D05.02 Big Data Interoperability Analysis
### Document Metadata

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>2018-05-31</td>
</tr>
<tr>
<td>Status</td>
<td>Accepted</td>
</tr>
<tr>
<td>Version</td>
<td>1.00</td>
</tr>
<tr>
<td>Authors</td>
<td>Jens Scheerlinck – PwC EU Services</td>
</tr>
<tr>
<td></td>
<td>Frederik Van Eeghem – PwC EU Services</td>
</tr>
<tr>
<td></td>
<td>Nikolaos Loutas – PwC EU Services</td>
</tr>
<tr>
<td>Reviewed by</td>
<td>Daniel Brulé – PwC EU Services</td>
</tr>
<tr>
<td></td>
<td>Makx Dekkers – AMI Consult</td>
</tr>
<tr>
<td></td>
<td>Susanne Wigard – European Commission</td>
</tr>
<tr>
<td></td>
<td>Fidel Santiago – European Commission</td>
</tr>
<tr>
<td>Approved by</td>
<td>Susanne Wigard – European Commission</td>
</tr>
</tbody>
</table>

**This study was prepared for the ISA² Programme by:**

*PwC EU Services*

**Disclaimer:**

The views expressed in this report are purely those of the authors and may not, in any circumstances, be interpreted as stating an official position of the European Commission.

The European Commission does not guarantee the accuracy of the information included in this study, nor does it accept any responsibility for any use thereof.

Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Commission.

All care has been taken by the author to ensure that s/he has obtained, where necessary, permission to use any parts of manuscripts including illustrations, maps, and graphs, on which intellectual property rights already exist from the titular holder(s) of such rights or from her/his or their legal representative.
Table of Contents

1. INTRODUCTION 1
   1.1. OBJECTIVES OF THIS REPORT ................................................................. 1
   1.2. CONTEXT OF THIS REPORT .............................................................................. 1
       1.2.1. What is big data?....................................................................................... 2
       1.2.2. What is data analytics?.............................................................................. 2
       1.2.3. What is interoperability?.......................................................................... 2
   1.3. STRUCTURE OF THE REPORT ......................................................................... 3

2. COMMON INTEROPERABILITY CHALLENGES WITH BIG DATA 4
   2.1. COMMON DATA INTEROPERABILITY ISSUES .................................................... 4
   2.2. RUNNING EXAMPLE ....................................................................................... 7

3. GOOD PRACTICES FOR ACHIEVING INTEROPERABILITY IN A BIG DATA ENVIRONMENT 8
   3.1. LEGAL INTEROPERABILITY ............................................................................ 8
       3.1.1. Legislation affecting data and interoperability ........................................... 8
       3.1.2. Considerations when processing personal data .......................................... 9
   3.2. ORGANISATIONAL INTEROPERABILITY .......................................................... 10
       3.2.1. Breaking down data silos within your organisation .................................... 10
       3.2.2. How to deal with various data licences ..................................................... 13
       3.2.3. Aligning data requirements with data usage ............................................. 17
   3.3. SEMANTIC INTEROPERABILITY ..................................................................... 18
       3.3.1. Closing the semantic gap between datasets ................................................ 19
       3.3.2. (Semi-)automated vocabulary creation ...................................................... 26
       3.3.3. Optimising performance and interoperability by choosing the right data serialisation format .................................................................................................................. 28
   3.4. TECHNICAL INTEROPERABILITY ................................................................... 29
       3.4.1. Choosing between data integration patterns .............................................. 29

4. CONCLUSION 34

ANNEX I. APPROACH AND METHODOLOGY 35

ANNEX II. GLOSSARY 36

ANNEX III. BIBLIOGRAPHY 38

Table of Figures

Figure 1: Architecture of the fictional case study .......................................................... 7
Figure 2: Information governance knowledge areas according to DAMA DMBOK .................. 12
Figure 3: An example vocabulary for user feedback on the Better Regulation Portal ............... 22
Figure 4: Case study architecture diagram with data serialisation ..................................... 29
Figure 5: Sliding (top) and hopping (bottom) windowing types .......................................... 30
Figure 6: Case study architecture diagram including a virtualisation layer .......................... 32
1. **INTRODUCTION**

This report was written as part of Action 2016.07, SEMIC (Semantic Interoperability Community), of the ISA² Programme and is one of the deliverables for task 5 - Capacity Building on Information Governance – for specific contract 508.

1.1. **Objectives of this report**

The purpose of this report is to analyse interoperability challenges in a big data environment, focusing on the requirements for data analytics across different publishers and across domains.

To this end, we first collect and analyse challenges based on available literature. Subsequently we examine existing standards and specifications used for data and metadata and provide guidelines on how these can be applied to deal with interoperability challenges in a big data context. This study looks at the interoperability challenges which may prevent organisations from adopting data-driven decision-making. It aims to assess interoperability challenges from a business perspective, so that officials in the EU institutions and Member States’ administrations can increase their understanding of the challenges to be tackled when integrating data for analytical purposes and the role of both technical and data standards to improve interoperability.

1.2. **Context of this report**

The European Commission, as part of its Digital Single Market strategy, has identified the need to make sense of ‘big data’ as a key driver leading to innovations in technology, development of new tools and new skills [1]. As big data presents opportunities to boost growth and jobs in Europe, as well as improve the quality of life for Europeans [2], big data is an area of interest for public administrations at every level of government.

Public administrations can benefit from big data technologies and analytical techniques to: make public services offered to citizens and business more efficient, improve policy making through data-driven decisions and prevent fraud [3] [4]. For example, in 2017 the Dutch Ministry of Internal Affairs finalised a proof of concept to demonstrate how big data technologies applied to the Dutch legal corpus can be used to provide citizens and legal experts with more relevant search results in the legal domain [5]. More recently, in January 2018, Romania’s Ministry of Communications and Information Society launched two “big data” projects to improve tax collection and prevent fraud [4].

In the last decade, the terms big data, analytics and interoperability have been defined and interpreted in different ways. As a guidepost for readers and to set the scene for our analysis, we provide consensus-based definitions for these terms in sections 1.2.1 through 1.2.3.

---

2 The ISA² Programme of the EU: [http://ec.europa.eu/isa/index_en.htm](http://ec.europa.eu/isa/index_en.htm)
1.2.1. What is big data?

Since its first use in the 1990s [6], the term “big data” has been associated mostly with datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable timeframe [7]. In a 2016 article of the journal Global Knowledge, Memory and Communication, researchers put forward a consensual definition of big data based on its essential features, stating “big data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value” [8]. While this definition evolves around the original “three Vs” of big data and the creation of value, practitioners often refer to additional Vs such as veracity, variability and visualisation [9], ranging to as much as ten [10].

In this report, we will focus on the dimensions variety and veracity to analyse the interoperability related challenges of big data. **Variety** is defined as dealing with unstructured, semi-structured and structured data from different sources. **Veracity** directly refers to the accuracy of data, which can be diminished by data inconsistence and data quality problems. Both variety and veracity are aspects of big data that are increasingly being recognised when trying to extract value and make big data operational [11].

1.2.2. What is data analytics?

We refer to data analytics as the process of research into massive amounts of data to reveal hidden patterns and secret correlations [12]. Data analytics allows organisations to acquire new insights by combining and enriching traditional data with new kinds of internal or external data. Where data is typically stored in silos per department or business unit, big data (technologies) facilitate a move towards analytics based on ‘zero latency’ operational data, integrating historical and real time data from different sources. This in turn allows for more accurate predictive analytical models and machine learning techniques to be applied.

Although many are convinced of the added value data analytics can bring, PwC’s 2016 Global Data and Analytics Survey found that 61% of respondents acknowledged their companies could rely more on data analysis less on intuition [13]. Furthermore, the study found that just 21% of the surveyed government senior executives consider their organisation to be highly data-driven.

1.2.3. What is interoperability?

Interoperability is the ability of organisations³ to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems [14]. Interoperability influences an organisation’s performance and is a complex challenge for organisations deploying big data architectures, because of the heterogeneous nature of the data used by them [15]. It influences how data can be stored, integrated and used as a single

---

³ ‘Organisations’ here means public administration units or any entity acting on their behalf, or EU institutions or bodies.
entity. The ISA\textsuperscript{2} programme of the European Commission has defined the European Interoperability Framework (EIF)\textsuperscript{4}, which defines interoperability across four layers:

- **Legal**: “ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together”.
- **Organisational**: “refers to the way in which public administrations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals”.
- **Semantic**: “ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties”.
- **Technical**: “covers the applications and infrastructures linking systems and services”.

Although most fundamentally (big) data interoperability revolves around the layer of **semantic** interoperability, all layers can have an impact on interoperability in a big data context. **Legal** constraints can impose restrictions on which data can be stored and analysed, **organisational** constraints can limit access to certain valuable datasets or dictate certain data licences that prevent reuse of the data and **technical** constraints can lead to limited data portability. Furthermore, the EIF also defines an overarching background layer ‘**interoperability governance**’, referring to decisions on interoperability frameworks, institutional arrangements, organisational structures, roles and responsibilities, policies, agreements and other aspects of ensuring and monitoring interoperability at national and EU levels.

