Maria Claudia Bodino

Challenges and enablers in using open government reference data in a data-driven public sector

Master Thesis

Supervisor: Prof. dr. ir. Joep Crompvoets

Presented by: Maria Claudia Bodino
Cuneo, Italy
mariaclaudia.bodino@gmail.com

Date of Submission: 2021-08-06

Acknowledgements

This thesis is just the last mile of a wonderful and challenging journey of two years' hard work trying to combine full-time work and study.

I'm very grateful to a number of people who support me in getting this programme and thesis to completion. Some, however, deserve a special mention.

I extend my gratitude to my mother, who taught me to never give up and keep pursuing my passion and always being there for me.

I want to say a huge thank you to Franco, who believed in me, for his constant source of inspiration, advice and continue support.

My warmest appreciation goes to my friends Sujani and Charlene for their special care and energy to proceed in the program. They were always there to make this journey much easier together.

A special word of gratitude to my friends Marianna, Romina and Rossella for showing confidence in my passion and work.

I want to express my sincere gratitude to my supervisor, Prof. Dr. Joep Crompvoets, who constantly encouraged me along the programme and provided his guidance and valuable feedbacks in writing this thesis.

Special thanks go to Luca Attias, formerly Italian Government Commissioner for the Digital Agenda. He encouraged me to follow this path and allowed me to work and study between Rome and Leuven during one of the best working experience of my life, the Italian Digital transformation team. Thanks to my previous colleagues Daniela Battisti and Michele Melchionda for their support and encouragement in starting this master.

I would also like to thank you Roberto Barcellan, my head of unit at the European Commission, for his support in writing this thesis and Magnisalis Ioannis and Jacopo Grazzini, my colleagues for their trust in my skills and the opportunity to learn more about the topic.

Finally, a wonderful thank you to all the Pioneers students, you definitely made my balance between work and study more enjoyable.

Abstract

Nowadays, public sector needs to move from hype to impact to concretely unlock the power of data. The ability to easily discover and access the relevant information and interlink the right datasets is a prerequisite to enable the creation of value in a data-driven public sector. However, these tasks are time consuming and challenging for data analysts. Open government reference data and controlled vocabularies in the form of taxonomies, code lists, authority tables and thesauri, can support data analysts in bringing diverse data together and are crucial building blocks to boost the creation of insight and intelligence in public sector. Their adoption fosters harmonisation of standard classifications and consequently enhances high semantic and technical interoperability, providing automatic mapping of data from different sources and sectors. Controlled vocabularies and open government reference data are a specific research topics under the major research area of interoperability in the open government data (OGD) research taxonomy. In this exploratory research, impediments and enablers of the discoverability and acquisition phase of the user process of controlled vocabularies have been investigated.

Semi-structured interviews have been conducted with data analysts working in the European Commission and domain experts in the field of open government data, reference data management, data governance, linked data technology and Semantic web. The analysis of the exploratory insights provided empirical findings to describe different types of barriers in the user process of identifying and acquiring controlled vocabularies relevant at EU level and in a DDPS context. In an international context, like the one of the European Union institutions, using controlled vocabularies is even more relevant because of the multi-level governance and multilingualism peculiarities. The findings indicate that users face several impediments in the process of identifying and acquiring relevant controlled vocabularies. The research confirms that the main common impediments related to discoverability and acquisition of open government data are similar with the ones in the area of OGD. The identification phase seems to present more elements of friction compared to the acquisition phase. In the identification phase, the obstacles related to the fragmentation of resources dispersed in many online locations, lack of documentation, metadata, multilingualism and bad UX, are aligned with the OGD data barriers. The lack of versioning management and mechanisms to verify authenticity, avoid ambiguities and map the resources over time, are impediments specifically related to the area of interest of the study. Regarding the acquisition phase, findings are more limited and related to technical aspects. Finally, the research provides a series of recommendations to overcome impediments in discovering and accessing controlled vocabularies. The study offers new empirical data and practical guidelines for future research on this specific topic under the area of OGD and from a user perspective, where very limited research is currently available.

Keywords: Data-driven public sector (DDPS), Open government data (OGD), Discoverability, Accessibility, Controlled vocabularies, Reference data, User process, Re-use, Barriers, Impediments, Enablers.

Content

Figures

Tables

Abbreviations

API	Application programming interface
BI	Business intelligence
CAP	Common agricultural policy
CSV	Comma-separated value
CV	Controlled vocabularies
DAC	Development Assistance Committee
DCAT-AP	DCAT Application profile for data portals in Europe
DDPS	Data-driven public sector
DGs	Directorates-General
EC	European Commission
ELISE	European Location Interoperability Solutions for e-Government
EU	European Union
EURIO	European Research Information Ontology
EuroSciVoc	European Science Vocabulary
H2020	Horizon 2020
INSPIRE	Infrastructure for Spatial Information in Europe
ISA2	Interoperability solutions for public administrations, businesses and citizens
JRC	Joint Research Centre
JSON	JavaScript Object Notation
LIFO	Location Interoperability Framework Observatory
MS	Member States
KPI	Key performance indicator
NACE	Statistical classification of economic activities in the EU
NLP	Natural language processing
NUTS	Nomenclature of territorial units for statistics
OGD	Open Government Data
OP	The Publications Office of the European Union
SDG	Sustainable development goals
SDMX	Statistical Data and Metadata eXchange
SKOS	Simple Knowledge Organisation System
TARIC	Integrated tariff of the European Union
URI	Uniform Resource Identifier

1 Introduction

The challenges our society currently faces, from the current pandemic to environmental threats, are of a scale and complexity that traditional policy tools cannot always address. From climate change to politics, there is growing recognition that some of the most intractable problems of our era are information problems. These problems need to be addressed with explorative data-driven collaborative approaches between different sectors and stakeholders (Susha et al., 2017).

On the other hand, today's digital revolution offers untapped opportunities for accelerating knowledge sharing and collaboration to unlock "wicked" problems for public good (Janssen, Konopnicki, et al., 2017) . The creation of a public sector intelligence can be accelerated by proper and shared usage of data and information knowledge management across the institutions, in order to improve the design and delivery of robust public policies and services to shape a better future (Ooijen et al., 2019).

Nowadays, organisations are investing significant resources in improving data governance to unlock the potential of data. To unlock a data-driven public sector (DDPS) and its benefits, first organisations need to truly understand the revolutionary value of data and data analytics, then prepare the necessary conditions to concretely embrace and adopt "a broad whole-of-governments data vision" (Ooijen et al., 2019, p. 57). The European Union is using new models to improve the use of data for decision making and to drive the public sector to become a data-driven public sector (European Commission, 2020b). Starting on what has already been achieved, measures and investments are planned for the coming years with the aim to increase the re-use and demand for data and data-enabled products and services (European Commission, 2020b). Therefore new data governance mechanisms, based on a holistic approach to unlock tangible and evident benefits in accessing and re-using the public sector information, need to be quickly set up and assessed in order to monitor their impact (European Commission, 2020b).

Leveraging the use of data and adopting a DDPS culture brings many opportunities for institutions to deliver better services and public value, but at the same time many barriers and challenges may arise as well (Ooijen et al., 2019).

From a user perspective, easily discovering and accessing relevant public sector information are mandatory factors to enable value creation in a DDPS context. Unfortunately, public sector information is spread across many different places and available in many different formats (Cavanillas et al., 2016). Data analysts and data experts that support institutions in the context of the DDPS, spend significant amounts of

time and resources in re-using and linking public data in order to finally extract public value (Cavanillas et al., 2016).

A relevant and valuable part of the public sector information accessed, used and generated in the context of a DDPS is called Open Government Data (OGD). According to Charalabidis et al., "the most important and socially beneficial OGD research can be conducted by using them as a basis of multidisciplinary research on important societal problems and challenges that modern societies face" (p. 56) that is one of the main goals a DDPS tries to address (Charalabidis et al., 2016). OGD is defined as public sector information freely available in machine-readable format without any legal impediment (M. Janssen et al., 2012). Many institutions are addressing initiatives to exploit the value of OGD and overcome barriers and challenges data users face in re-using public data (European Commission, 2020b; Eurostat, 2020; Publications Office of the European Union, 2020c; Toots et al., 2017).

An important pillar of OGD is open government reference data, public data used for consistent classifications managed by organisations and institutions such as the ISO country code list or the statistical classification of commercial activities. The adoption of consistent and harmonised reference data unlocks the linking and mapping from different datasets and enables interoperability at different levels between datasets from different domains. In the context of the different opportunities of a DDPS, accessing and re-using open reference data are everyday activities for data analysts involved in combining data to link different pieces of information (Cavanillas et al., 2016). Charalabidis et al. position the research topic of Open Reference data, precisely of semantic assets such as controlled vocabularies and code lists preservation, under the major research area of interoperability in their OGD Research Taxonomy (Charalabidis et al., 2016). The term controlled vocabulary usually indicates many common types of term management: taxonomy, code list, authority table and thesaurus (Hedden, 2010).

Nowadays, open reference data used and produced at EU level and from other many international organisations are published in many locations and in many formats with possible implications in easily re-use them (European Commission, 2020a).

This thesis motivation is focused on the need to understand the main barriers and impediments data analysts, working in a DDPS, face in reusing open government reference data. The focus of the research is on the user process and in particular on the identification and acquisition phases of open government reference data. The research aims to provide additional empirical findings on barriers and an initial set of enablers to foster a better re-use of open government reference data. It constitutes a small

contribution to the development of more effective and efficient data governance on public sector information, aligned with the main goals of the European data strategy.

In order to collect insights and relevant information about the challenges users working in a DDPS environment face, the author decided to focus on the context of the European Commission. More precisely, this thesis wants to focus on the barriers data analysts working in the European Commission face in discovering and accessing reference data, such as code lists and controlled vocabulary, relevant at the EU level, in order to perform activities for policy making evaluations and analysis in the context of a DDPS.

The EC is investing significant resources in order to transform the institution into a data-driven organisation, defining a clear data ecosystem supported by corporate data governance and policies (European Commission, 2020c). The need to remove obstacles in discovering, accessing, sharing, combining and re-using data assets to improve information discovery and support decision making processes across different policy areas and institutions is paramount. A corporate commitment to transform the EC into a proper DDPS is the goal of the Commission's data strategy, which is in place from 2019 and many initiatives are taking place at different levels. In this context, a *reference data management group* has been set up to investigate and adopt a common approach to manage reference data assets to maximize their re-use, overcoming numerous challenges in data discoverability, interoperability and exchange (European Commission, 2020a).

Therefore, the following research question was formulated to guide the study:

RQ: "*What are the challenges that data analysts and experts face identifying and acquiring open government reference data (focusing on controlled vocabularies) at EU level in the context of a data-driven public sector? What are the key enablers to enhance them?"*

To answer the research question, the author adopts an exploratory research design and uses semi-structured qualitative interviews with data analysts working in the European Commission and domain experts in OGD, reference data and semantic web.

This research includes the following sections. First, a comprehensive overview of the main concepts of the research questions such as the meaning of a data-driven public sector, open government data, and of reference data and controlled vocabularies are explained.

Then, a theoretical conceptualization and framework of the main opportunities of a data-driven public sector, followed by the definition of the OGD user's process, looking at the identify and acquire phase and the COBIT enablers, are provided. Accompanied is a

presentation of the research design methodology that the author followed to answer the research question and analyse the results. Following the results are presented. First, a summary of interviewees' responses is presented according to the interview structure and the main topic of the research question. Then, discussion and analyses of findings, as well as indicating the theoretical and practical implications of this study, is presented.

The research ends with conclusions, followed by limitations and suggestions for future research. Presented in appendix there is the interview template used for data analysts and experts.

2 Background and Conceptual framework

In this chapter, key concepts that support and define the scope of the research are introduced and described for the readers to easily understand the analysis of the paper.

2.1 Literature review

This chapter aims to focus on the context of the research area and the identification of the research problem. An overview of the main aspects, definition and conceptualisation of the terms covered by the research questions are provided. First, the meaning of a data-driven public sector will be explained. Second, the importance of Open Government Data and its relevance in the context of a data-driven public sector will be presented. Third, the scope of the current research in the scientific field of Open Government Data is narrowed down to the topic of open reference data. Finally, a focus on code lists and controlled vocabularies, as enablers for unlocking the potential of data, and as a key ingredient to connect and link information from different sources and sectors is provided.

2.1.1 Data-driven public sector

Over the last decade, the amount of data available has grown dramatically; governments and institutions publish data on many topics, in many formats and in numerous repositories (Lněnička et al., 2021). Unlocking the power of such heterogeneous and scattered information to effectively design inclusive and better policies is a priority on many digital transformation action plans and agendas and to become a data-driven public sector (DDPS) is the priority and the new mantra of the public organisations (Ooijen et al., 2019).

The concept of a DPPS promotes the power of data as a strategic asset to transform policy making and service delivery to increase social inclusion, trust in governments and rethink the democratic sphere (Ooijen et al., 2019).

Data can speed up digital transformation in the public sector if decision-makers can easily access the right information at the right time, removing any form of obstacles and simplifying access to information (Ooijen et al., 2019). Governments and Institutions need information from many sources. At the same time, they can benefit from a growing number and variety of information to evaluate policy implications and support policy development (Janssen, Konopnicki, et al., 2017). However, the public sector faces many barriers and challenges in implementing a proper DDPS, concerning different aspects. On the one hand, it is generally extremely difficult to quickly find and access consistent raw data because datasets are often published in heterogeneous format and dispersed in many

platforms. On the other hand, the public sector also faces a lack of internal capabilities and the data skills needed to properly unlock the potential of datasets (Kalampokis et al., 2013; Ooijen et al., 2019).

The creation of public sector intelligence is supported by the government data value cycle, where different stakeholders interact in different stages and activities (Cavanillas et al., 2016; Janssen, van der Voort, et al., 2017; Ooijen et al., 2019). In order to facilitate this mission, many organisations from diverse sectors collaborate together to extract fresh insights and use their joint expertise to address policy challenges (Ooijen et al., 2019).

Multiple actors are involved in supporting the achievement of a DDPS. There are the data producers or providers that represent the source of data, for example the citizens through their interaction with the public sector in the context of so-called " life events" related to study, work or leisure. Then, the entity of the data collectors, who has the responsibility to manage and secure the data. In addition, the data analysts that try to answer questions linking datasets to each other to enrich, aggregate, and analyse them to create knowledge and value-adding applications. The last actors are ultimately the policymakers that, based on the input provided by the data analysts, can deliver better policies and services supported by a data-driven and evidence-based approach (Ooijen et al., 2019)

The ability to easily discover and quickly access the relevant information, find the right datasets and their metadata, combine, and enrich them, are prerequisites to create value in data-driven decision-making tools and application (Máchová et al., 2018). These tasks are even more challenging and time-consuming in the current data era. Nowadays, governments and institutions struggle with data variety from emerging decentralized data platforms and data catalogues, and spend significant resources in finding, accessing, cleaning, transforming and semantically linking them (Cavanillas et al., 2016). On the other hand, turning data into valuable information and knowledge is supported by the modern analytics methods and tools (Ooijen et al., 2019). Additionally, the proliferation of self-service analytics tools represents a new opportunity to leverage and enable more users to perform further analysis and investigations (Jacquin et al., 2020). As a result, a DDPS can become a reality thanks to the rapid evolution of ICT and the decrease of costs for technologies and techniques adapted to store and analyse data (Chatfield & Reddick, 2018).

Because of this, data analyst jobs and skills are in high demand in the public sector to address these challenges and speed up the creation of a proper DDPS ecosystem crossing departments and institutions boundaries to make more effective, data-informed decisions (OECD, 2020). Many national and international organizations, such as the European Commission, the OECD and the UK's Government Digital Service, hire data specialists

and foster data literacy amongst their employees (European Commission, 2020b; Nolan, 2021; Ooijen et al., 2019). Data analysts and data experts apply analytics methods and use technologies to support evidence-based policy-making. Their adoption and exploitation are encouraged in all aspects of the policy cycle (GDS, 2017).

2.1.2 Open government data

Data to boost a proper DDPS comes from numerous sources, from the public sector, the private one, and citizens and academia (European Commission, 2020b). The public sector generates and collects an enormous amount of information every day through different public bodies, sectors and activities (European Commission, 2020b). Managing payments and pensions, collecting tax, and monitoring the pandemic's spread through health and mobility data are just some few concrete examples (European Commission, 2020b).

A significant part of the information generated by the public sector is public. It can be published for free in an open machine-readable format as open data to the citizens without any legal or privacy impediments (Janssen et al., 2012). This information is also referred to as OGD, a subset of Open data, that represents a valuable resource for delivering better informed decision-making, creating data-based services and fostering public debate (Janssen et al., 2012). In this thesis, OGD is defined according to Janssen et al. as "*non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distributio*n" (Janssen et al., 2012). Citizens, policymakers, public servants, researchers, and companies can benefit from having this information freely available and re-usable (Janssen, 2011).

OGD represents one important pillar of the open government ecosystem (Janssen, Konopnicki, et al., 2017; McBride et al., 2019). The main goals of OGD are to boost a proper DDPS, enforce transparency and accountability, promote engagement and participation from public and private stakeholders, and accelerate the data economy for companies and start-ups (Charalabidis et al., 2018; European Parliament and Council, 2019; Open Knowledge Foundation, 2019). Furthermore, recently OGD is getting a more and more relevant role in driving the co-creation of new data-driven public services at different scales, involving heterogeneous stakeholders and target groups in creating public value and DDPS (McBride et al., 2019).

One essential ingredient at the core of a successful DDPS is the ability to access and obtain useful data in an easy way avoiding a silo based approach (Janssen, Konopnicki, et al., 2017; Kučera et al., 2013; Ooijen et al., 2019). OGD plays a crucial role in this direction, thanks to increased awareness of the use of data as a pillar for contributing in extracting new insights and innovative solutions to face societal and policy problems (

Janssen, van der Voort, et al., 2017; Ooijen et al., 2019; Susha et al., 2017). Another important element to mention is that the amount of OGD opened to the public has grown incrementally in the last decade and, many public organisations have started publishing their raw data through single points of access such as OGD portals and catalogues (Kučera et al., 2013; Ooijen et al., 2019). To conclude, as Kalampokis et al claim:

> *"the real value of OGD will unveil from performing data analytics on top of combined statistical datasets that were previously closed in disparate sources and can now be linked in order to provide unexpected and unexplored insights into different domains and problem areas.*" (Kalampokis et al., 2013, p.100).

