Open Statistical Data Interoperability Framework

www.cef-interstat.eu

D1.2 - Report on the overall technical scenarios in Europe in relation to the provisioning and reuse of (Linked) Open Statistical Data

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Project full title
INTERSTAT - Open Statistical Data Interoperability Framework

Grant Agreement No.
INEA/CEF/ICT/A2019/2063524

Project Document Number
Deliverable 1.2 (Activity 1)

Project Document Delivery Date
12.04.2021 (v2.0)

Deliverable Type and Security
Report – Public
This document is licensed under a Creative Commons Attribution 4.0 International License

Authors
ENG: Francesco Arigliano, ISTAT: Carlo Vaccari, Paolo Francescangeli

Contributors
INSEE: Franck Cotton

Reviewers
ENG: Martino Maggio
Stéphane Dufour and Davide Taibi (External Expert Advisory Board)

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Table of Contents

List of figures

List of tables

Executive Summary

The scope of this document is to provide an overview about the technical scenarios in Europe in the domain of the (Linked) Open Statistical Data. In particular the document includes a brief survey and analysis of the main technological initiatives but also standard and technologies with the final aim of identifying the possible technical barriers for the provisioning and access to LOSD.

The chapter 1 will cover some relevant initiatives related to the “standard” open data provisioning in Europe: after a brief recap about the technical maturity of the European countries in the context of open data, it is presented a set of key standard and technologies related to open data provisioning publication and interoperability.

Chapter 2 has been focused more on the specific field of Linked Open Statistical Data, presenting, initiatives and projects related to the LODS dissemination and interoperability, but also standards, tool and methodology for LOSD governance and modelling.

In the last chapter are reported some final considerations, resuming the results of the analysis performed, about the most relevant technological barriers that can prevent a large and simple provisioning and usage of the linked open statistical data: for each of these topics, that represent the starting point for the identification of the requirements for the INTERSTAT framework, it is also mentioned how INTERSTAT project is going to find specific technical solutions to face them.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

1 Open Data technical background in EU

1.1 Overall scenarios of open data publication and usage in Europe

This section addresses the analysis of different aspects related to the open data adoption and provisioning maturity starting from the results of the Open Data Maturity Assessment [1] that is the European Data Portal's annual benchmark study on the development in the field of open data in the European Union's Member States (EU27) and the European Free Trade Association (EFTA) countries Liechtenstein, Norway, and Switzerland. The assessment evaluates the degree of maturity in relation to four dimensions: *policy, impact, portal, and quality*. We will give an overview of the report approach, focusing in particular on the technical implications it contains.

Europe is achieving the goals set at European level about open data and making it available for all the citizens. European countries present a strong increase in their maturity levels (Figure 1). Scores have increased in all dimensions since last year. Convergence of countries at the higher end of the performance levels is also evident (Figure 2). 2020 has also led to a renewed emphasis on the importance of systematically collecting and making data available to the public due to the COVID-19 pandemic that underlined the crucial importance of data availability. The need to face the emergency has led many countries to publish related data and implement applications and dashboards to produce data easier to understand.

In the last years the most advanced countries, in terms of open data provisioning shifted their attention from the amount of data available to ensuring its quality as well. Furthermore, quality is not viewed as an independent aspect, but as an enabling factor for interoperability: the ability to collaborate within countries and across borders, facilitating the exchange of data between information systems.

Another key factor taken in consideration in the report is the impact creation: producing a positive impact on the society and the economy through the publication of open data has always been the ultimate goal of the data initiatives in Europe. Estimating impact is a hard task and there is still no agreement of the best way to do it. Several European countries are making successful efforts to understand and capture the extent to which open data is reused and how value is created by interacting with communities of re-users.

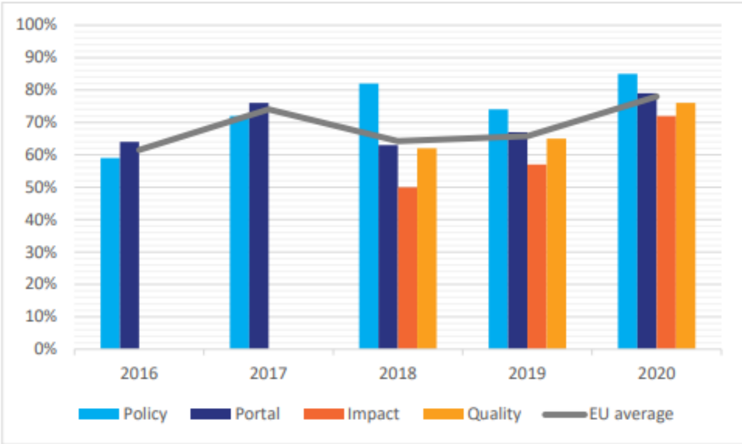


Figure 1: The Open Data Maturity scores of the European countries. [1]

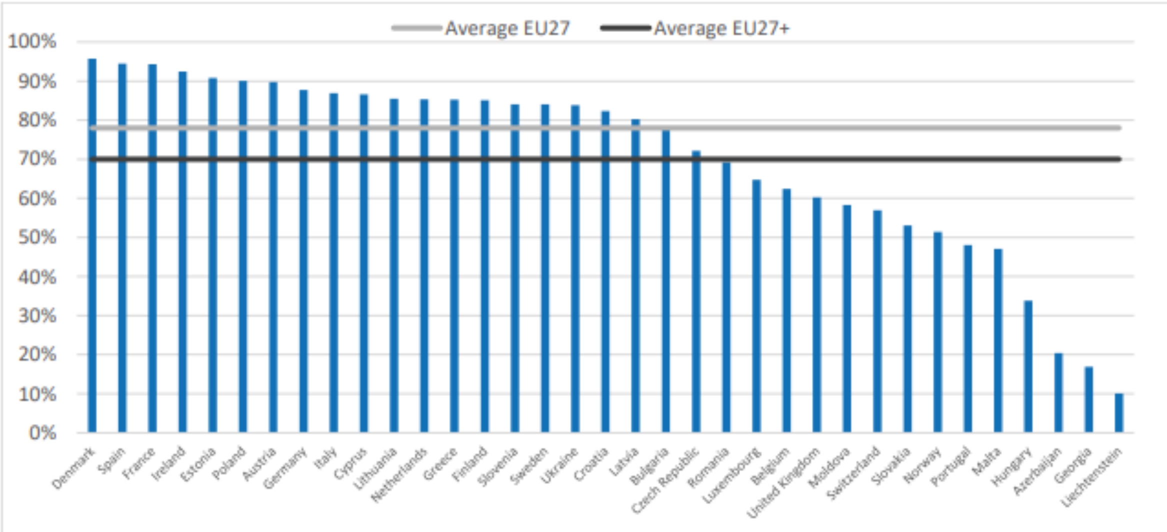


Figure 2: The overall open data maturity scores of the 2020 assessment

1.1.1 Dimensions of the assessment

As already mentioned, the four dimensions considered in the assessment are: policy, impact, portal, and quality. Each dimension has its own set of metrics and indicators (Figure 3).

Open Data Policy analyse the presence of specific policies and strategies to promote open data at the national level. The dimension also checks the existence of governance structures that allow the participation of private and third sector actors, as well as implementation measures that allow open data initiatives at the national, regional and local levels. In addition, the dimension takes into account training programs that improve the data capabilities of public officials who work with data and collection mechanisms that facilitate the discovery of all open data available in the country.

Open Data Impact examines the activities executed to control and evaluate open data reuse and the resulting impact. The impact dimension focuses on four areas of sectorial impact: political, social, environmental and economic. Within these areas, the questionnaire examines to what extent is tracked to document the reuse of published open data in these fields, to what extent applications, products and services have been developed to address challenges in these fields.

Open Data Portal dimension looks at advanced portal features that allow experienced and novice users to access open datasets through the national portal and capabilities that improve interaction between publishers and re-users (through community tools). In addition, the dimension assesses the extent to which portal administrators use web analysis tools to better understand the needs and behaviour of their users and update the functionality of the portals based on the information obtained from these results.

Open Data Quality focuses on the actions undertaken by portal administrators to ensure systematic collection of metadata from sources through the country, as well as evaluation of available metadata and, if possible, actual data, compliance monitoring of the DCAT-AP [2] metadata standards, as well as the quality of the distribution of published data. The fourth dimension empowers portal administrators and policy makers to enable good quality open data publishing at all levels, using high quality, machine-readable, open data formats suitable for a connected data approach. Data quality is a very relevant dimension regarding technical impacts.

Dimension	Metrics
Open Data Policy	Policy framework
	Governance of open data
	Open data implementation
Open Data Impact	Strategic awareness
	Political impact
	Social impact
	Environmental impact
	Economic impact
Open Data Portal	Portal features
	Portal usage
	Data provision
	Portal sustainability
Open Data Quality	Currency
	Monitoring and measures
	DCAT-AP compliance
	Deployment quality and linked data

Figure 3: Open Data Maturity dimensions and dimension-specific metrics [1]

1.1.2 Technical aspects related to metrics

In this section are highlighted some technical-related aspects to be taken in consideration in the degree of open data maturity:

Access to real-time and dynamic data

Real-Time data is information whose nature makes it most useful when, without delay, it is intended to be reused after it is collected. It has become more prevalent with the popularity of the Internet of Things, sensors in smart devices, and social media analytics. An example would be a weather system, which automatically retrieves real-time data from weather stations to continually improve and update its forecasts, or a public transport application that not only informs passengers of schedules, but also of location as well as the estimated time of arrival. The application programming interface (API) is the most widely used technology for real-time data delivery and a key topic of the new open data policy. A particular emphasis on obtaining real-time data is also evident in the policies of many countries.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Harvesting data from regional and local data sources

All national portals in Member States, where applicable, collected data from local and regional portals. For instance, in countries like Cyprus and Malta, where there are no regional portals, all data is published directly on the national open data portals. 67% of Member States report that most of all datasets are collected from local and regional sources. A shared approach based also on the creation of federated meta-catalogue, at national and regional level is therefore essential to ensure quality and updated data.

Up-to-date metadata

78% of Member States have a pre-defined approach in place to ensure that the metadata is kept up to date. Up-to-date metadata on the national portals is critical for users to obtain the correct information about the data that it describes. The ability to keep the metadata up to date can be reached in different ways, but a key factor is to ensure that metadata is obtained from its source automatically.

DCAT-AP compliance

DCAT is a W3C standard designed to facilitate interoperability between data catalogues published on the web. DCAT-AP (DCAT Application Profile) is an extension of DCAT that was developed by the European Commission to improve interoperability and facilitate the availability and reuse of open data in European catalogues. Compliance with DCAT-AP is increasingly recognized among Member States. The DCAT-AP Compliance Indicator examines the extent to which metadata conforms to the DCAT-AP standard for describing datasets and how much effort has been spent to help data publishers to ensure DCAT-AP compliance. 89% of Member States provide data suppliers with documentation on DCAT-AP.

Structured and Machine-readable format

51% of Member States offer more than 90% of the published data available in a structured format in addition to having an open licence. 44% of Member States offer more than 90% of the published data available in a machine-readable format in addition to having a structured format and an open licence.

1.2 European Data portal and EU Open Data Portal

This section describes the two main portals at European level for the provisioning of Open Data: European Data Portal (EDP) [3] and EU Open Data Portal (ODP) [4]

The difference between the European Data Portal (EDP) and EU Open Data Portal (ODP) is that ODP is the Open Data portal of the European Union containing datasets that are collected and published by the European Institutions. EDP is a European portal that harvests metadata from public sector portals throughout European countries. In addition, EDP also harvests metadata from ODP. The European Commission is currently exploring how to bring those two portals closer together.

1.2.1 European Data Portal

EDP harvests metadata from Public Sector Information available on public data portals across European countries. Portals can be national, regional, local or domain specific. EDP connect portals in the EU Member States, the European Economic Area (EEA), the European Free Trade Area (EFTA), the

EU accession countries, and countries involved in the EU's neighbourhood policy. In addition, by harvesting the European Union Open Data Portal, also cover European institutions.
The strategic objective of the European Data Portal is to improve accessibility and increase the value of Open Data addressing the whole data value chain, from data publishing to data re-use.