### 1.3. Structure of the report

Section 2 of this document explores the common interoperability issues in a big data context through literature review and the experience of the authors. Section 3 outlines a set of good practices to deal with these interoperability issues and how these can be implemented in a case study serving as an example. In section 4, we summarise the good practices and draw conclusions.

2. COMMON INTEROPERABILITY CHALLENGES WITH BIG DATA

In this section, an overview of common interoperability issues in the context of big data is presented. Section 2.1 gives an overview of the issues and their impact in a big data environment in terms of the effort required to resolve the issue. In section 2.2, a fictional case study incorporating the different challenges is presented.

2.1. Common data interoperability issues

In this section, the key issues are elaborated, taking into account the different layers of the European Interoperability Framework [16]. Subsequently, in section 3, the issues are translated into a number of specific business challenges for which good practices will be defined.
### Table 1 - Overview of common data interoperability issues

<table>
<thead>
<tr>
<th>Issue/challenge</th>
<th>Description</th>
<th>EIF Interoperability Layer</th>
<th>Impacted Vs of big data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor data quality</td>
<td>One or multiple data sources may exhibit poor data quality, rendering the whole data unreliable. Data quality affects interoperability when the fields required to link different data sources together are missing or inconsistent [17].</td>
<td>All</td>
<td>Veracity</td>
</tr>
<tr>
<td>Data protection considerations</td>
<td>The nature of big data makes the application of traditional principles of personal data protection challenging, such as purpose limitation or data minimisation. We must ensure that when big data involves the processing of personal data, the persons affected can exercise their personal autonomy and their rights to control their data [18].</td>
<td>Legal</td>
<td>Variety</td>
</tr>
<tr>
<td>Different data licences apply to the data sources.</td>
<td>The terms of use for the data content differ between the data sources used. This may include restrictions on (commercial) reuse and modification, and requirements such as attribution and share-alike [19].</td>
<td>Organisational</td>
<td>Variety</td>
</tr>
<tr>
<td>Decoupling between data producer and data user/scientist</td>
<td>When data is collected, this is often with a specific purpose in mind. Later, the same data may be used in different contexts, where each context can have different requirements towards data quality.</td>
<td>Organisational</td>
<td>Veracity</td>
</tr>
<tr>
<td>Complex and time-consuming data integration process.</td>
<td>The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration, to an extent that traditional Extract, Transform, Load (ETL) methods no longer suffice [17].</td>
<td>Semantic Technical</td>
<td>Variety</td>
</tr>
<tr>
<td>Schema-level conflicts between data sources</td>
<td>Schema-level conflicts - such as the use of different names for the same concept or describing the same concept with different attributes - prevent information systems to automatically interpret the information exchanged in a meaningful and accurate way because of differences in logical structures and/or inconsistencies in metadata [20, 21]. In a big data context this</td>
<td>Semantic</td>
<td>Variety</td>
</tr>
</tbody>
</table>
### Issue

- **Data-level conflicts between data sources**
  - Data-level conflicts are caused by differences occurring in data domains due to multiple possible representations and interpretations of similar data [22, 21]. For example, different units of measure, date formats, code lists, etc. may be used for the same data attribute.
  - The demand for analytics on streaming data, on-demand integration and self-service business intelligence applications means that there is less time, both in terms of human and machine processing time, for applying post-processing steps to harmonise the data [23].

- **Lack of interfacing mechanisms between systems.**
  - The inability of two or more systems to communicate with each other caused by the use of different or incompatible communication protocols and data formats [24]. Furthermore, when adopting technologies that are platform-dependent and only work with certain proprietary hardware or software components, there is a risk that it becomes difficult or even impossible to interface with technologies from other vendors or the open-source community. This affects both the interoperability and portability of the big data solution [25].

<table>
<thead>
<tr>
<th><strong>Data-level conflicts between data sources</strong></th>
<th><strong>Semantic</strong></th>
<th><strong>Variety</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-level conflicts are caused by differences occurring in data domains due to multiple possible representations and interpretations of similar data [22, 21]. For example, different units of measure, date formats, code lists, etc. may be used for the same data attribute.</td>
<td>Semantic</td>
<td>Variety</td>
</tr>
<tr>
<td>The demand for analytics on streaming data, on-demand integration and self-service business intelligence applications means that there is less time, both in terms of human and machine processing time, for applying post-processing steps to harmonise the data [23].</td>
<td>Technical</td>
<td>Velocity</td>
</tr>
<tr>
<td><strong>Lack of interfacing mechanisms between systems.</strong></td>
<td><strong>Technical</strong></td>
<td><strong>Variety</strong></td>
</tr>
<tr>
<td>The inability of two or more systems to communicate with each other caused by the use of different or incompatible communication protocols and data formats [24]. Furthermore, when adopting technologies that are platform-dependent and only work with certain proprietary hardware or software components, there is a risk that it becomes difficult or even impossible to interface with technologies from other vendors or the open-source community. This affects both the interoperability and portability of the big data solution [25].</td>
<td>Technical</td>
<td>Variety</td>
</tr>
</tbody>
</table>
2.2. Running example

The challenges and related good practices for big data interoperability will be studied in the light of a fictional case study related to informed EU policy making.

The case study aims at identifying gaps in EU policy and ideas for effective measures and actions based on existing policies, by bringing to the surface the opinions and concerns of citizens and other stakeholders towards the issues and actions in question. To realise this, two analytical models are proposed: One employing text analysis of related contributions posted on the Better Regulation portal and on Twitter (identification and analysis of policy gaps), and another focusing on the analysis of data from Eurostat’s socio-economic datasets and existing EU policies and legislation. The outcomes from applying these analytical models can be used as input to a policy benchmarking exercise in order to derive specific policy recommendations.

We assume that the data used in the analysis will be stored in a data lake, a storage repository containing raw data, and pulled from the following sources:

- Better regulation portal website\(^5\)
- Twitter’s Search API\(^6\)
- Eurostat public database\(^7\)
- CELLAR\(^8\) SPARQL Endpoint\(^9\)
- Flash Eurobarometer 234: Citizens' perceptions of EU Regional Policy\(^10\)

These sources expose their data in a variety of formats, using various licenses and through a range of delivery channels, from data dumps to web services.

---

\(^5\) [https://ec.europa.eu/info/law/better-regulation/](https://ec.europa.eu/info/law/better-regulation/)


\(^7\) [http://ec.europa.eu/eurostat/data/database](http://ec.europa.eu/eurostat/data/database)

\(^8\) CELLAR stores and disseminates all content and metadata created or disseminated by the Publications Office. It also drives the Publications Office’s major portals, including EUR-Lex and OP Portal.


3. **GOOD PRACTICES FOR ACHIEVING INTEROPERABILITY IN A BIG DATA ENVIRONMENT**

This section provides good practices and guidelines for those stakeholders that will be faced with the challenges defined in section 2.1. The challenges are structured according to the interoperability layers of the European Interoperability Framework. To provide a tangible example, the good practices and guidelines are applied to the fictional business case defined in section 2.2.

3.1. **Legal interoperability**

Legal interoperability is defined by the European Interoperability Framework as “ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together”. As part of this layer, it is proposed that legislation should undergo a ‘digital check’, including the identification of any barriers to digital exchange. In this section, we examine the challenges of dealing with different legislation that applies to data sources, intellectual property rights and the implications of data protection regulation on big data projects.

3.1.1. **Legislation affecting data and interoperability**

As part of its Digital Single Market strategy, the European Commission has reviewed and revised the directive on the re-use of public sector information (Directive 2003/98/EC, known as the 'PSI Directive')\(^{11}\). The PSI Directive provides a common legal framework for government-held data (public sector information) in the European market. It states that: "**All content that can be accessed under national access to documents laws is in principle re-usable beyond its initial purpose of collection for commercial and non-commercial purposes**”\(^{12}\). Moreover, the data should be made available to everyone, with charges not exceeding the marginal cost of reproduction, provisioning or dissemination. Member States were obliged to transpose Directive 2013/37/EU into national law by 18 July 2015.

Moreover, useful data to include in analyses, such as personal data, spatial data, information about businesses, statistics, legislation and procurement data, is all covered by Directives, Regulations and national legislative instruments. To foster legal interoperability across the European single market, the European Commission has made and updated multiple Directives, going beyond the PSI Directive. For example, Directive 96/9/EC (known as the ‘Database Directive’)\(^{13}\), aims to harmonise how copyright law is applied to databases across the European single market. Another example is Directive 2007/2/EC (known as the ‘INSPIRE Directive’)\(^{14}\) which aims to ensure the interoperability of geospatial information across Europe, by defining a set of technical implementation rules to allow for smooth cross-border data integration across 34 spatial data themes\(^{15}\). At the same time, the General Data Protection


\(^{12}\) Exceptions to this principle exist, the PSI directive states that "content held by museums, libraries and archives is only re-useable if it is made available by the institutions for re-use". Another exception exists in the context of 'personal data'. More information on this can be found in section 3.1.2.