2.1.3 Open government data in the EU

Institutions and governments are taking up many initiatives to grasp the potential of the data age and improve access to public sector data (European Commission, 2020b). Improving the discoverability of data and cooperation across public sector organisations, for developing and providing access to data inventories, is a common challenge for open data policies and data management strategies. Findable, accessible, interoperable, and high-quality data are must-have prerequisites for trusted and reliable analytics to enable evidence-based policymaking at an inter institutional level (European Commission, 2020c).

At a European level, the topic of OGD and the debate on the need to open up public sector data is not something new. The European Union (EU) started to stimulate the re-use of public government data since the end of the 1980s both with legislative and non-legislative measures (K. Janssen, 2011). Recently, the EU launched the European data strategy that aims to accelerate the role of the EU in a data-driven society, in the interest of citizens, business, research and the public administrations (European Commission, 2020b).

An important pillar of this strategy related to the context of the current research and OGD is the Directive (EU) 2019/1024 on open data and the re-use of public sector information (PSI) approved on 16 July 2019 (European Parliament and Council, 2019). The PSI Directive fosters the re-use of open government data information "open by default and by design" keeping in mind transparency and fair competition and the focus is on quality, interoperability and availability of high value datasets (European Parliament and Council, 2019).

On the non-legislative side, EU institutions are investing in new actions and initiatives to foster re-use, interoperability and better data-driven policy-making processes.

The urgency to adopt harmonised approaches to facilitate and speed up interconnections of knowledge across the EU institutions, and break out data silos to ensure information availability for a data-driven public sector is a specific objective that drives their strategic plans for the coming years (European Commission, 2020b; Eurostat, 2020; Publications Office of the European Union, 2020c). Increasing the exploitation of OGD, thanks to the implementation of standards to facilitate the interlinking and interoperability of data assets, is a priority at every level of government (European Commission, 2020c). At the EU level, the main bodies are working and collaborating closely in this direction. The Publications office of the European Union (OP), the official publisher of the EU institutions, agencies and bodies, has the mission to provide all institutional bodies accessible and reusable OGD to facilitate the support of a DDPS between them. Many initiatives are already in place, such as the EU Data portal that provides EU institutions' and Member States' OGD to be re-used for free by citizens, academia, the private sector and the public as well. In April 2021, the new official portal for European data, data.europa.eu, was released (Publications Office of the European Union, 2021). It aims to replace the two separated websites, the EU Open Data Portal and the European Data Portal. It improves accessibility, providing a central catalogue for open data from international, EU, national, regional, local and geo data portals (Publications Office of the European Union, 2021). Similarly, Eurostat, the official point of reference for statistics in the EU, is constantly fostering harmonisation with standards and international norms to improve the data availability and quality (Eurostat, 2020).

2.1.4 Open government reference data

The scientific field of OGD covers many research topics and it is important to identify a clear focal point to limit the scope of the current research. Based on the OGD Research Taxonomy proposed by Charalabadis et al., there are four major research areas: OGD Management and Policies, OGD Infrastructures, OGD Interoperability, and OGD Usage and Value (Charalabidis et al., 2016).

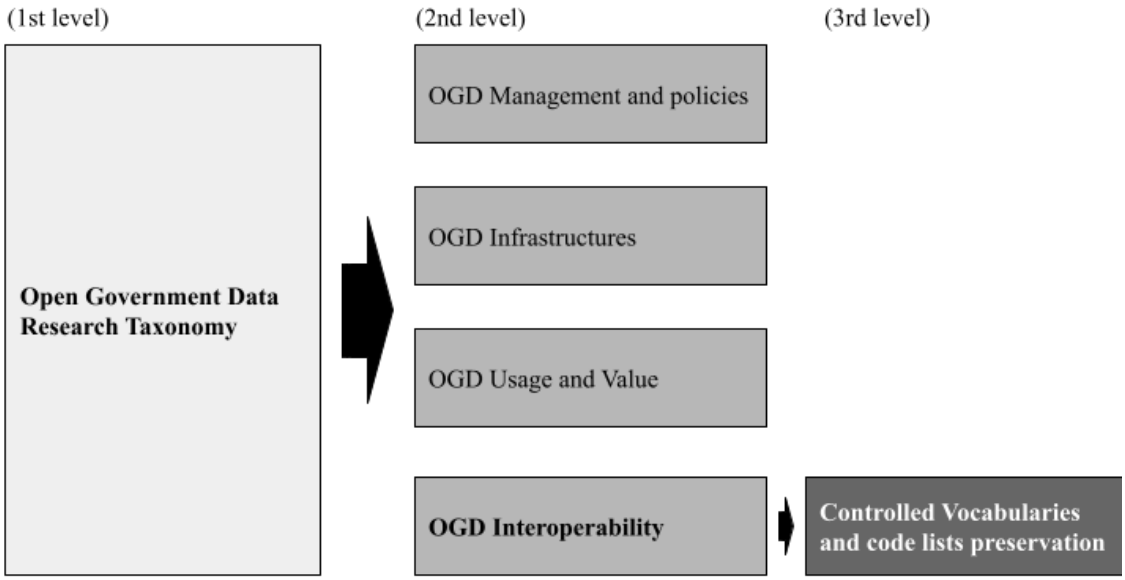


Figure 1: OGD Research Taxonomy (adapted from by Charalabadis et al., 2016)

The author, as already anticipated in the research questions, aims to focus the research on open government reference data, more specifically on controlled vocabularies. This research topic is clearly identified in the OGD Research Taxonomy proposed by Charalabadis et al. under the major research area of Interoperability (Charalabidis et al., 2016).

An important factor that affects the quality of a DDPS, even more nowadays that big data processing is increasingly becoming a relevant activity, is the ability to collect information from many places inside and outside the organisation (Janssen, van der Voort, et al., 2017). As previously mentioned, data analysts spend a lot of time doing manual work to find, access and re-use data. The automation of data analytics processes will still take more time to become completely automatic because of the lack of standardisation and interoperability (Janssen, van der Voort, et al., 2017). Interoperability is the fundamental enabler to unlock the potential of data, the key ingredient to connect and link information from different sources and sectors and to overcome barriers imposed by a data silos approach and a divergent interpretation of data (Charalabidis et al., 2018). In the EU context, the topic of interoperability has been considered of primary importance in the creation of a DDPS since 1995 and a lot of researches has been done and frameworks have been proposed such as the European Interoperability Framework (EIF) (Charalabidis et al., 2018). More specifically, the EIF framework identifies four layers of interoperability: legal, organisational, semantic and technical and underlines with a list of recommendations the importance of these features (European Commission, 2017).

The ability to connect data from different domains and levels of detail, is often based on the usage of "lookup" data and tables such as classification and categorisation (European Commission, 2014). These building blocks of information are alternatively called reference data.

The ISA programme from the European Commission (2014) defines reference data as: "small, discrete sets of values that are not updated as part of business transactions but are usually used to impose consistent classification. Reference data normally has a low update frequency. Reference data is relevant across more than one business system belonging to different organisations and sectors." (European Commission, 2014). In similar words, reference data is a special type of data, essentially codes and labels, that are finite and quite static, usually managed and published by standards organisations and institutions (Malcolm Chisholm, n.d.).

The relevance of using common reference data is twofold: first the adoption of common values and classifications for describing data limit the semantic interoperability issues, and second the adoption of common classifications across different domain avoid the effort, often manual, in managing and creating mappings (European Commission, 2014). Therefore, the adoption of official code lists and classifications enable maximum re-usability and plays a crucial role in linking datasets from different sectors and domains.

Examples of this include:

- postal codes
- NACE, the statistical classification of economic activities in the EU
- NUTS, the nomenclature of territorial units for statistics
- the classification of health care functions
- the statistical classification of products by activity
- ISO country codes

These are examples of code lists, a type of controlled vocabulary and reference data, used in cross-domain contexts.

These data represent fundamental building blocks for facilitating linking of datasets, achieving interoperability and enabling a DDPS based on innovative data-based services

A typical scenario in which it is easy to understand the relevance of reference data is represented by the CORDIS project. The initiative, funded under the Horizon 2020 (H2020) framework programme, provides OGD datasets with the list of projects founded. Every row of these datasets is composed of numerous columns with different pieces of information related to the project itself. Many of them such as countries, organisation types, programmes topics and funding schemes, are reference data fields coming from code lists defined at the EU level (Publications Office of the European Union, 2020b). For example, it would be of some interest to perform a data-driven analysis of country participation in the H2020 programme and link the dataset to the other datasets based on the programmes topics to analyse calls that aim to address societal challenges such as transport innovations. Furthermore, based on social data analysis through these OGD, it is possible to understand the connections between the actors in the network, which countries mostly control the flow of resources and the relationships with universities and research institutes (Bralić, 2017).

The building blocks that unlock similar types of analysis and allow a high level of interoperability, are the above mentioned reference data, in this case controlled vocabulary relevant at the EU level. Reference data make data analysts' tasks to classify and combine data with other data easy, and enable a more holistic data-supported picture of different scenarios (Cavanillas et al., 2016).

Nowadays many independent projects and several initiatives on reference data exist, limiting the re-use of open reference data. Reference data, such as controlled vocabularies, are published in many different locations. Consequently, open reference data at the EU level are not easy to find, as they are scattered in many places and in many formats (European Commission, 2020a). There is not an official single one stop shop for reference data that is publicly available and produced by the different EU bodies and institutions.

The Statistical Office of the European Union (Eurostat), provides hundreds of standard code lists available through RAMON, the Eurostat's metadata server (Eurostat, n.d.). The Publications Office of the European Union (OP), disseminates reference data such as controlled vocabularies, code lists, ontologies, data models through the EU Vocabularies websites, providing different formats, alignments with other controlled vocabularies, versioning information and descriptive documentation, especially for expert users (Publications Office of the European Union, 2020c).

In order to provide a consistent classification, ensure continuity and quality of service and reduce interoperability issues, a proper reference data management and governance need to be guaranteed and shared amongst the institutions in charge of managing these types

of data assets (European Commission, 2014). The lifecycle of reference data management includes different phases, from the data design, to the change management, harmonisation and implementation (European Commission, 2014). Additionally, the documentation process is an important horizontal task that needs to be present in every phase and contributes to limit misinterpretations and foster the re-use facilitating the searching and accessing phase (European Commission, 2014).

In 2019, the European Commission, decided to prioritize and invest in the definition of a common governance for reference data management (European Commission, 2020a). A reference data management board has been created and a first draft policy has been published (European Commission, 2020a). The objective of fostering the accessibility and reusability of reference data assets across the European Commission, the European institutions, the Member States as well as the private and public sector is a main priority (European Commission, 2020a).

2.1.5 Controlled vocabularies

A controlled vocabularies (CV) is a selection of terms or words usually regarding a scoped area, for descriptive cataloguing and classification of knowledge (Hedden, 2010). The term “controlled” is mainly used for two reasons: firstly, because the list of words and concepts available in the vocabulary is limited to the area of interest, and secondly because the CV is managed under the control of the owner and domain specialist that, through a review process can update and modify it (Hedden, 2010).

The adoption of controlled vocabularies have many purposes: on one side to harmonize the adoption of concepts in order to foster technical and business interoperability between institutions, on the other side to facilitate and improve machine-readable dissemination, automatic discovery and linking of data assets available in the semantic web (Publications Office of the European Union, 2020a). Additionally, in the context of the EU which is highly characterized by linguistic diversity, with 24 official languages and numerous language combinations, controlled vocabularies play an essential role in the translation requirements of the different concepts (Publications Office of the European Union, 2020a).

The term CV is commonly used mainly for the following types of term management: taxonomy, code list, authority table and thesaurus (Hedden, 2010).

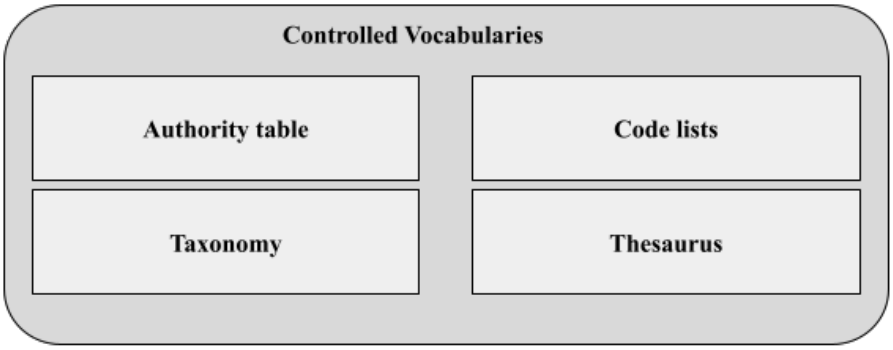


Figure 2: Types of Controlled vocabularies (adapted from Hedden, 2010)

The term *taxonomy* is commonly used for categorisation systems and hierarchical classification. Originally it meant the science of studying classification, more precisely the classification of living organisms according to their natural relationship (Voultsiadou et al., 2017). In a taxonomy, all the items are mapped in a structure similar to a tree and linked by parent/child or broader/narrower relationships (Voultsiadou et al., 2017). The use of taxonomies is something old and essential in order to communicate acquired knowledge along the time (Voultsiadou et al., 2017). The Greek philosopher Aristotle (384-322 BC) is considered the first father of taxonomies; he introduced for the first time the key concept of taxonomy as it is practiced today. His contribution to the classification of marine animals by type and binomial definition is still used and attracts the interests of marine scientists (Voultsiadou et al., 2017). An example of *taxonomy* at the EU level is EuroSciVoc, a multilingual taxonomy that describes the main fields of science based on the information from the CORDIS database, the European Commission's open data repository for EU-funded research projects, and organised through a semi-automatic process using Natural Language Processing (NLP) techniques (Publications Office of the European Union, n.d.-a).

An *authority table* or authority file, is a specific type of CV where for every term, there are one or many possible synonyms that are useful to enable cross-references and harmonise and standardise inter-institutional information exchange (Hedden, 2010; Publications Office of the European Union, 2020a). In addition, authority tables guide users to find the "preferred term" through different variants and synonyms of the same concepts (Hedden, 2010). Examples of authority tables at EU level include the authority table with the possible statuses of the EU budget preparatory actions and pilot projects, and the Court type authority table with the classification of EU courts with the general three-level classification of national courts (Publications Office of the European Union, 2020a).

The other popular type of controlled vocabulary is represented by *code lists*. A code list, also called "pick list" or lookup table, is the easiest type of controlled vocabulary and it is represented by a flat and finite list of codes and meaning (Hedden, 2010). Examples of code lists include: Transport mode, Classification of health care functions, European Socio-economic groups, Pesticide groups, River Basin Districts, Territorial typology, and Work intensity. CVs are often easily recognisable within drop-down boxes or check-boxes on online forms.

Ultimately, the term *thesaurus* has a structure similar to a controlled and structured vocabulary, where labels are used to describe concepts and for every label there is a list of possible synonyms and antonyms (Publications Office of the European Union, 2021a). Additionally, for every concept a thesaurus can provide the level of alignment with other correspondences based on different scopes (for example using relationships such as: has exact match, has close match, has broad match, has narrow match and has related match). At the European level, a relevant example of a thesaurus is EuroVoc, a multilingual, multidisciplinary thesaurus covering the activities of the EU in the 23 EU languages plus three languages of candidate countries to enter in the EU.

The current research aims to focus specifically on the different types of controlled vocabularies as described above and with relevance in the EU context. The resources the research is interested in are those available in the public domain and available as open government data.

Type of Controlled vocabulary	**Examples**
Taxonomy	*Digital Competence Framework:* the list identifies the key components of digital competence in 5 broader, and a total of 21 narrower areas (Publications Office of the European Union, 2021c). *EuroSciVoc:* is a multilingual taxonomy that represents all the main fields of science that were discovered from the projects funded under the EU's research and innovation Framework Programmes (Publications Office of the European Union, n.d.-a)
Code list	*Currency code:* the list of global currencies codes (ISO, n.d.-b). *Language code:* the list of language code(ISO, n.d.-a)
Authority table	*Administrative territorial unit:* a controlled vocabulary that lists concepts associated with various

	administrative territorial units of current and past EU Member States (Publications Office of the European Union, 2021b). *Procurement procedure type:* a list of activities leading to the conclusion of public contracts used in public procurement according to the legislation (Publications Office of the European Union, n.d.-c)
Thesaurus	*Agrovoc*: represent anything in food and agriculture, such as maize, hunger, aquaculture, value chains or forestry. Tnt. It consists of more than 39.100 concepts and 844.000 terms in up to 40 languages (FAO, n.d.). *Eurovoc*: the EU's multilingual and multidisciplinary thesaurus. It contains 21 domains and 127 sub-domains which are used to describe the content of documents in EUR-Lex (Publications Office of the European Union, n.d.-b).

Table 1: Examples of types of controlled vocabulary

2.2 Conceptual framework

This chapter presents the conceptualization of the main categories relevant to the research questions and useful to analyse data.

First, the author summarized the different areas of activities and opportunities of a DDPS to support the institutions. Second, the relevant phases for this research related to the OGD users' process framework of Crusoe and Ahlin are summarized (Crusoe & Ahlin, 2019). Third, the definition of the activities covered by the "identify" and "acquire" phase are presented in order to clearly scope the focus of the research question. Finally, the COBIT framework and enablers useful to derive findings to improve the discoverability and acquisition of open reference data in a DDPS context will be discussed.

