';

GEMET and Eurovoc us single form without capitalisation, e.g. '*protected area*';

Agrovoc uses plural forms without capitalisation, e.g. '*protected areas*'.

In a graphical user interface, this causes typically not many problems. But for automated services, the differences in the representation of the same concepts must be solved. Lemmatisation of the

search words, which reduces a word to its basic form, can help tackle this in an automated approach. In the *Synonyms finder*, this is (partially) implemented together with a careful design of the queries used in the services. It is a complex task because different data sources use different search algorithms.

### 5.5.4 Spatial object labels

The INSPIRE feature concept dictionary contains terms related to spatial object types. As mentioned before, sometimes, the spatial nature of the concept is indicated by specific words such as 'area' or 'node'. But sometimes, this spatial indication is left out. As an effect, some labels do not indicate a spatial object. Some examples are '*railway type*' or '*railway use'*. These labels indicate more an attribute of a spatial object than the object itself. This nuance has to be considered when deciding if a match is correct. Such labels most often result in alternatives that cannot be considered synonyms. Instead, they have a more generic 'is related to' relation.

On the other hand, some labels are very generic, and the exact meaning can only be derived from the INSPIRE context it is used in. An example is the INSPIRE concept '*Site*'. The knowledge that this concept is defined within the data model of Agricultural facilities is needed to have a complete interpretation of the label name.

### 5.5.5 General applicability of the methodology

The feasibility of the methodology is demonstrated in three use cases related to INSPIRE. In these use cases, synonyms for geospatial objects are searched. Although some INSPIRE specific data sources were considered, they were not retained in the end. None of the sources used in the final results is specifically geospatial sources. As a result, it can be concluded that **the developed methodology and the *Synonyms finder* can also be used for generic, non-geospatial terms**.

The proposed methodology can therefore break interoperability barriers in general application areas, much broader than INSPIRE alone.

# 6 Use of the results and recommendations

Once synonyms are harvested, the question is ***how* the harvested information can be used**.

This chapter provides different alternatives for using the results of the proposed methodology. Only the harvested synonyms are used as alternative labels in the simplest form. The most advanced use fully exploits also the semantic output of the *Synonyms finder* tool as a starting point to align INSPIRE with other resources.

A second part discusses several possible enhancements of the procedure itself. Finally, a third part formulates recommendations to fully benefit the procedure results.

Optimal exploitation of the results of this study focuses more on the collected semantic information than on the synonyms themselves.

## 6.1 Use of the results

The results of a search for synonyms can be used in different ways. These are presented here with progressive use of semantic data available in the input data and sources where synonyms are found. The more elaborated methods of using the results use the synonyms themselves and semantic data obtained during the harvesting process.

**The more semantic data is preserved, the more added value is created for the linked sources**. This positive effect is enhanced if the process can start from semantically structured input.

During the *Synonyms Webinar*[61]: *Using synonyms to improve geospatial data discovery*, the remark was made that synonyms might go against the current use of fixed code lists for tagging datasets. Directly tagging datasets with synonyms would indeed not be the correct use of the results of this study.

It is recommended that synonyms are used to enrich the data specifications and the description of the data model. The alternative terms are not intended to tag separate datasets or instances of objects.

### 6.1.1 Use without exploiting semantic information

The most straightforward way to use the synonyms is by adding a list of keywords to each feature concept and adding that list to the search fields (besides the label and description used in the current tool). In this way, the object will be found when one of the keywords is entered in the search. However, the search engine will not be able to provide additional information.

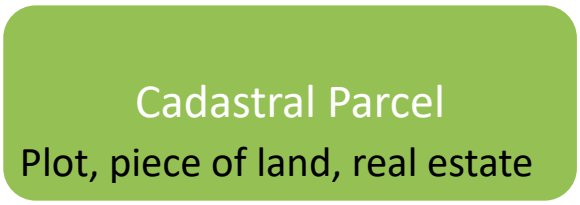



**Figure 41:** Synonyms, hypernyms, hyponyms as keywords

[61] https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/presentation-using-synonyms-improve-discovery-geospatial-data

A user will know if a spatial object is related to his search term but not know what the relation is. He/she will have to explore the INSPIRE documentation to see if both exactly match (synonym) or only partly overlap (hyponym or hypernym).