\(^{15}\) Overview of the different INSPIRE themes: [https://inspire.ec.europa.eu/index.cfm/pageid/2/list/7](https://inspire.ec.europa.eu/index.cfm/pageid/2/list/7)
Regulation (GDPR) (Regulation (EU) 2016/679) specifies how to process personal data while respecting individual’s fundamental rights.

**Good practice:** when reusing data, especially in a cross-border context, it is paramount to check which legislation applies to the domain in question. For European Union Law, consult EUR-Lex\(^{16}\) or use the machine-readable versions offered through CELLAR\(^{17}\).

### 3.1.2. Considerations when processing personal data

In the context of big data, dealing with personal data raises much concern. The use of big data analytics may reveal certain personal details when analysing customer behaviour, for example that people with a specific ethnicity, living in a certain area and with a given marital status are willing to pay more for a certain product \([26]\) [27]. Furthermore, analytical data models built on historical data may result in automated bias or discrimination \([28]\).

These risks have not gone unnoticed to lawmakers. Since 2016, multiple guidelines and resolutions have been issued by the European Data Protection Supervisor, the Council of Europe and the European Parliament, among which:

- European Data Protection Supervisor Opinion 8/2016 on coherent enforcement of fundamental rights in the age of big data \([30]\).
- The establishment of a Digital Clearinghouse by the European Data Protection Supervisor to bring together agencies from the areas of competition, consumer and data protection willing to share information and discuss how best to enforce rules in the interests of the individual \([31]\).
- The guidelines on the protection of individuals with regard to the processing of personal data in a world of big data published by the Council of Europe \([32]\).
- The European Parliament resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement \([33]\).

In addition, the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) \([34]\), enforceable from 25 May 2018, aims to "strengthen citizens’ fundamental rights in the digital age and facilitate business by simplifying rules for companies in the digital single market”.

From the point of view of interoperability, the GDPR homogenises the rules by which data controllers and data processors in the European market have to play, avoiding data interoperability conflicts caused by different legal frameworks. Furthermore there is the concept of data portability, defined by the GDPR as “The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been stored.”

---


To fulfil this requirement, data controllers need to a common data standard to structure export and import formats for the transmission of personal data. Existing initiatives, such as the ISA Core Vocabularies\(^\text{18}\) and Schema.org\(^\text{19}\).

### 3.2. Organisational interoperability

Data is increasingly used for purposes other than those initially foreseen. As discussed in section 3.1 this can pose legal challenges when processing personal data, but also from an organisational point of view, the new use cases for data induced by big data technologies put forward challenges. In section 3.2.1 we investigate how existing data silos can be broken down, section 3.2.2 looks at how to deal with various data licences and in section 3.2.3 we discuss how data requirements can be aligned within an organisation.

#### 3.2.1. Breaking down data silos within your organisation

The communication to the Commission on “Data, information and knowledge management at the European Commission”\(^\text{20}\) recognises the strategic use of data, information and knowledge as a key element for improving the current way of working in the Commission. Emphasising teamwork, overcoming silo mentalities, and harnessing synergies between portfolios. European Union institutions and public administrations in the EU show a high interest in learning and sharing good practices on information governance and management, seeking to increase the value of their own data and information assets [36].

When data is stored within organisational silos, unbeknown to other teams, departments or institutions, the data cannot not be utilised for other purposes. As a result, a lot of the data collected by organisations (both public and private) is currently underutilised. For example, the 2017 Open Data Barometer found through a survey of 115 public administrations that when it comes to using open data to increase overall government efficiency and effectiveness, the average government only scores 1.20 out of 10 [37].

**Good practice:** to break down data silos, organisations can invest in open (available for anyone on the web) or closed (availability within the organisation) data marketplaces or portals, where datasets can be published and discovered by potential users. Such a marketplace or ‘data portal’ can [38]:

- support interoperability through a shared metadata model, like the Data Catalog Vocabulary (DCAT)\(^\text{21}\);
- create a single point of discoverability for data scientists and;

---

18 A set of simplified, reusable and extensible data models created by the ISA\(^2\) programme of the European Commission, more information: [https://ec.europa.eu/isa2/solutions/core-vocabularies_en](https://ec.europa.eu/isa2/solutions/core-vocabularies_en)

19 Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet. More information: [http://schema.org/](http://schema.org/)


21 [https://www.w3.org/TR/vocab-dcat/](https://www.w3.org/TR/vocab-dcat/)
- improve data quality consistency through mutually agreed service-level agreements.

Existing open data portals, available for both data publishers and data scientists, include:

- European Union Open Data Portal\textsuperscript{22}: contains thousands of datasets about a broad range of topics in the European Union.
- European Data Portal\textsuperscript{23}: harvests metadata from public sector portals throughout Europe (e.g. from national open data portals).
- Google’s Public Data Directory\textsuperscript{24}
- Amazon Web Services Public Datasets\textsuperscript{25}: contains a wide range of dataset among which the human genome project.

However, to ensure optimal reuse of this data, there is a need for holistic information governance within the organisation. Information governance defines roles and responsibilities for information management, and deals with information quality, security and data architecture.

**Good practice:** successful information governance requires active support from the organisation’s highest executive level as it can touch upon many functional aspects within one organisation.

Based on the mandate given by the executive level to an individual or an organisational body for the implementation of information governance, policies and procedures can be created and enforced across different domains of information governance. Figure 2 provides an overview of the domains defined by the DAMA Data Management Body of Knowledge\textsuperscript{26}.

\textsuperscript{22} https://data.europa.eu/euodp/data/
\textsuperscript{23} http://www.europeandataportal.eu/
\textsuperscript{24} https://www.google.com/publicdata/directory
\textsuperscript{25} https://aws.amazon.com/datasets/
\textsuperscript{26} https://dama.org/content/body-knowledge
Good practice: use existing information governance frameworks such as DAMA DMBOK\(^\text{27}\) as a basis for the implementation of information governance within the organisation.

A 2017 study, conducted in the context of ISA\(^2\)'s SEMIC Action, on 'Information governance and management for public administrations' [36] whose aim is to raise awareness among the Member States of the importance and the benefits of information governance, provides a set of high-level good practices based on several relevant case studies:

1. Information governance is based on the principle that data, information and knowledge should be shared, accessible and reusable as widely as possible. Generally, information should be regarded as a valuable asset.

2. The governance structure combines roles on both a strategic and tactical level. The most common roles on a tactical level are:
   - **Information steward**: the person from the business who is ultimately responsible for an information asset\(^\text{28}\);
   - **Information manager**: the person who has technical or operational control of an information asset. Information managers define concepts, the content of their attributes, and the quality requirements;
   - **Information architect**: responsible to bridge the business and IT sides.

3. The management of reference data, metadata and master data follows a structured and formalised approach. In most of the analysed cases, the management is centralised.

\(^{27}\) https://www.dama.org/content/body-knowledge

\(^{28}\) An information asset is a body of information, defined and managed as a single unit so it can be understood, shared, protected and exploited efficiently: http://www.nationalarchives.gov.uk/documents/information-management/information-assets-factsheet.pdf
4. Information quality requirements are defined with specific quality criteria identified based on principles and in-line with organisations’ strategies. Specific processes to address quality issues are established.
5. Information classification schemes are defined to assess the appropriate information security measures to apply.
6. Organisations successful in implementing information governance put in place ongoing activities to train staff and develop capabilities on data, information and knowledge management.

3.2.2. How to deal with various data licences

The absence of clarity regarding data licences is a first hurdle to overcome when looking to reuse (external) data. The restrictions that apply may include limitations on (commercial) reuse, modification and requirements such as attribution and share-alike [19]. Creative Commons, a non-profit organisation offering predefined licences to content creators, defines four categories of licensing conditions which may apply:

1. Attribution: the requirement that credit is given to the people or organisations whose work is reused.
2. Share-alike: if the work of others is reused, the modified work has to be published under the same licensing terms that apply to the source(s).
3. Non-commercial: reuse is only permitted for non-commercial purposes.
4. No-derivatives: the source may not be modified.

Even though a dataset contains the perfect data for a certain project or analysis, any of the above licensing conditions may prevent its reuse. In a big data context, when dealing with multiple sources, making sure all applicable licences are compatible poses a serious challenge.