2.2.1 Identifying activities of a Data-driven public sector (DDPS)

There are many activities and opportunities to boost the leverage of a data-driven public sector. Different forms of data-driven initiatives, depending on the purposes, can be identified to support public organisations in delivering their goals (Ooijen et al., 2019). According to the OECD framework by Ooijen et al. (2019), there are three main areas where organisations can use a data-driven approach to produce better value; these areas can be conceptualized as the following:

- Anticipatory governance
- Design and delivery
- Performance management

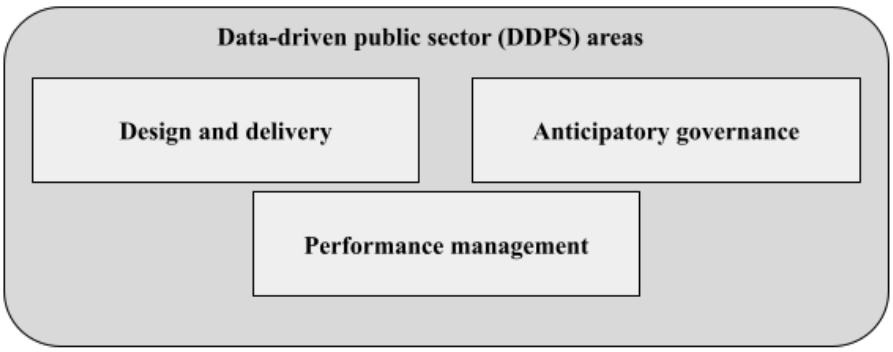


Figure 3: DDPS areas (adapted from Ooijen et al., 2019)

Anticipatory governance

The "Anticipatory governance" category mainly covers the activities of forecasting and foresight for a better planning of the future (Ooijen et al., 2019). Anticipating governance and using evidence-based approaches to design policy through a forward looking approach, can improve organisations by allowing better allocation of resources and investments and consequently, allowing for better policies and services on emerging issues and needs (Ooijen et al., 2019).

The forecasting activity consists of using existing data and time series regarding the past to model and predict future trends and outcomes. In this context, predictive techniques can be used to forecast social and economic trends, and measure policy impacts (Ooijen et al., 2019). Then, the foresight activity is not interested in predicting the future but in providing different scenarios and exploring emerging patterns and trends in the present, and it is useful to collect insights for designing interventions for the future based on different approaches. The efforts of replacing old protocols and experience with data-driven approaches and techniques, to adopt proactive instead of reactive attitudes are the main approaches covered by the "Anticipatory governance" (Ooijen et al., 2019).

Design and delivery

The second category is about data-driven design and delivery of policies and services. Using data to assess, discover and validate user needs to design more inclusive and effective public services is the main goal of this second pillar. Enabling different

stakeholders and citizens in designing public services, combining data from different sources and with a highest level of granularity, to provide better tailored services are the main goals of this DDPS approach (Ooijen et al., 2019). Examples of these services concretely simplify service design and life cycle of public services with a more agile approach and, at the same time, help in reducing the gap between citizens and institutions and improve public trust (Ooijen et al., 2019).

Performance management

Finally, the last opportunity covered by a DDPS is named "Performance management" and includes the use of data-driven techniques for a better management of human and economic resources (Ooijen et al., 2019). Looking at the performances and productivity at different levels through the lens of data can help organisations to improve their efficiency in spending and managing financial resources. In addition, activities such as auditing and risk assessment can benefit from data-driven techniques (Ooijen et al., 2019). Furthermore, adopting data-driven processes and governance, helps in reducing the time needed to complete procedures and processes and at the same time enable a more consistent and efficient usage of resources. A concrete implementation covered by this pillar, is the Once Only principle that aims to reduce administrative burdens on citizens, it aims in requesting just one time data to citizens and fostering the internal sharing and re-use of data avoiding multiple submissions from the citizen (European Commission, n.d.-b).

Anticipatory governance
• Forecasting to proactively identify developments and future needs • Foresight to prepare for multiple plausible alternative outcomes
Design and delivery
• Better predicting policy solutions • Engaging with citizens as co-value creators • Responding better to citizens' needs
Performance management
• More efficient use of resources • Increase of resources • Higher quality and evaluation • Continuous improvement

Table 2: Opportunities of a DDPS (adapted from Ooijen et al., 2019)

2.2.2 Identifying OGD user's process

The public sector faces many challenges in implementing a proper DDPS, concerning different aspects, OGD adoption is one of them (Ooijen et al., 2019). Even though OGD is promising, its potential has not yet been fully exploited and there is still limited availability of empirical data to assess the impact and effects of OGD initiatives (Charalabidis et al., 2018; Crusoe & Melin, 2018; Donker & van Loenen, 2017; Janssen et al., 2012; Janssen, Konopnicki, et al., 2017; Lněnička et al., 2021; Lourenço, 2015; Wang & Lo, 2016). Nowadays, it is extremely difficult to quickly find consistent raw data because datasets are often published in heterogeneous ways and dispersed on many platforms (Kalampokis et al., 2013; Xiao et al., 2019). Furthermore, often similar datasets are published from different organisations, and it is not clear to users which one is the official and trusted dataset to be used (Crusoe & Melin, 2018).

In general, data analysts, and data experts spend significant effort and time in numerous manual tasks such as discovering, accessing and linking OGD (Kalampokis et al., 2013). Consequently, there is an urgent need to clearly identify these barriers and make tangible ways to overcome these challenges to address the lack of ability to find and access the right information (Machova et al., 2018).

As a result, there is an emerging interest for research on understanding more about motivations of successful OGD adoptions from different perspectives and many authors have tried to clearly identify processes, barriers and impediments (Crusoe & Melin, 2018; Donker & van Loenen, 2017; Janssen et al., 2012; Kalampokis et al., 2013).

In recent years, many researchers have started studying more the user's activities, in order to conceptualize challenges and barriers in the OGD user process and consequently address them to maximize the potential of OGD (Crusoe et al., 2019; Crusoe & Ahlin, 2019; Crusoe & Melin, 2018; Lněnička et al., 2021; van Loenen, 2018; van Loenen et al., n.d.). For example, in the study "Investigating Open Government Data Barriers. A Literature Review and Conceptualization", Crusoe and Melin tried to systematize the OGD barriers, providing different research focuses on the topic and a mapping of the barriers depending on the OGD processes (Crusoe & Melin, 2018).

In 2019, Crusoe and Ahlin, in the research "Users' activities for using open government data", tried to summarize previous contributions and empirical findings on the OGD user process (Crusoe & Ahlin, 2019). Recently, Crusoe and co-workers investigated impediments of OGD lack of use, collecting and analysing the user's process. The authors identified the different phases of the user process in using OGD in a demand-driven context. They described three main building blocks, steps that are performed iteratively

by the users when he/she looks for data to answer questions. The three main activities are the following:

1. Identify
2. Acquire
3. Enrich

Because the research question aims to investigate the challenges users face in identifying and acquiring open government reference data, the author will use the first two phases of the OGD user process defined by Crusoe et al. (2019). The "identify" and "acquire" phases will be investigated, while the "enrich" phase is not in the scope of the current research.

Identity phase: explore and assess

The identify phase starts when users, based on their context, need to identify the data useful and valuable for their tasks and assess if the data are the right one to be used (Crusoe & Melin, 2018). The identify phase is composed of two steps: *explore* and *assess*. These two steps are characterized by technical and social interactions (Crusoe & Melin, 2018). For example, users may interact with domain specific communities using forum and social media or contact experts, at the same time they can use the internet, web search engines and dedicated online resources (Crusoe & Melin, 2018).

Explore

First, users need to explore and find data relevant for their activities through different resources and infrastructure managed by different actors (Crusoe & Melin, 2018). The exploration activity is the result of the actions performed by the user to discover and find data. According to Crusoe and Melin, findability can be defined as "the ability of the user to discover and identify the data, which to a degree can be solved with an open data portal" (Crusoe & Melin, 2018). The research topic of findability, the ability to quickly discover and access the relevant data, has been investigated by many authors in the field of the research related to OGD, and it is considered one of the essential preconditions to unlock the potential of OGD (Janssen et al., 2012; Kučera et al., 2013; Lourenço, 2015; Máchová et al., 2018; van Loenen, 2018; van Loenen et al., n.d.). Examples of tasks under this activity are users looking for data in OGD portals, contacting publishers or engaging in online forums (Crusoe & Ahlin, 2019).

Assess

After spending time exploring and finally finding the information, users need to assess and evaluate if the acquired information meets their expectations and can contribute to the mission of their activities in terms of relevance (Crusoe & Ahlin, 2019). Assessing is the user's activity to check and validate if the content meets the expectation in terms of qualities, properties and delivery methods (Crusoe & Ahlin, 2019). Examples of tasks under this step are metadata and data model analysis to verify the data and the interactions with domain experts.

Acquire phase: access and delivery

Consequently, the acquire phase is composed of two main steps that are highly interrelated: *access* and *delivery*. The user performs this phase after a successful identification phase, when he/she has already explored and assessed the data and wants to acquire it and then move to the enrich phase (Crusoe & Ahlin, 2019).

Access

First, users need to perform a list of steps to enable and set up the condition and infrastructure to gather data. Then, they physically have to prepare to connect to the data in order to boost the delivery of their services or products. Data infrastructures put in place by the data provider and methods available to retrieve information, play a crucial role in this phase to speed up and facilitate the acquisition of raw data (Crusoe & Ahlin, 2019). Users need to read documentation to understand how to download or transfer data; then they implement the actions needed to acquire the data either manually or in an automated way (Crusoe & Ahlin, 2019). Tasks under this phase include reading the technical specification to access the API, clicking to start the bulk download of a CSV or Excel file, registering to have an API key and developing the code needed to consume the data through API.

Finally, after the access to the correct information, the data should be concretely transferred from the provider to the users.

Delivery

The delivery phase can take place in different ways and the transfer of data can be automated or manual. Users can transfer data in several ways. For example, he/she can fetch data through an API, scrape a web page or a PDF or download a csv-file. In this phase, data needs to be stored and processed. At this stage, users spend time and need

resources by preparing the environment that can host the data to then move at the enrich phase (Crusoe & Ahlin, 2019).

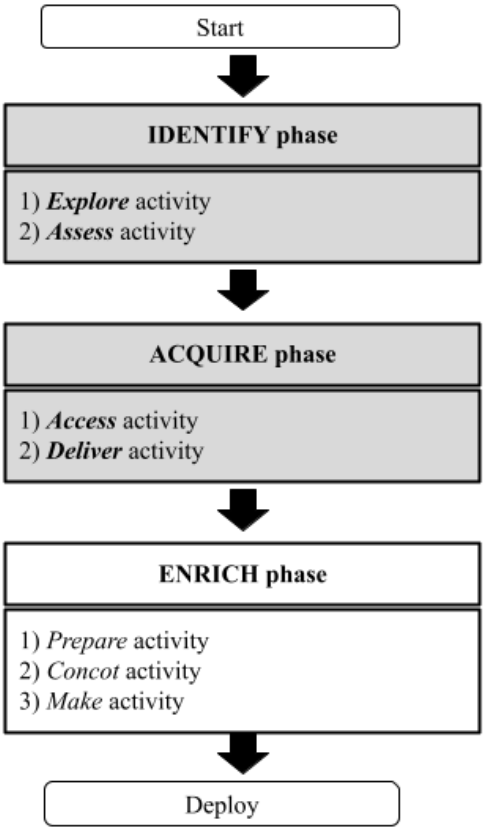


Figure 4: OGD user process (adapted from Crusoe et al., 2019)

2.2.3 Defining enablers

To properly implement efficient and effective data governance, the public sector, similar to any other organisation, needs to adopt a holistic approach and pay attention to processes, structure and people. Therefore, many contributions and guidelines from academia have been focused on how an organisation can achieve a robust IT governance, implementing structures, processes and relationships to deliver business value from IT investments (De Haes & Van Grembergen, 2015).

In this context, the Control Objectives for Information and Related Technologies framework (COBIT) is highly adopted and considered a comprehensive toolkit to support organisations in creating value managing the alignment between IT and business. COBIT is a management generic framework developed by the ISACA, released in its first version in 1996 and currently in its fifth edition published in 2012 (Jochen, 2019). The framework provides a clear approach to communicate goals, objectives and results between all the

stakeholders involved (Jochen, 2019). Furthermore, it aims to support organisations of any type, in developing, organizing and implementing strategies around information management and governance making a clear distinction between them (ISACA, 2012). Hence, COBIT 5 helps enterprises create optimal value from IT by maintaining a balance between realizing benefits and optimizing risk levels and resource use.

Unlocking a holistic approach and view on governance and management of information inside the organisation is the main goal of the COBIT framework (ISACA, 2012). According to COBIT, there are different categories of enablers that can drive and influence the success of such collective views. Enablers are defined as "*actors that, individually and collectively, influence whether something will work – in this case, governance and management of enterprise IT* " (ISACA, 2012). These enablers framework can be applied in practical situations and can be used to implement effective and efficient information governance and data management of IT governance. The enablers indicated in the framework help organisations in addressing real business needs and issues, adopting an holistic governance (ISACA, 2012).

Furthermore, COBIT enablers have four common dimensions: stakeholders, goals, life cycle and good practices.

Stakeholders are actors and participants with interest in the enabler. They can be internal or external to the organisation and their needs have to be translated into goals for the organisation (ISACA, 2012). Then, every enabler has clear goals to reach the expected outcomes and drive the creation of value for the organisation. In addition, every enabler goes through a number of phases defined by its life cycle and follows a set of good practices to achieve the above mentioned goals (ISACA, 2012).

According to COBIT, there are seven interconnected categories of enablers. The categories are the following:

- Principles, Policies, and Frameworks
- Processes
- Organizational Structures
- Culture, Ethics, and Behaviour
- Information
- Services, Infrastructure, and Applications

- People, Skills, and Competencies

Principles, Policies, and Frameworks

The Principles, Policies, and Frameworks enabler is considered a governance factor. It covers the actions and mechanisms adopted to translate in practice vision and mission to the stakeholders (ISACA, 2012). Principles, Policies, and Frameworks are tools to provide a clear overview and easy access to the rules, governance objectives and organisational values for all the stakeholders (ISACA, 2012).

Processes

The second important factor is related to processes. A process is a collection of actions and activities in the organisation to create specific and defined outputs to reach IT related goals and meet stakeholders needs. (ISACA, 2012). Processes are highly interdependent with the other enablers; they need and at the same time produce information, they require services capabilities such as infrastructure and applications and ultimately, they need clear policies and procedures to be aligned with the organisational structures (ISACA, 2012).

Organizational Structures

The third enabler is related to organisational structures, systems in charge of decision making in the organisation with a clear mandate and operating principles. Organisational structures should have defined roles, purpose and level of involvement with clear inputs and expected outputs. Concrete examples of organisational structures are the Chief executive officer (CEO), the highest-ranking person in an organisation, ultimately responsible for taking decisions regarding the total management or the Chief data officer (CDO) responsible for the governance and utilization of every data and information assets inside the organisation (‘Chief Data Officer’, 2021).

Culture, Ethics, and Behaviour

Culture, Ethics, and Behaviour are powerful and often underestimated enablers to realize proper governance in the organisations. It refers to the individual and collective behaviours that together shape the organisational ethics and values. They are highly dependent on external factors such as geography and socio-economic background. These behaviours influence different attitudes such as the risk-prone, the willingness to comply with policies, and resilience in facing and mitigating negative outcomes in the organisation (ISACA, 2012).

Then, the three remaining enablers are considered as well enterprise resources and need to be managed and governed as well.

Services, Infrastructure, and Applications

The Service enabler is related to the IT resources such as technologies, infrastructure and applications. The Services, infrastructure and applications enablers are defined by the IT ecosystem in its different components that provide services and procedures. In the context of an organisation, a successful service delivery is possible just with a proper architecture of infrastructure and applications. Good practices for these enablers include the definition of different viewpoints based on the needs of the actors involved, in the case of the current research based on the needs of the data users. In addition, the adoption of different architecture principles are important drivers to implement scalable and efficient IT components and solutions. Examples of these principles are agility, re-use and openness (ISACA, 2012).

People, Skills, and Competencies

The second enabler is represented by people, skills and competences. People with an appropriate level of technical and knowledge skills, experience and competences, and an appropriate number of resources, play an essential role in providing efficient management and delivery of all activities and goals in the organisation (ISACA, 2012).

Information

Finally, the information enabler covers the different types of information, structured and unstructured, and informal inside the organisation. Business processes through IT create and generate data that needs to be transformed in information and knowledge to ultimately create value for the entire organisation (ISACA, 2012). The frameworks identify different stakeholders involved with this enabler. The category of information and data users, that represent the one on which the current thesis focuses, is clearly identified in the list of the stakeholders. Furthermore, an important role in the context of this enabler is the quality of the information. Aspects such as the accessibility and availability of the information and, the ease of manipulation and the understandability are criteria used to assess the quality of the information available in an organisation (ISACA, 2012). In addition, the framework identifies different phases in the information life cycle: plan, design, build and use. The phase in which the information is used, includes as well the sharing phase.

To summarize, the application and implementation of the COBIT enablers, bring positive outcomes to the organisation. Focusing on questions such as “Are stakeholder needs addressed?” and “Are stakeholders facing barriers in accessing information?” are

concrete examples to assess what the stakeholders expect and understand which are the challenges they are facing in delivering value.

As COBIT 5 covers specifically the topics of data and knowledge, the author decides to adopt the seven enablers identified in the framework and presented above, to map and operationalize the insights and recommendations expressed during the interviews. In particular the aspects related to the information quality assessment and the activities related to the accessibility, availability, use and sharing of data are relevant in the context of the current research objectives.

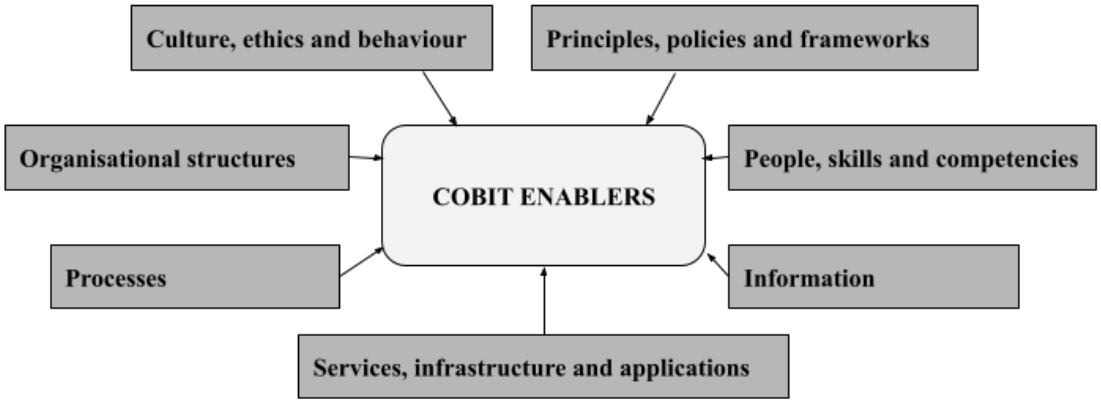


Figure 5: COBIT enablers, (ISACA, 2012)

3 Methodology

This chapter outlines the overall research design methodology followed by the author to answer the research question and drive the results of the thesis. The design of the research, how data have been collected, as well as the process of data analysis, are discussed.