A more structured implementation is shown in

**Figure 42** and uses different lists of terms for synonyms, hypernyms and related terms. Suppose it is communicated in which list the search word is found. In that case, the user can better evaluate the validity of the found data object for his/her use case. The user knows if the match is exact (synonym), if the geospatial type also contains objects not related to the search term (search term is a hyponym) or if the geospatial type only contains part of the searched objects (the search term is a hypernym).

The technical implementation of this method in the catalogue is again not complex. The search engine will indicate the relation between the keyword and INSPIRE object. For example, it can suggest that '*police station*' is only one of the objects provided by the INSPIRE object '*Governmental services*'. In analogy with the terminology used in RDF, 'synonym' could be replaced by 'alternative label'.

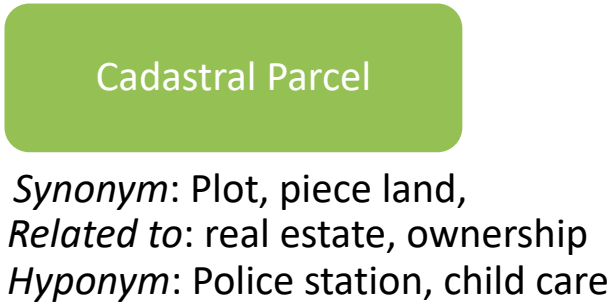



**Figure 42**: Synonyms, hypernyms, hyponyms as intelligent keywords

### 6.1.2 Taking advantage of semantic information and relations

More advanced integration of the alternative terms completely preserves the semantic information from the thesauri or ontologies from where the terms originate. This semantic information would allow the user (both human and machine) to explore further the term's semantic connections in its original domain. It would also expand the search to related concepts in that original domain. In other words, this would allow another search for related terms in a way very similar to what is applied in this study. One can argue that in that case, a list of synonyms is not needed because it can entirely be created on the fly. This is true if all connections are based on semantic relations. However, for connections based on lexical relations, too much manual interaction is needed, e.g. to solve the issue of different meanings for the same word.

On the other hand, there are not enough relations present between INSPIRE FCD and the other thesauri to initiate the process from scratch. In most cases, the first link, the alignment between the FCD and other thesauri, is missing. Another method, like the lexical matching used in this study, must be used to create these first links. Therefore, the combined lexical and semantic results from

the semi-automatic process developed in this study are needed to initiate the process. The provision of the results in RDF format facilitates this process.

| |
|---|
| The value of preserving and using semantic information becomes completely clear if the semantic relations on the INSPIRE side are also fully explored. For this, the different INSPIRE registries should be used to build a real INSPIRE ontology. |

The above process introduces an alignment between the INSPIRE ontology and the sources used to harvest the synonyms.