As a data scientist reusing external data, it is imperative to make sure that the applicable data licence complies with the Open Definition, "knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness." [31]. A list of reusable licences that comply with this definition, and are thus compatible, include:

- Creative Commons CCZero (CC0)
- Open Data Commons Public Domain Dedication and Licence (PDDL)
- Creative Commons Attribution 4.0 (CC-BY-4.0)
- Open Data Commons Attribution License (ODC-BY)
- Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)

---

29 https://dl.acm.org/citation.cfm?id=2887378
30 https://creativecommons.org/share-your-work/licensing-types-examples/
31 http://opendefinition.org/od/2.1/en/
32 http://opendefinition.org/licenses/
33 http://opendefinition.org/licenses/cc-zero
34 http://opendefinition.org/licenses/odc-pddl
35 http://opendefinition.org/licenses/cc-by
36 http://opendefinition.org/licenses/odc-by
37 http://opendefinition.org/licenses/cc-by-sa
• Open Data Commons Open Database License (ODbL)\textsuperscript{38}

Furthermore, the European Commission has created the first European Free/Open Source license, the European Union Public Licence (EUPL)\textsuperscript{39}. This licence is certified by the Open Source Initiative\textsuperscript{40} as complying with the Open Source Definition\textsuperscript{41}, making it especially convenient for use by public administrations.

Even when no specific licence is provided, a dataset may still be covered under copyright law. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases states that under the *sui generis* right\textsuperscript{42}, a property right to recognise the investment that is made in compiling a database, even when this does not involve the ‘creative’ aspect that is reflected by copyright, the creator of the database may prohibit reuse for fifteen years from the end of the year the dataset was made public \textsuperscript{[39]}. Therefore, in case of doubt, it is best to not use the dataset before reaching an agreement with the publisher.

**Good practice for data scientists:** validate whether the datasets planned to be reused are published under a license that complies with your objectives and ideally with the Open Definition\textsuperscript{43}. If no license information can be found, contact the owner of the dataset. A full re-user’s guide to open data licensing can be found on the Open Data Institute’s website\textsuperscript{44}.

**Case study example – a re-user’s perspective**

In our case, data from five different data sources is reused. In this example, the reuse conditions that apply for three of those are analysed:

1. **Better regulation portal**

   The data collected through the better regulation portal\textsuperscript{45} is published via the website. The website’s legal notice\textsuperscript{46} states that reuse is authorised, provided the source is acknowledged. As no specific data licence is referenced, it is advisable to contact the data controller as listed on the website\textsuperscript{47} before reusing the data.

2. **Twitter**

   As part of its Developer Agreement and Policy\textsuperscript{48}, Twitter provides users of its Twitter Application Programming Interface (API) with a proprietary license allowing:

   - Use the Twitter API to integrate Content into your Services or conduct analysis of such Content;
   - Copy a reasonable amount of and display the Content on and through your Services to End Users, as permitted by this Agreement;
   - Modify Content only to format it for display on your Services; and

\textsuperscript{38} [http://opendefinition.org/licenses/odc-odbl](http://opendefinition.org/licenses/odc-odbl)

\textsuperscript{39} [https://ec.europa.eu/info/european-union-public-licence_en](https://ec.europa.eu/info/european-union-public-licence_en)

\textsuperscript{40} [https://opensource.org/licenses/EUPL-1.1](https://opensource.org/licenses/EUPL-1.1)

\textsuperscript{41} [https://opendefinition.org/od/2.1/en/](https://opendefinition.org/od/2.1/en/)


\textsuperscript{43} [https://opensource.org/osd](https://opensource.org/osd)

\textsuperscript{44} [https://theodi.org/guides/reusers-guide-open-data-licensing](https://theodi.org/guides/reusers-guide-open-data-licensing)

\textsuperscript{45} [https://ec.europa.eu/info/law/better-regulation/have-your-say_en](https://ec.europa.eu/info/law/better-regulation/have-your-say_en)

\textsuperscript{46} [https://ec.europa.eu/info/legal-notice_en](https://ec.europa.eu/info/legal-notice_en)

\textsuperscript{47} [https://ec.europa.eu/info/law/better-regulation/specic-privacy-statement_en](https://ec.europa.eu/info/law/better-regulation/specific-privacy-statement_en)

As a data publisher, it is possible to stimulate reuse of your datasets by applying clear and consistent licensing conditions to them. When making datasets available, it is important to make sure all rights necessary to do so are fulfilled. As a publisher or provider, you may not be the original author of the contents of the dataset; this is the case for many open data portals. If this is the case, permission from the rightholder(s) must first be secured before publishing the dataset. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases specifies the database author as the rightholder, stipulating that "the author of a database shall be the natural person or group of natural persons who created the base or, where the legislation of the Member States so permits, the legal person designated as the rightholder by that legislation".

As the datasets relevant for our analysis are all provided by Eurostat, these can be reused in the analysis, provided that source attribution is given.

### 3. Eurostat socio-economic data

Eurostat’s policy on free reuse of data states that: “Eurostat has a policy of encouraging free reuse of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that the source is indicated as Eurostat and when reuse involves modifications to the data or text, this must be stated clearly to the end user of the information.”.

As the scope of the analysis entails performing text analysis on Tweets, the data made available through the Twitter API can be reused given that reference to Twitter as the source is provided in the final output of the work.

---

As a data publisher, it is possible to stimulate reuse of your datasets by applying clear and consistent licensing conditions to them. When making datasets available, it is important to make sure all rights necessary to do so are fulfilled. As a publisher or provider, you may not be the original author of the contents of the dataset; this is the case for many open data portals. If this is the case, permission from the rightholder(s) must first be secured before publishing the dataset. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases specifies the database author as the rightholder, stipulating that "the author of a database shall be the natural person or group of natural persons who created the base or, where the legislation of the Member States so permits, the legal person designated as the rightholder by that legislation”.

In order to make it easier for data scientists to discover datasets and the licensing conditions that apply, a licence can be published in a machine-readable format, allowing potential re-users to automatically determine whether a dataset is suitable for reuse in a certain project. Popular open licences, like Creative Commons through its License Chooser tool, provide code snippets with machine-readable license information that can be embedded in web pages.

**Good practice for data publishers:** the licence under which your dataset is published should be clear both in human-readable content and as machine-readable data. A full publisher’s guide to open data licensing can be found on the Open Data Institute’s website.

---

49 http://ec.europa.eu/eurostat/about/policies/copyright
50 https://creativecommons.org/choose/
51 https://theodi.org/guides/publishers-guide-open-data-licensing
Case study example – publishing machine-readable license information

In Code snippet 1 a machine-readable version of the dataset “Flash Eurobarometer 234: Citizens' perceptions of EU Regional Policy” is presented, obtained through the EU Open Data portal\textsuperscript{52}, including machine-readable license information expressed according to the Resource Description Framework (RDF)\textsuperscript{53}.

The specified license is Creative Commons Attribution-ShareAlike 4.0 International\textsuperscript{54}, described according to the Creative Commons Rights Expression Language\textsuperscript{55}.

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix cc: <http://creativecommons.org/ns#> .

#Metadata concerning the dataset
  a dcat:Dataset ;
  dc:title "Flash Eurobarometer 234: Citizens' perceptions of EU Regional Policy"@en ;
  dc:theme <http://eurovoc.europa.eu/4214> ;
  dc:license <https://creativecommons.org/licenses/by-sa/4.0/> .

#Metadata concerning the license referenced by the dataset
<https://creativecommons.org/licenses/by-sa/4.0/>
  a dc:LicenseDocument, cc:License ;
  dc:title "Attribution-ShareAlike 4.0 International" ;
  dc:identifier "CC BY-SA 4.0" ;
  cc:permits cc:Reproduction, cc:Distribution, cc:DerivativeWorks ;
  cc:requires cc:Attribution, cc:Notice, cc:ShareAlike .
```

Code snippet 1: machine readable license information

\textsuperscript{53} The Resource Description Framework is a W3C standard for the interchange of data on the Web, more information: https://www.w3.org/RDF/
\textsuperscript{54} https://creativecommons.org/licenses/by-sa/4.0/
\textsuperscript{55} https://creativecommons.org/ns
### 3.2.3. Aligning data requirements with data usage

Data is typically collected by organisations with a specific purpose or use case in mind; to support a business process or as input for a (management) decision. Depending on the use case, different requirements apply to the data quality and data collection methods. In a big data context, data is often reused as input for use cases that were not considered when defining the data requirements. In data quality literature, this concept is referred to as ‘data relevancy’: the degree to which the data is relevant for a certain task \[17\] \[11\]. The data reused in a big data project should therefore be ‘fit for use’, but how can these unforeseen use cases when defining data requirements be foreseen?

The people, teams or departments responsible for the collection and maintenance of datasets often lack feedback concerning how the data is being used outside of the initial use case. This makes it challenging to adapt current data collection processes to accommodate for those data usage purposes that were not foreseen. This feedback may come from stakeholders within your organisation or — when publishing open datasets on the Web — from anywhere in the world. This chasm between dataset curators and users can be bridged by implementing feedback and usage loops in our data management processes.

**Good practice:** to counter the discrepancy between the initial requirements for a dataset and its applications in practice, organisations should foresee in a structured manner to provide feedback on how data is being used and the related usage experiences. To structure this feedback, the Data Usage Vocabulary\[56\] published by W3C’s Data on the Web Best Practices Working Group can be used.

#### Case study example

The Flash Eurobarometer 234: Citizens' perceptions of EU Regional Policy, one of the data sources for our analysis, collects views of citizens on the European Cohesion Policy. In the analysis, data to help identify ideas for effective measures and actions for EU policy making is used. Making this use case known to the dataset's publisher allows the publisher to take it into consideration when working on a new version of the dataset. For example, they may opt to include additional questions related to the demographics of the respondents or provide alternative distribution formats that facilitate reuse.

In code snippet 2, user feedback and usage information is presented in a machine readable format using the Data Usage Vocabulary. By structuring the feedback, different data portals can automatically harvest the user feedback and display it alongside the dataset's metadata.