3.1 Research Design

This research is a cross-section exploratory research that investigates barriers and enablers data analysts encounter in identifying and accessing open government reference data relevant at the EU level. This study investigates the following the research questions:

"What are the challenges that data analysts and experts face identifying and acquiring open government reference data (focusing on controlled vocabularies) at EU level in the context of a data-driven public sector? What are the key enablers to enhance them?"

The aim of the current research is to ask questions and try to assess and clarify the understanding of a problem at a specific point of time (Crompvoets et al., 2019) . The author, mainly based on qualitative research strategy, investigates the barriers data analysts face in the process of identifying and acquiring government reference data relevant at the EU level in the context of a DDPS. The study aims to contribute with empirical findings from data analysts and experts working in a DDPS context to the research topic of OGD user's process. Specifically, the research intends to contribute to the scientific field of open government reference data, clearly identified under the OGD Research Taxonomy presented by Charalabadis et al and explained in chapter 2 (Charalabidis et al., 2016).

In order to structure and guide the author to move from the research question to the answer of the problem statement, the model of the research "onion" suggested by Saunders et al. is followed (Saunders et al., 2009). Saunders' method clearly identifies six different aspects or layers that drive the author in structuring the thesis, defining the questions to be addressed and which can be summarized as "what", "why", "how", and "when" (Walliman, 2017). The identifies levels, from the more higher theoretical one, concerning the research perspective to the more practical one, on the techniques used, can be summarized as follow: 1) research philosophy; 2) research approach; 3) research strategy; 4) choice of method; 5) time horizon; 6) technique and procedures (Saunders et al., 2009).

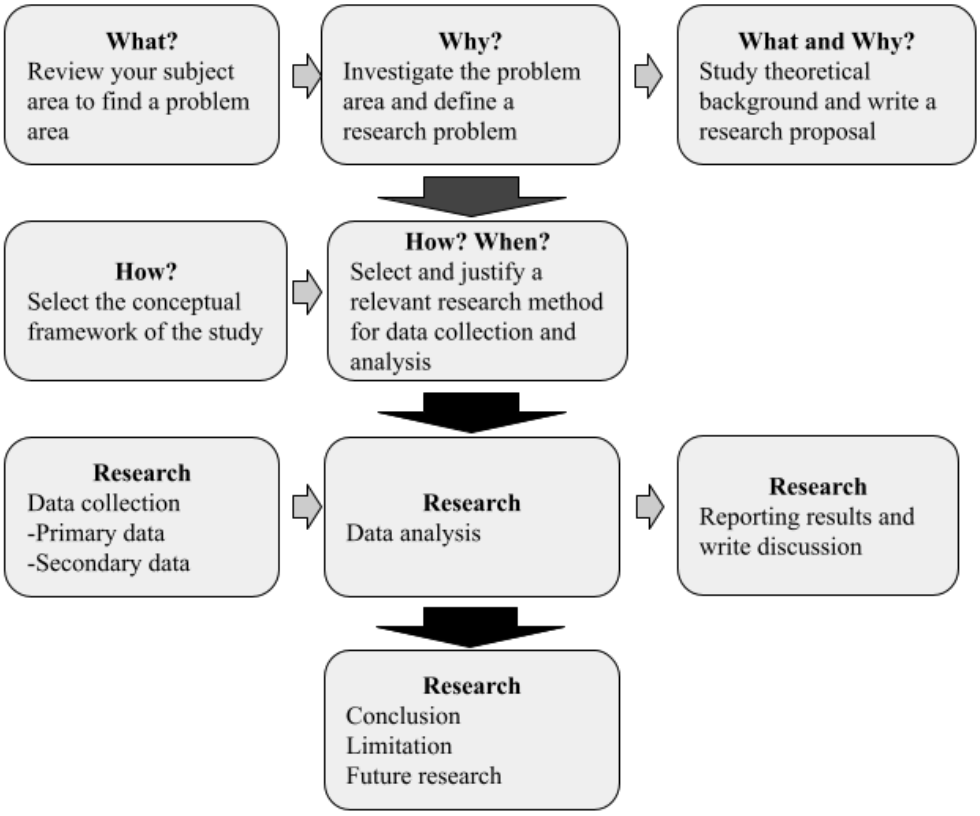


Figure 6: Research process used, suggested by Walliman (2017)

3.2 Data collection

Literature review, previous research, analysis of reports and material in the fields covered by the current research has been analysed to collect relevant secondary data. In addition, semi-structured interviews with data analysts working on a DDPS context and domain expert have been conducted to collect primary data. The following sections will provide more details on the data collection strategies used by the author.

3.2.1 Primary data

Semi-structured interviews have been conducted to obtain deeper insights from the user's perspective. The study uses semi-structured in-depth qualitative interviews to collect primary data from nine data analysts working in the European Commissions and five domain experts with knowledge in OGD, semantic data and reference data management.

Semi-structured qualitative interviews are a common method to investigate the experiences and explore the view of individual participants (Gill et al., 2008). This approach has been selected from the author for two main reasons. Firstly, it can improve the understanding of the specific challenges and needs users might face in the OGD

process and the identification of enablers to support them. Second, there is limited research specifically on the topic of re-using open government reference data from the user's perspective.

This method helps in guiding the areas of interest, and provides the opportunity to catch and discover information from the interviewer that may not have been covered before (Gill et al., 2008). In the context of the current research, semi-structured interviews have been used in order to provide a more structured and flexible approach with the participants. In addition, as Sauders et al. (2009) stated, semi-structured qualitative interviews allow flexibility during the conversation and the researcher can modulate and limit questions depending on the level of interaction and progress (Saunders et al., 2009).

Interviewee Selection

The author expects to collect different insights on user's barriers and enablers related to the identifying and acquiring phases of open reference data in the context of a DDPS. As described in the previous section 3.2, the DDPS context in which the research is focused is at the European level with a specific focus on the EC as a data-driven public organization.

The interviewees were chosen based on their relevance to the research question; they should be data analysts working in the context of a DDPS at EU level or experts about OGD, reference data management, and semantic web to provide and contribute with a more complete perspective. Therefore, two distinct groups were created to conduct the in-depth interview: namely data analysts and domain experts.

The participants selected for the study, had been approached by invitation via mail sent in May 2021. In total the emails with the invitation were sent to fifteen data analysts and eight experts. Finally, nine data analysts and five experts were interviewed in May and June 2021. All participants contributed on a voluntary basis. Due to the COVID restrictions, the interviews took place online using Microsoft Team. The interview lasted from 40 to 60 minutes. Two interviewees didn't want to be recorded. All the recordings are accessible and stored in a personal cloud-based folder.

Interview groups and profiles

The first group, data analysts, is the fundamental part of the data collection because the research focuses on the user perspective and especially on the barriers data analysts face in discovering and accessing open reference data in the context of a DDPS. In their daily activities, data analysts combine data from disparate sources and domains to identify

evidence or evaluate policy interventions in a DDPS, and invest huge amounts of time and resources in finding and accessing data assets (Janssen, van der Voort, et al., 2017). Therefore, an important prerequisite for a better understanding of the research question, was to involve data analysts working on multidisciplinary research on societal challenges on data-driven public organisations where reference data usage is highly relevant (Charalabidis et al., 2016).

Because of this, the researcher decided to target experts working as data analysts and advising policymakers at the EC. More precisely, the respondents contribute to a DDPS and have a role where data analysis related activities are central and relevant for their position. They currently work for Directorates Generals and departments, focusing on "Policy making and implementation" where policy makers engage with data analysts to design and implement better decision-making (European Commission, n.d.-a).

The research investigates the first two main phases of the OGD user process: the identifying and acquiring open reference data in the user process. Based on the research of Lnenicka and Nikiforova, the categories of "learn and explore" together with "search and filter" are considered the most important from users with basic level skills, according to the International Certification of Digital Literacy (ICDL) profiles (Lnenicka & Nikiforova, 2021). Hence, it is important to underline that:

- these tasks are expected to be performed from users with a basic level of digital skills
- the interviewers, based on the job profile and skills, should be categorised as advanced users and more advanced than basic users

Consequently, the level of data skills that can affect the validity of the research question specifically tailored on data analysts, should not represent a limitation.

Furthermore, the reliability of the findings related to the research question should not be affected by an inadequate level of skills from the user perspective.

ID Analyst	**Function**	**Department**	**Interview Date**
D1	Information Systems Analyst	Directorate-General for Communications Networks, Content and Technology (EC)	28/05/2021

D2	Policy Officer - Data Analyst	Directorate-General for Taxation and Customs Union (EC)	31/05/2021
D3	Intelligence Analysis Assistant - Planning, Budget, Reporting	Directorate-General for International Partnerships (EC)	08/06/2021
D4	Data analyst on EU budget and EU fiscal policy	Directorate-General for Budget (EC)	10/06/2021
D5	Economic and Market Analyst - Senior Investment programmes management	Directorate-General for Economic and Financial Affairs (EC)	11/06/2021
D6	Business Analysis Assistant and Financial reporting	Directorate-General for Budget (EC)	11/06/2021
D7	Knowledge Management Assistant	The Publications Office of the EU	17/06/2021
D8	Senior policy analyst	Directorate-General for Agriculture and Rural Development (EC)	22/06/2021
D9	Lead team on monitoring and indicators	Directorate-General for Regional and Urban policy (EC)	22/06/2021

Table 3: Data analysts interviewees

The second group that were targeted are domain experts with academic and professional experience in the domain of data governance, OGD, data interoperability, standardisation, reference data and semantic web. The participants contacted are currently working and researching topics related to OGD, reference data management, linked data technologies and semantics. They support companies in the private sector, collaborate with public institutions at international and national level. Some of the precipitants are specifically currently involved in activities related to the implementation of the different pillar of the European data strategy. Therefore, this group could provide more inputs and a more complete and detailed overview about barriers and enablers in the two phases of the user process and specifically to the governance and technical dimension. Consequently, experts may provide interesting insights and complement and confirm findings on the barriers and enablers from the users, based on their direct experience in working and possibly solving these issues.

ID Expert	Function	Institution / company	Why relevant	Interview Date
E1	Knowledge Management Officer	Publications Office of the European Union (OP)	Data curation. Ontology and taxonomies definition	01/06/2021
E2	Programme Manager at the EC	Directorate general for informatics (EC)	Interoperability. Standards. Core vocabularies. Metadata quality	07/06/2021
E3	Project Officer - Scientific	Joint Research Center (EC)	INSPIRE and ELISE interoperability ISA2 Action. Member of the Corporate Reference Data Board of the EC	23/06/2021
E4	Team leader on Data and metadata services Standards	Eurostat	Statistical metadata production and standardisation. Member of the Corporate Reference Data Board of the EC.	21/06/2021
E5	Open Government Data expert	Open data Expert	Contributor of the DCAP profile. Member of the Corporate Reference Data Board of the EC. Private sector.	23/06/2021

Table 4: Experts interviewees

Interview Questions

The conceptual framework presented in section 2.2 has been adopted to guide the development of the questionnaire to properly answer the research question. The main concepts of interest have been reflected while structuring the interview in fifth main sections, to cover all the important aspects of the current research and keywords in the research questions. The same structure is then used to examine the results of the interviews. The template with the interview guide is available in Appendix and represented in table 5.

After collecting some general data about the participants on their background and working experience, a second section is dedicated to understand better how a strategic use of data drives the context in which the participants work and perform their activities.

The two questions about the use of data and analytics, investigate if the environment where the participants work is data-driven and in which actions they contribute implementing a DDPS context (Ooijen et al., 2019).

The third section starts to go deeper in the topic of OGD and in particular of open reference data relevant in the EU context from a user's perspective, to gather info about the awareness and usage of them in the interviewer's activities. In this section, the questions aim to establish if and how participants know open reference data. The questions start from a generic one to check the awareness of open reference data in the participant's activities and then become more specific in order to gather information about the type of open reference data used, how the participants re-use them and for which reasons. This section, framed into the data-driven context assessed in the first part of the interview, shifts the focus of the research on the OGD user process researched by J. Crusoe and Ahlin (Crusoe & Ahlin, 2019).

Then, in the two last sections the author derived questions based on the conceptual framework of the user process, looking at the identity and acquire phases as identified by J. Crusoe and Ahlin (Crusoe & Ahlin, 2019).

In the first part, the questions regarding the identification phase investigates the activities performed by the participants in order to explore and assess open reference data (Crusoe & Ahlin, 2019). One open question prompts participants how they identify reference data. This question aims at assessing the tasks performed by the users when he/she wants to find data that should fit in the exploring and assessing activity (Crusoe & Ahlin, 2019).

Then, the questions shift the focus to the central aspects covered by the research questions: barriers and enablers in the user's activity of identifying OGD reference datasets.

In the last section, a set of questions address the main aspects covered in the second phase of the OGD user process, when the data analyst, after discovering the data, wants to set up the necessary steps and environment to retrieve the data from the publisher. In this phase, users first access then transfer the data. The first part of questions presented in this part establishes how participants access data, if they spend time and effort accessing and transferring open government reference data and the formats they usually use. Finally,

the last two questions investigate barriers and enablers in the acquire phase, which consists of the two main activities namely access and delivery (Crusoe & Ahlin, 2019).

The same questionnaire has been used both for the data analysts and the experts for comparative reasons, in order to provide a more complete overview from different perspectives, not just limited to the user's one but enriched by the experience of experts that work on these topics.

#	**Guiding questions**	**Important aspects for the research**
1	How do you use data and analytics in your activities?	DDPS context, opportunities,
2	What type of information and datasets do you use?	DDPS, data driven dimension of the context
3	Are open reference data produced at EU level and in the form of controlled vocabularies (taxonomies, code lists, authority tables and thesauri) relevant in the development of your analysis? If yes, why?	OGD, Open reference data, EU as DDPS context
4	What type of reference data do you use? How do you use them? For which purpose?	Open reference data, EU
5	How do you identify reference data such as controlled vocabularies?	Open reference data - OGD user process, Identification phase, exploring and accessing activities
6	What type of barriers did you face while trying to identify reference data, such as controlled vocabularies, provided at EU level relevant for your tasks?	Open reference data - OGD user process, barriers Identification phase, exploring and accessing activities
7	What do you think are key enablers in general to facilitate the identification of open reference data?	Open reference data - OGD user process, enablers Identification phase, exploring and accessing activities
8	Are reference data such as controlled vocabularies easy to acquire?	Open reference data - OGD user process, enablers Acquire phase

9	How do you access them?	Open reference data - OGD user process, enablers Acquire phase, access activity
10	Do you spend time and effort in accessing them?	Open reference data - OGD user process, enablers Acquire phase, access activity
11	Which format do you usually expect and use?	Open reference data - OGD user process, enablers Acquire phase, delivery activity
12	What type of barriers did you face while trying to acquire reference data such as controlled vocabularies, provided at EU level relevant for your tasks?	Open reference data - OGD user process, enablers Acquire phase, barriers
13	What do you think are key enablers in general to facilitate the acquisition of open reference data?	Open reference data - OGD user process, enablers Acquire phase, enablers

Table 5: List of interview questions and related aspects for the research

3.2.2 Secondary data

Numerous reports, documents and online resources, governmental documentation and policies produced by the institution of the EU and by many international organisations such as the OECD have been collected and reviewed.

An extensive systematic literature review was performed by collecting relevant information from various academic sources, government published data and grey literature. Thus, information for the literature review was collected online through Google Scholar, academic databases including KU Leuven's Limo article search, ProQuest, ResearchGate, and other online academic sources.

The literature review started defining clusters of keywords related to the main concepts covered in the theoretical framework from the research question (DDPS, OGD and reference data) and then filtering the results.

In order to collect information about the topic of a DDPS, several keywords have been used such as "public sector information", "data", "analytics", "data-analysis", "data-

driven" and "governance". Then, other keywords have been used to gain information more specifically on the topic of open government data combining "open data" or "open government data" or "OGD" with "public sector". Finally, to search relevant contributions on the topic of reference data, keywords such as "reference data", "lookup tables", "controlled vocabularies", "categorisations", "codelists" have been used.

OGD User's process and barriers

Next, in order to conceptualize the first part of the research question related to the exploration of the OGD user's process and barriers, a literature research was conducted to conceptualize users' impediments in the process of using OGD. The author used keywords combining "open data" or "open government data" with "findability", "accessibility", "barrier", "risk", "challenge", "impediment" and "user".

Initially, the author started analysing the FAIR principles and their alignment to be used as a theoretical framework to conceptualize barriers. Based on the FAIR principles, in order to boost and maximize knowledge generation and discovery, data should be Findable, Accessible, Interoperable and Reusable to the greatest extent possible by humans and machines (Wilkinson et al., 2016). In 2016, the FAIR principles were published and they are still nowadays a mainstream reference for data-producers and publishers to optimize the re-use of data (Wilkinson et al., 2016). The FAIR principles are a set of concepts rather than strict rules and technical specification, that define a set of independent practices to enable data findability, accessibility, discoverability and interoperability (Research Data Alliance FAIR Data Maturity Model Working Group, 2020). However, after extensive research, the author has not adopted the FAIR principles because the theoretical conceptualization of these principles does not cover the peculiarity of the OGD user's process.

Consequently, a deep analysis on open data user's perspective and on the conditions for fostering re-use of data have been performed. The author analysed many articles and identified the research guided by van Loenen et all in their research "How to assess the success of the open data ecosystem?" and the adoption of the Backx model. (Backx, 2003; Donker & van Loenen, 2017; van Loenen et al., n.d.). The concentric framework from Backx has been assessed, it describes "the open data supply from a user perspective" and provides a series of steps useful to verify if data are in good shape to be concretely findable, accessible and reusable. The framework concretely lists factors to be assessed from a user perspective but does not focus specifically on barriers.