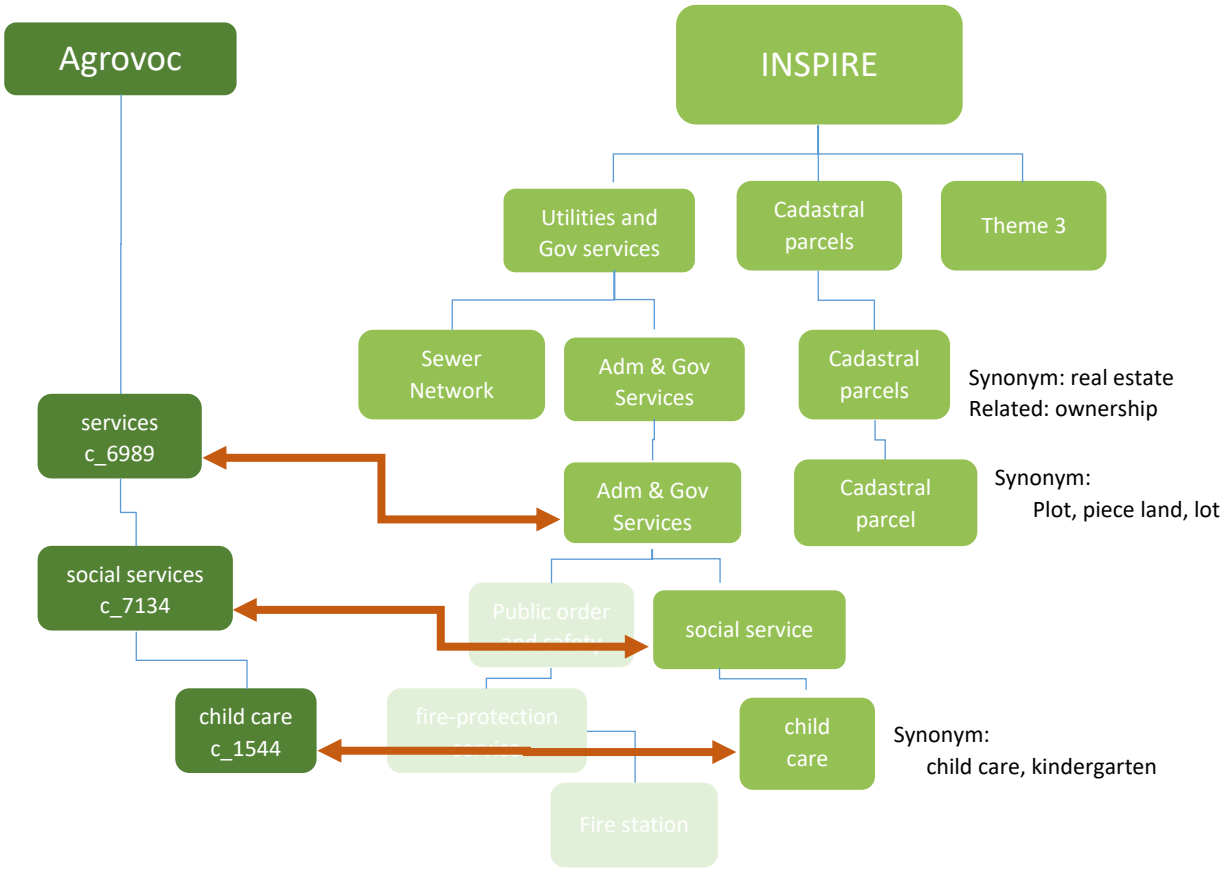



**Figure 43:** Alignment of an INSPIRE ontology with Agrovoc by linking similar concepts

| |
|---|
| The user of a catalogue that implements this semantic model will be able to completely see the context in which his search terms are related to the geospatial object type. |

In the example of

**Figure 43** the user can explore how service objects are organised in INSPIRE and in Agrovoc. The semantic relations guide this exploration.

**Once this semantic alignment is in place, additional information can be reached through the semantic richness of the combined resources**.

In **Figure 44**, the INSPIRE layer *Fire Station* can be linked to the OSM *tag:amenity=fire_station* because both are linked to the Wikidata concept *Fire station*. This shows the strong potential of

using semantically linked data. The original search terms can be linked to concepts in additional sources with minimal effort.