Within the Portal there are sections dedicated to searching datasets. Categories have been established to structure the metadata harvested from the various countries following the DCAT Application Profile and Eurovoc Thesaurus. Providing Data section gives an insight into understanding Open Data from the perspective of a data provider. Instructions are provided for those who wish their data portal to be harvested by the European Data Portal. Using Data section detail how Open Data is being used, as well as the economic benefits of Open Data.
EDP provide also eLearning modules about Open Data as well as training guides and a knowledge base referencing publication around Open Data projects.
When data catalogues of origin contain datasets with errors the Portal checks these errors and reports them back to the owners of the catalogues improving the quality of the metadata and data. Metadata contains the information made available with the data set by the initial publisher. Different licences, or absence of licence may occur and re-uses are invited to check with the owners/publishers of the data what terms and conditions apply to the data reuse.

1.2.1.1 EDP Architecture overview

From a technical perspective, EDP is very complex and uses several state of the art components and open technologies. In Figure 4 the overall architecture is depicted.
EDP is accessible both through a machine-readable API and a human readable web site (GUI). The API enables to search, create, modify and delete metadata on the portal. The GUI is basically built on two components: CKAN [5] and DRUPAL [6]. CKAN manages and provides metadata. DRUPAL provides the Portal's Home Page with editorial and links to an Adapt Framework based training platform. In addition, it offers extended functionalities to registered users via EU Login (former ECAS) [7].

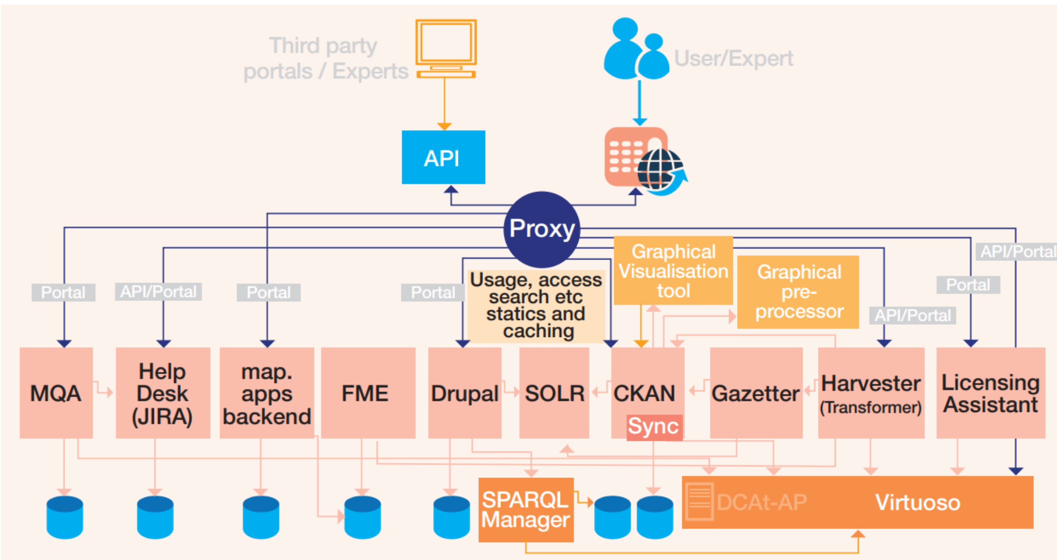


Figure 4: EDP Overall Architecture. Source [8]

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

The portal uses the SOLR [9] search engine in order to separately search for editorial content in DRUPAL and for datasets in the CKAN repository. The SPARQL Manager component allows to run SPARQL queries on the Virtuoso linked data repository.
In the following table you can find a summary of all components with a brief description

Component	Description
API	Machine-readable (SOAP / REST) API
GUI Access	Portal website graphical user interface
CKAN	Portal's central metadata (dataset) repository
DRUPAL	Portal's Home Page managing editorial content
ECAS (EU LOGIN)	European Commission Authentication System used for user registration and login in order to provide extended functionalities of the Portal
Adapt Framework	Online platform used for Portal Training Modules
Proxy	Routing of (HTTP(S)) user requests to corresponding components
SOLR Portal Search engine	used for searching portal editorial content
SOLR Dataset Search engine	used for searching and filtering datasets in the CKAN Search metadata repository
Licensing Assistant	Component to provide legal information on (re-)usage of specific datasets
SPARQL Manager	Query editor allowing to run SPARQL queries on linked data in the Virtuoso repository
Virtuoso	Linked data quadruple store that is synchronized with the CKAN repository
map.apps: Geo-spatial Data Visualisation	application to visualize geo-spatial data and information using Data Visualisation geo-maps. It comes with different tooling and thematic focus, a graphical configuration interface, supports responsive web-design, i18n internationalization files, client side implementation of the OSGI specification
Graphical Data Visualisation	Recline.js/D3.js JavaScript libraries to visualize (statistical) data in tables and graphical charts
Pre-processor	RESTFUL web API, running on a Node.js server which analyses and transforms XSL/XLSX files into CSV format in order to be used from the Visualisation tool
Harvester	Single entry point component for harvesting data from multiple data sources in different formats and from different APIs
Gazetteer	Component providing geo-spatial data and information
FME	Component used by the Gazetteer as a universal spatial ETL tool (Extract-Transform-Load) that supports the accessing, processing and outputting of, all spatial file / database formats and that is used for harvesting the sources for geographical names
smart.finder	Component used by the Gazetteer and simplifying searches for spatial data, services and documents. It enables fast and structured access to extensive, distributed and heterogeneous data stores
Helpdesk	Portal offers a user request/feedback form via the JIRA-API and generates JIRA tickets for follow-up by helpdesk
Metadata Quality Assistant (MQA)	Component to report on the quality of the harvested metadata and to alert helpdesk in case of issues
Monitoring	PIWIK component that provides Traffic Analytics of portal usage
Multilingual Support	Web pages + core editorial content + dataset descriptions available in all 24 official EU languages

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

MT@EC	Machine Translation Services of the European Commission used for translation of the metadata into all of the supported languages by the portal

Table 1 - EDP Components

The EDP provides specifications [10] about interfaces that data providers (e.g. national portals, public data portals in the Member States, portals from international organizations etc.) must have in place for being harvested by the European Data Portal. The interfaces supported by EDP are:

- **DCAT-AP / CKAN** for classical open data portal
- **INSPIRE** Catalogue Services (for geospatial datasets)
- **OpenSearch** (GEO/EOP) (for geospatial datasets)

Information about these standard/technologies will be provided in the next sections.

1.2.2 EU Open Data Portal

The EU ODP was created in 2012, following Decision 2011/833 of the European Commission on the reuse of Commission documents [11]. All EU institutions are invited to make their data publicly available whenever possible. This means that the data should be reused free of charge and without copyright restrictions in order to foster economic development and transparency within the EU institutions.

The data provided by EU ODP include:

- geographic, geopolitical and financial data
- statistics
- election results
- legal acts
- data on crime, health, the environment, transport and scientific research.

All these data are freely available. They can be reused in databases, reports or projects. A variety of digital formats are available from the EU institutions and other EU bodies.

1.2.2.1 ODP Architecture

From a technical perspective ODP has same pillars of EDP, the portal is built using open source solutions such as the Drupal content management system and CKAN as data catalogue software. It uses Virtuoso as an RDF triple store and has a SPARQL endpoint. ODP metadata catalogue applies international standards such as Dublin Core and the data catalogue vocabulary DCAT-AP.

1.2.3 data.europa.eu

Recently the two above mentioned initiatives (EU Open Data Portal and the European Data Portal) have been replaced by a brand-new portal "data.europa.eu" [1] that provides access to open data from international, EU, national, regional, local and geo data portals.

[1] https://data.europa.eu/

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

1.3 Main standards and technologies

In this section we are going to list and describe the main technologies and standards adopted in the open data provisioning in Europe. The scope is to highlight the main role of the technologies, the degree of adoption, main features and limitation.

1.3.1 Metadata standards

1.3.1.1 DCAT Application Profile

The DCAT Application Profile for data portals in Europe (DCAT-AP) [2] is a specification based on the Data Catalogue Vocabulary (DCAT) developed by W3C. This application profile is a specification for metadata to meet the specific needs of data portals in Europe. DCAT-AP adoption enable the achievement semantic interoperability with other applications on the basis of reuse of established controlled vocabularies (e.g. EuroVoc) and mapping to existing metadata vocabularies (e.g. Dublin Core, SDMX, INSPIRE metadata, etc.).

DCAT-AP also provides a common specification for describing public sector datasets in Europe to enable the exchange of datasets descriptions among data portals, allowing content aggregators, such as the European Data Portal, to aggregate such descriptions into a single access point in order to allow the consumers to easily find datasets.

DCAT-AP has the extension GeoDCAT-AP for describing geospatial datasets, dataset series and services. Another extension, StatDCAT-AP, provides specifications and tools that enhance interoperability between descriptions of statistical datasets within the statistical domain and open data portals. Current versions are: DCAT-AP v2.01, GeoDCAT-AP v1.01, StatDCAT-AP v1.01.

DCAT-AP concepts and terminology

- An **Application Profile** is a specification that re-uses terms from one or more base standards, adding more specificity by identifying mandatory, recommended and optional elements to be used for a particular application, as well as recommendations for controlled vocabularies to be used.
- A **Dataset** is a collection of data, published or curated by a single source, and available for access or download in one or more formats.
- A **Data Portal** is a Web-based system that contains a data catalogue with descriptions of datasets and provides services enabling discovery and re-use of the datasets.
- A **Catalog** is a curated collection of metadata about resources.
- A **Distribution** is A specific representation of a dataset. A dataset might be available in multiple serializations that may differ in various ways, including natural language, media-type or format, schematic organization, temporal and spatial resolution, level of detail or profiles.

The Application Profile reuses terms from various existing specifications. Classes and properties specified have been taken from the following namespaces:

```
adms: http://www.w3.org/ns/adms#
dcat: http://www.w3.org/ns/dcat#
dcatap: http://data.europa.eu/r5r/
dct: http://purl.org/dc/terms/
```

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

```
foaf: http://xmlns.com/foaf/0.1/
locn: http://www.w3.org/ns/locn#
owl: http://www.w3.org/2002/07/owl#
odrl: http://www.w3.org/ns/odrl/2/
prov: http://www.w3.org/ns/prov
rdfs: http://www.w3.org/2000/01/rdf-schema#
schema: http://schema.org/
skos: http://www.w3.org/2004/02/skos/core#
spdx: http://spdx.org/rdf/terms#
xsd: http://www.w3.org/2001/XMLSchema#
vann: http://purl.org/vocab/vann/
voaf: http://purl.org/vocommons/voaf#
vcard: http://www.w3.org/2006/vcard/ns#
time: http://www.w3.org/2006/time#
```

In Figure 5 a UML diagram of all classes and properties included in the DCAT Application Profile.