When looking at barriers and user process, the work conducted in 2018 by Crusoe and Melin in "Investigating open government data barriers: A literature review and conceptualization" has been examined by the author (Crusoe & Melin, 2018). Because of

the clarity of their systematized approach in investigating OGD barriers and tentative to conceptualize an OGD user's process, the authors decided to contact Crusoe and Melin in order to discuss further the finding of their research and new works on the topic. They provided two additional useful and relevant papers: "Users' activities for using open government data – a process framework" from 2019 (Crusoe & Ahlin, 2019) and "The Impact of Impediments on Open Government Data Use: Insights from Users" from 2019 (Crusoe et al., 2019).

After analysing these articles, a decision was made to use the work from the above research (Crusoe et al., 2019; Crusoe & Ahlin, 2019; Crusoe & Melin, 2018) as the foundation of the conceptual framework because of its value in synthesizing previous research on process and barriers on OGD from a user perspective with an empirical and practical approach.

Enablers

In order to address the second part of the research question, related to enablers to foster the re-use of open reference data at EU level, the author decided to explore the application of COBIT's 5 enablers. Primary information regarding the COBIT 5 framework was obtained directly from ISACA.org through material provided during the semester in Leuven in the class of Business Information Management (BIS). Additional material regarding the framework was obtained through academic research databases with keywords "COBIT," "ENABLERS"." IT," and "governance."

In this thesis, the author proposes to use COBIT enablers as part of the theoretical framework to categorize the different enablers that participants will propose. In order to organize the different inputs from the interviewees regarding the second part of the research question on the key enablers to improve the discoverability and accessibility of open reference data, the author will use the COBIT enablers as guidance.

COBIT 5 covers specifically the topics of data and knowledge that are the main building blocks behind this research. Consequently, the author decides to adopt the seven enablers identified in the framework to derive the information collected from data analysts and experts about the enabler to improve the process of identification of open reference data at EU level.

3.3 Data analysis

The primary data were initially processed using inductive coding to allow the authors to identify categories, concepts and themes using a bottom-up approach. The inductive approach, differently from the deductive approach, is not based on predefined codes and

categories and allows the narrative to emerge from raw data (Creswell & Poth, 2016). The interviews were analysed processing the information with initial coding, a first summarization for each interview has been recorded in a data memo. Then, the author skimmed the interviews section by section and question by question, respectively for the data analysts and then for experts. Keywords, relationships and important sentences were highlighted and saved in a second data memo, divided by the interviewees. The data collected in this second memo has progressively analysed using coding that allowed the categorisation of the findings for every section covered.

At this stage, a mixed approach has been used to move categories to themes. More precisely, at this phase the categories of activities related to the DDPS from Ooijen et al., and described in the conceptual framework in section 2.2.1, helped the author in grouping the findings related to how users interact with data.

The same action has been performed for the data related to the two questions on enablers. The seven categories of COBIT enablers have been used to map the findings. Then, the activities covered in the identify and acquire phase, the explore and assess steps and the access and deliver steps, have been used to map categories to themes.

All coding was conducted in the same cloud-based document, the same where the recordings of the interviews were stored.

4 Results and Discussion

In total, nine data analysts working in the EC and five experts on OGD, reference data management, linked data and semantic participated in the research.

The structure of the following chapter reflects the way the interview had been formulated, going step by step through the main concepts covered by the research question. After some general questions prompted the participants in order to introduce themselves, the first part of the results contains findings related to the DDPS in which interviewees are involved. Then, the main information related to open reference data, their use and relevance accordingly to the participants are summarized. Next, based on the contributions, the aspects related to the identification and acquisition phase of the user's process, specifically focusing on points representing the barriers and enablers in re-using open reference data are presented.

For every section, first the main results from the data analysts and then from the experts group will be presented separately, followed by a single discussion for every section, in order to have a clear structure and convenient format for a comprehensive approach in reading the document.

4.1 Data-driven public sector

After collecting some general information about the interviewees, the discussion started by asking them about their activities and experience related to data and data analytics. In order to gather insights related to the DDPS, questions such as "How do you use data and analytics in your activities?" and "What type of information or datasets do you use?" helped to gain rich and wide answers on the peculiarity of using data and analytics in their work. These questions helped the author understand the magnitude and relevance of the data dimension and the data-driven environment in which the participants contribute. After the presentation of the main results, the author discusses the findings on how data and analytics are used as a strategic enablers to transform policy making and provide better services, focusing on the type of data, area of interests and the different thematic categorisation used.

4.1.1 Use of data and analytics - Results

Data analysts

In general, many participants mentioned their tasks as part of the policy designs, evaluation and monitoring lifecycle. The main part of the participants engages every day with data and analytics activities, to answer different needs that mainly depend on the thematic and business goals of their directorate-general and department. In addition to

this, some interviewees manage generic tasks of the data analytics lifecycle such as data cleaning, data harmonisation and integrations, and data classification and data quality check. The majority of respondents are in charge of specific and limited activities related to data, at the same time a small portion is involved in different phases of the data journey at the same time. Interviewees D1 and D2 write and develop code by themselves to perform these tasks and seem to not highly rely on external solutions or tools.

Different needs and areas of interest drive the activities of the interviewers. For example, many of them contribute in reporting financial performance and monitoring Key Performance Indicators (KPI) (D3, D8, D9) using Business intelligence (BI) and corporate tools. Some other interviewees use more advanced data techniques to create predictive analytics and forecasting models, and provide real time insights and analysis.

Interviewer D1 is specifically focused on NLP and semantic technologies to perform entity recognition, data classification and knowledge linking in order to collect and structure information from unstructured and dispersed public data into a knowledge graph to use as input to decision making.

In addition, the interviewer D8 engages in activities more broadly related to data governance, defining and organizing the different data assets covered by the unit in order to create an inventory to facilitate the findability and reusability of knowledge inside and outside the EC. The same interviewer has been the only one mentioning specific and dedicated internal activities related to the creation of reusable and interoperable data to fuel the EU data economy.

Similarly, D7's activities are related to curation, harmonisation and analysis of metadata related to the data assets published by the MSs in the European open data catalogue.

Many of the participants contribute in data dissemination activities, promoting data-driven approaches with various formats and through different communication channels such as publications, dashboards, infographic and data storytelling. Interviewer D9 also argues the importance of telling stories with data in a simple, effective and interactive way in order to disseminate information to different users using different formats.

Experts

As for the use of data and analytics, the experts are involved in many different projects with many stakeholders. In these projects, experts shared a common set of aspects in which they contribute such as *data curation, data quality and standardization* of semantic assets in order to boost discoverability, *interoperability* and finally re-use of public sector

data across the MS and institutions. All the interviewees repeatedly stress the efforts in activities related to *metadata* management and their importance.

The majority of respondents contribute to activities to developing *semantic representations*, aligning with semantic web standards to maximize and achieve semantic and technical interoperability between different domains inside and outside the EU institutions. Experts E1 and E2 are particularly involved in the *definition, curation and unification of ontologies, taxonomies, core vocabularies and knowledge graph*, to interlink objects, concepts and entities. Particularly, one expert is focused on defining and improving multilingual the EurosSciVoc taxonomy related to fields of science, in order to categorize CORDIS projects, and improve their discoverability and standardisation using linked data and NLP technologies.

Expert E3 is in charge of assessing and monitoring the implementation of good practise along the MS on *location data interoperability* and usage of metadata. Finally, expert E4 contributes mainly in specific activities related to production and dissemination of *statistical data and metadata* at EU level based on SDMX.

4.1.2 Use of data and analytics - Discussion

The findings suggest the users' focus and their context in concretely fostering the adoption of data-driven approaches and recognising the value of data as a strategic asset.

All the data analysts showed themself familiar with data and data analytics usage in their daily activity with differences in terms of level of engagement, knowledge and expertise. The findings confirm that different phases of the analytics lifecycle are necessary in the definition of a DDPS and, at the same time different actions are needed to concretely boost a data-driven ecosystem. After analysing the results of the interview conducted with the data analysts, the author suggests two distinguished phases seem to emerge in the implementation of a DDPS. On one hand, users are still contributing in the phase of planning and preparing preliminary activities necessary to frame the data-ecosystem to enable data-driven opportunities. On the other side, users rely on a data-driven context that is already properly set up and ready to use to answer their needs. In general, a strong need for strategic use of data, information and knowledge emerges as a clear goal from the data analysts and experts.

Experts confirmed the need of a set of enablers and transformations to properly boost a successful DDPS through data. They provided a series of best practices finalized to the realization of a proper multi-level data governance and data analytics lifecycle. The findings both from experts and data analysts confirm the uniqueness and peculiarities of

the EU, where various actors such as the EU institutions, the MS, the regional and local authorities cooperate and share information. Consequently, there is a paramount need to properly address a consistent multi-level data governance and enable data interoperability at different levels. Findings related to semantic and technical interoperability and activities aimed in improving data discoverability through metadata, data harmonisation and data curation are aligned with previous research from Máchová et. al. (2018) and Cavanillas et al. (2016) on prerequisites and needs to implement a DDPS.

This confirms that data analysts and the context in which they contribute, are aware of the main priorities to address and are committed to implementing internal processes and practices to foster findable, accessible, interoperable data.

Looking at the data-driven activities and projects from the data-analysts and according to Ooiijen et al., the adoption of a DDPS seems to be prevalent under the areas of Anticipatory governance with a minor presence of activities related to the Performance management area.

Findings identify activities focused in forecasting future needs at EU level and comparing scenarios, in these contexts financial and economic indicators seems prevalent. In parallel, actions to improve the use of data to monitor financial and budgeting resources and focus on performance management are also reflected in the data collection. It is important to note that the findings demonstrate that data analysts, as well as experts, are aware of the lack of data-related skills and of the importance of data dissemination and data literacy activities to promote a data-driven culture as the previous research from Kalampokis et al.s (2013) and Ooijen et al.'s (2019) indicate.

4.1.3 Data and type of information - Results

Data analysts

Depending on the various needs and provenience of the interviewees, an heterogeneity of level of data access, types and data domains has been observed.

Mostly, information with different levels of access is used in the DDPS context described by the participants. All the participants except two access and use *public sector data with restricted and internal access*. This information can come from different sources, internal processes can generate it in the unit in which they are working, as well as from other DGs inside the EC or from external agencies, MS and other EU institutions such as the European Bank.

At the same time, all the interviewees use *OGD*, as well as reports and publications, published from internal and external international bodies such as the OECD, Eurostat and

the United Nations. Just one interviewer mentioned using open data not directly created inside the public sector but specifically collected from citizens and semi-structure public data from wikibase.

Two out of nine use private sector paid data (D2, D6), more precisely they need to access data related to companies, such as company reports, financial indicators and ownership information, that are not available in the Business registries used at MS level. Because of this, they have to pay for access to *private data* from a company database named ORBIS (Bureau van Dijk, n.d.)

The information used by the data analysts covered specific thematic areas. Mostly are interested in *economic and financial data* (D2, D3, D4, D6, D8, D9) related to different topics such as public procurement, borders duties, custom declarations, cohesion policies in the EU, the EU's common agricultural policy (CAP), indicators related to the development sustainable goals (SDG) and development co-operation efforts. Under this category, some interviewees (D4, D6) expressed to use data specifically related to budget lines at EU and MS level. In addition, D2 and D8 are mainly focused on data and information concerning the importation and exportation of goods and coming from the integrated Tariff of the European Union (TARIC), a multilingual database containing measures on custom tariff, agricultural and commercial legislation. D3 at the same time uses financial data to analyse and monitor development co-operation, policies and implementations in developing countries related to the Development Assistance Committee (DAC) mandate coming from different sources such as the UN and the OECD. D4 as well, uses different internal and external data sources to provide insights on the EU budget, such as the Annual macro-economic database of the European Commission's Directorate General for Economic and Financial Affairs (AMECO) database and data from the EU data portal on the European structural and investments funds.

Finally, one interviewer, differently from the others, does not cover specific thematic data but focuses specifically on *metadata* and controlled vocabularies defined in the DCAT_AP and DCAT_AP_OP. The metadata are used in the context of the European data portal and by the MS and EU institutions.

Experts

Experts interact with the topic of *metadata* and *data* from different and multiple angles. *Multilingualism* is a common feature of the metadata and data which all the experts mentioned. *OGD* is another recurrent one. E1 contributes to the definition and curation of the EuroSciVoc taxonomy, a multilingual taxonomy representing the main fields of science with 1000 categories in 6 languages (English, French, German, Italian, Polish and

Spanish). Additionally, E2, E3 and E5 contribute and support interoperability solutions for the public sector, specifically focused on *metadata* aspects such as the DCAT-AP profile for data portals, based on the W3C's Data Catalogue vocabulary for describing public sector datasets. Then, E2 mentions the needs of common definition and *standardisation* in the context of the implementation of the Single Digital Gateway (SDG) regulation, which aims to standardize access to procedures, information to citizens and companies across the MS. In addition, E3's attention is more on location and *geographical data and metadata* coming from the different MS in the context of the Location Interoperability Framework Observatory (LIFO) and the European Location Interoperability Solutions for e-Government (ELISA).

4.1.4 Data and type of information - Discussion

Findings confirm that data analysts and consequently the DDPS, can benefit from a variety of information generated from different sources and on different topics, according to similar empirical studies conducted previously (M. Janssen, van der Voort, et al., 2017). Related to the theme and category of data, economic and financial data seems to be relevant and highly used in different contexts. This finding can be partially explained by the fact that the participants contribute in areas where economic data represents one of the main information assets.

Interviewees clearly elaborated on the need of using data with different levels of access. Finding confirms the important role of OGD as one of the main sources for creating and delivering data-based products. These findings are consistent with previous results on the benefit of OGD (Charalabidis et al., 2018; M. Janssen, Konopnicki, et al., 2017; Kalampokis et al., 2013; Ooijen et al., 2019; Toots et al., 2017).

This aspect provides a valuable interesting insight: in this specific case, the public sector itself is at the same time the producer and the user of OGD. These examples of re-use provide empirical data on the implementation of the OGD principles that consider the public sector itself one of the main beneficiaries of OGD. Consequently, these results could support the investigations on the limited availability of empirical data to assess the impact and effects of OGD initiatives (Charalabidis et al., 2018; Crusoe & Melin, 2018; Donker & van Loenen, 2017; M. Janssen et al., 2012; M. Janssen, Konopnicki, et al., 2017; Lněnička et al., 2021; Lourenço, 2015; Wang & Lo, 2016).

Additionally, looking at the different types of data access, it is relevant to mention that some interviewees need to access data that are sold from private companies such as the information related to the business registry of the MS. This category of information is

clearly classified by the 2019 EU Open Data Directive as a High Value dataset that needs to be accessible as OGD because of its numerous benefits and potential of OGD re-use.

The importance of metadata, as findings revealed, has been mentioned both from experts and data analysts. Metadata are prerequisites that enable and ensure data findability and re-use as numerous previous empirical and theoretical studies claimed (Charalabidis et al., 2018; Crusoe et al., 2019; Lněnička et al., 2021). In particular, the aspect of multilingualism of metadata in the EU context is extremely relevant to enable a proper re-use of data produced at different levels.

4.2 Open government reference data and user process

After a general discussion on the DDPS context, the interview moved more to the central topic of the research: open government reference data. Initially, the questionnaire presented open-ended questions about open reference data, aimed to collect input and general discussion on open reference data. Then, in the scope of this research, the *identity* and the *acquire* phase of the OGD user's process related to open reference data have been presented and investigated.

4.2.1 Open government reference data - Results

In the first part, the interviewees were asked three main questions which were more specific to start investigating this topic. Both data analysts and experts were asked open-ended questions as follows: "Are open government reference data produced at EU level and in the form of controlled vocabularies (taxonomies, code lists, authority tables and thesauri) relevant in the development of your analysis? If yes, why?", "What type of reference data do you use?", "How do you use them? For which purpose?".

Data analysts

All interview participants have experience with using open government reference data. Just two of them were initially asked to provide some examples of controlled vocabularies to confirm their definition of reference data and use a common definition. The overall response to the question of why open reference data are relevant is because of the European Union peculiarities, where the *need to collect, harmonize and standardise information* of various types from different entities such as the MS and external institutions is crucial. Many participants express the need to *decrease ambiguity, interlink data* and provide reliable and *interoperable data*. The *need for standardized classification and controlled vocabulary* in order to connect data from different sources and cross MS has been clearly expressed by many of the participants (D1, D2, D3, D6, D8, D9). D1, in particular, underlined the importance of using common controlled vocabulary in internal

as well as external classification, arguing that in some contexts, controlled vocabulary used in the EC are *not aligned* with external resources.

Most of the participants, except D4 and D5, clearly listed some controlled vocabulary especially in the form of *code lists* used in their daily activities. D2, D8 and D9 explained in detail how the usage of controlled vocabularies enable *cross country* collection and analysis of data.

D7, because of the peculiarities of the activities related to portal data.europa.eu, defined the adoption of controlled vocabularies related to the DCAT_AP profile, which is crucial and necessary for the collection and categorisation of OGD from the MS.

When asked about the different types of reference data used, most participants answered with examples of controlled vocabularies without clearly expressing if they were taxonomies, code lists, authority tables or thesauruses.

Many interviewees use *code lists* and a minority of them *thesauruses* produced both at the EU level and international level. D1, D7 use the multilingual and multidisciplinary Eurovoc and controlled vocabularies managed from the OP. The NUTS nomenclature, which classifies the EU in territorial units and is managed by Eurostat, is used by many participants to collect and analyse socio-economic indicators and regional policies. Interviewees D2 and D8 mentioned several times the usage of many code lists from the TARIC database, such as the goods nomenclature and the agricultural components.

Other reference data produced at EU level and used by the participants are the statistical classification of economic activities (NACE) from Eurostat, the combined nomenclature (CN) for classifying goods based on common customs Tariff for EU's external intra trade statistics.

D6, D8 and D9 use additional reference data produced at the EU level to identify the intervention fields such as the ESIF cohesion categorisation and the different regional programs, and many categorisations used in the EU budget.