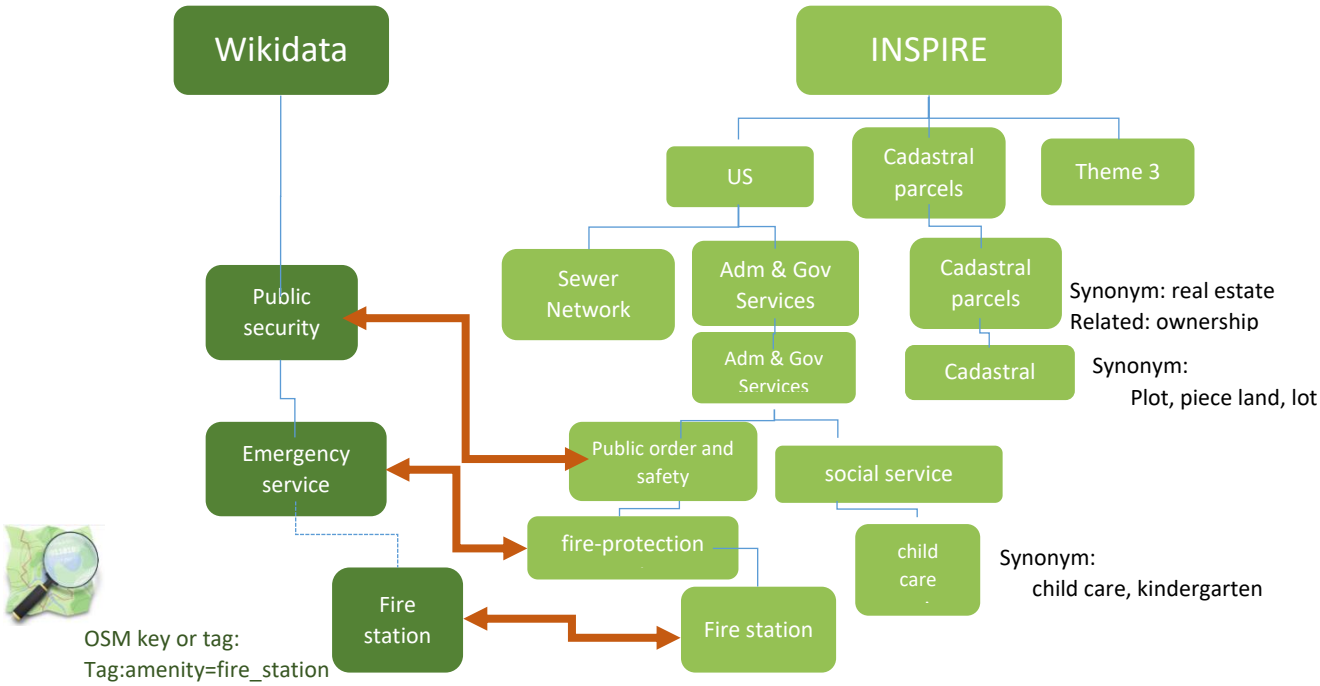



**Figure 44**: Wikidata to provide OSM keys or tags for INSPIRE concepts

It is important to mention that this process is not as easy as presented here. The methodology proposed in this study is only a starting point to develop the semantic system pictured here.

But these examples illustrate the added value of the system that can be developed in this way.

Most of the used sources are enriched with semantic information in the search for synonyms, and this information supports the search for synonyms and other related terms. This semantic information can only be fully exploited if the original data, the list of words to find synonyms for, is semantically linked. Both the search for synonyms and the application of the results can benefit from it.

## 6.2 Enhancements for the proposed methodology

The proposed methodology and the *Synonyms finder* tool are developed and tested in the context of this study. Although directly applicable in its current form, several enhancements can be formulated already.

### 6.2.1 Vocabularies and search

The methodology used in this study is applied to a limited number of existing vocabularies. A logic extension integrates additional data sources, especially if **additional vocabularies** exist related to the domain of the starting list of words. For each vocabulary, it is essential to fine-tune the queries to get optimal results.

The **fine-tuning of queries** for the already integrated vocabularies can still be optimised after a deeper investigation of the search engines used for the different sources. Better strategies to handle compound words might also be developed; these might be different for different sources.

When semantic links are sparse or completely missing, a lexical approach is only used on the English labels. Many vocabularies provided labels in several languages, and this multilingualism opens the option to also search for **matches in these other languages**. This also could allow linking to local vocabularies that are not available in English.

The *Synonyms finder* could also be extended to use simple file-based synonym sources, for example, in simple CSV lists. Using file base sources would allow easy integration of data sources like the data mapping information presented in section 5.2.3.2.

More advanced, instead of only lexical matching, **semantic comparison of concepts** might be feasible by integrating existing semantic matching software libraries. It would allow returning weighted results based on concept definitions and existing semantic relations. Weighted results would guide the user in the approval of proposed synonyms. Experience in the LusTRE project indicates that user input, especially expert knowledge, is still needed for final approval of results, even with semantic matching. In most cases, the (short) definitions of concepts are not enough to trigger an automated acceptance. Semantic comparison is not further explored in this study.