```mermaid
@prefix duv: <http://www.w3.org/ns/duv#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix dct: <http://purl.org/dc/terms/> .
```

\[56\] [https://www.w3.org/TR/vocab-duv](https://www.w3.org/TR/vocab-duv)
3.3. Semantic interoperability

In big data, it is common to combine a large number of different data sources. Achieving meaningful semantic interoperability of data from heterogeneous sources is a challenging issue for both private enterprises and the public sector [22]. This section explores how the concepts of ontology engineering, the Semantic Web and linked data can reduce this variability.

Traditional big data technology stacks based on for example Hadoop — which focus on manipulating metadata when processing information because this is far more efficient than moving around the data itself — embed the metadata into the program code used to integrate and analyse data and thus render data integration with other information systems more complex [40]. Or, in other words, these traditional big data implementations often lead to the creation of new data silos to replace the ones that were present previously [41]. Using Semantic Web standards has the advantage

---

57 Apache Hadoop is an open-source software framework used for distributed storage and processing of datasets of big data using the MapReduce programming model: [http://hadoop.apache.org/](http://hadoop.apache.org/)
that the semantic enrichment provides metadata which is machine-readable and linked to the data independently of any system that uses the data.

The term “Semantic Web” refers to a set of standards and technologies to realise W3C’s vision of the Web of linked data[^58]; where the Web not only contains hyperlinks from one document to another, but where those documents also contain structured data interpretable by machines, thus creating a collection of interlinked data stores on the Web. Standards such as the Resource Description Framework (RDF)[^59] and the Web Ontology Language (OWL)[^60] offer a way to specify semantic definitions and relations. These enable the use of inferencing to help with data integration, by automatically discovering new relationships or analysing the data contents[^61].

Section 3.3.1 outlines how Semantic Web standards and technologies can be used by data publishers to “close the semantic gap”. Next, section 3.3.2 points to techniques to build and reuse ontologies to structure data as a data scientist. Lastly, section 3.3.3 looks at how to optimise performance by choosing the right data format.

### 3.3.1. Closing the semantic gap between datasets

When integrating data from a variety of sources, each source may contain a different description and interpretation of the same concept or resource. As such, one dataset may describe a ‘novel’ written by ‘writer A’, another dataset may describe a ‘book’ authored by ‘author B’. In these two datasets, writer A and author B may both describe the exact same person, e.g. Alexandre Dumas, and respectively these datasets may refer to his work “The Count of Monte Cristo” as either a ‘novel’ or a ‘book’.

A human can easily interpret that in this context, a ‘writer’ and ‘author’ or a ‘novel’ and ‘book’ refer to the same thing. For a machine, however, these considerations are not as straightforward. Closing this semantic gap by manually instructing machines how to process this data is a time-consuming activity which is not scalable in a big data context [42].

Semantic gaps can be divided into two levels following the classification of Peristeras, Loutas, Goudos, & Tarabanis [43]:

- **Schema-level conflicts** are caused by differences in logical structures and/or inconsistencies in metadata. Examples include the use of homonyms or synonyms in the naming of concepts, the use of different identifiers for the same object, generalisation and aggregation of attributes.

- **Data-level conflicts** are caused by differences occurring in data domains due to multiple possible representations and interpretations of similar data. For example, the use of different controlled vocabularies as values for the same attribute, using different representation formats (e.g. for dates), the use of different units of measurement or a difference in data precision.

**Good practice:** A good practice to avoid semantic gaps lies in defining how the information should be understood. This can be done on two levels. The simplest solution for this is to build shared metadata repositories that describe the content.

[^58]: https://www.w3.org/standards/semanticweb/
[^59]: http://www.w3.org/TR/rdf-schema/
[^60]: http://www.w3.org/TR/owl-primer
[^61]: https://www.w3.org/standards/semanticweb/inference
and intent of data stored in the various information systems. Another solution is to build an ontology to support interoperability. This is more challenging but allows for smooth interoperability once in place [24].

**As a data publisher**, semantic interoperability of datasets can be improved by publishing the semantics together with the data itself. While it is possible to create and publish your own vocabulary or ontology^62^, many vocabularies already exist, possibly describing similar or complementary domains to the one described by the dataset you wish to publish. Therefore, interoperability between vocabularies becomes essential to limit the schema-level conflicts and take advantage of the semantic web [42].

**Good practice:** Before defining a proprietary vocabulary, search whether there are existing, community supported vocabularies that can be reused for a specific domain of interest. As a starting point, the following resources can be consulted:

- ISA² Core Vocabularies^63^
- Linked Open Vocabularies portal^64^  
- List of W3C endorsed vocabularies^65^  
- Eurovoc, the EU's multilingual thesaurus^66^  
- European Publication Office’s Named Authority Lists^67^  

Specifically for the statistical domain, the Data Cube Vocabulary^68^ provides a standard for representing statistical datasets and the observations contained therein.

It is likely that existing vocabularies will not fit your dataset 100%. For new datasets, existing vocabularies can be used as a starting point and, if necessary, extend upon, ensuring a minimum set of semantically interoperable descriptors for data. For existing datasets, the data model can be mapped to existing vocabularies, serving as a bridge to a common foundational data model.

**Good practice:** existing data models used for the publication of datasets can be mapped to vocabularies using the semantic relationships^69^ defined by the Simple Knowledge Organisation System (SKOS). A template to create such a mapping as well as examples for the different ISA² Core Vocabularies can be found in the Core Data Model Mapping Directory^70^.

^62^ Ontology and vocabulary are interchangeable terms. Although in practice there is a trend to use "ontology" for a more complex or formal collection of terms.


^64^ [http://lov.okfn.org/dataset/lov/](http://lov.okfn.org/dataset/lov/)

^65^ [https://www.w3.org/standards/techs/rdfvocab](https://www.w3.org/standards/techs/rdfvocab)


^68^ [https://www.w3.org/TR/vocab-data-cube/](https://www.w3.org/TR/vocab-data-cube/)

^69^ An overview of the possible semantic relationships: [https://www.w3.org/TR/skos-primer/#secrel](https://www.w3.org/TR/skos-primer/#secrel)

^70^ [http://mapping.semic.eu/](http://mapping.semic.eu/)
When extending an existing vocabulary, or developing a new one from scratch, with the aim of creating semantically interoperable information systems, it is important to first establish semantic agreements with stakeholders in order to prevent schema-level conflicts down the road.

**Good practice:** consult (potential) reusers of your data to build consensus regarding a canonical domain model, including definitions for the terms and concepts used within the domain model. Serving as a blueprint, a process and methodology for developing semantic agreements has been formalised by the ISA Programme\(^\text{71}\), based on their experience with the Core Vocabularies.

In order to avoid data-level conflicts, one must also specify which controlled vocabularies and data types should be used as values for attributes in a vocabulary.

**Good practice:** when compiling a vocabulary, consider and decide which controlled vocabularies to use as a standardised set of values for data attributes. The Publications Office, for example, has published a list of reusable controlled vocabularies dubbed ‘Named Authority Lists’ (NALs), including ones for country, language and gender\(^\text{72}\). Another source for controlled vocabularies is the EU’s multilingual thesaurus Eurovoc\(^\text{73}\). If a controlled vocabulary is too limiting, specify a data type such as Boolean, integer, string, date, etc. based on, for example, XML Schema definitions\(^\text{74}\). Further guidance can be found as part of the ISA\(^2\) Guidelines for the Use of Code Lists\(^\text{75}\).

These good practices to close the semantic gap, however, will only be useful in a big data context when these semantic agreements are published in a way machines can interpret.

**Good practice:** Using the principles of the semantic web, publishing data in a RDF format and providing users with access to the accompanying vocabulary described using OWL will improve the semantic interoperability of a dataset, allowing re-users to link it to other compatible data sources. We refer to W3C’s Best Practices for Publishing Linked Data as an implementers guide\(^\text{76}\).

**Case study example**

To allow the contributions on the Better Regulation portal to be more easily linked with other data, they can be represented in a Linked Data format using existing

---

\(^\text{71}\) Process and methodology for the development of common models and specific data entities that can be used as a basis for interoperability among information systems: [https://joinup.ec.europa.eu/sites/default/files/document/2015-03/Process%20and%20methodology%20for%20developing%20semantic%20agreements.pdf](https://joinup.ec.europa.eu/sites/default/files/document/2015-03/Process%20and%20methodology%20for%20developing%20semantic%20agreements.pdf)


\(^\text{74}\) XML Schema Datatypes In RDF and OWL: [https://www.w3.org/TR/swbp-xsch-datatypes/](https://www.w3.org/TR/swbp-xsch-datatypes/)


\(^\text{76}\) [https://www.w3.org/TR/ld-bp/](https://www.w3.org/TR/ld-bp/)
vocabularies like the Core Person Vocabulary\(^77\) and RDF Review Vocabulary\(^78\). Figure 3 provides a schematic representation using a UML class diagram, a code example using RDF is given in Code snippet 3.

![UML class diagram](image)

Figure 3: An example vocabulary for user feedback on the Better Regulation Portal

```turtle
#Core Person Vocabulary
@prefix person: <http://www.w3.org/ns/person#> .