The majority of respondents mentioned the use of referenda data produced and maintained from *international organisations* external to the EU institutions, such as the country code lists, the Harmonized System (HS) with the type of goods classifications and international trade managed by the World Customs Organization.

Interestingly, D3 acknowledged that the country code list used in his activities is provided by the European external actions servers (EEAS) from the EU's diplomatic service instead of the authoritative table provided by the OP and based on the ISO 3166 international

standard. This decision is because some countries are currently in a transition phase, such as the case of Palestine or Kosovo.

Experts

For comparative reasons, as already explained in chapter three, the same questions used for the data analysts have been prompted to the expert group even if the focus of this initial section on open reference data was more oriented towards the user's perspective covered by the data analyst group.

Nevertheless, all experts agreed on the importance and relevance of open reference data produced both at the EU and international level, mainly for *interoperability and standardisation needs*. Based on their expertise and experience in the fields, many of them mentioned as well the existence of reference data commonly used inside the organisation that is not publicly available because they are managed and created without proper reference data governance. All experts underlined as well the crucial role of *metadata* related to open reference data and OGD in general. Many of the experts worked in activities related to interoperability and focused their activities on the topic of metadata. Some of the experts from EU institutions mentioned that they are currently working in activities related to the definition of a reference data governance across organisations because there is a growing recognition about the importance of having reusable and accessible reference data across them and with the MS.

4.2.2 Open government reference data - Discussion

Data analysts show themself aligned on their response about the relevance of open reference data in their activities, providing a set of common motivations. The majority had no difficulties in explaining and recognizing the importance of controlled vocabularies in the context of their data related activities. However, when asked about the type of reference data used, according to the classification provided by Hedden (2010), participants did not list the different types of controlled vocabularies but examples of controlled vocabulary. They probably are not aware of the different classifications but focus on the specific type and on the name of the resources. Consequently, findings suggested data analysts do not clearly distinguish the different types of controlled vocabularies. On the other side, experts provided information on the different types of resources, mentioning code lists and thesaurus as examples. Additionally, open reference data produced at the international level and non from EU-institutions, as results revealed, represent an additional important source of reference data. This element suggests the need of an international data governance on open reference data in which all the main institutional players should have a clear role.

It is important to note that both data analysts and experts confirmed the relevance of open reference data claimed by (Charalabidis et al., 2018). (2018) and the European Commission (2014), in fostering interoperability and data linkage thanks to the adoption of common classification across different domains and languages.

Furthermore, data revealed the benefits of multilingual classifications provided by open reference data produced at EU level to enable the re-use of different sources of data from the MS. The multilingual aspect is relevant as well for the definition of reference data used in the metadata domain such as the one defined by the DCAT_AP profile and available in the different EU languages.

4.2.3 Identification phase - Results

Next, after the last part of the interview aimed to collect insights on the level of awareness of open reference data, the interviewees were asked questions about the *identification* phase, the first phase of the OGD user's process in the scope of this research. According to the conceptual framework, the identify phase is composed of two steps: *exploration* and *assessment*. The phase starts when users look for data and ends when they identify or are not able to find the information they were looking for (Crusoe & Ahlin, 2019). Both data analysts and experts were asked open-ended questions as follows: "How do you identify reference data such as controlled vocabularies?", "What type of barriers did you face while trying to identify reference data, such as controlled vocabularies, provided at EU level relevant for your tasks?" and "What do you think are key enablers in general to facilitate the identification of open reference data?"

Identifying open government reference data

Data analysts

According to the interview, data analysts use different and heterogeneous approaches simultaneously to identify open government reference data they are interested in.

All of them reported using *online resources*. Many data analysts mentioned using public *search engines* such as Google as a first step when they are looking for reference data. They think it is the quickest way to retrieve relevant results from many providers at the same time and discover new platforms and resources they were not aware of. Another group of data analysts mentioned relying directly on specific *public online repositories*. It is the case of one data analyst that uses GitHub repositories where controlled vocabularies are published in the specific context of the DCAP_AP initiative. Other participants used additional *public websites* as well to identify reference data controlled vocabulary such as Wikidata or Geonames to gather geographical information.

Four data analysts reported the use of *public sector EU online resources* from the EU institutions such as Ramon from Eurostat and the EU Vocabularies web site from OP as their first place to identify reference data.

Additionally, three users interact with *internal corporate tools* such as the reporting and business intelligence system to identify reference data they are looking for and that are already available internally.

The second recurrent approach, reported by six participants out of nine, consists of having *social interactions* with a quick phone call or email, with colleagues or experts from the same domain or organisation. Calling colleagues on the phone to gather information to identify reference data, is perceived as one of the safest solutions in order to reduce ambiguity and errors and identify reliable information.

Experts

The ways experts use to identify relevant reference data are partially similar to those expressed by the data analysts. Many of them rely on a network of colleagues working in the field and use social interactions to instantly have confirmation or receive information on specific reference data needed. One mentioned as well the usage of Google as a search engine to have quick results and updates about new potential resources from external providers. Finally, two experts expressed the usage of public sector EU online resources such as Ramon and the ones developed by the OP.

Barriers in the exploring activity

Data analysts

Findings reveal that resources are spared in many websites and platforms and there is not a unique point of access where users can easily identify them. Related to these barriers, this is the reason why some data analysts use web search engines as a first step to identify resources. Open Reference data produced at EU level are perceived as *silos* by many data analysts. They face issues in identifying resources that should be horizontally shared between institutions and commonly used, such for example the countries list that seems to be not unique as one referred. Many participants reported barriers in identifying the right resources between many *ambiguities* and easily understanding how this reference dataset can be linked to others.

Additionally, one participant mentioned that he discovered reference data by chance: *accidentally*, in the process of looking for some other resources, he found out the existence of open reference data produced at EU level that he was not aware of. On the

other hand, he reported the fact that in his opinion, just a minor part of controlled vocabularies relevant at EU level are online and available to the users as OGD.

Related to this, one interviewee stated that the reference data available as OGD at eu level are *incomplete*, in some cases he was not able to identify open reference data and then he discovered that these resources did not exist online and were not present in any portal or public repository.

The majority of respondents reported a general *lack of proper user and technical documentation*, and necessary conditions to guide the user in identifying, understanding, and re-using the content. Participants D5 argued that information provided to the users in the web portals and repository is confusing and not effective, he mentioned that in specific cases, even if the provider publishes long text, at the end of the reading it is not clear the content of the resource and how this resource can be linked to third party dataset. Additionally, D2 reported that in some EU portals where reference data are available, normal users need *support from experts* in order to understand the content of the resources. The documentation provided is too technical and difficult to analyse, it is not possible to quickly verify if the content corresponds to what the user is looking for. Many of the data analysts reported a general *lack of metadata* that can support their exploratory phase. D9 mentioned the difficulties of finding *multilingual* reference data that are essential for the task he needs to perform linking information from the MS.

Finally, three participants mentioned difficulties in interacting with the *User interface (UI)* and *a bad user experience (UX).*

Experts

Experts provided many inputs on barriers faced in the process of identifying open reference data at EU level, both in regards to the exploratory and assessment steps.

Experts mentioned numerous barriers in the activities performed to explore the open reference data they are looking for. Participants reported many impediments in the area of information, the *high level of fragmentation* and the lack of a *central repository* at the EU level has been mentioned by all the participants. E1 mentioned that to explore and successfully find the expected resources, users need to *have a high level of expertise* in the field of reference data and, these tasks can be really challenging for normal users such as citizens, journalists and not domain experts. According to E1, even domain experts do not have a complete and clear *overview* and *documentation* of the resource available and where to find them. Similarly, interviewee E4 argues that because open reference data can be freely copied, many online resources report the same information and it is confusing and difficult to understand when the resource is the official one and which is

the *single source of truth.* Almost two-thirds of the experts commented that the lack of clear technical and not technical *documentation* represents an important barrier in this phase. On these impediments, E1 expressed his concern about the *lack of expertise* in people involved in the documentation process, who are often more policy and less technically oriented.

The majority of experts reported impediments related to a poor *user experience (UX)* and low usability of the *user interface (UI)* of the institutional provider of open reference data. They argue that the user's needs are not put at the centre and many times, they need to spend time in understanding how to query the system and understand if the dataset they are looking for is available in the portal.

Barriers in the assessing activity

Data analysts

All the data analysts reported barriers related to a *lack of user and technical documentation* in the assess phase, as already mentioned in the exploratory phase. For example, users have difficulties understanding precisely the information covered in the resources and access the most appropriate resource when multiple providers provide similar concepts. More specifically, three participants mentioned a *scarcity of metadata* that are published together with the resource. Metadata related to the thematic category covered by the resource, the data owner, the date of the last update, the license are not present or difficult to access. Always related to the documentation, a small minority reported the lack of information related to the *contact point* from the data providers without mentioning the concept of metadata. The information related to the contact point, according to the participants, represents important information that supports the data analyst in assessing the content of the resources. Six out of nine participants associated the lack of proper information about the *versioning* and the last update date for a resource as a barrier to identifying the correct reference dataset. The lack of version management of an open reference data set that can vary along the time, such as the case of the code lists with the EU funding programmes, is a recurrent barrier inside the data analysts.

Related to the *UX and UI experience***,** two data analysts mentioned that in some cases, after the release of the new UI, they were not able to access resources that were accessed in the past. For example, reference data, easily accessible before, became unavailable and they were not able to find them anymore. Additionally, D4 referred to the obstacle of assessing the resources because some portals force the user to go through a *registration process* to access reference data instead of having accessible data by default as defined by the open data principles.

Experts

Experts shared different barriers they and users may face in the access phase. Some of them reported impediments related to a lack of metadata, for example, related to the data owner as well as the updated date of the resources. Many experts mentioned the impediments in easily verifying that the resources accessed are updated and represent tha last version available. Additionally, E1 reported the difficulty in accessing similar reference data and querying them by using alignments.

Two experts have expressed the barriers related to a lack of proper technical documentation. In particular, they highlighted that some institutional data providers do not publish the resources using formats easy to access such as csv-files. Some organisations provide access to their reference data using advanced formats, e.g., linked data, and this can represent a barrier if clear and exhaustive documentation to support the user is not available. In regards to the UI and UX aspects, three experts mentioned the lack of a proper UI that supports users in accessing resources, for example, it is difficult to find the web pages where there are direct links to download the dataset in different formats.

<table>
<tr><th colspan="2">Barriers in the Identification phase</th></tr>
<tr><td colspan="2">Barriers in the exploring activity</td></tr>
<tr><td>Users</td><td>Experts</td></tr>
<tr><td>Fragmentation of resources
Data providers similar to silos
Ambiguities in the description of the resources
Lack of user and technical documentation
Complicated technical documentation
Lack of metadata
Lack of multilingualism
Bad UI and UX</td><td>Fragmentation of resources
Verification of single source of truth
Lack of user and technical documentation
Complex technical documentation, difficult to understand
Lack of technical expertise
Bad UI and UX</td></tr>
<tr><td colspan="2">Barriers in the accessing activity</td></tr>
<tr><td>Users</td><td>Experts</td></tr>
<tr><td>Lack of user and technical documentation</td><td>Lack of metadata
Lack of technical documentation</td></tr>
</table>

• Lack of metadata • Contact point • Lack of version management • Difficulties with the UI in accessing the resources • Registration process	• Difficulties with the UI and bad UX

Table 6: Barriers in the Identification phase

Enablers for the Identification phase

Data analysts

Based on the challenges and impediments users reported, the participants were asked to suggest and provide key enablers to improve the identification phase.

Looking at the organizational structure and principles, policies, and frameworks enablers, many data analysts suggested the implementation of an effective data governance, defined not just at EU level but at supra-international level including as well other international organisations that contribute in the management of reference data. As D7 proposed, this enabler should improve the alignment between similar semantic concepts defined in the controlled vocabularies defined at the international level. Additionally, some participants suggest the definition of clear roles and responsibilities to ensure consistent and coherent feeding of data.

Related to processes, D1 and D2 mentioned the adoption of new processes to foster automatic findability and discoverability of reference data assets inside the EC that should be published as OGD.

Under the area of Services, Infrastructure and Application, four users proposed creating a single central repository for managing open reference at the EU level. D3 in addition suggested enabling this central repository with the functionality of delivering content with APIs, to directly discover and automatically import reference data in BI tools used to create reports and dashboards. D1 supports the adoption of OpenAPI specification to describe API and facilitate the discoverability and usability of open reference data. Finally, three users advised to invest in improving the UI and UX of the existing web applications where EU reference data are available to provide a better experience based on users' needs and maximize the discoverability of resources at the moment not easy to find.

Many other inputs under the information enablers have been suggested during the interviews with the data analysts. All of them proposed to invest in better and multilingual

documentation. Many data analysts promote the adoption of an extensive usage of metadata to support users in the findability phase. D9 suggests using data storytelling, using open reference data, to easily inform the users about the content of the data and how to use them in different contests. D7 and D9 proposed to measure user satisfaction and promote surveys to collect user feedback and suggestions. Additionally, D2 suggested the implementation of automatisms that trigger notifications to the users as soon as new resources or new versions of reference data are available and published.

Experts

After assessing barriers from the expert's perspective, the participants have been asked to share improvements and key enablers based on their relevant knowledge and expertise in the domain.

Under the area of *organisational structure* and *principles, policies, and frameworks,* all experts agreed on the need to promote and implement proper *data governance* on reference data management at the EU and international level. Another one expressed the need to define a strategic view and clearly *define roles and responsibilities* at the different levels of the organisations. Some of them suggested investing in concrete initiatives to foster synergies and collaboration between the different actors involved in reference data management, and *create a network* including institutions external to the EU context. One expert underlined the need to involve more technical people in the definition of policies and implementations related to data management in general. This set of recommendations should enable better governance and consequently improve discoverability of open reference data.

Looking at *processes,* E2 suggested analysing and redesigning processes related to the publication of reference data. Based on his experience, current processes do not scale and are too slow and consequently impact the creation of value and discoverability of reference data. He suggested learning and taking inspiration from successful implementations and use cases of public administrations that embrace more *agile* and *collaborative processes and approaches*, as for example the integration of Wikibase technologies (Use Cases for Wikibase, n.d.). The participant provided examples of collaborative knowledge creation that seem to be quite common in the digital humanities field. In particular, he suggested fostering the exchange of good practise as input for new solutions and integration with current tools adopted in reference data management. He recommended promoting pilot projects and prototypes inside the European institution to explore and assess new collaborative ways to manage open government data.

Related to the *People, Skills, and Competencies* enablers, one expert provided a strong recommendation that can be extended not just to the identification phase but also at a general level. The participant underlined the importance of *fostering knowledge* between the users and investing in data literacy and the development of data competencies across the institutions. In his opinion, the lack of technical and specific skills in data management is the main reason behind the lack of data governance inside the different institutions. He expressed the need to provide roadmap of competencies relevant to data literacy required for specialist and non-specialists.

Experts addressed numerous suggestions that can be categorized under the area of the *information* enabler. Many of them recommended investing in *multilingual* metadata and data in order to facilitate the discoverability of reference data relevant at the EU level to all the MS. Many of them express the need to invest in *data standardisation* and *semantic harmonisation* between the different organisations in order to reduce and limit the heterogeneity of similar concepts and definition, and uniform the way these are expressed. Some of the experts strongly recommend the adoption of *persistent identifiers* to reference open reference data, in this way, the resource can be constantly accessed regardless of changes. Two experts suggest specifying and adopting *common licenses* used to publish open reference data; according to them licenses are sometimes not expressed and not easy to understand. Some experts underlined the importance of fostering the adoption of *semantic technologies* to link different resources from different providers and facilitate the discoverability of reference data based on knowledge graph technologies.

Other two recommendations have been collected from the experts under the information enabler. One expert suggests using a more comprehensive and less technical *terminology* that supports users in the process of identifying resources. According to the participant, terms such as controlled vocabularies or reference data need to be defined in more detail with list of examples. Non-expert users may not know that categorisations of concepts are called reference data. It may help to communicate and describe the role and power of controlled vocabularies with use cases and explanations in order to foster knowledge and understanding of reference data, uses and applications.

Finally, another expert mentioned investing in *communication and branding activities* to raise awareness and promote the discoverability of open reference data.

The participants expressed numerous recommendations related to the improvement of *services, infrastructure and applications.* Many of them are related to *UX and UI* and address recommendations to improve the usability of the applications. E2 suggested enriching the user interface with *visualizations and graphical widgets* that visualize the

content of the reference data in order to facilitate the discoverability phase. Thanks to these features, users can easily assess the data. In addition, he proposed to redesign the UI so that the application can *guide the users* step by step with a *bottom-up approach* in order to better understand what they are looking for and their needs. Furthermore, two experts recommend performing a usability test and avoid as much complexity as possible. One expert shared the Amazon example, where a minimal search interface without any complexity perceived from the users, provides a list of products based on the researched term. Thus, even if the user does not provide the right term of the object taxonomy, the search engine is able to provide a list of products of that category.

Enablers for the Identification phase
Users
Organizational structure and principles, policies, and frameworks • Supra-international data governance with clear role and responsibilities at different level *Processes* • New processes to improve automatic findability *Services, Infrastructure and Application* • Single central repository • API adoption • Better UI and UX *Information* • Multilingual documentation • Metadata enrichment • Automatic Notification to inform users about new published resources • Data storytelling to explain how to use the resources • Collect user feedbacks
Experts
Organizational structure and principles, policies, and frameworks • Supra-international data governance with clear role and responsibilities at different level • Network of institutions involved in reference government data definition *Processes* • Processes based on collaborative approaches and on successful use cases from the public sector

Services, Infrastructure and Application
- Better UI and UX
- Data visualizations and graphical widgets
- Bottom-up user experience

Information
- Multilingual metadata
- Metadata enrichment
- Semantic harmonisation and standardisation
- Semantic technologies and knowledge graph
- Persistent URIs
- Uniform licenses for dataset
- Avoiding jargon
- Communication and branding activities

People, Skills, and Competencies
- Data literacy and competencies

Table 7: Enablers for the Identification phase

4.2.4 Identification phase - Discussion

Findings on the identification phase of the OGD user's process on open reference data are consistent with previous results by Charalabidis et al. (2018), Crusoe et al. (2019), and Xiao et al. (2019).