The *Synonyms finder* should exploit this structure when input data is structured to facilitate the search. Hyponyms from the source data could be added automatically if relations are clearly defined. These hyponyms can then be used as additional input terms, as done in the tests with INSPIRE code lists.

### 6.2.2 *Synonyms finder* generalisation

The *Synonyms finder* is a proof of concept tool developed to support this study. It is a valuable tool for searching related terms over different resources, and some possible generalisations can enhance its usability.

The current synonyms tool has the service connections to the used vocabularies built-in. Programming code changes are needed to connect to additional sources. Adding additional sources without changing the code in a subsequent iteration should be possible. Implementing this flexibility should be feasible, especially for those services that support SPARQL.

The status of search results and how these are translated to RDF is also hardcoded in the tool. This translation could be parameterised to provide more flexibility. One might, for example, incorporate additional approval status values for the results. These additional values might be translated to more specific RDF relations in the RDF output, especially if the initial list of terms originates from an ontology or thesaurus that already contains semantic relations.

## 6.3 Recommendations

### 6.3.1 Provide structured input data.

**Recommendation 1**: First structure the input data. With unstructured input data, only isolated terms are linked. If structured input data is used, the process turns terms into (the start of) a semantic data alignment.

The value of structured data, consisting of concepts enriched with semantic links, is shown in this test. In the INSPIRE use case, the use of code list values as hyponyms of the original input term was

sometimes the only way to obtain results. When this semantic information is available through services, it can be integrated into the *Synonyms finder*.

Particularly for INSPIRE, the INSPIRE Registry consists of several separate files. The entries in the different files often contain links to parent objects in other registry files. But this information could be enriched, and the information could be provided through services. The Registry could be turned from a list of files into an ontology served through a SPARQL endpoint optimal format. As demonstrated in this study, concepts in the ontology can then be enriched with synonyms obtained by processes.

### 6.3.2 Share alignment results

**Recommendation 2:** Share alignment results. This can be done by sharing the information on found relations. Another option is to integrate additional information in existing data sources.

Lexical matching, as done in this study, is a laborious exercise, and each proposed link has to be analysed by the user. **For technical concepts, it is expected that domain experts must do validation,** which is a time-consuming job. To maximise the return for these efforts, it is recommended to preserve connections found and publicly share them. This consolidation of results can be done in different ways.

Semantic links can be created if the input data source concepts have persistent identifiers (long-lasting, unique references). In that way, the alignment to concepts in other vocabularies are consolidated. If the input data is semantically structured, this matching of concepts supports the alignment of the vocabularies.

The alignment itself might be context-specific. But this should not stop sharing the alignments publicly. As long as the context is explained, the provided relations will help others to find data.

During the Synonyms Webinar, several participants indicated having done linking and alignment exercises themselves, with mixed success. Unfortunately, the results of these efforts are often not found online. **As a positive example, the results of this study are available online in Joinup**[62], not only in the tool-specific CSV format but also in the standard RDF format. The CSV format can be used to check the results of this study or as a starting point to test the *Synonyms finder*. The RDF file provides valuable labels and semantic data that can be directly used in other projects.

Still, the data might be difficult to discover by potential users.

It would be beneficial to many if the well-known thesauri managers created an open repository where users can share and document their alignment efforts. This would make these alignments easier discoverable and reusable for other users.

In specific cases, one might go a step further and ask if an existing term should not be added to an existing vocabulary. For authoritative vocabularies, this might imply following strict procedures. But this might be worth it because it enhances the visibility of the terms used.

For crowdsourced vocabularies, adding the information is more manageable. In Wikidata, one might create a new concept around the term. Another option is to add a term as an alternative label for existing terms. Finally, one might create specific tags to link the user's terms to existing concepts.

[62] https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-inspire-spatial-objects

This last option opens doors to other vocabularies because most terms in Wikidata already have many connections to different sources, as mentioned before. For example, Wikidata INSPIRE objects can be linked to OpenStreetMap feature keys and tags. Wikidata can be seen as a hub, providing connections between different thesauri.