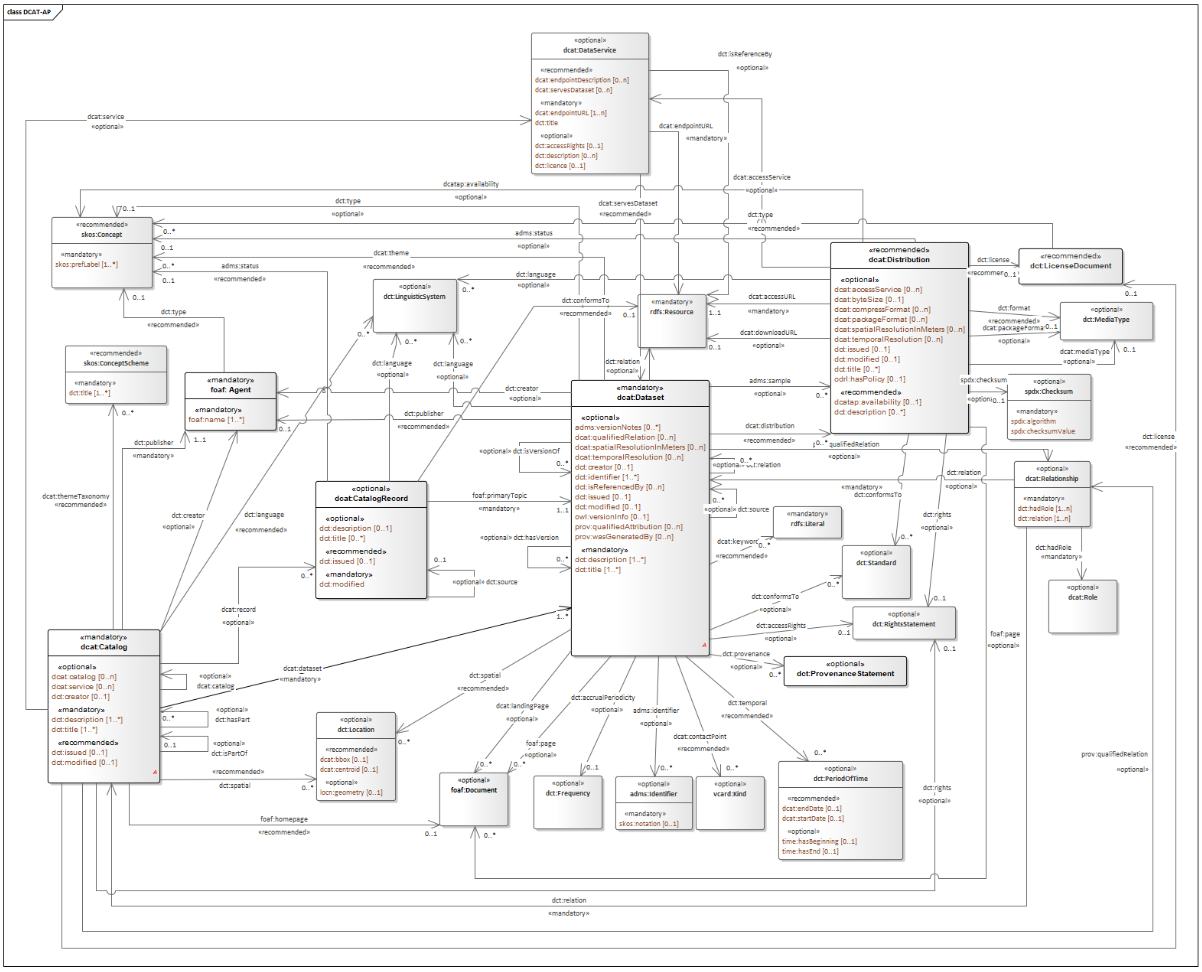


Figure 5: DCAT-AP UML Classes diagram. Source [2]

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

1.3.1.2 INSPIRE

The INSPIRE Directive aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. This European Spatial Data Infrastructure (SDI) has the objective of enabling the sharing of environmental spatial information among public sector organisations making simple the access to spatial information in Europe.
According to Article 5 of INSPIRE Directive 2007/2/EC [12], "Member States shall ensure that metadata are created for the spatial data sets and services corresponding to the themes listed in Annexes I, II and III (see Figure 6) , and that those metadata are kept up to date". The Directive addresses 34 spatial data themes needed for environmental applications. The directive does not include specific standards or technologies to be used, but states basic requirements. The technical guidelines are provided in the document INSPIRE Implementing Rules on Metadata [13].

Within this context, under the EU ISA Program, and taking into account the INSPIRE metadata profile and relevant ISO standards on which it is based, the community developed a geospatial profile of DCAT-AP the GeoDCAT-AP.

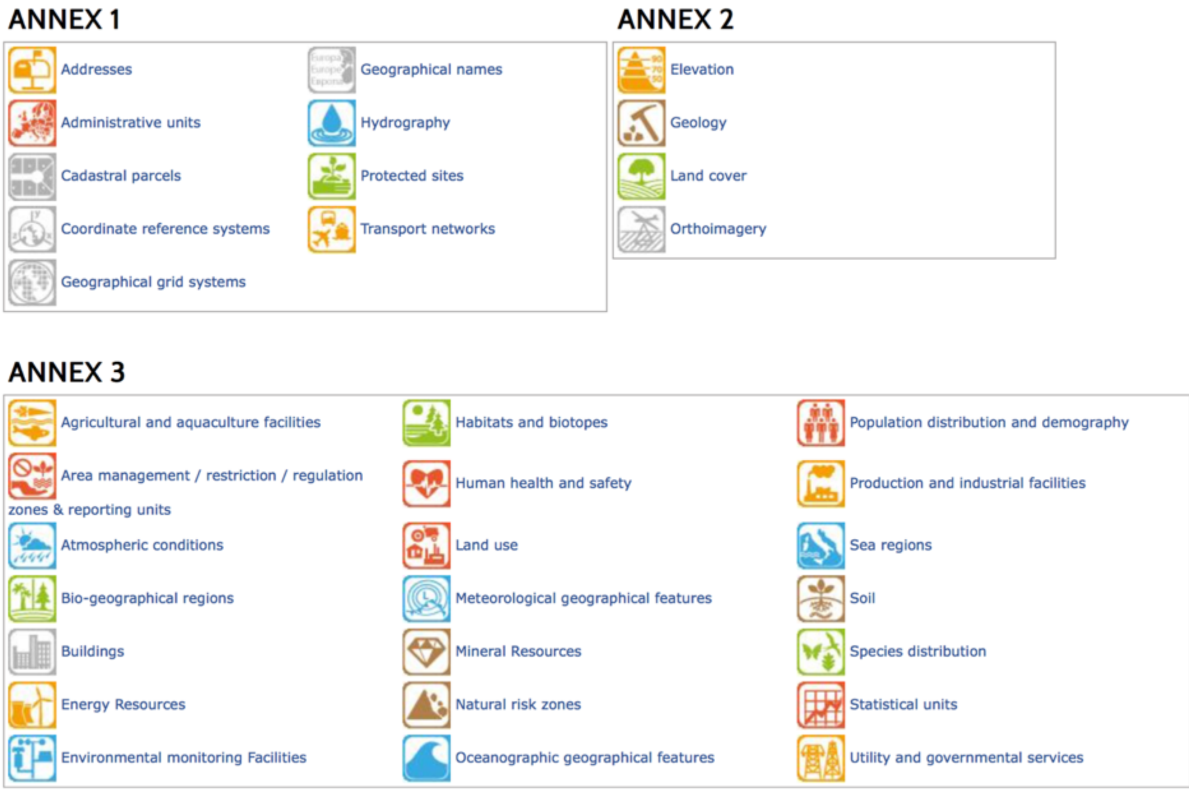


Figure 6 - INSPIRE spatial data themes

1.3.1.3 GeoDCAT-AP

GeoDCAT-AP is an extension of DCAT-AP created to describe geospatial datasets, dataset series and services. Its scope is to provide a RDF syntax binding for the union of metadata elements defined in

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

the core profile of ISO 19115:2003 and those defined in the framework of the INSPIRE Directive. GeoDCAT-AP is used to allow an easy provisioning and cross-domain search of spatial datasets in generic data portals through the exchange of descriptions of data sets among data portals.

In particular GeoDCAT-AP covers the datasets provided by the INSPIRE Geoportal, that aggregates metadata for over 100k datasets across Europe. In this way the metadata stored on this portal is structured according to the INSPIRE Metadata Technical Guidelines and can be also accessible through the generic open data portal including EU ODP and EDP. GeoDCAT-AP also simplifies the integration between Spatial Data Infrastructure (SDI) and data portals compliant with DCAT-AP metadata.

The GeoDCAT-AP is maintained following the Change and Release Management Policy for DCAT-AP in which are involved the main stakeholder that defined and supported this profile form the beginning (the stakeholders (ISA² Committee, ISA² Programme Management Team, the Operational Team and the GeoDCAT-AP Working Group).

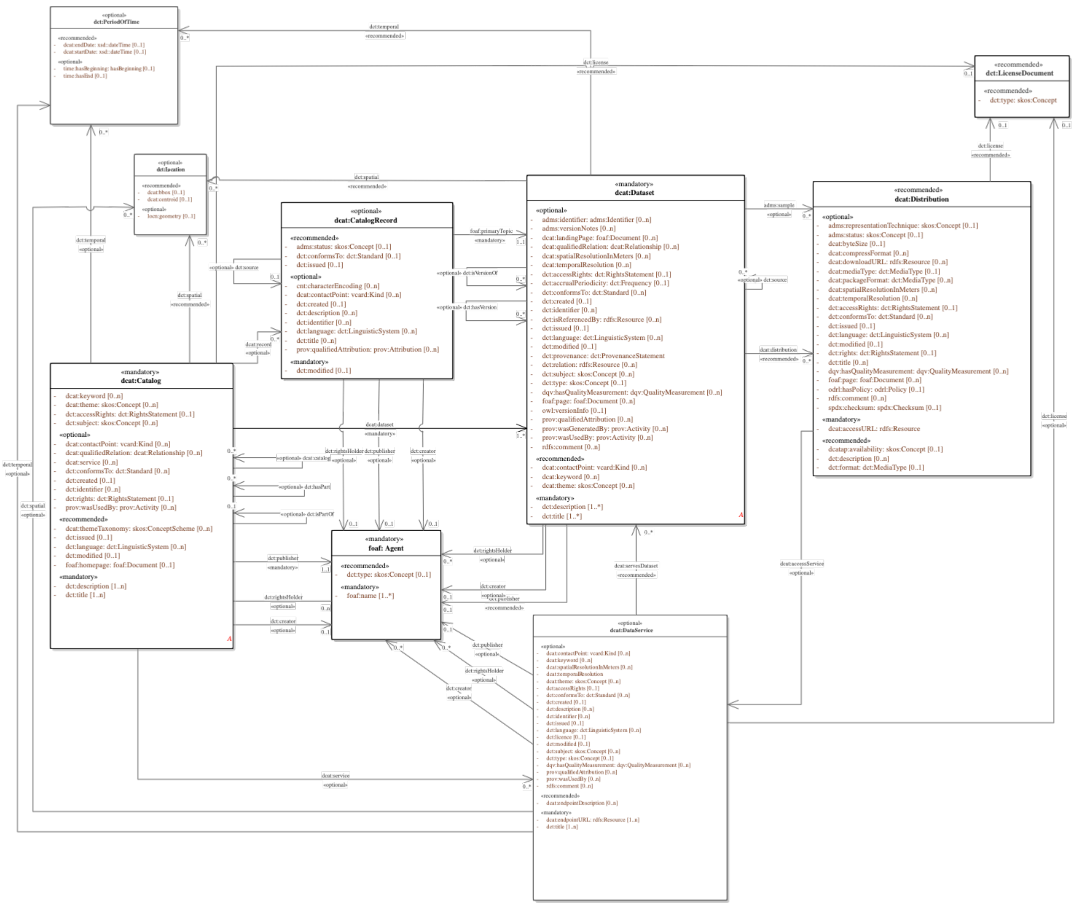


Figure 7 - GeoDCAT-AP UML Class Diagram

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

1.3.2 Open data portals

1.3.2.1 CKAN

The Comprehensive Knowledge Archive Network (CKAN) [5] is a web-based system for storing, cataloguing and distributing data, such as csv, spreadsheets or database content. The CKAN base code is maintained by the Open Knowledge Foundation. The system is used by various public administrations as part of their open data publishing strategy. Its strengths are represented by the high customization offered, the simple management of complex datasets and by the excellent ability to catalogue data that can be referenced via URI. It is an open source product, which through its intuitive and simple graphical interface accompanies Public Administrations in all phases related to the publication of Open Data, from physical storage to metadata, from logical organization to processing, from loading to archiving. Furthermore, once the data has been entered, it is indexed and classified in order to simplify the search, reuse and reprocessing operations, not only by humans, but also by search engines. In particular, the search function is managed through keywords and the search results, as well as providing the data corresponding to the keyword, include additional information such as the format, the license to use or the date of publication. In addition, it also offers geo-referencing functions, which allow, for example, to know the data of a digital map, or to take advantage of graphing functions, which allow data to be displayed in different formats. Finally, thanks to the integration module, CKAN can be extended with any other type of additional and customized function.