The *identification* of open government reference data is performed in many ways by experts, and data analysts and findings clearly showed that there is not a single and trusted one-stop shop for reference data at the EU level. Findings curiously demonstrate that both groups prefer to rely on and use non-institutional search engines such as Google to discover resources quickly. Furthermore, resources such as data portals and websites from the EU institutions as well as other international resources seem to be used when users already know existing resources instead of accessing them to discover new controlled vocabularies. On the other hand, as of interesting findings, many participants mentioned using social interactions with colleagues and domain experts to identify and discover reference data. This action seems not to be mentioned by the previous research on OGD and suggests that the topic of open reference data is something specific and requires peculiar expertise and ability.

Regarding the *barriers* faced while trying to identify reference data relevant at the EU level, the participants outlined several challenges of different types to various extend. All the data analysts and experts reported barriers of different nature to find the data they are looking for, both in the exploration and assessment steps.

In terms of barriers in the *exploring phase*, experts and data analysts were aligned and outlined a fragmented information landscape in open government reference data produced by the organisation at the EU and worldwide level. These results confirm the contributions on the topic of OGD barriers and in particular the findings from Crusoe et al. (2019). Furthermore, the different activities performed by the users in this phase are similar to those identified from Crusoe et al (2019) related to OGD. Interestingly, findings from data analysts and experts show clearly the need to grasp information on the existence of controlled vocabulary using different approaches. There is no proper process for discovering open government reference data at the EU and international level, which can be considered an important impediment for future research.

Following the observations on the barriers in the *assessing phase*, users confirmed the main aspects mentioned by the group of experts. Data analysts can encounter many barriers related to different aspects such as the lack of metadata, the lack of proper documentation and multi-language translation and difficulties accessing the user interface and navigating the online resources.

It is interesting to note that findings clearly identify the barrier related to the versioning of reference data. Users mention the fact that they face many ambiguities related to versions of data and mapping over times. It is difficult to distinguish between the new and previous versions of code lists and how to map them for backwards compatibility.

The challenges users face in discovering and accessing the right version of a specific controlled vocabulary, could provide additional empirical inputs for further implementations to overcome barriers on OGD and open reference data.

Another remarkable finding is related to the barrier in assessing the source of truth for controlled vocabulary. Because of open data principles, it is common to find copies of the same resource in different online locations. This barrier seems to be relatively not covered by the previous research and literature and could represent an interesting aspect to investigate for future studies.

Looking at the enablers for better data discoverability, experts confirm the inputs from the data analysts that are consistent with previous findings from the European Commission (2019). The definition and implementation of horizontal data governance on reference data management between internal and external institutions at EU level seems to be an important enabler to overcome many barriers mentioned above.

4.2.5 Acquire phase - Results

Following the section related to the identification phase of the OGD user's process, the interview continued with the second phase in the scope of the research: the acquire phase. Based on the conceptual framework, the acquire phase is composed of two steps: access and deliver. The phase starts after finding the data and then wants to access and transfer it (Crusoe & Ahlin, 2019). In this part, the participants were asked open-ended questions as follows: “Are reference data such as controlled vocabularies easy to acquire?”, “Do you spend time and effort in accessing them?”, “How do you access them?” and “Which format do you usually expect and use?”.

Then, in the last section: “What type of barriers did you face while trying to acquire reference data such as controlled vocabularies, provided at EU level relevant for your tasks?” and “What do you think are key enablers in general to facilitate the acquisition of open reference data?

Acquiring open reference data

Data analysts

Regarding the ease of acquiring open reference data and the effort needed, the most remarkable result is that it seems extremely difficult to acquire *historical versions of open reference data*. This task has been reported by five out of nine participants and, many of them reported that often they need to look for previous versions of reference data and understand how to join different labels and codes. In particular, D2 mentioned that retrieving old reference data such as categorisation is extremely challenging, and he needs to spend time and effort in updating the code. On the other hand, two participants said they do not perceive difficulties acquiring reference data because they have direct access to internal repositories or BI tools where reference data are ready to be used.

Users access and deliver open reference data in two main ways: *automatically* or *manually*. Some users reported that they need to write code to *scrape* web pages because the reference data is not available in a convenient format. Another user said that he had to *copy and paste* specific code lists because he was not able to find downloadable resources. On the other side, two data analysts directly access reference data inside their BI tools. Just one user mentioned the use of *API* and to prepare the code necessary to query them.

In terms of the format expected and used, many users reported using *CSV* or *MS Excel files*, a minority *PDF* resources.

D1 mentioned that he uses *API* when available, and in general, he would have the opportunity to use them more extensively to retrieve data. Additionally, he is used to gathering data using SPARQL query accessing *SPARQL endpoint* via a web interface as well as D7. Finally, D7 mentioned that her favourite way to access and check controlled vocabularies is based on the Simple Knowledge Organisation System (*SKOS*). Using SKOS it is easy to understanding the structure and how controlled vocabularies are defined and linked but, she is aware that probably this is not the easiest way for no-expert users.

Experts

Many experts are aligned on the importance of promoting *linked data and semantic technologies* to share and access reference data. Some experts underlined the importance of exposing *API* endpoints to deliver open reference and documenting the APIs with *OpenAPI specification* to describe how to access them easily. Furthermore, one expert mentioned that the number of reference data providers offering API endpoints is still limited.

Another expert suggested promoting the *JSON* format instead of XML, and he said that the *SKOS* format that usually is offered by some European providers is not easy to access for non-technical users.

Barriers in the Access activity

Data analysts

The main barriers reported from the data analysts in the access phase are related to different aspects: they reported impediments because of *lack of documentation, lack of metadata, lack of machine-readable format* and difficulty in interacting with *legacy systems* or *specific technologies.*

Regarding the lack of documentation, users expressed this concern specifically for reference data that are available through SPARQL endpoint and based on linked data format. According to the interviewees, these *technologies are advanced but not easy to be used* by non-expert users, poor documentation can impact significantly on the accessing phase. On the other hand, poor *metadata* can represent another impediment, one data analyst specifically mentioned the case of the metadata on the *license*: if the license is not expressed or not adequate, users can not access and use the resources. Finally, other two users reported challenges in preparing the environment to access reference data stored in *legacy systems* used internally in their units and DGs.

Experts

Experts reported several barriers related to the access phase. First, one expert lamented the lack of technical expertise from providers' side working at the development of solutions. In his opinion, better UI and easier ways to display, navigate and then download data should be implemented. At the same time, many of them reported that the technical documentation available on the website of the main EU reference data provider is not accurate, clear and does not support the users in how to access data and re-use it.

Other experts mentioned a lack of metadata adoption, specifically related to the licence, information that needs to be clearly mentioned in order to resume the data. Then, related to interoperability, experts mentioned barriers at the technical and semantic level. At semantic level, one expert reported that it happened many times to find resources related to similar concepts but labelled with different names. On the other hand, he mentioned that there is some confusion in defining concepts, and sometimes concepts related to different categories are defined in the same way. To address these barriers he suggested fostering the adoption of ontologies and marked the concepts used.

Finally, some experts reported technical barriers in accessing reference datasets. One of them, reported that based on his experience, the decision to publish reference data resources using SKOS can be a barrier for many not expert users, who do not know how to easily navigate and validate the content of the reference data as a controlled vocabulary.

Barriers in the Deliver activity

Data analysts

The barriers mentioned by the users are *technical barriers* and they can fit in three main categories: *no machine-readable format, restrictions* in automatizing the delivering process and *broken links* because resources change locations.

Regarding technical interoperability, two data scientists reported that no machine-readable format such as code lists stored in PDF or just display in text in the webpages represents an impediment in retrieving reference data at eu level. In this scenario, they need to engage with colleagues and sometimes ask for support in order to write or modify code that can be used to gather the data. Similarly, users need to use ad-hoc procedures to import and use reference data that are not available in a machine-readable format, with consequently risks on the quality of service delivered.

Another technical barrier reported by one of the participants is represented by technical restrictions in downloading reference data such as the Comtrade dataset from the UN where there are *limited API calls* for guest users. Finally, two data scientists reported the

effort in updating code because resources did not use a persistent identifier and happened to have *broken links.*

Experts

Experts reported some barriers in the delivery activity. One of them argued that for many providers the only way to use data is to *download bulk resources* in a file or set of files instead of using API. Additionally, many experts mentioned that *persistent identifiers* are rarely used and these two elements together are a big obstacle in building automated and repeatable processes to acquire the resources.

Barriers in the Acquire phase	
Barriers in the *access* activity	
Users	Experts
• Lack of machine readable format • Lack of documentation • Legacy systems • Lack of documentation • Difficulties in querying semantic web and technologies	• Lack of technical expertise • Lack of comprehensible technical documentation • Lack of metadata • Technical interoperability • Semantic interoperability
Barriers in the *deliver* activity	
Users	Experts
• No machine readable format (es .PDF) • Broken links • Limited API calls	• Bulk downloads • No persistent URIs

Table 8: Barriers in the Acquire phase

Enablers for the Acquire phase

Data analysts

Data analysts have been asked to suggest and propose key enablers to overcome challenges they face in acquiring open reference data. Based on the COBIT framework, all the suggestions collected from the users have been categorized under different enablers. Users recommended many improvements under *services, infrastructure and application*. Many of them proposed improving the adoption of *metadata* and assessing them through *quality check mechanisms*. Two data analysts requested to introduce *backward compatibility mechanisms* to overcome the impediments related to retrieving historical reference data. In terms of *technical interoperability,* many users mentioned the

importance of investing in general improvements such as maximizing the providers using *API* and removing no machine-readable resources such as PDF.

Regarding the *information* enabler, four users suggested improving the *documentation* to better support the delivery action, specifically in case the data provider exposes an API or SPARQL endpoint. The documentation should include concrete and real *examples* and how-To to retrieve the data and support the users in accessing the data. In addition, a participant requested to implement communication channels to inform users when systems are under *maintenance*.

Experts

Experts have been asked to provide enablers to improve the acquire phase, based on their expertise and the barriers they mentioned. The recommendations have been categorized in the different enablers identified by the COBIT framework.

Many suggestions belong to the *information* enabler. First, beautiful and interactive *documentation* needs to be developed according to many experts. Reference data providers should illustrate clear examples, using real cases and explain them in natural language avoiding jargons. In addition, when APIs are available, experts mention documenting them using *OpenAPI specification.*

In regards to the *Services, Infrastructure, and Applications* enabler, experts provided several recommendations.

First, many of them suggest promoting data access through *API*. Second, two experts mentioned the importance of using *persistent URIs* to provide more reliability to the users. Then, one expert recommended data providers to offer data in *different formats* including csv-file format that is the most common and easy to acquire for non-expert users. He said that it is important in investing in linked data technologies but to not forget that many users do not have the necessary skills to query a SPARQL endpoint.

Experts recommend investing in the *People, Skills, and Competencies* enablers and introducing training about implementing and using new technologies such as API and Linked data for public servants .

Finally, related to the one *Organisational structure* enabler, experts suggested fostering *synergies* between the different EU institutions in reusing common platforms, technology and tools in order to maximise the investment and at the same time the adoption of common approaches.

Enablers for the Acquire phase
Users
Services, Infrastructure and Application • Metadata adoption • Quality check mechanism • Backward compatibility mechanism • API adoption *Information* • Better documentation with concrete examples and how-to • Metadata enrichment • Communication channels for informing about system maintenance
Experts
Services, Infrastructure and Application • API adoption • Persistent URIs • Variety of format, from CSV to linked data to meet different level of expertise *Information* • Better technical documentation with clear and useful examples • OpenAPI specification *People, Skills, and Competencies* • Data literacy and training about linked data and how to use API *Organisational structure* • Foster reuse of common platforms and tools between the institutions

Table 9: Enablers for the Acquire phase

4.2.6 Acquire phase - Discussion

The results and information collected from the data analysts and the experts on the acquire phase seem to be more limited compared to the one on the identification phase. Findings on the acquire phase of the OGD user's process on open reference data confirm the previous empirical findings by Crusoe et al. (2019), Xiao et al. (2019) and Attard et al. (2014). The main challenges and obstacles in the *acquire* phase of open reference data are mainly related to technical aspects. Experts confirm the main impediments expressed by the users, such as barriers at semantic and technical level. The importance of clear technical documentation, adoption of metadata and availability of machine-readable

format are important factors to facilitate the acquisition process. These findings provide additional support and are consistent with previous research from Crusoe et al. (2019) on OGD user's process. However, the current findings on the acquire phase show that data analysts do not always recognize distinctly the different two steps identified by the framework from Crusoe et and Ahlin (2019) that clearly separates the access and delivery phase. The distinction is more clear and evident for the groups of experts.

As of interesting findings, the need of acquiring different historical versions of controlled vocabulary constitutes a peculiar insight. The lack of managing different versions of controlled vocabularies and providing clear information on the resource version could be identified as a new type of impediment that seems not to be covered specifically by earlier findings (Crusoe et al., 2019).

Furthermore, findings from experts and data analysts confirmed the need to foster the adoption of API and publishing controlled vocabularies using OpenAPI. This result is aligned with the Directive (EU) 2019/1024 on open data that promotes the uptake of API for publishing OGD.

Looking at the enablers, the findings from the experts confirmed the ones from the data analysts. Both are firmly convinced that information enablers are necessary. In addition, proper technical documentation, usage of metadata, adoption of persistent URIs, and machine-readable formats, especially based on API, are key enablers to overcome the current impediments.

5 Conclusion

This exploratory research has explored and investigated user barriers and enablers in discovering and accessing open government reference data in the form of controlled vocabularies relevant at the EU level in a DDPS context.

The author used different dimensions to frame the context and address the research question. The research started investigating the DDPS context and areas of activities where data analysts and experts contribute, then the relevance of open government reference data and the impediments they face in the identification and acquisition phases. Finally, the research has identified a set of enablers useful to derive findings to improve the identification and acquisition of open government reference data.

For this purpose, empirical data describing the user's and expert's perspective have been gathered through semi-structured interviews.Nine data analysts working in different Directorates Generals and department departments in the EC, focusing on "Policy making and implementation" and five domain experts working on OGD, semantic data and reference data management, provided more understanding of impediments and enablers on re-using open government reference data in a DDPS context.

The study showed that data analysts working in the organisation benefit and combine information from various sources and different domains. Data is the main ingredient of a DDPS and OGD plays an important role in defining and creating data-driven products to enable and address policy evaluations and analysis.

The research has highlighted the relevance of the subject and the importance of the OGD research topic on controlled vocabularies and reference data. They play an important and concrete role in boosting interoperability and facing the societal challenge on language divide and lack of cross-communities communication. The study confirms how the multilevel governance within the EU and its multilingualism make adopting controlled vocabularies an essential precondition to foster interoperability and enable interlinking of information from different level, sources and sectors.

Furthermore, the research suggests that open government reference data represents an important enabling factor to leverage public sector intelligence and streamline decision making processes, specifically in the area of anticipatory governance and performance management.

Open government reference data and controlled vocabularies produced from the EU institutions and other international organisations are used by the data analysts in their activity. This is interesting because it confirms that the public sector itself is, as expected,

is one of the direct beneficiaries of the open data policies that foster the re-use of public sector information. The analysis has indicated that users are not aware of the peculiarities of the different types of controlled vocabularies, but they use many vocabularies in their tasks, mostly in the form of code lists.

In the research, the author has investigated the main challenges users experience in the identification and acquisition phase of controlled vocabularies relevant at EU level in the DDPS context.

The first main element of evidence from the study suggests that the main common impediments related to discoverability and accessibility of open government data are similar to the ones recognised to affect the more generic area of OGD.

Looking at the identification phase, evidence point towards the high fragmentation of resources, lack of horizontal governance between institutions, proper documentation and technical impediments being the key sources of friction. The absence of a single, complete and trusted one-stop shop for reference data at the EU level had been reported in the study. Different EU online resources are accessed to explore open government reference data, but the use of public search engines and social interactions with colleagues or domain experts seem to be the favourite way to discover resources. This element suggests that discovering controlled vocabularies is a time-consuming task for users. There is no defined streamlined process to easily discover open government data relevant at the EU level. Data analysts pointed out that many ambiguities and difficulties arise as well when exploring controlled vocabularies relevant at EU level. Documentation may be not clear or complete, metadata is not always accessible in different languages and properly used, and a low usability in the UI of web portals with a consequent bad UX experience affects the users. An interesting impediment related to the specific domain of controlled vocabularies is the lack of version management, that is crucial in linking historical datasets valid within a specific time frame.

Regarding the acquisition phase, findings are more limited and less specific compared to the identification phase. The study suggests that impediments are mainly limited to technical and interoperability factors and not specific to the peculiarities of controlled vocabularies, but more aligned with the common barriers of OGD. In more detail, impediments faced in this phase concern the lack of data in machine readable formats, interoperability issues with legacy systems, broken links and lack of technical documentation.

The research suggests that the discoverability phase is the most challenging and the findings provide valuable practical inputs specific to the topic of discovering and acquiring open government reference data relevant at the EU level.

As mentioned before, the research has aimed to propose actions to overcome impediments based on the recommendations given by the data analysts and the group of experts. According to the COBIT categories, the main enablers are related to information and governance, including the definition of principles, policies, and frameworks. The need to implement a coherent and horizontal data governance on reference data management internally and between the EU institutions and international organisations, with clear roles and responsibilities and processes, has been suggested by experts and users.

On the information side, the main recommendations are related to improving technical aspects for data findability such as metadata, semantic harmonisation and persistent URIs. Quickly discovering the different versions of resources, and identifying the official owner of the controlled vocabulary, are two important aspects to consider and improve to support the users. Another important emerging enabler seems to be the adoption of APIs and OpenAPI specifications to facilitate and speed up the discoverability and accessibility of controlled vocabularies and integration with data analysis tools.

Furthermore, the research seems to underline the importance of investing in effective documentation to make the resources findable, making the users' needs and experience central.

At the same time, investing in skills and competencies, both on the users' and data providers' sides, seems to be an important enabler to facilitate the re-use of open government data and boost the contribution of open government reference data in the DDPS context.