#Friend of a Friend Vocabulary
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

#RDF Review Vocabulary
@prefix rev: <http://purl.org/stuff/rev#> .

#RDF Schema Vocabulary
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://example.org/BetterRegulationPortal/feedback/1234> a rev:Review ;
    rev:text "Lorem ipsum dolor sit amet" ;

<http://example.org/contributor/5678> a person:Person ;
    foaf:name "John Smith" ;
    person:citizenship <http://example.org/jurisdiction/France> .
```

Code snippet 3: RDF code example in Turtle syntax

**Case study example**

For our analytical model employing text analysis of related contributions posted on the Better Regulation portal and on Twitter, we need to be able to merge both datasets. In this example, we will use the JSON-LD syntax\(^79\) to make the outputs of the Better Regulation portal and on Twitter interoperable. JSON-LD is a RDF serialization that remains compatible with JSON and allows to embed the semantics into web service responses. The code snippet below show a simplified JSON output of the Twitter API\(^80\), which in its current form is not interoperable with our RDF based output from the Better Regulation Portal in the previous example.

---

\(^77\) [https://joinup.ec.europa.eu/solution/core-person-vocabulary](https://joinup.ec.europa.eu/solution/core-person-vocabulary)

\(^78\) [http://vocab.org/review/](http://vocab.org/review/)

\(^79\) [https://www.w3.org/TR/json-ld/](https://www.w3.org/TR/json-ld/)

Using a "JSON-LD context" to add semantic annotations to this output (either as a data publisher when building the API or by the data scientist when integrating the different datasets) helps a machine to automatically integrate data from different sources by replacing the context-depended keys in the JSON output (e.g. "text", "user", "name" in the example above) with URIs pointing to semantic web vocabularies.

### Code snippet 4: Example Twitter API output

```json
{
  "created_at": "Sun Feb 25 18:11:01 +0000 2018",
  "text": "Lorem ipsum dolor sit amet",
  "user": {
    "id": 11348282,
    "name": "John Smith",
  }
}
```

### Twitter API output supplemented with a JSON-LD context:

```json
{
  "@context": {
    "Review": "http://purl.org/stuff/rev#Review",
    "Person": "http://www.w3.org/ns/person#Person",
    "text": "http://purl.org/stuff/rev#text",
    "user": "http://purl.org/stuff/rev#reviewer",
    "name": "http://purl.org/stuff/rev#name"
  },
  "@type": "Review",
  "created_at": "Sun Feb 25 18:11:01 +0000 2018",
  "text": "Lorem ipsum dolor sit amet",
  "user": {
    "@type": "Person",
    "id": 11348282,
    "name": "John Smith"
  }
}
```

### Example Better Regulation Portal RDF output serialized as JSON-LD:

```json
{
  "@context": {
    "foaf": "http://xmlns.com/foaf/0.1/",
    "rev": "http://purl.org/stuff/rev#",
    "person": "http://www.w3.org/ns/person#"
  }
}
```
As a data scientist, the first step is to look at the available metadata concerning the datasets which are planned to be integrated. When reusing datasets external to your organisation and published on the web, metadata may be provided following one of multiple existing metadata standards. These standards are often tailored to a specific domain, examples include ‘Statistical Data and Metadata eXchange’ (SDMX)\(^2\) in the statistical domain and ISO 19115\(^3\) for metadata about geographic information.Metadata improves the findability of datasets and allows a data scientist to evaluate whether the dataset will be useful and compatible with other data sources without having to look at the data itself. However, this does not necessarily guarantee a smooth data integration path.

**Case study example**

In our running example, to identify gaps in EU policy and ideas for effective measures and actions based on existing policies, we want to create an analytical model focusing on the analysis of data from Eurostat’s socio-economic datasets and existing EU policies and legislation. One of the data sources in our case study

---

\(^1\) [https://json-ld.org/playground/](https://json-ld.org/playground/)

\(^2\) [https://sdmx.org/?page_id=5008](https://sdmx.org/?page_id=5008)

\(^3\) [https://www.iso.org/standard/53798.html](https://www.iso.org/standard/53798.html)
is Eurostat's socioeconomic statistics. Eurostat offers multiple datasets in the socioeconomic domain, three datasets are selected to include in the analysis:

- Income and living conditions\(^{84}\)
- Net earnings and tax rates\(^{85}\)
- Consumption expenditure of private households\(^{86}\)

Each dataset contains metadata records following the Euro SDMX Metadata Structure, Eurostat's metadata standard for describing statistical datasets. A small selection of the metadata fields provided is evaluated:

- Statistical concepts and definitions
- Statistical unit
- Reference area

**Table 2** provides an overview of the evaluation. For the sake of clarity, only the relevant parts of the field’s values have been listed. The ‘statistical unit’ row gives an idea of the concepts described within the dataset. For two out of three datasets, these are ‘households and household members’. Even though they do not share the exact same definition (cf. row ‘statistical concepts and definitions’ in **Table 2**), we evaluate them as semantically equivalent because both definitions are based around the criteria of people living together and sharing costs. Our third dataset (net earnings and tax rates) defines the concept of ‘the family’ on the basis of marital status, number of workers and number of children, which is semantically different from the concept of a ‘household’ in the other two datasets.

Furthermore, we note that also the geographical scope (cf. row reference area in **Table 2**) differs between our datasets. Before using them to train our analytical models, the datasets will have to be filtered by area.

**Table 2: Evaluation of dataset compatibility**

<table>
<thead>
<tr>
<th>Statistical unit</th>
<th>Income and living conditions</th>
<th>Net earnings and tax rates</th>
<th>Consumption expenditure of private households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Households and household members.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The family</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Households and household members.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Statistical concepts and definitions   |                             |                            |                                               |
| A 'private household' means "a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint |                             |                            |                                               |
| The data refer to an average worker at national level for different illustrative cases, defined on the basis of marital status (single vs. married), number of workers (only in the case of couples), number of |                             |                            |                                               |
| The definition of the household for the purpose of the HBS is based on the two following criteria: co-residence and sharing of expenditures. |                             |                            |                                               |

<table>
<thead>
<tr>
<th>Reference area</th>
<th>provision of the essentials of living”</th>
<th>dependent children, and level of gross earnings, expressed as percentage of the average earnings of an average worker (AW).</th>
<th>European Union and neighbouring countries. Aggregates: European Union, Euro Area, EEA and EFTA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU-Member States, Iceland, Norway, Switzerland, Turkey</td>
<td>EU Member States, Turkey, Iceland, Norway, Switzerland, Japan and the USA.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although our example, based on a small sample of datasets and metadata fields, shows how metadata on a dataset level can help to close the semantic gap, this is a laborious and interpretative process, which is not scalable in a big data context. Doing this on the data level is even more laborious, therefore we look at how to create semantic mappings in a (semi-)automated way as part of section 3.3.2.

### 3.3.2. (Semi-)automated vocabulary creation

As outlined in section 3.3.1, linked data relies on a relatively small set of conventions, one of which is the use of vocabularies, created using a few formally well-defined languages such as RDF and OWL [45]. In practice, however, many vocabularies are created to describe a domain for which one or more vocabularies already exist. Maintaining interoperability between ontologies is therefore also essential to help with data integration.

If various vocabularies are used across the data sources looking to be reused by a data scientist, a good practice is to map them to each other. A mapping establishes correspondence rules between concepts of two vocabularies, facilitating data integration. Once the mapping is finalised, data structured according to one vocabulary can easily be transformed to the other. In the domain of ontology engineering, this technique is referred to as “ontology mapping” [24] [42]. The same mapping technique as presented in 3.3.1, based on The Simple Knowledge Organisation System (SKOS) [47], can be used to define semantic relationships between the different concepts in two vocabularies.

#### Case study example

[45] [47] [24]

---

87 [https://www.w3.org/TR/skos-primer/](https://www.w3.org/TR/skos-primer/)
Building on the example vocabulary developed in 3.3.1 and presented in Figure 3, our vocabulary based on the Core Person Vocabulary is mapped to another existing one in the same domain; vCard – for describing people and organisations.

<table>
<thead>
<tr>
<th>Core Person Vocabulary</th>
<th>Semantic mapping relationship</th>
<th>vCard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Close match[^89]</td>
<td>Individual</td>
</tr>
<tr>
<td>Name</td>
<td>Exact match[^90]</td>
<td>Name</td>
</tr>
<tr>
<td>Citizenship</td>
<td>No match[^91]</td>
<td></td>
</tr>
</tbody>
</table>

Manually creating these mappings for a multitude of data sources and using this to create a canonical information model for your data integration solution is a useful but time-consuming process that requires expertise in the knowledge engineering domain.

**Good practice:** reuse or develop a vocabulary that can serve as a common baseline to map different data sources to. This vocabulary will serve as the canonical information model for the integration of different data sources.

To aid with this process, multiple open source tools are available that can assist with data integration by (semi-)automatically deriving these semantic mappings. Examples include:

- **Chimaera[^92]:** semi-automated vocabulary generation by merging different source vocabularies. The end user is presented with decisions that affect the end result.
- **PROMPT[^93]:** a plugin for Protégé (an open source ontology editor)[^94] for automatically mapping different vocabularies. The user has to resolve remaining conflicts that cannot be mapped or merged automatically.
- **DOG4DAG[^95]:** a Protégé plugin using natural language processing techniques to build vocabularies from an array of source documents (currently supports English and German). It also automatically maps discovered terms to existing ontologies.

[^88]: https://www.w3.org/TR/vcard-rdf/
[^89]: In SKOS, a “close match” assertion indicates that two concepts are sufficiently similar that they can be used interchangeably in applications that consider the two concept schemes they belong to.
[^90]: In SKOS, an “exact match” indicates semantic similarity—it is a sub-property of “close match”. However, it denotes an even higher degree of closeness: the two concepts have equivalent meaning, and the link can be exploited across a wider range of applications and schemes.
[^91]: No match indicates no corresponding term is present in the target vocabulary.
[^92]: http://www.ksl.stanford.edu/software/chimaera/
[^93]: https://protegewiki.stanford.edu/wiki/PROMPT
[^94]: https://protege.stanford.edu/
[^95]: https://protegewiki.stanford.edu/wiki/Ontology_Generation_Plugin_(DOG4DAG)
Good practice: To enable (semi-)automated semantic mapping during data integration, it is useful to also store the available metadata of datasets when harvesting data from different sources. This metadata can be used to derive insights that help with data integration and may improve the accuracy of the analytical model.

3.3.3. Optimising performance and interoperability by choosing the right data serialisation format

The good practices presented in section 3.3 to improve semantic interoperability are based on the Resource Description Framework. From a technical perspective, typical text (Turtle, JSON-LD, etc.) or XML (XML/RDF) based RDF formats incur a performance cost when parsing and writing, as do traditional formats like CSV. In a big data setting, these formats provide limited scalability and therefore decrease the overall processing speed. Binary formats on the other hand are compact and suitable for large-scale, high-performance systems. RDF Binary\textsuperscript{96} and Semantic Annotations for Linked Avro Data (SALAD)\textsuperscript{97} provide a basic encoding for RDF terms built respectively on the Apache Avro and Apache Thrift binary encoding. These frameworks like Avro and Thrift also specify a protocol for high-performance communication within a big data architecture and towards data consumers. Furthermore, standalone initiatives exist such as RDF HDT (Header, Dictionary, Triples), which also provide a binary serialisation format for RDF terms.

Good practice: choose an encoding format to store the collected data that balances scalability and performance with semantic annotations to support interoperability.

The conversion to a binary format should be performed before processing and serving the data to consumers. The data can be stored in the database in this binary format, or stored virtually in a cache. In benchmarking exercises, the RDF HDT binary format has proven to be up to 14 times faster than other RDF serialisation for downloading and querying data\textsuperscript{98}.

Case study example

In our example case study architecture, data serialisation can be performed as part of our virtualisation layer using Apache Avro to transform our data into a binary format for further processing.

\textsuperscript{96} https://jena.apache.org/documentation/io/rdf-binary.html
\textsuperscript{97} http://www.commonwl.org/draft-3/SchemaSalad.html
\textsuperscript{98} For the full details of the benchmarking exercise we refer to: http://www.rdfhdt.org/technical-specification/#numbers
3.4. Technical interoperability

In section 3.3 techniques that can aid with data integration on a semantic level were explored. In practice, the integration of multiple data sources also requires a unified set of technical interfaces for accessing data and an architecture for efficiently storing and processing the data. In this section, we zoom in on the technical interoperability challenges related to data integration patterns and avoiding vendor lock-in.

3.4.1. Choosing between data integration patterns

In a big data setting, multiple integration patterns exist, including [46]:

- **Data consolidation**: data is collected from multiple source systems and integrated into a single data store, using traditional Extract, Transform and Load (ETL) procedures.
- **Data federation**: the data remains stored in the source system and is only pulled on-demand to provide a virtual unified view. The metadata can be harvested and stored centrally.
- **Data propagation**: updates in the source systems are pushed automatically to a target system in a synchronous or asynchronous manner.

Based on the organisational setup and the availability of technical resources, one approach may be preferable over the others. For example, in a setting where all data sources are managed by the same organisation, it may be feasible to implement automated data propagation. The decision to go with synchronous or asynchronous propagation will largely depend on whether real-time analytics are required (if yes,
synchronous propagation is preferable) and the availability of system resources like processing power and memory (synchronous propagation will require more system resources). In environments where a central data store is not available, data federation may be needed. For a heterogeneous set of data sources managed by different organisations, data consolidation using ETL techniques may be the only way to obtain and integrate the data, requiring the least interoperability between sources but the highest processing effort. Lastly, these different integration patterns can be combined into hybrid approaches, tailored to a specific environment and use case [46].