CKAN Architecture

CKAN architecture is modular and extensible through plugins. It exposes the repository data via APIs and enables efficient searches of both data and metadata (Figure 8)

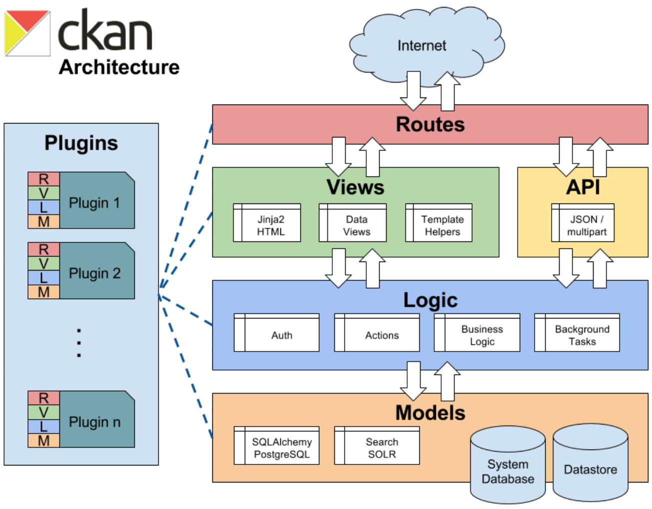


Figure 8: CKAN Logical Architecture. Source [5]

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Routes define the connection between CKAN URLs and views that process requests and provide responses. Views process requests by reading and updating data with action function and return a response by rendering templates. Plugins define new views by adding or updating routes.

1.3.2.2 Socrata

Socrata is a commercial Open Data Platform [14] with a Community edition [15] released under an open source license. Socrata hosts several catalogues for governments, non-profits, and NGOs around the world. SODA 2.0 is the name of the API defined in Socrata that could be used across any open data service. They provide an interface to update data, as well as a SODA Query Language (SoQL) for querying data.[2]

Socrata Architecture

The architecture (Figure 9) includes a layer that could sit above many types of data: SOCRATA introduces the concept of a Durable Store that is updated by users, as well as more performative Secondary Store that is optimized to faster queries and lookup operations. In this way read and write paths are split, so transactional semantics is provided to writers (data owners) through the Update APIs, and it is also provided to the user a fast query data access through the Read APIs. The mechanisms needed to provide synchronization between the transactional Durable Store and the Secondary store it is critical in order to always maintain data consistency.

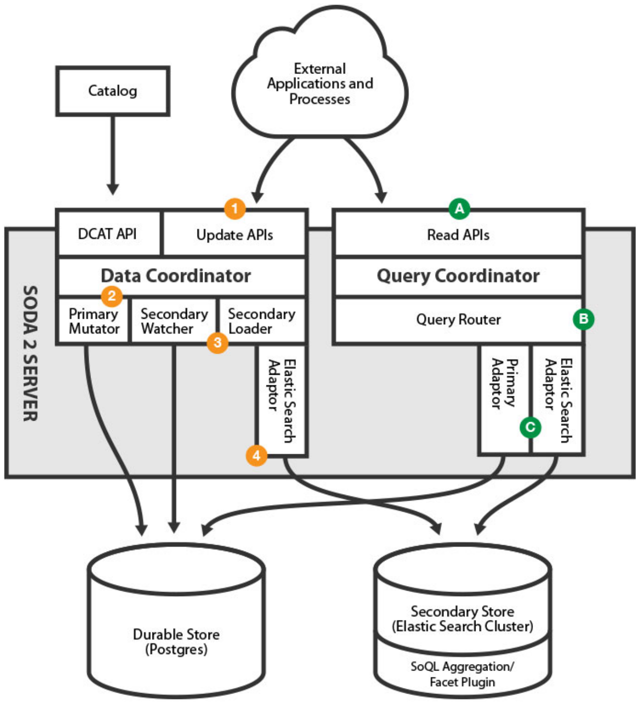


Figure 9: Socrata technical architecture. Source [15]

Write APIs

2 http://dev.socrata.com/

When an application or process starts sending insert, update, or delete operation to the Server, the request will be turned into a set of more granular operations called *mutations* and passed into the Data Coordinator which will then run then over the Durable Store in a transactional way. After the Data Coordinator writes the data, the Secondary Watcher trigger a synchronization check of the secondary store. The adaptor for the secondary store, in this case Elastic Search, then imports the data from the primary. The mechanism is built so that it will only get notifications for the records that have changed.

Read APIs

When an application sends a SoQL request to the server this will be parsed, the Query Coordinator will make a determination about which store the request should go to. The Query Coordinator will then hand off the query to the appropriate adaptor, which will execute the query against the correct store and return the appropriate C-JSON payload.

1.3.3 Linked Open Data technologies

Linked Data paradigm and technologies are becoming very relevant in the field of open data provisioning and access. When we talk about Linked Data (LD) we refer to a set of principles, technologies and practices that facilitate data integration outlined by Tim Berners Lee [16], that can be resumed in the following topics:

1. use URIs to name things and relationships among things;
2. use HTTP URIs so those names can be looked up (a technique called dereferencing);
3. return useful information upon lookup of URIs, using open standards such as RDF;
4. include links to other URIs, so more things and relationships can be discovered organically.

In [17] was performed a quality assessment of the adoption and quality of Linked Data in the context of the European Data Portal where Linked Data is used as a means to improve the re-usability and the interoperability of data assets within the European Union. From this assessment is clear that RDF is still a minority format. Also, the most of open data publishers that provided RDF versions of their datasets, take advantage of specific web portals capabilities to convert CSV or XML datasets into RDF. However, this conversion cannot be considered a real linked data provisioning because often datasets are provided without links among them and no well-known vocabularies are used. In the Figure 10 it is clearly shown the different presence of native file formats for all the dataset analysed in the report, highlighting the very low adoption of RDF.

In sections 2 of this document will be provided more information about Linked Data technologies in relation to open statistical data

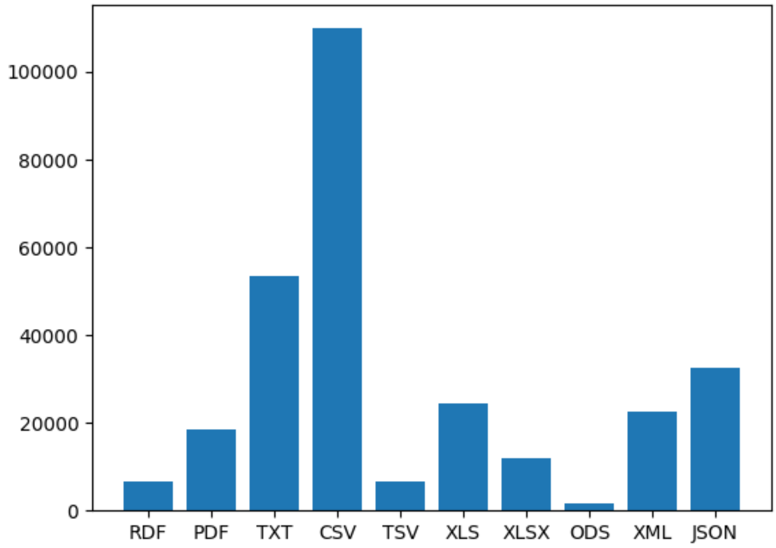


Figure 10: Number of datasets with at least one distribution on each format. Source [17]

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

1.4 Interoperability and harmonisation

This section is focused on the standard and technologies that can foster the achievement of data interoperability across different system and applicative domains. In particular, among the different initiatives already in place in Europe, following are presented two key assets that will be largely reused in INTERSTAT framework: the ETSI NGSI-LD specifications for the description and access to the context information and the CEF Context Broker, a building block of the CEF initiative that aims at managing context information being a concrete implementation of NGSI-LD.

1.4.1 ETSI CIM NGSI-LD

NGSI-LD is an evolution of the former OMA NGSI 9 and 10 interfaces [18] and FIWARE NGSIv2 [19] to incorporate the latest advances from Linked Data. The NGSI-LD Information Model prescribes the structure of context information that shall be supported by an NGSI-LD system specifying the data representation mechanisms to be used by the NGSI-LD API. Furthermore, the structure of the Context Information Management vocabularies to be used are defined. *"The Context Information Management API NGSI-LD allows users to provide, consume and subscribe to context information in multiple scenarios and involving multiple stakeholders. It enables close to real-time access to information coming from many different sources (not only IoT data sources)* [20]." So with the term "Context Information" we refer to a generic data representation of entities of the real word that can be (but not limited to) provided and accessed in real-time manner.
The NGSI-LD Information Model is defined at two levels: the foundation classes which correspond to the Core Meta-model and the Cross-Domain Ontology. On top of these two levels, domain-specific ontologies or vocabularies can be developed.

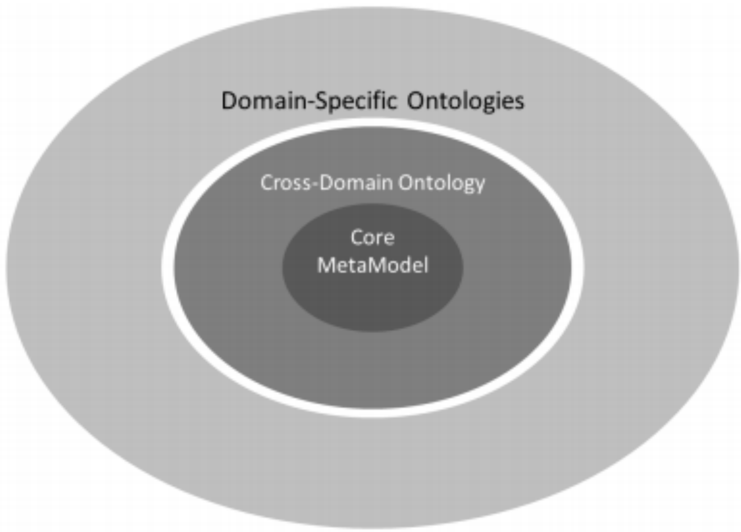


Figure 11:NGSI-LD Information Model Structure Source: [20]

1.4.1.1 NGSI-LD Meta model

Figure 12 provides a graphical representation of the NGSI-LD core Meta-Model in terms of classes and their relationships.

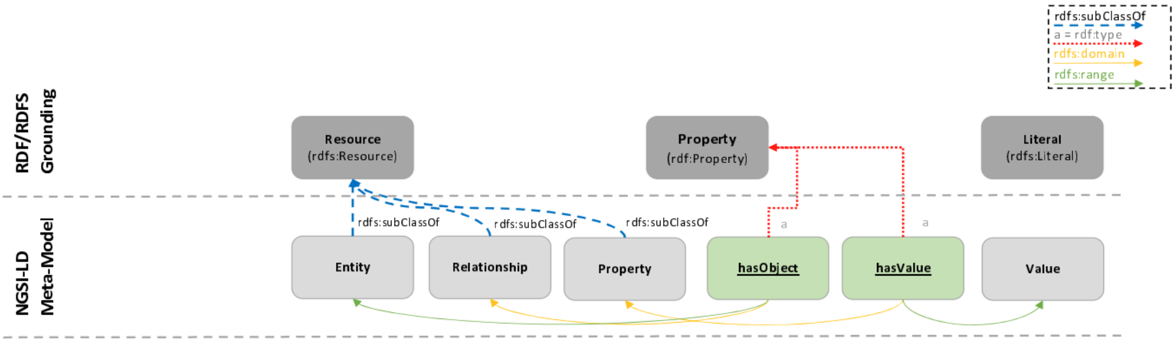


Figure 12: NGSI-LD Core Meta-Model. Source: [20]

Implementations shall support the NGSI-LD Meta-model as follows:

- An NGSI-LD Entity is a subclass of rdfs:Resource .
- An NGSI-LD Relationship is a subclass of rdfs:Resource .
- An NGSI-LD Property is a subclass of rdfs:Resource .
- An NGSI-LD Value shall be either an rdfs:Literal or a node object (in JSON-LD language) to represent complex data structures .
- An NGSI-LD Property shall have a value, stated through hasValue, which is of type rdf:Property .
- An NGSI-LD Relationship shall have an object stated through hasObject which is of type rdf:Property