Limitations

The exploratory study was subject to several limitations. The research has been conducted using a qualitative approach and not based on quantitative data. Therefore, its findings may be affected by bias and subject to interpretation (Stebbins, 2001). Then, even if they provide useful empirical data to understand better barriers and enablers in the user process of discovering and accessing controlled vocabularies, they are not statistically relevant and useful to measure the impact on the OGD user process. Assessing the perceived difficulties was not a goal of this research but further research could use quantitative insights on these aspects as Crusoe et al. conducted in their analysis (Crusoe et al., 2019)

to gain a better understanding on the severity of the impediments. The same could apply to the perceived usefulness of the enablers.

Furthermore, there are other limitations related to the sample's representativeness. The number and type of participants and their selection process may have limited the ability to understand the diversity of barriers related to open government reference data usage. A first limitation was the sample size, represented by nine users and five experts. Second, the specific sample consisted of a group of data analysts working in the European Commission and experts with a high level of digital literacy. Thus, the sample is not representative for the full population of OGD and open reference data users. Alternative studies involving users with different levels of digital literacy and from different sectors such as companies and academia could be useful.

Thirdly, alternative research could be useful to assess barriers and enablers at different levels of governance and DDPS context. Using samples with users working in the MS and at regional levels may contribute additional findings.

A fourth limitation is that data analysts involved in the research seemed to focus more on their respective financial and economic areas of interest. Consequently, additional research with users working on different perspectives, themes and context could bring additional value on assessing barriers and enablers for specific categories of controlled vocabularies.

Future research and recommendations

Despite these limitations, this study has offered new empirical data and practical guidelines into a relevant topic under the area of OGD that seems to have been the subject of very limited research.

This paper has offered valuable insights for future research that might be beneficial to increase and improve the re-use of controlled vocabularies. It has contributed to the research related to challenges public sector organisations face when implementing a DDPS. The exploratory research has focused on the OGD user process and has contributed to the specific area of controlled vocabularies and code lists preservation where limited literature is available. The results have provides contributions to the increasing literature on impediments of OGD initiatives in a DDPS context at EU level, specifically from the user perspective. Specifically, the research has contributed with empirical findings from an international perspective such as the European Commission's and has offered practical contribution to the ongoing discussion on reference data management and its role in the definition of an effective multi-level data governance.

Findings on barriers and enablers based on practical recommendations can encourage institutions to adopt concrete measures to further improve discoverability and acquisition of open government reference data, so that they become more reusable and adoption and value creation increases.

References

Backx, M. (2003). *Gebouwgegevens redden levens [Building information saves lives]*. MSc thesis TU Delft, in cooperation with DataLand.

Bralić, A. (2017). *Social Network Analysis of Country Participation in Horizon 2020 Programme*. 7.

Bureau van Dijk. (n.d.). *Orbis*. Bvd. Retrieved 3 July 2021, from https://www.bvdinfo.com/en-gb/our-products/data/international/orbis

Cavanillas, J. M., Curry, E., & Wahlster, W. (Eds.). (2016). *New Horizons for a Data-Driven Economy*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3

Charalabidis, Y., Alexopoulos, C., & Loukis, E. (2016). A taxonomy of open government data research areas and topics. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 41–63. https://doi.org/10.1080/10919392.2015.1124720

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). *The World of Open Data: Concepts, Methods, Tools and Experiences* (Vol. 28). Springer International Publishing. https://doi.org/10.1007/978-3-319-90850-2

Chatfield, A. T., & Reddick, C. G. (2018). Customer agility and responsiveness through big data analytics for public value creation: A case study of Houston 311 on-demand services. *Government Information Quarterly*, *35*(2), 336–347. https://doi.org/10.1016/j.giq.2017.11.002

Chief data officer. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Chief_data_officer&oldid=1020159455

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Crompvoets, J., Laenen, L., & Fobé, E. (2019, October). *Integrated Research Seminar: Part I Research Design & Data Collection*.

Crusoe, J., & Ahlin, K. (2019). Users' activities for using open government data – a process framework. *Transforming Government: People, Process and Policy*, *13*(3/4), 213–236. https://doi.org/10.1108/TG-04-2019-0028

Crusoe, J., & Melin, U. (2018). Investigating Open Government Data Barriers: A Literature Review and Conceptualization. In P. Parycek, O. Glassey, M. Janssen, H. J. Scholl, E. Tambouris, E. Kalampokis, & S. Virkar (Eds.), *Electronic Government* (Vol. 11020, pp. 169–183). Springer International Publishing. https://doi.org/10.1007/978-3-319-98690-6_15

Crusoe, J., Simonofski, A., Clarinval, A., & Gebka, E. (2019). The Impact of Impediments on Open Government Data Use: Insights from Users. *2019 13th International Conference on Research Challenges in Information Science (RCIS)*, 1–12. https://doi.org/10.1109/RCIS.2019.8877055

De Haes, S., & Van Grembergen, W. (2015). *Enterprise Governance of Information Technology: Achieving Alignment and Value, Featuring COBIT 5*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14547-1

Donker, F. W., & van Loenen, B. (2017). How to assess the success of the open data ecosystem? *International Journal of Digital Earth*, *10*(3), 284–306. https://doi.org/10.1080/17538947.2016.1224938

European Commission. (n.d.-a). *Departments and Executive agencies*. https://ec.europa.eu/info/departments_en?field_core_topics_target_id_entityreference_filter=All&field_core_ecorganisation_value_i18n=Directorate-General&field_department_tasks_tid_entityreference_filter=178

European Commission. (n.d.-b). *Once Only Principle*. CEF Digital. Retrieved 27 June 2021, from https://ec.europa.eu/cefdigital/wiki/cefdigital/wiki/display/CEFDIGITAL/Once+Only+Principle

European Commission. (2014, June). *Semantic Interoperability Courses, Reference Data Management*.

European Commission. (2017, February 16). *The New European Interoperability Framework* [Text]. ISA^2 - European Commission. https://ec.europa.eu/isa2/eif_en

European Commission. (2020a). *Corporate Reference Data Management in the European Commission—Draft policy*.

European Commission. (2020b). *Strategy for Data, Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/strategy-data

European Commission. (2020c). *Data governance and data policies*. https://ec.europa.eu/info/sites/default/files/summary-data-governance-data-policies_en.pdf

European Parliament and Council. (2019). *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information* (Vol. 172). http://data.europa.eu/eli/dir/2019/1024/oj/eng

Eurostat. (n.d.). *RAMON, Eurostat's Metadata Server*. RAMON, Eurostat's Metadata Server. Retrieved 25 April 2021, from https://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC

Eurostat. (2020). *Strategic Plan 2020-2024*. https://ec.europa.eu/info/system/files/estat_sp_2020_2024_en.pdf

FAO. (n.d.). *AGROVOC*. Retrieved 24 May 2021, from http://www.fao.org/agrovoc/about

GDS. (2017, July 20). *Building capability and community through the Government Data Science Partnership*. https://gds.blog.gov.uk/2017/07/20/building-capability-and-community-through-the-government-data-science-partnership/

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Methods of data collection in qualitative research: Interviews and focus groups. *British Dental Journal*, *204*(6), 291–295. https://doi.org/10.1038/bdj.2008.192

Hedden, H. (2010). Taxonomies and controlled vocabularies best practices for metadata. *Journal of Digital Asset Management*, *6*(5), 279–284. https://doi.org/10.1057/dam.2010.29

ISACA. (2012). *COBIT 5: A Business Framework for the Governance and Management of Enterprise IT*. ISACA.

ISO. (n.d.-a). *ISO 639, Language codes*. ISO. Retrieved 30 July 2021, from https://www.iso.org/iso-639-language-codes.html

ISO. (n.d.-b). *ISO 4217, Currency codes*. ISO. Retrieved 30 July 2021, from https://www.iso.org/iso-4217-currency-codes.html

Jacquin, S., Pawlewitz, J., & Doyle, A. (2020). Democratizing Data with Self-Service Analytics. *Day 4 Thu, May 07, 2020*, D041S055R007. https://doi.org/10.4043/30685-MS

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, *28*(4), 446–456. https://doi.org/10.1016/j.giq.2011.01.004

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Janssen, M., Konopnicki, D., Snowdon, J. L., & Ojo, A. (2017). Driving public sector innovation using big and open linked data (BOLD). *Information Systems Frontiers*, *19*(2), 189–195. https://doi.org/10.1007/s10796-017-9746-2

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338–345. https://doi.org/10.1016/j.jbusres.2016.08.007

Jochen, D. W. (2019, November). *Information systems, strategy and governance, Lecture 3: IS governance*. Business Information Systems, Leuven.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Linked Open Government Data Analytics. In M. A. Wimmer, M. Janssen, & H. J. Scholl (Eds.), *Electronic Government* (Vol. 8074, pp. 99–110). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40358-3_9

Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open Government Data Catalogs: Current Approaches and Quality Perspective. In A. Kő, C. Leitner, H. Leitold, & A. Prosser (Eds.), *Technology-Enabled Innovation for Democracy, Government and Governance* (pp. 152–166). Springer. https://doi.org/10.1007/978-3-642-40160-2_13

Lněnička, M., Machova, R., Volejníková, J., Linhartová, V., Knezackova, R., & Hub, M. (2021). Enhancing transparency through open government data: The case of data portals and their features and capabilities. *Online Information Review*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/OIR-05-2020-0204

Lnenicka, M., & Nikiforova, A. (2021). Transparency-by-design: What is the role of open data portals? *Telematics and Informatics*, *61*, 101605. https://doi.org/10.1016/j.tele.2021.101605

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, *32*(3), 323–332. https://doi.org/10.1016/j.giq.2015.05.006

Máchová, R., Hub, M., & Lnenicka, M. (2018). Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective. *Aslib Journal of Information Management*, *70*(3), 252–268. https://doi.org/10.1108/AJIM-02-2018-0026

Malcolm Chisholm. (n.d.). *Top Quadrant Reference Data Management Whitepaper*. Retrieved 1 April 2021, from https://www.topquadrant.com/docs/whitepapers/TopBraid_ReferenceDataManagement Whitepaper-3-18-15.pdf

McBride, K., Aavik, G., Toots, M., Kalvet, T., & Krimmer, R. (2019). How does open government data driven co-creation occur? Six factors and a 'perfect storm'; insights from Chicago's food inspection forecasting model. *Government Information Quarterly*, *36*(1), 88–97. https://doi.org/10.1016/j.giq.2018.11.006

Nolan, L. (2021, April 8). *500 not out – building data science capacity for the future*. https://datasciencecampus.ons.gov.uk/500-not-out-building-data-science-capacity-for-the-future/

OECD. (2020). *Building digital workforce capacity and skills for data-intensive science* (OECD Science, Technology and Industry Policy Papers No. 90; OECD Science, Technology and Industry Policy Papers, Vol. 90). https://doi.org/10.1787/e08aa3bb-en

Ooijen, C. van, Ubaldi, B., & Welby, B. (2019). A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance. *OECD*. https://doi.org/10.1787/09ab162c-en

Open Knowledge Foundation. (2019). *Missed opportunities in the EU's revised open data and re-use of public sector information directive*. https://blog.okfn.org/2019/07/09/missed-opportunities-in-the-eus-revised-open-data-and-re-use-of-public-sector-information-directive/

Publications Office of the European Union. (n.d.-a). *European Science Vocabulary (EuroSciVoc)* -. Retrieved 6 June 2021, from https://op.europa.eu/it/web/eu-vocabularies/euroscivoc

Publications Office of the European Union. (n.d.-b). *EuroVoc*. Retrieved 30 July 2021, from https://eur-lex.europa.eu/browse/eurovoc.html?locale=en

Publications Office of the European Union. (n.d.-c). *Procurement procedure type*. Retrieved 30 July 2021, from https://op.europa.eu/en/web/eu-vocabularies/dataset/-

/resource?uri=http://publications.europa.eu/resource/dataset/procurement-procedure-type

Publications Office of the European Union. (2020a). *Controlled vocabularies*. https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies

Publications Office of the European Union. (2020b). *CORDIS - EU research projects under Horizon 2020 (2014-2020)*. https://data.europa.eu/data/datasets/cordish2020projects?locale=en

Publications Office of the European Union. (2020c). *Strategic Plan 2020-2024*. Publication Office of the European Union. https://op.europa.eu/documents/10530/8415767/Strategic+Plan+2020-2024.pdf/

Publications Office of the European Union. (2021a). *Thesauri*. https://op.europa.eu/en/web/eu-vocabularies/thesauri

Publications Office of the European Union. (2021b, January 27). *Administrative territorial unit*. https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/atu

Publications Office of the European Union. (2021c, September 4). *Digital Competence Framework*. https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/digital-competence-framework

Research Data Alliance FAIR Data Maturity Model Working Group. (2020). *FAIR Data Maturity Model: Specification and guidelines*. https://doi.org/10.15497/RDA00050

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Pearson education.

Stebbins, R. A. (2001). *Exploratory research in the social sciences* (Vol. 48). Sage.

Susha, I., Janssen, M., & Verhulst, S. (2017). Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. *Hawaii International Conference on System Sciences 2017 (HICSS-50)*. https://aisel.aisnet.org/hicss-50/eg/open_data_in_government/4

Toots, M., McBride, K., Kalvet, T., & Krimmer, R. (2017). Open Data as Enabler of Public Service Co-creation: Exploring the Drivers and Barriers. *2017 Conference for E-*

Democracy and Open Government (CeDEM), 102–112. https://doi.org/10.1109/CeDEM.2017.12

Use Cases for Wikibase. (n.d.). Learning Wikibase. Retrieved 11 July 2021, from http://learningwikibase.com/usecases/

van Loenen, B. (2018). Towards a User-Oriented Open Data Strategy. In B. van Loenen, G. Vancauwenberghe, & J. Crompvoets (Eds.), *Open Data Exposed* (Vol. 30, pp. 33–53). T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-261-3_3

van Loenen, B., Crompvoets, J., & Poplin, A. (n.d.). *Assessing geoportals from a user perspective*. 9.

Voultsiadou, E., Gerovasileiou, V., Vandepitte, L., Ganias, K., & Arvanitidis, C. (2017). Aristotle's scientific contributions to the classification, nomenclature and distribution of marine organisms. *Mediterranean Marine Science*, 468. https://doi.org/10.12681/mms.13874

Walliman, N. (2017). *Research Methods: The Basics: 2nd edition*. Routledge. https://doi.org/10.4324/9781315529011

Wang, H.-J., & Lo, J. (2016). Adoption of open government data among government agencies. *Government Information Quarterly*, *33*(1), 80–88. https://doi.org/10.1016/j.giq.2015.11.004

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Xiao, F., He, D., Chi, Y., Jeng, W., & Tomer, C. (2019). Challenges and Supports for Accessing Open Government Datasets: Data Guide for Better Open Data Access and Uses. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 313–317. https://doi.org/10.1145/3295750.3298958

Appendix

Interview Discussion Guide

"*What are the challenges that data analysts and experts face discovering and accessing open government reference data (focusing on controlled vocabularies) at EU level in the context of a data-driven public sector? What are the key enablers to enhance them?"*

1.General data

1. Date
2. Name
3. Can you tell me about your background? (Studies, work experience)
4. What are the main tasks of this position?
5. How long have you been working here?

2. Data driven public sector

1. How do you use data and analytics in your activities?
2. What type of information and datasets do you use?

3. Open Reference Data

1. Are open reference data produced at EU level and in the form of controlled vocabularies (taxonomies, code lists, authority tables and thesauri) relevant in the development of your analysis? If yes, why?
2. What type of reference data do you use? How do you use them? For which purpose?

3.1 Identification phase

1. How do you identify reference data such as controlled vocabularies?
2. What type of barriers did you face while trying to identify reference data, such as controlled vocabularies, provided at EU level relevant for your tasks?
3. What do you think are key enablers in general to facilitate the identification of open reference data?

3.2 Acquisition phase

1. Are reference data such as controlled vocabularies easy to acquire?
2. How do you access them?
3. Do you spend time and effort in accessing them?
4. Which format do you usually expect and use?

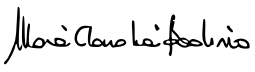
5. What type of barriers did you face while trying to acquire reference data such as controlled vocabularies, provided at EU level relevant for your tasks?
6. What do you think are key enablers in general to facilitate the acquisition of open reference data?

Declaration of Authorship

I hereby declare that, to the best of my knowledge and belief, this Master Thesis titled "Challenges and enablers in using Open Government Reference Data in a data-driven public sector" is my own work. I confirm that each significant contribution to and quotation in this thesis that originates from the work or works of others is indicated by proper use of citation and references.

Cuneo, 08 August 2021

Maria Claudia Bodino



Consent Form

for the use of plagiarism detection software to check my thesis

Name: Bodino
Given Name: Maria Claudia
Student number: 0781974
Course of Study: Public Sector Innovation and eGovernance
Address: via Monsignor Peano 24. Cuneo, Italy
Title of the thesis: Challenges and enablers in using Open Government Reference Data in a data-driven public sector
What is plagiarism? Plagiarism is defined as submitting someone else's work or ideas as your own without a complete indication of the source. It is hereby irrelevant whether the work of others is copied word by word without acknowledgment of the source, text structures (e.g. line of argumentation or outline) are borrowed or texts are translated from a foreign language.

Use of plagiarism detection software. The examination office uses plagiarism software to check each submitted bachelor and master thesis for plagiarism. For that purpose the thesis is electronically forwarded to a software service provider where the software checks for potential matches between the submitted work and work from other sources. For future comparisons with other theses, your thesis will be permanently stored in a database. Only the School of Business and Economics of the University of Münster is allowed to access your stored thesis. The student agrees that his or her thesis may be stored and reproduced only for the purpose of plagiarism assessment. The first examiner of the thesis will be advised on the outcome of the plagiarism assessment.

Sanctions. Each case of plagiarism constitutes an attempt to deceive in terms of the examination regulations and will lead to the thesis being graded as "failed". This will be communicated to the examination office where your case will be documented. In the event of a serious case of deception the examinee can be generally excluded from any further examination. This can lead to the exmatriculation of the student. Even after completion of the examination procedure and graduation from university, plagiarism can result in a withdrawal of the awarded academic degree.

I confirm that I have read and understood the information in this document. I agree to the outlined procedure for plagiarism assessment and potential sanctioning.

Cuneo, 6/08/2021

Maria Claudia Bodino