# 7 Conclusions

Starting from the use case of the INSPIRE Catalogue of Objects, this study proposes a convenient methodology to find synonyms and other related terms for given input terms. The proposed methodology answers three questions:

— Where can synonyms be found?

— How can they be harvested efficiently?

— How can the harvested information be applied?

Several online resources are identified that might provide such synonyms. First, these sources are analysed, exploring if each source contains enough information to propose synonyms for a given term? This information can consist of alternative labels present in the data source, and semantic links between concepts can provide synonyms, hyponyms or hypernyms. These semantic links can be internal, inside the source, and external, providing links between different sources. The sources identified can be divided into domain-specific resources that provide more technical language related to a specific domain. On the other hand, natural language and crowdsourced resources provide a more generic dataset.

Access to these sources is technically heterogeneous, making it difficult to harvest synonyms from them efficiently. Therefore a *Synonyms finder* tool is developed to allow a more straightforward combination and integration of results from different sources.

When semantic relations exist between terms within a source or between different resources, they provide a natural path for the synonyms process. When semantic information is lacking, an approach of lexical comparison is used. Several issues can create barriers hindering lexical comparison. The use of domain-specific language in the input is the most important barrier to finding lexical matches. In the INSPIRE use cases, especially using domain-specific collective terms and geospatial specific words are the most important examples of this.

The use of additional information available in INSPIRE as the source of the input terms helps enhance lexical matching. Implicitly this comes down to exploiting semantic information of the input data source.

When applying lexical matching, input from the user is needed to evaluate the matches found. Lexical matching provides the first connections to a source, and from there, semantic links within that source or alignments with other sources provide rich additional information.

The methodology is tested on three geospatial use cases: agriculture, noise and water. The test input is INSPIRE objects related to these use cases, and semantic links between INSPIRE and the used sources are missing. Even in these non-optimal conditions, combining lexical and semantic matching, the applied methodology provides alternative terms for 43 out of 68 input terms (63%). This is a good result, effectively breaking the lexical barriers mentioned before. The full use of information on the input side, like the INSPIRE code lists, was crucial to reach this result to minimise lexical barriers. Without the additional information, less than 30 per cent of input terms lead to lexical matches.

These results demonstrate the importance of structured input to start the synonyms procedure. Semantic information in the input data significantly assists the process. Semantic links between different sources facilitate the process further and enhance the synonyms results. When semantic information is missing, lexical matching provides an alternative to creating initial connections with the different sources.

The three pillars of the proposed approach are integrated into the *Synonyms finder* tool developed for this study:

- Use structured data sources (if possible also on the input side)
- Exploit the semantic information provided in the data sources
- Use lexical matching when semantic information is missing

. The tool combines the three pillars and provides efficient access to the different data sources. Lexical matching needs user validation and can be time-consuming. The test cases prove that this validation can be done efficiently in the *Synonyms finder*. If it is executed correctly, Lexical matching with validation creates new valuable semantic links between data sources. It is highly recommended to publicly share the semantic results of the process to facilitate the exercise for other users.

Although the initial goal of the study and the developed procedure is to facilitate data discoverability by providing synonyms, it is clear that the semantic information collected in the process is at least as valuable as the synonyms themselves. Therefore this semantic information is integrated into the output of the *Synonyms finder*. If users share this semantic information, it can foster the breaking down of interoperability barriers between the different data sources as ELISE aims.

The proposed methodology and the *Synonyms finder* tool presented in this study provide good results to the geospatial use cases tested, but it can equally be applied to non-geospatial data. Despite the good results, further enhancements can be performed.

- **Best results are obtained if the data sources and the input data are semantically structured and exploitable**. An exploitable semantic structure is missing in the current INSPIRE Registers, and it is recommended to integrate the registries in one ontology. The semantic relations provided by the *Synonyms finder* can then be used to align that INSPIRE ontology with other data sources. This alignment is an important step in breaking down interoperability barriers.
- **For the procedure itself and the tool, several enhancements are already identified**. (Flexible) integration of additional vocabularies, implementation of multilingualism, fine-tuning search criteria and alternatives for the lexical matching of concept labels can provide more and better results and at the same time reduce the human interaction needed to reach these results.
- **Better disclosure of the semantic results is needed** to consolidate the interoperability gain provided by the methodology. A repository can be created to allow sharing and documenting semantic alignment results. Sometimes, integrating concepts in external vocabularies can provide another way to align sources better. Integration in external vocabularies is especially to be considered for crowdsourced datasets like Wikidata.