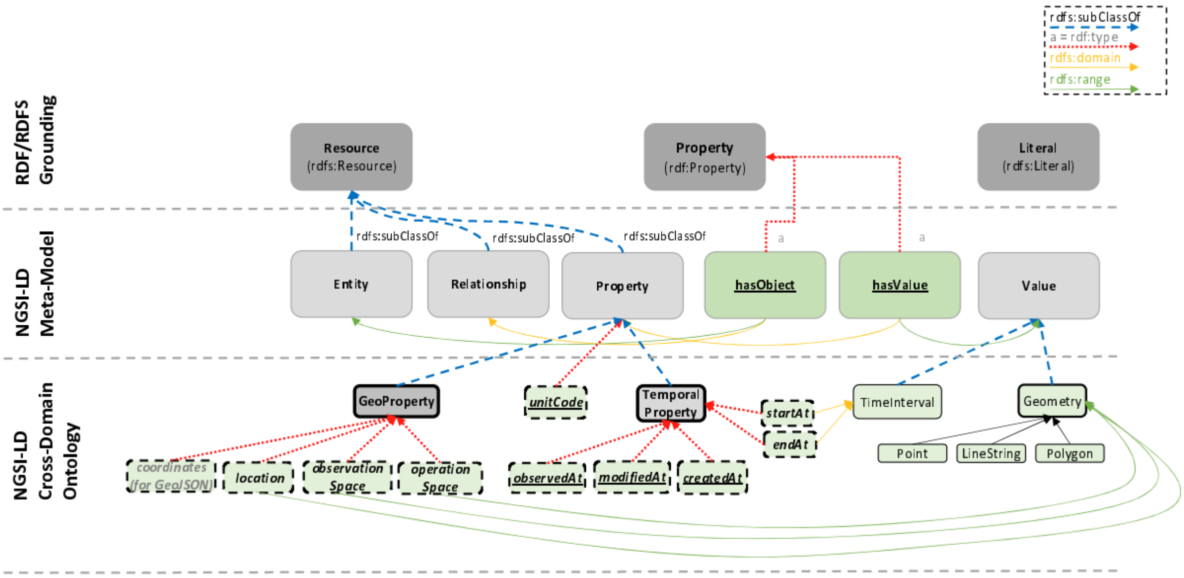


Figure 13: NGSI-LD Core Meta-Model plus the Cross-Domain Ontology. Source: [20]

Figure 13 describes the concepts introduced by the NGSI-LD Cross-Domain Ontology, which shall be supported by implementations as follows:

- Geo Properties are intended to convey geospatial information.

- Temporal Properties are non-reified Properties (represented only by its Value) that convey temporal information for capturing the time series evolution of other Properties.
- unitCode Property provides the units of measurement of a value.
- Geometry Values are a special type of NGSI-LD Value intended to convey geometries corresponding to geospatial properties.
- Time Values are a special type of NGSI-LD Value intended to convey time instants or intervals representations.

1.4.1.2 NGSI-LD API overview

In this paragraph we will give an overview of the operations provided by the API as reported in the specifications. All resource URIs shall have the following root:

```
{apiRoot}/{apiName}/{apiVersion}/
```

Figure 14 shows the main resource URIs and method specification in the NGSI-LD API, Figure 15 shows the complete tree of the resource URI structure.

Resource Name	Resource URI	HTTP Method	Meaning
Entity List	/entities/	POST	Entity creation
		GET	Query entities
Entity by id	/entities/{entityId}	GET	Entity retrieval by id
		DELETE	Entity deletion by id
Entity Attribute List	/entities/{entityId}/attrs/	POST	Append entity Attributes
		PATCH	Update entity Attributes
Attribute by id	/entities/{entityId}/attrs/{attrId}	PATCH	Attribute partial update
		DELETE	Attribute delete
Subscriptions List	/subscriptions/	POST	Subscription creation
		GET	Subscription list retrieval
Subscription by Id	/subscriptions/{subscriptionId}	GET	Subscription retrieval by id
		PATCH	Subscription update by id
		DELETE	Subscription deletion by id
Entity Types	/types/	GET	Available entity types
Entity Type	/types/{type}	GET	Details about available entity type
Attributes	/attributes/	GET	Available attributes
Attribute	/attributes/{attrId}	GET	Details about available attribute

Figure 14: Resources and HTTP methods defined on them. Source: [20]

in the following tables are shown respectively an example of an entity in JSON-LD format representing a Point of Interest and the call to the API to create the entity in an ETSI CIM compliant system. For the complete definition of API see the reference documentation in [20]

```
{
    "id": "urn:ngsi-ld:PointOfInterest:RZ:MainSquare ",
    "type": "PointOfInterest",
    "category": {
        "type": "Property",
        "value": ["113"]
```

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

```
    },
    "description": {
        "type": "Property",
        "value": "Beach of RZ"
    },
    "location": {
        "type": "GeoProperty",
        "value": {
            "type": "Point",
            "coordinates": [-8, 44]
        }
    },
    "@context": [
        "https://schema.lab.fiware.org/ld/context",
        "https://uri.etsi.org/ngsi-ld/v1/ngsi-ld-core-context.jsonld"
    ]
}
```

Table 2 - NGSI-LD Entity representation example

```
curl -X POST \
  http://broker.example.org:3000/ngsi-ld/v1/entities/ \
  -H 'Content-Type: application/ld+json' \
  -H 'Content-Length: 903' \
  -d '{
    "id": "urn:ngsi-ld:PointOfInterest:RZ:MainSquare ",
    "type": "PointOfInterest",
    "category": {
        "type": "Property",
        "value": ["113"]
    },
    "description": {
        "type": "Property",
        "value": "Beach of RZ"
    },
    "location": {
        "type": "GeoProperty",
        "value": {
            "type": "Point",
            "coordinates": [-8, 44]
        }
    },
    "@context": [
        "https://schema.lab.fiware.org/ld/context",
        "https://uri.etsi.org/ngsi-ld/v1/ngsi-ld-core-context.jsonld"
    ]
}
```

Table 3 -Entity creation example

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

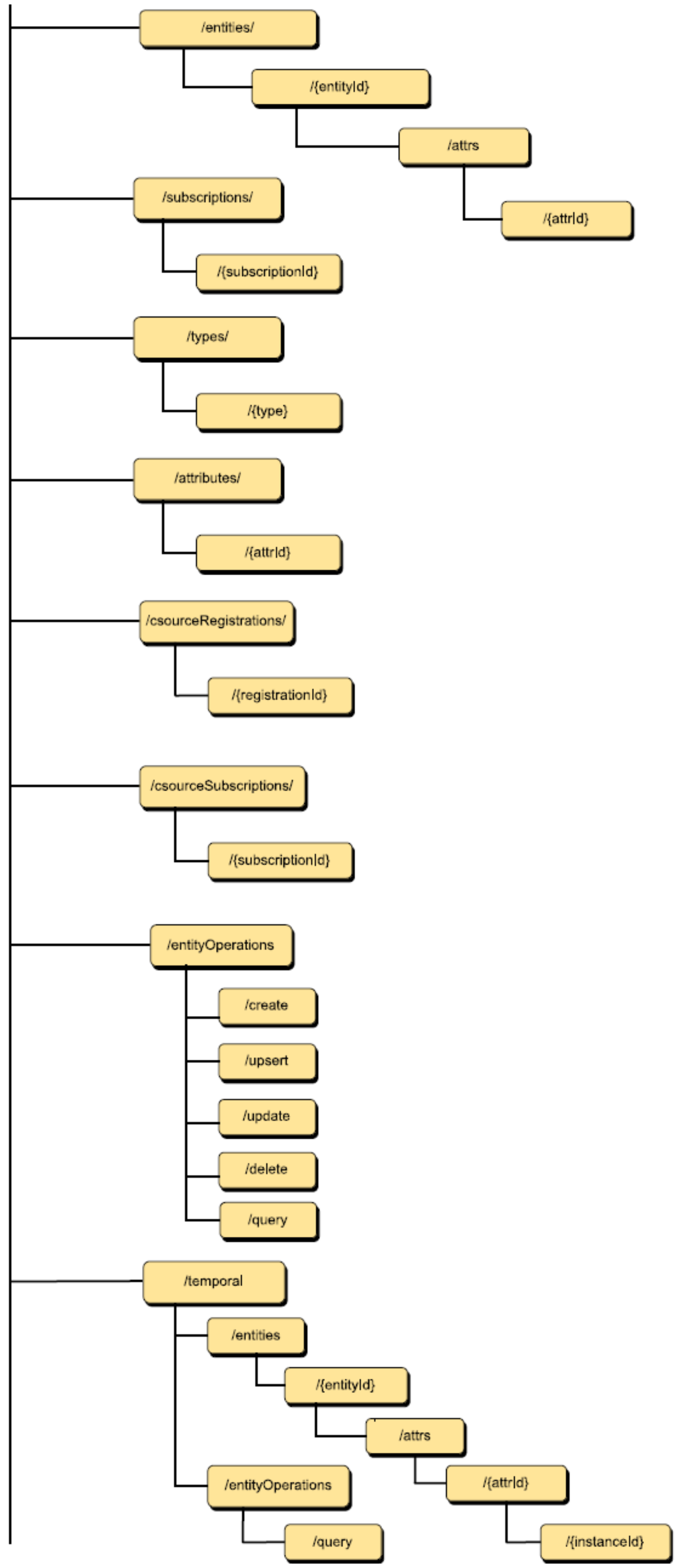


Figure 15: Resource URI structure of the NGSI-LD API. Source: [20]

1.4.2 CEF Context Broker

NGSIv2 specification and in particular the new NGSI-LD have been adopted by the European Commission in the Connecting Europe Facility [21] initiative as the specification for the CEF Context Broker, a component that aims at providing integration and interoperability across different systems and domains. The CEF Context Broker [21] (that is implemented by FIWARE Orion Context broker) supports the publication of context information (data that represents real world entities) by Context Producers, that is available to Context Consumers, which are interested in this data (Figure 16). Applications or even other platform components may play the role of Context Producers, Context Consumers or both. On the other hand, updates on context information are considered as events that can be handled by applications or platform components which subscribe to those events.

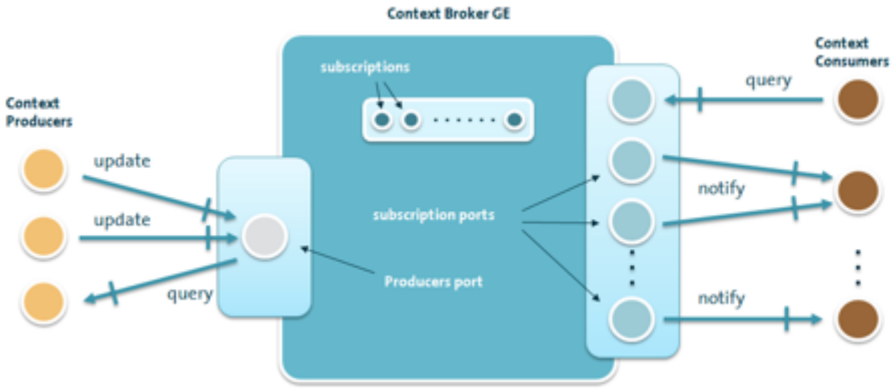


Figure 16: CEF Context Broker. Source: [21]

Context Producer with a minimal or very simple logic may continuously push the context information into the Context Broker. Context Consumers can pull the context information from the Context Broker (query mode), or the Context Broker can notify on updates Context Consumers interested in such updates (subscription mode). A fundamental principle supported by the Context Broker is the total decoupling between Context Producers and Context Consumers.
The use of API and standard meta-models combined with the use of shared vocabularies improves the interoperability of data and applications making the Context Broker a fundamental component to increase the maturity of European open data adoption.

1.4.3 Data exchange formats

In this section the most used data exchange formats in the context of open data production and provisioning are briefly presented. The support of these formats can be crucial in both interoperability achievement among different systems and user friendly provisioning of open data for the final users.

1.4.3.1 XML

XML (eXtensible Markup Language) [22]. It is the W3C standard for markup languages used for annotating documents. Furthermore, this language is used as a basis for the creation of further languages for the description of structured documents. It is based on a complex syntactic mechanism, through which the meaning of the elements contained in a document is defined.

1.4.3.2 PDF

PDF (Portable Document Format) [23]. It is an open format created by Adobe to represent documents composed of both text and images. The format can be read by any platform regardless of the operating system installed, the hardware or the program used for reading. It is also a format that is part of the ISO 19005-4 [24] standard and there are different versions based on their content: for example, the PDF / A standard is used for processing, the PDF / H standard is used for health documents, etc.