As part of the data integration strategy, when opting for consolidation or propagation, it is imperative to also consider different data storage solutions. In broad terms, storage solutions in a big data context can be classified as either data warehouses or data lakes [47]. A **data warehouse** contains information accumulated from different sources in a structured, pre-processed manner, ready to be consumed by its users. A **data lake** on the other hand, is built on the notion that when collecting data, it is not always known or clear upfront how the data will be used or analysed. Therefore, data lake solutions store the data in a raw, unprocessed format (provided that the data lake solution supports the various data formats used in the analysis). It is up to the data scientist to process the data for the desired use case after retrieving it from the data lake.

Considering the increased demand for real-time analytics [23], one also needs to consider how to integrate **streams of unbounded data**. Unbounded data is an infinite stream of data collected by sensors (e.g. continuous monitoring of air quality) or events generated by a system (e.g. user clicks on a high-traffic website). A common technique is to perform windowing on the stream of data. With this technique, data points are grouped in a count or time-based manner and the window can move over time by sliding or hopping, as presented in Figure 5.

![Figure 5: Sliding (top) and hopping (bottom) windowing types](image)

In practice, for each of the data integration patterns presented, high system heterogeneity, the use of legacy systems and the use of different data formats can hinder interoperability.

**Good practice:** A good practice to deal with data integration in an environment with system incompatibility due to high system heterogeneity, legacy systems and
various data formats consist of working with web services and service-oriented architectures. As stated in [24] and references therein, these represent the state of the art for building integrated and interoperable enterprise systems.

A service-oriented architecture (SOA) is a design paradigm where functionality is provided through reusable service components over a network. Such an architecture may include data services providing ready-to-consume (processed) information, activity services which consume the data services to perform a specific task (e.g. an algorithm) and workflow services to trigger activities in a time or event driven manner. A SOA intrinsically prioritises interoperability over custom integrations, as all services are conceived to be reused by multiple other services on the same or a different architectural layer [48].

**Case study example**

To integrate the data coming from the different sources presented in Figure 1, we opt for a data consolidation strategy. All data sources are considered to be external to our organisation and the data is obtained through the publicly available distributions. All data is imported into a data lake, which will allow us to perform different data processing steps for our two analytical models as proposed in 2.2.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Distribution/format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better Regulation Portal</td>
<td>Data dump in CSV format</td>
</tr>
<tr>
<td>Twitter</td>
<td>REST JSON API</td>
</tr>
<tr>
<td>Eurostat socio-economic data</td>
<td>Data dump in TSV format</td>
</tr>
<tr>
<td>CELLAR</td>
<td>SPARQL Endpoint providing results in CSV, Turtle, XML, JSON, etc.</td>
</tr>
<tr>
<td>Flash Eurobarometer 234: Citizens' perceptions of EU Regional Policy</td>
<td>Data dump in XLS format</td>
</tr>
</tbody>
</table>

The data dumps in Comma Separated Values (CSV) or Tab Separated Values (TSV) format, as well as the output from the CELLAR SPARQL Endpoint, can be imported directly into our data lake without the need for any pre-processing. As the Flash Eurobarometer 234 results are distributed as Excel files, we will either have to make sure our data lake supports this format or transform our Excel dataset to CSV (e.g. by exporting to CSV directly from Microsoft Excel, LibreOffice Calc or using a programming language like R with an open source transformation library99).

To integrate Twitter data, an interface between our data lake and the Twitter API is set up, which provides access to a continuous stream of Tweets, pulling in information following a sliding window containing all Tweets between one year ago and today.

99 [https://cran.r-project.org/web/packages/rio/vignettes/rio.html](https://cran.r-project.org/web/packages/rio/vignettes/rio.html)
Next, to serve the data as input for our analytical models, a virtualisation layer is introduced on top of our data lake. This virtualisation layer will process the data on-demand, for example by transforming it to a structure adhering to a common baseline vocabulary, as outlined in section 3.3.2. By performing these transformations virtually (on-the-fly), the source data does not have to be altered, allowing us to process the data for different use cases as well as make quick adaptations to our transformation process without the need to store multiple versions of the source data. Furthermore, the processed data will be offered as data services to our analytical model, and any future model that would like to make use of this information.

![Case study architecture diagram including a virtualisation layer](https://ec.europa.eu/isa2/publications/european-interoperability-framework-eif_en)

Vendor lock-in occurs when adopting technologies that are platform-dependent and only work with certain proprietary data formats, hardware- or software components. In doing so, there is a risk that interfacing with or switching to technologies from other vendors or the open-source community becomes difficult, costly or even impossible. This affects both the interoperability and portability of the big data solution [25]. One mitigation strategy is to select technologies that are based on open standards. The EIF defines open standards using the following criteria:

- The standard is adopted and will be maintained by a not-for-profit organization, and its ongoing development occurs on the basis of an open decision-making procedure available to all interested parties (consensus or majority decision etc.).
- The standard has been published and the standard specification document is available either freely or at a nominal charge. It must be permissible to all to copy, distribute and use it for no fee or at a nominal fee.

• The intellectual property - i.e. patents possibly present - of (parts of) the standard is made irrevocably available on a royalty-free basis.

• There are no constraints on the re-use of the standard.

**Good practice:** assess providers’ technology implementations for potential areas of vendor lock-in. For example, Big Data solutions are available from different vendors, including Cloudera, HortonWorks, Amazon, Microsoft, IBM, etc. While all are based on the same basic open-source technology (Hadoop), they vary in technologies and versions for data integration, processing, etc.

Furthermore, setting up a service-oriented architecture as pointed out in 3.4.1 can also help to reduce the risks related to vendor lock-in, as custom integrations will be avoided in favour of reusable services. This implies that changes in technology solutions will only impact specific parts of the backend architecture and not the interfaces with other solutions.
4. Conclusion

This study has analysed interoperability in a big data context following the four layers of the European Interoperability Framework, with the aim of providing users with a set of good practices to deal with challenges affecting different dimensions of big data, with a focus on variety and veracity.

To ensure interoperability in a legal context (section 3.1), complying with legal frameworks, policies and strategies, we have found that:

- As a data producer, it is imperative to provide consumers with the means to access their data in a usable, machine-readable format to ensure data portability.
- When processing personal data to build predictive analytical models, organisations should be transparent about the decision criteria used by the analytical algorithm.
- Individuals should be provided with the means to access and manage the data that is being collected about them.

On an organisational level (section 3.2), to align business processes, responsibilities and expectations to achieve data interoperability beyond existing silos, we have stated the importance of:

- Making data available outside of existing organisational silos by publishing it on data marketplaces or portals to improve discoverability, data quality and interoperability, while also providing easy to find, clear, human and machine-readable license information.
- As a data scientist, checking the licensing conditions that apply to the data to be reused.
- Capturing feedback from data users to be able to align data requirements with actual use cases.

From a semantic interoperability point of view (section 3.3), this study has explored:

- Different scenarios to avoid or close semantic gaps, based on Semantic Web technologies, when integrating data from different sources.
- How to build a canonical information model based on existing standards to serve as a common baseline for semantic interoperability in a big data project.
- The selection of appropriate encoding formats which balance scalability, performance and interoperability.

Lastly, technical considerations (section 3.4) have been presented regarding:

- Different data integration patterns using service-oriented architectures and.
- The assessment of different technology solutions to avoid vendor lock-in.

These good practices aim to increase the understanding of officials in the EU institutions and Member States’ administrations on how to tackle different interoperability challenges when integrating data for analytical purposes in a big data context. The challenges and good practices presented in this study can help with the development of appropriate project plans and the selection of appropriate software vendors.
Annex I. Approach and Methodology

Literature review
This study started with a review of contemporary, peer-reviewed academic literature related to big data interoperability challenges, through the consultation of online resources, including Google Scholar, publications of the NIST big data working group, Harvard Business Review and other relevant sources. In total between 20 and 30 relevant papers have been read and analysed for big data interoperability challenges and best practices.

Identifying key interoperability
The challenges derived from the literature review were distilled into a number of key business challenges related to big data interoperability. For each challenge a short definition was provided as well as an indication of the significance of the impact in a big data context. The impact was assessed based on literature review and the experience of the authors. Based on the available body of research, the challenges defined aimed to cover the different interoperability layers of the European Interoperability Framework [16]. Furthermore, a fictional case study incorporating the different challenges was created. This allowed the challenges to be pinpointed to a practical example.

Defining good practices for achieving interoperability in a big data context
Based on the literature review and the experience of the authors, good practices were defined and applied to the fictional case study. These good practices were linked to specific challenges and the expected outcome if applied.
Annex II. GLOSSARY

Anonymisation | The process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous.\textsuperscript{101}

Data controller | The natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data [34], article 4

Data lake | A data lake is a collection of storage instances of various data assets additional to the originating data sources. These assets are stored in a near-exact, or even exact, copy of the source format. The purpose of a data lake is to present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).\textsuperscript{102}

Data marketplace | A data marketplace is a specific venue created for the buying and selling of data. This idea relies in large part on the rapid advance of technology, which has results in a data-rich environment where enormous amounts of data are routinely collected by many different parties\textsuperscript{103}.

Data portability | Data portability is a concept to protect users from having their data stored in "silos" or "walled gardens" that are incompatible with one another, i.e. closed platforms, thus subjecting them to vendor lock-in. Data portability requires common technical standards to facilitate the transfer from one data controller to another, thus promoting interoperability.\textsuperscript{104}

Data portal | A list of datasets with pointers to how those datasets can be accessed\textsuperscript{105}.

Data processor | A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller [34], article 4.

Data warehouse | A data warehouse is a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs.\textsuperscript{106}

\textsuperscript{101} https://en.wikipedia.org/wiki/Data_anonymization  
\textsuperscript{102} https://www.gartner.com/it-glossary/data-lake  
\textsuperscript{103} https://www.techopedia.com/definition/30322/data-marketplace  
\textsuperscript{104} https://en.wikipedia.org/wiki/Data_portability  
\textsuperscript{105} https://blog.ldodds.com/2015/10/13/what-is-a-data-portal/  
\textsuperscript{106} https://www.gartner.com/it-glossary/data-warehouse
<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OWL</td>
<td>The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit.¹⁰⁷</td>
</tr>
<tr>
<td>Pseudonymisation</td>
<td>Pseudonymisation is a procedure by which the most identifying fields within a data record are replaced by one or more artificial identifiers, or pseudonyms.¹⁰⁸</td>
</tr>
<tr>
<td>RDF</td>
<td>The Resource Description Framework is a W3C standard for the interchange of data on the Web, more information: <a href="https://www.w3.org/RDF/">https://www.w3.org/RDF/</a></td>
</tr>
<tr>
<td>SEMIC</td>
<td>Semantic Interoperability Community, Action 2016.07 of the ISA² Programme.</td>
</tr>
<tr>
<td>SOA</td>
<td>A service-oriented architecture (SOA) is a design paradigm where functionality is provided through reusable service components over a network.</td>
</tr>
</tbody>
</table>

¹⁰⁷ [Source: https://www.w3.org/OWL/](https://www.w3.org/OWL/)
¹⁰⁸ [https://en.wikipedia.org/wiki/Pseudonymization](https://en.wikipedia.org/wiki/Pseudonymization)
Annex III. BIBLIOGRAPHY


[51] ISA Programme of the EU, D03.05 Report on the real-life implementation of ISA Action 1.1 specifications, Brussels: European Commission, 2016.