The use cases prove that the presented methodology and *Synonyms finder* are already valuable tools to enrich portal users' search experience and enhance data interoperability. By providing the results in standard semantic RDF format, the output can efficiently be used by online data managers (especially vocabulary managers) or users of existing online resources.

The executed test cases reveal the potential to further develop the methodology and tool into powerful instruments that support the goals of ELISE and Knowledge transfer in general. Knowledge transfer is the complex process of disseminating knowledge from one individual, team or

organisation to another to solve problems, foster innovation, or increase efficiency. Providing synonyms and other semantic information makes the transferred knowledge more understandable it also allows to better link and integrate new and existing knowledge. The results already partially integrate the different data sources used in the tests, and this integration increases efficiency and opens new possibilities for combined knowledge. For example, further exploring the created links allows to connection INSPIRE object types with Open Streetmap map features through the connection with Wikidata. This effectively breaks the barrier between vertical silos and opens possibilities to combine Open Streetmap and INSPIRE data.

In all places where data is to be shared or combined, semantic information is imperative. Regarding data sharing in the EU, this work can facilitate connecting the different upcoming European data spaces. To allow that, the *Synonyms finder* must use vocabularies related to those different data spaces. And as demonstrated in the use cases, the additional use of natural language resources facilitates the alignment of the (partially) different languages spoken by public bodies, businesses and citizens.

As identified in this work, semantics not only implies aligning internal data with external resources. Building an internal semantic data structure is just as important, and it will improve the alignment with external resources. Sharing the internal semantics and the external alignments is a final step to real interoperability.

## References

### Directly related to this study


ELISE Webinar announcement: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/event/elise-webinar-using-synonyms-improve-discovery-geospatial-data

ELISE Webinar slides: https://joinup.ec.europa.eu/sites/default/files/document/2020-12/ELISEWebinar_presentation-synonyms_FINAL.pdf

ELISE Webinar recording: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/presentation-using-synonyms-improve-discovery-geospatial-data

*Synonyms finder*. Available on Joinup at: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-finder

Synonyms for INSPIRE spatial objects. Available on Joinup at: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-inspire-spatial-objects


### Other references


An Architecture for Ontology-based Discovery and Retrieval of Geographic Information, in: Toppen, F. & Prastacos, P. (Eds.): 7th Conference on Geographic Information Science (AGILE 2004):179-188. Available at: https://dl.gi.de/bitstream/handle/20.500.12116/28701/GI-Proceedings.51-118.pdf?sequence=1&isAllowed=y

DBPedia. Available at: https://wiki.dbpedia.org/

eENVplus, eEnvironmental services for advanced applications within INSPIRE. Available at: http://www.eenvplus.eu/

Eurovoc thesaurus. Available at: https://op.europa.eu/en/web/eu-vocabularies

FAO AGROVOC thesaurus. Available at: http://www.fao.org/agrovoc/about

GEMET. General Multilingual Environmental Thesaurus. Available at: https://www.eionet.europa.eu/gemet/en/about/

GeoWordNet: A Resource for Geo-spatial Applications. In: Arroyo L. et al. (eds) The Semantic Web: Research and Applications. ESWC 2010. Lecture Notes in Computer Science, vol 6088. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/978-3-642-13486-9_9

IKEA webshop. Avaiable at: https://www.ikea.com/nl/en/search/products/?q=closet

INSPIRE Data Specifications. Available at: https://inspire.ec.europa.eu/data-specifications/2892

INSPIRE Geoportal. Available at: https://inspire-geoportal.ec.europa.eu/

INSPIRE Registry. Available at: https://inspire.ec.europa.eu/registry

Klien, E., Einspanier, U., Lutz, M., & Hübner, S. (2004): An Architecture for Ontology-based Discovery and Retrieval of Geographic Information, in: Toppen, F. &Prastacos, P. (Eds.): 7th Conference on Geographic Information Science (AGILE 2004):179-188.