1.4.3.3 CSV

CSV (Comma Separated Values) [25]. It is an open format used to represent documents composed mainly of text, but in tabular form with no additional elements. It is particularly used to export data tables extracted from spread sheets or databases. In this format, the rows of the database table correspond to the lines of text which are in turn divided into fields. The fields separated by commas represent the columns of the database. Although CVS is not a standard it is considered to be a de-facto one.

1.4.3.4 Open Document

ODT (Open Document Text), ODS (Open Document Spreadsheet) and ODP (Open Document Presentation) [26]. These are open standard formats used to represent text documents, spreadsheet and presentation documents respectively and are based on XML. The standard they are part of is the OASIS Open Document Format for Office Applications. Their importance stems from the fact that they are considered the main open formats used in the most used open office suites such as OpenOffice.org and LibreOffice. Furthermore, this file formats are interoperable with other platforms, such as Microsoft Office, Google Drive and IBM Lotus.

1.4.3.5 RDF

RDF (Resource Description Framework) [27] was created by the W3C to represent and use structured metadata and to ensure interoperability between different resources that share information on the Web. Today, however, it is the model par excellence of Open Data published on the Web, it allows the simple representation of data semantics and the information encoded in RDF can be easily manipulated by agents and automatic procedures. RDF is also a framework that allows to represent everything in reality as a resource, modelled through statements of the form subject–predicate–object (known as triples), which can always be accessed on the Web because it is identified by a unique URI.

1.4.3.6 RDFS

RDFS (RDF Schema) [28] is the extension that deals with defining simple schemes to represent data and exposes the syntax to define the vocabularies for metadata. For example, resources can be represented by class instances and properties and constraints by subclasses, types and collections. In this way, the RDF triples can be connected through simple constructs, such as inheritance and extension relationships between classes, while common types of constraints can be those of domain and range.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

1.4.3.7 SHAPEFILE

SHAPEFILE [29] is the open standard used for the representation of data from geographic information systems (GIS) in vector mode. Usually, this data type is in .shp format, which is the file that contains the geometric shapes. Furthermore, this file is often accompanied by other files, such as the .dbf file, which contains the attributes of the geometric shapes, the .shx file, which contains the index of the geometric shapes and finally the .prj file, which contains the reference system settings.

1.4.3.8 OWL

OWL (Web Ontology Language) [30] is a further extension of RDF, the W3C standard, which takes care of enriching it with further formalisms dictated by formal semantics and also allows the use of automatic agents for the so-called descriptive logics. In this way the created schemes are even more complex, detailed and complete. In particular, OWL is considered the set of languages for the representation of information through ontologies on the Web.

1.4.3.9 JSON

JSON (JavaScript Object Notation) [31] is an open format that serves for the simple representation of data in JavaScript language, so that they are also readable by humans. It is widely used because it is a very flexible and compact language. JSON format was specified in the early 2000s, and it was first standardized in 2013, as ECMA-404, JSON was also standardized as ISO/IEC 21778:2017.

1.4.3.10 GeoJSON

GeoJSON [32] [33] is an open standard format designed for representing simple geographical features, along with their non-spatial attributes. It is based on the JSON format. The features include points, line strings, polygons, and multi-part collections of these types.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

2 Overview of European technical (L)OSD scenarios

In this section, similarly to what we introduced in the previous chapter, we give elements, from a technological point of view, about the situation of statistical open data and linked open data provisioning in Europe. A source of information on this subject is given by the outcomes of the DIGICOM project [34]. DIGICOM that is one of the implementation projects of the Vision 2020 [35] strategic plan of the European Statistical System.

2.1 Dissemination of metadata and data by Eurostat

DIGICOM recommended that Eurostat feed the Portal daily with data set descriptions conforming to the DCAT standard, and to the StatDCAT-AP profile [2]. StatDCAT-AP is an extension of the DCAT-AP for describing datasets within the statistical domain. It basically adds information about the dimensions and attributes of the statistical dataset according to the SDMX (see section 2.3.4) information model, as well as quality metadata represented using the DQV vocabulary [36]. According to DIGICOM's conclusions, the transition to StatDCAT-AP should be realized in 2021. Eurostat is also active in the publication of structural metadata according to semantic web standards. For example, the NUTS classification (Nomenclature of territorial units for statistics), a hierarchical system for dividing up the economic territory of the EU, is available as RDF on the Eurostat web site. Eurostat has also collaborated with the Publications Office of the European Union in order to publish several classifications as SKOS or the International Standard Classification of Education. However, Eurostat's most prominent structural metadata are disseminated using more classical data formats via the RAMON metadata server [37]. RAMON provides classifications and code lists as HTML, CSV or XML. On the other hand, structural metadata used for SDMX dissemination, like DSDs (data structure definitions), code lists, concepts or category schemes, etc., are available from SDMX registries, in particular the SDMX Global Registry and the Euro SDMX Registry.
Regarding the dissemination of open data, Eurostat has not yet implemented linked data technologies, although third parties have worked on possible ways of doing so [38]. The current solutions are quite representative of what can be found in most statistical organizations. Data is available for bulk download in TSV (tab separated values) and SDMX formats and can also be accessed through a navigation tree and via thematic and alphabetical indexes. Data can also be viewed via dedicated online tools like the Data browser, which allows filtering of dimensions, and customized views and provides additional download formats like Excel, and the Data explorer, an interface designed for dynamic selection in multi-dimensional tables (see here and here for examples on producer price indexes).

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Eurostat data is provided under a license[3] which is complicated by various exceptions and not explicitly related to any standard data distribution standard license. As a consequence, this license is not machine-processable.
Attached to Eurostat's statistical data is another type of metadata: explanatory texts documenting the data and providing summary information useful for assessing data quality and the statistical production process in general. This type of metadata conforms to a dedicated set of standards (see [39]for detailed information). In addition to the possibilities listed above, Eurostat provides web services to retrieve data, and more specifically SDMX web services, in conformance with the SDMX Technical Specifications (*Figure 17*). Both SOAP and REST services are offered, with SOAP covering versions 2.0 and 2.1 of SDMX and REST limited to version 2.1 (note that SOAP should be deprecated in upcoming version 3.0 of SDMX). The SDMX web services page on Eurostat's web site gives guidelines and examples on how to create and submit SDMX queries and describes potential use cases.

Eurostat's web services are powered by the SDMX Reference Infrastructure, a set of tools that facilitate the publication of statistical data through SDMX web services as symbolized in the following figure. The SDMX RI is published as open source under an EUPL license.

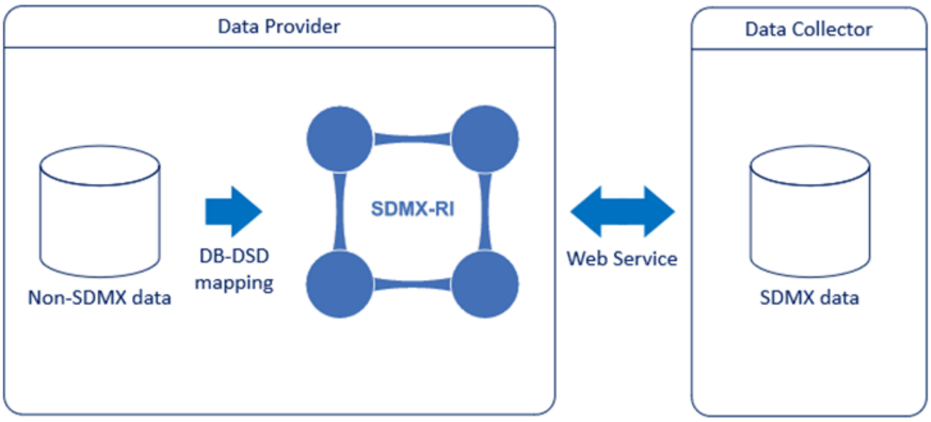


Figure 17: Eurostat SDMX web service

The Reference Infrastructure is also used in the Census Hub, a data dissemination software specially created for statistics based on the national census exercises undertaken in EU Member States and EFTA countries during 2011. Though the user interface allows to produce customised tabulations and thus reminds of the data explorer pattern, the back-office is actually a quite complicated distributed system based on a collection of SDMX RI instances installed by each data provider. This organization prefigures a long-vision conceived by Eurostat where a central Data Hub would distribute SDMX queries to national data providers and aggregate results for the user.

3 https://ec.europa.eu/eurostat/about/policies/copyright

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

2.2 Other European and International Organizations

Insee recently launched an extensive overhaul of its data dissemination process. On this occasion, it conducted a study of the service offer in matter of data dissemination in different statistical organizations, by comparison with what can be found at Eurostat's web site, which was detailed previously. The following table gives an overview of the results.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Organization	Country	URL	Web services / API	Data exploration	Data visualisation	Catalogue	Bulk download	LOD
Eurostat	International	https://ec.europa.eu/eurostat/en/data/database	X	X	X	X	X	
OECD	International	https://data.oecd.org/	X	X	X			
Istat	Italy	http://dati.istat.it/?lang=en http://datiopen.istat.it/	X	X	X			Geography 2011 Census
Statistics Canada	Canada	https://www150.statcan.gc.ca/n1/en/type/data	X	X			X	
CSO	Ireland	https://data.cso.ie/	X	X	X		X	2011 Census
Census Bureau	USA	https://data.census.gov/cedsci/	X				X	
Bureau of Economic Analysis	USA	https://www.bea.gov/API/signup/	X				X	
Bureau of Labor Statistics	USA	https://www.bls.gov/data/		X			X	
INE	Spain	http://www.ine.es/dyngs/INEbase/listaoperaciones.htm	X	X	X	X	X	
INEGI	Mexico	http://www.inegi.org.mx/sistemas/bie/	X			X	X	
Statistics Finland	Finlande	http://www.stat.fi/org/index_en.html http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/	X	X		X		
Destatis	Allemagne	https://www.destatis.de/EN/Home/_node.html	X		X	X		
Statistics Denmark	Denmark	https://www.dst.dk/en	X	X	X	X		
CBS	Netherlands	https://www.cbs.nl/en-gb https://opendata.cbs.nl/statline/#/CBS/en/	X	X	X		X	

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

2.3 Key Standards and technical initiatives

2.3.1 New European Interoperability Framework

An overview of the New European Interoperability Framework (EIF) [40] provides a glimpse of the actions undertaken to promote and support an Open data strategy at European level. In order to face interoperability challenges, a first version of the framework was adopted in 2010 and revised in 2017 [41], after a stakeholder consultation. The update also helped the alignment with related EU regulations (e.g. the revised Directive on the reuse of Public Sector Information, the INSPIRE [42] directive and the eIDAS Regulation [43]), and other initiatives (such as the EU eGovernment Action Plan 2016-2020 [44], the Single Digital Gateway [45]) fostering the development of a European Digital Single Market with common policies and networks of interoperable solutions. The ISA programme, was established[4] to promote and implement the EIF strategy, assessing the achieved results by key performance indicators. The main goal of EIF is to provide guidance at national and domain level, for the design and development of public services according to the following general requirements:

- Digital by default, namely provision through digital channels
- Cross-border by-default, that is access for all European citizens
- Open-by-default, to promote reuse and transparency.
- The implementation of the actions suggested by the EIF is based on three main conceptual components: the Conceptual model, the EIF principles and the Interoperability levels. Although these elements are independent one another, their combination covers the different interoperability aspects.

For each component, a set of recommendations [46] complements the guidance, while the Interoperability Action Plan reports the priorities to support the EIF implementation over a four-year period. The Interoperability Action Plan [47] aims at overcoming interoperability issues, focusing on five key areas related to:

- Interoperability governance
- Development of organisational interoperability solutions
- Stakeholders' involvement and awareness of interoperability benefits
- Development, management and promotion of key interoperability enablers
- Development, management and promotion of supporting interoperability tools.

Member States are required to align with the actions described in the Interoperability Action Plan, also adding additional initiatives, in order to achieve coherence between the national and the European strategies. In the EIF, the conceptual model is based on the principle of 'interoperability by design' that aims at developing interoperability functions in the design of public services. Particularly, according to the model shown in Figure 18, the reuse of available information and services, both internal and external to Public Administrations, is one of the main drivers to enhance interoperability.

[4] Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015

Therefore, Public Administrations are recommended to identify and adopt a shared approach fostering the interconnection of public service components and the adoption of interoperable formats to allow the retrieval of services and information provided by internal and external sources.

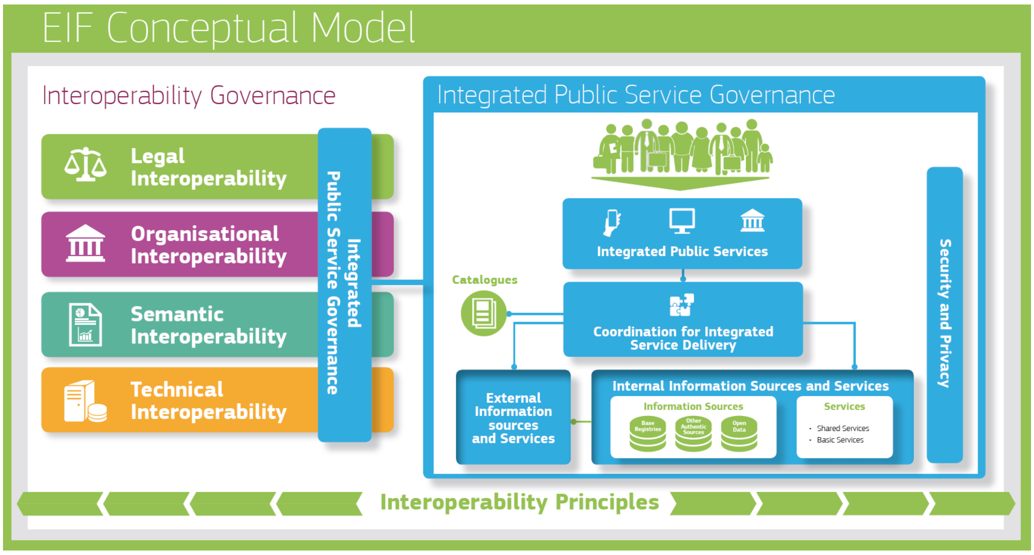


Figure 18: EIF conceptual model[5]

The model is composed by:

- An 'Integrated Service Delivery', built on a 'Coordination function' to increase user-friendliness
- A 'no wrong door' service delivery policy, to deliver the service through several options, included the digital channels (digital-by-default)
- Reuse of data and services to increase efficiency, service quality and interoperability
- Catalogues for describing available services and other resources and facilitating their access and reuse
- Integrated public service governance
- Security and privacy.

The basic components of the model are:

- **Coordination function**, to identify needs and suitable sources and services to integrate for the provision of a European public service. This function can be automated or manual.
- **Internal information sources and services**. These sources and services, available in the internal environment, as well as the external ones can be combined to integrate public services, in order to implement service-oriented building blocks. This approach promotes access and reuse of implemented public services.
- **Base registries**, the foundation of European public service delivery. These components are trusted sources of primary information, providing reliable descriptions of data items.
- **Open data**. Based on a common legal framework that fosters the reuse of Public sector data, this component supports the release of machine-readable data, reducing the barriers to publish and access open data.

[5] Source: https://ec.europa.eu/isa2/eif_en

- **Catalogues**, to facilitate the search of reusable assets, such as services or data models. Open data portals, metadata catalogues and catalogues of standards are a few examples of catalogues. The European Interoperability Cartography (EIC) is a particular type of catalogue, storing solutions compliant with re-usability and interoperability criteria described in the European Interoperability Reference Architecture (EIRA).
- **External information sources and services**, related to public services delivery by third parties, or reuse available information collected by external providers (e.g. data collected through the Internet of Things).
- **Security and privacy**, to ensure the adoption of privacy-by-design and security-by-design approach and the compliance with data protection regulation.

The revised version of the conceptual model includes also the interoperability model (Figure 18, left side), to highlight the relevance of the interoperability by design approach. The main elements of the interoperability model are:

- Four interoperability layers
- A cross-cutting component, 'Integrated public service governance' to promote service integration and reuse, seamless execution and implementation of new services and 'building blocks'
- A background layer, 'Interoperability governance', concerning the strategic decisions about interoperability frameworks, and other organisational aspects (e.g. activities, policies, agreements, roles and responsibilities) to support and monitor national and EU interoperability.

The four interoperability layers address the following dimensions, respectively

- Legal Interoperability, to cope with legal frameworks differences, preventing cross-border service delivery. Member States may reach clear agreements, to establish how to overcome identified legal divergences. This layer relates also to the compliance between EIF principles and the public services requirements prescribed by law (organisational, semantic, and technical)
- Semantic Interoperability, to preserve format and meaning of information, during data exchanges between parties. This layer concerns to semantic and syntactic dimensions and standardization
- Organisational Interoperability, enabling Public authorities to learn more about each other's business processes, tasks and expectations, to set common goals and gain mutual advantages. The fulfilment of users' needs and the trust between the parties sharing information are other aspects covered by this layer
- Technical Interoperability refers to software solutions and infrastructures developed to connect Public services. This layer embraces all technical aspects enabling interoperability, such as data integration, or data exchange, including the protocols for secure communication.

2.3.2 LOSD Governance and Life Cycle

Legacy idea of data publishing and consume in a rigid one-way fashion from provider to final user doesn't suit modern use of information where data are being produced and consumed continuously in a feedback cycle where data are transformed and enriched. There are several steps in modern day data life cycle where data are produced, revised, validated, used and reproduced back again in an incremental loop model. But data are produced with a wide range of data formats and metadata aren't always clear and easy to access. Proprietary data formats also limit free data access and exclude a wide range of potential interested consumers. Consumers can also fail to find relevant information because of lack of interoperability and because data are often hidden beyond walls of corporate systems.

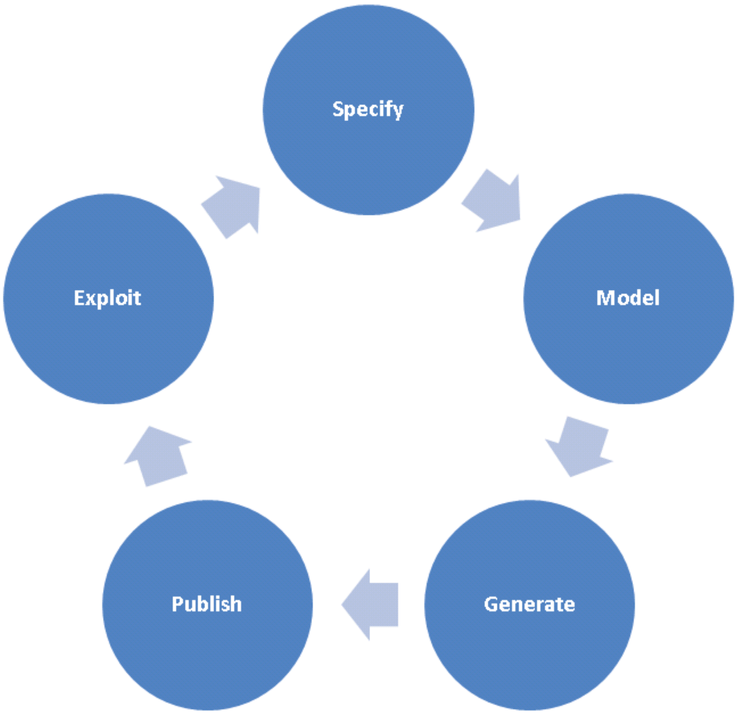


Figure 19: LOSD Life Cycle - https://www.w3.org/2011/gld/wiki/GLD_Life_cycle

As a key pillar of the European Data Strategy, a new way of data governance will increase trust in data sharing, strengthen mechanisms to increase data availability and overcome technical obstacles to the reuse of data [48]. The initiative aims to make more data available and facilitate data sharing across sectors and Member States in order to leverage the potential of data for the benefit of European citizens and businesses. Simple rules were addressed to sort out about optimal data interoperability, standardized data production and share points on the web for easy data access [49]. These rules are:

1. Model the Data
2. Use standard namespace and resource locators
3. Adopt high level standards and widen reuse of ontologies, vocabularies and taxonomies
4. Publish both human and machine readable descriptions.
5. Convert data to open format
6. Disseminate data with an appropriate license
7. Host the Linked Data Set on public and well-advertised sites for public domain availability.

The LOD standards support a human-centric way of exploring data sets by basing itself on triples (subject, object, predicate) known from natural languages. Publishing as linked-data offers a flexible, non-proprietary, machine programmable means for providing web addressable, advanced dynamic

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

querying and visualization capabilities allowing publishers and third parties to annotate and link to data and metadata. Reference to EU statistics data sets compiled by Eurostat are published on EU open data portal using stat-DCAT RDF specifications, data exposed as LOD can be flexibly combined across datasets. The statistical data becomes an integral part of the broader web of linked data. Following the effort for making LOD attractive for the public, High Value Dataset, aka Specific Area Domains have been proposed and Common Shared Workspaces have been created. As a matter of fact, first experiences show that full LOD standards take up on the web of data is relatively modest, with Public administrations and Research community being in the first row for data publications. Internet search engines supported reduced version of RDF: JSON-LD. Web use of LOD is shifting from simple data publication to data integration into distributed web applications. I.e. Google's view of the subject [50].

But general public hardly reap the benefits of LOD due to complexity of querying. Although LOD are simple at the conceptual level, generic users can't yet grasp all of the benefit LOD strategy brings, so enhancements in user experience by providers is needed. Data have been disseminated with the final user being human until now, so much effort was given to human readable interfaces. Dissemination shifted from documents to data sharing in open formats to achieve best availability. Linked open data bring machines in the mix, so data provided with this format seems less attractive to normal human users. On the other hand machine processability make so that data can be integrated into applications and web application seamlessly above the data presentation level.

Linked Open Data apparently makes it possible to link any kind of data, but there are pitfalls in this approach. Universal linkability can't be addressed efficiently when data aren't harmonized nor standardized. Having multiple dataset linked together without a clue about each dataset semantics can drive to meaningless results. Experiments in interlinking data provided by several NSI in latest ESSNET program demonstrated that linking is hard even when there is general agreement on technical details between publishers. That is to mean, a higher level of standardization and better decoupling from technical details is needed yet. So, even with the potential of data linking, the aspect of metadata standardization and data harmonization become a central topic to achieve wanted results

Early dataset design and early enforcement of semantic harmonization guarantees dataset are disseminated by several publishers, with matching metadata that guarantees best linkability and federation capabilities. On the other hand, Dataset are to be designed using well defined knowledge objects for unique addressing and better exploit. The goal is to define object of interest decoupling the data structure definition from the actual semantics associated to the area of interest.

I.e. The object of knowledge can be aggregated data in data marts, micro-data dataset, classification objects and the like. There are models to formally define and express object like GSIM or SDMX, and there's a suitable translation to Ontology structure model based on them. There is a catalogue for statistic use of object called statDCAT-AP an extension of DCAT-AP as already mentioned in the section 1 [2].

Secondary there is the need to model the object contents. Thematic area domain model data contents formally with emphasis on reuse and standardization. Dataset of strategic value need to be collected and made available for public consultation in a centralized way, using portals like High Value Dataset portal [3].

2.3.3 Standards and Shared Ontologies

The principal goal of Semantic Web is to give a semantic to resources using formal languages. Semantic interoperability is achieved this way at several levels: human to human, human to machine and machine to machine. The core goal is to let machines interpret data. In this context, as we see in web semantic layer of Tim Berners-Lee, ontologies play an important role because data and related metadata are managed together. An ontology is a formal representation, shared and explicit of a conceptualization – simplified view - of a domain of interest.
The term “formal” relates to the ontology being a logic language processable by computers. The term “shared” is defined by the consent of a plurality, the widest possible, of subjects. The term “explicit” avoids ambiguity and clarify all the assumptions made.