LusTRE: a framework of linked environmental thesauri for metadata management. Available at: https://link.springer.com/article/10.1007/s12145-018-0344-8

LusTRE, Linked Thesaurus fRamework for Environment. Available at: http://linkeddata.ge.imati.cnr.it/StartPage.jsp

Merriam-Webster Vocabulary and Thesaurus. Available at: https://www.merriam-webster.com/

Princeton University (2020). About WordNet. Available at: https://wordnet.Princeton.edu

UNESCO thesaurus. Available at: http://vocabularies.unesco.org/browser/thesaurus/en/

Web Ontology Service to facilitate interoperability within a Spatial Data Infrastructure: Applicability to discovery. Data & Knowledge Engineering, 63(3), 2007. Available at: https://doi.org/10.1016/j.datak.2007.06.002

Wikidata. Available at: https://www.wikidata.org/wiki/Wikidata:Main_Page

**Glossary**

**Abstract spatial object type: An object type in a data specification that groups common semantic properties of real object types but does not represent a real-world object itself.**

**Application schema: Conceptual schema for data required by one or more applications (ISO 19101)**

**Compound term: A term that is a composition of 2 or more words. E.g. 'social service', 'runway area.'**

**Collective term or concept: In the context of this report: a term defined to group several related terms or concepts.**

**Geospatial objects/geospatial object types: A geospatial object (type) is an object (type) that has explicit geographic information included within it in either vector or raster format**

**Lemmatisation: Reducing a word to its base form (e.g. rooms > room; writing > write)**

**Lexical matching: Comparing and matching words by comparing them as a string of tokens (letters)**

**Semantic matching: Comparing and matching words by comparing their meaning**

**Semantic interoperability: Semantic interoperability ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties, in other words 'what is sent is what is understood'. In the EIF, semantic interoperability covers both semantic and syntactic aspects[63]**

**SPARQL: SPARQL Protocol and RDF Query Language. A semantic query language for databases—able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. SPARQL is a W3C standard. List of boxes**

[63] https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers#3.5

## List of figures

## List of tables

## Annexes

### Annex 1. Results of the use cases in CSV and RDF format.

The three case studies (agriculture, noise and water) are publicly available for download. Two CSV files and one RDF file are available for each use case. The CSV files can be re-imported in the *Synonyms finder*. The **final_all.csv* files contain all information shown and used in the *Synonyms finder*. These files allow analysing how the information is gathered from the different sources. When re-imported in the *Synonyms finder*, these files are also a good starting point to explore the tool's functionality.

The **final_selected.csv* files contain only those rows from the data validated as synonym, hyponym or hypernym. These selected rows can point to alternative labels for the input or linked concepts available in one of the source thesauri.

The **_final.rdf* files contain the same information as **final_selected.csv* but in the standard RDF format with labels and concepts linked directly to the input. The RDF file used the standard SKOS relations.

The information is publicly available at the Joinup entry Synonyms for INSPIRE spatial objects:

https://joinup.ec.europa.eu/node/704085

The results files are downloadable from Joinup:

- INSPIRE Synonyms – CSV files
- INSPIRE Synonyms – RDF files

### Annex 2. *Synonyms finder*.

- The *Synonyms finder* tool developed within this study is available at Joinup: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/solution/elise-semantic-resources/synonyms-finder
- The quick guide to the tool is available here: https://joinup.ec.europa.eu/node/704135
- and the Synonym finder itself can be downloaded here: http://data.europa.eu/w21/3e1c3ffd-9aab-4676-8946-ed182f3b3a76

The quick guide contains installation information for the tool.

**GETTING IN TOUCH WITH THE EU**

**In person**

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

**On the phone or by email**

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

**FINDING INFORMATION ABOUT THE EU**

**Online**

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

**EU publications**

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see

KJ-01-22-100-EN-N

## The European Commission's science and knowledge service

Joint Research Centre

### JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

### EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/08796

ISBN 978-92-76-48660-2