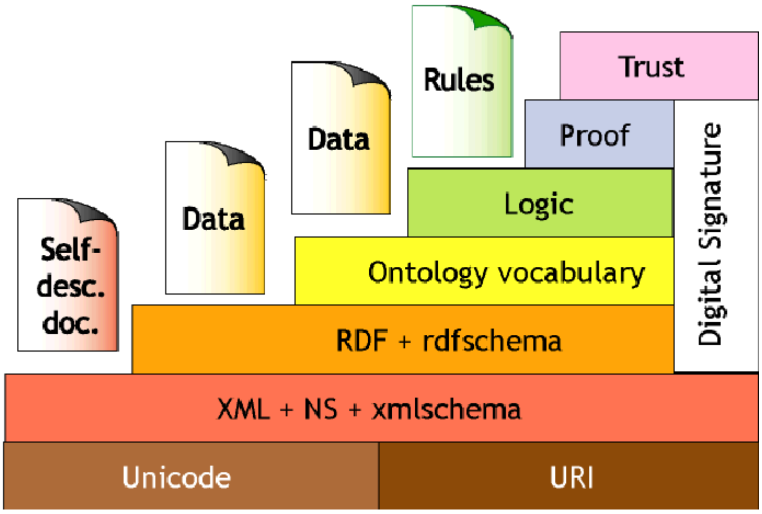


Figure 20: Semantic Web Layers -figure by Tim Berners Lee[6]

Public ontologies consist of two sets:

- **High Level Ontologies - Meta Ontologies**: Concern the more generic and abstract aspects that can be seen in the "reality" understood such as Dublin Core, Foaf, SKOS, XKOS, PROV, DCAT and GSIM-SUM.
- **Domain Ontologies:** Model a particular portion of reality such as people and families, labour. Ontopia Framework [51] the network of ontologies developed by AGID for Public Administration, based on the standards of the Semantic Web and aligned with Core Vocabulary of the ISA2 program of the European Commission, Istat CPV_AP-IT are examples of this type of ontology.

Both sets are useful during the modelling of a new ontology; the former represents a common way to modelling common and frequent pattern- modelling of statistical classification is a typical example of this meta ontologies-; the latter represent a source to retrieve and reuse existing ontologies. In both cases users may only need to use parts of the entire ontology, depending on their needs.

[6] Source: https://www.w3.org/2000/Talks/1206-xml2k-tbl/all.htm

Co-financed by the Connecting Europe Facility of the European Union
The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Examples of Meta Ontologies widely used for LOSD publication.

- Dublin Core (DC) [52] is a metadata standard to describe principal properties of digital resources such as title, creator, subject, description, publisher and contributor. It represents the base level to high level ontologies: it is common to find links to this standard in the ontologies that we will cite later. It's important to note that DC standard and in particular title, creator and subject property can be used to describe concepts and concept schemes.
- Friend of a Friend (FOAF) [53] is a simple vocabulary that describes People and their relation with other people. The core concept of this vocabulary is Person, the core role is "knows" and then there are some characteristics of person such as email, telephone number, web site and age. Furthermore FOAF defines the concept of Agent - person, group, software or physical artifact- and Organization that represents a kind of Agent corresponding to social institutions such as companies, societies etc.
- Simple Knowledge Organization System (SKOS) [54] vocabulary is focused on Concepts or conceptual resources. Concepts can be identified with URIs, labelled with lexical strings in one or more natural languages – only one label for each language, documented with various types of note- human-readable ("informal") documentation-, semantically related to each other in informal hierarchies and association networks and aggregated into concept schemes. SKOS represents one example of design pattern for statistical classification.
- One important ontology in statistical domain is Data Cube Vocabulary that is focused on multi-dimensional data and is based on SDMX (Statistical Data and Metadata eXchange), an ISO standard for exchanging and sharing statistical data and metadata among organizations. The Data Cube in turn builds upon the RDF vocabularies mentioned before: SKOS for concept schemes, Dublin Core Terms for metadata and FOAF for agents.
- The GSIM-SUM [55] ontology is the Istat profiling version of the GSIM - Generic Statistic Information Model developed by the HLG-MOS (UNECE)- ontology that was developed by Istat and the French Institute of Statistics -Insee. Istat and Insee specialized some concepts that in GSIM were too generic for their requirements. The GSIM-SUM add to GSIM Model other concepts such as, for example, the Data Content, the version of a classification, the variant of a classification and the variant structure of a classification.
- Provenance Vocabulary [56] describes information about entities, activities, and people involved in producing a piece of data or thing. Unlike other ontologies Provenance ontology is placed in "Trust" layer of semantic web but it is not enough for this layer. The trust layer of semantic tower is an unexplored issue: no efforts have been made in this regard. Without answers remain some question such as: Where does the information came from, how has been obtained and I can trust?
- Data Cube Vocabulary [57] (qb) provides a way to publish multi-dimensional data, such as statistics, on the web. In this way data and related metadata can be linked to external data sets and concepts. DQ (qb) describe at conceptual level SDMX Information Model; it describes Data Set, Data Structure Definition, Measure, Dimension and others.
  A Data Set (ds) is a collection of statistical data that corresponds to a defined structure. Data Set belongs to one of the following kinds: Observations- In a statistical table, the observations would be the values in the table cells, Organizational Structure, Structural metadata – metadata necessary to interpret observation such as unit of measurement - or Reference Metadata- is metadata that describes the dataset as a whole.
  Data Structure Definition (dsd) defines the structure of one or more datasets, it defines the dimensions, attributes and measures used in the dataset along with qualifying information such as ordering of dimensions and whether attributes are required or optional.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

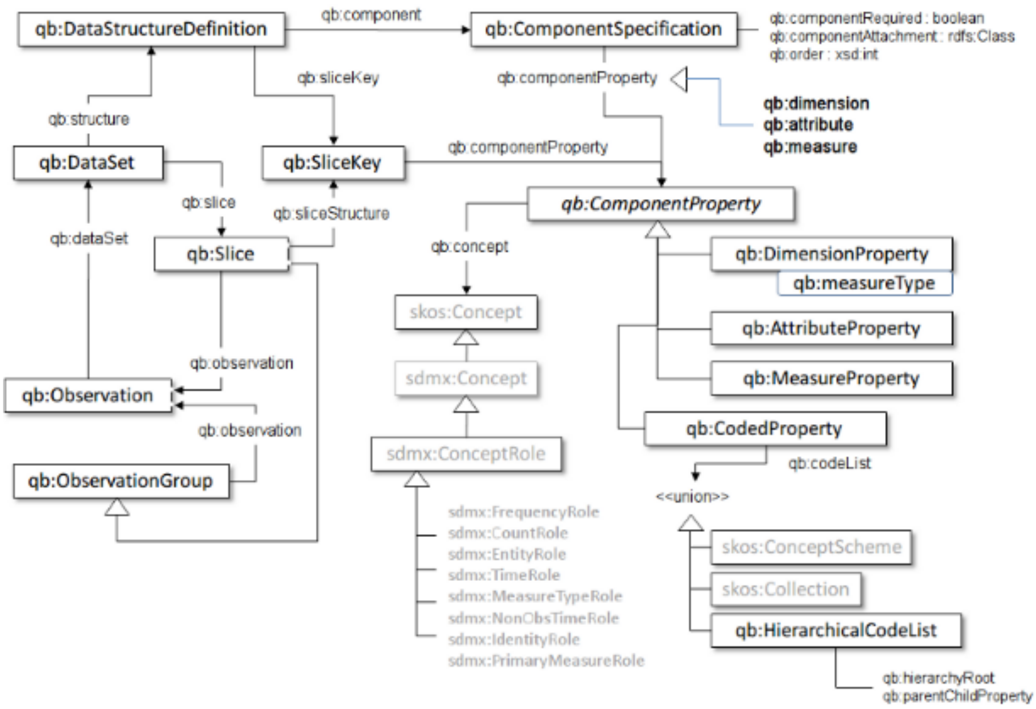


Figure 21: Data Cube Vocabulary- Pictorial summary. Source [57]

Starting from 2016, Agid (Agenzia per l'Italia Digitale) has encouraged the use of ontologies for the enhancement of the information asset. So Istat has adopted the guidelines promoting the use of ontologies. In particular Istat works at data modelling with ontologies in the context of Integrated System of Registers. Process of data modelling is complex and iterative: domain experts and ontology designers work together to achieve a validated version of the ontology. In the first phase- analysis- domain expert describe the domain of interest, in the second –modelling phase- ontology designer try to modelling the ontology and come back to expert for ontology validation-revision phase. If domain expert validates the ontology the process ends otherwise the process starts again. This process never ends at the first iteration.
First version of Families Ontology and Places Ontology have been published on Istat web page and their upgrade is on the work. Labour and Places Ontologies are yet in a development state.

2.3.4 SDMX

SDMX [58] [59] - Statistical Data and Metadata eXchange is a standard for the exchange of statistical data and metadata among international organisations. The most recent version, SDMX 2.1, has been approved in 2011. SDMX provides a common language to describe statistical dataset, a collection of related observations, organized according to a predefined structure.

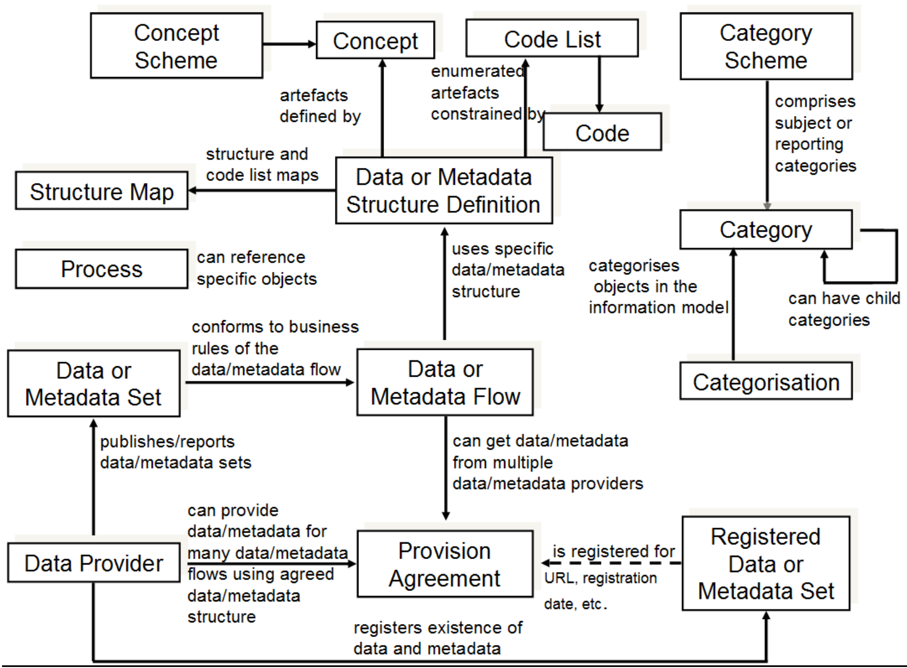


Figure 22: High Level Schematic of Major Artefacts [60]

SDMX consists of:

- Technical standards including the Information Model for the representation of statistical data and metadata
- Statistical guidelines
- IT architecture and tools

The core part for the LOSD lifecycle is the Information Model. This component has been developed to support statistics; it defines concepts such as dataset, data structure, dimensions, codelist and attributes. Data is organized into discrete sets, which include particular observations for a specific period of time. A Data Set is a collection of similar data, sharing a defined data structure. Each SDMX Data Set has a set of structural metadata. Specifically, Data Structure Definition includes information about how concepts are associated with the measures, dimensions, and attributes. SDMX Code lists enumerate a set of values to be used in the representation of dimensions, attributes, and other structural parts of SDMX. Codes can also be organized into hierarchies.

This model is also applicable to other organizational contexts involving statistical data and related metadata. The cube structure commonly used to process statistical data is similar to the Data Structure Definition idea in the SDMX Information Model. QB vocabulary is based on the SDMX Information Model, so for this reason SDMX can take part into LOSD lifecycle if we introduce a process of SDMX to RDF conversion.

In 2015 Istat described in the paper “Official Statistics meets the Semantic Web: How SDMX and RDF can live together" a project to integrate the SDMX dissemination architecture with the Semantic Web Standards [61].

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

2.3.5 Namespaces and Catalogues

In order to identify technical barriers to LOSD dissemination and increase cross-domain and cross-border interoperability, a relevant issue to consider concerns data and metadata standardization. Statistics dissemination usually relates to aggregated data, combined with explanatory structured information (metadata). On a conceptual level, such kind of information is represented by data cubes, described through a multidimensional model that combines measures, dimensions and hierarchies. Each cell of a data cube represents a numerical value measuring an observed phenomenon, while dimensions are the attributes of interest describing the context related to measures. The values of each dimension can be hierarchically ordered, thus providing information at different levels of granularity. In the statistical domain, dimensions can also correspond to official classifications or code lists. Capturing key metadata throughout the statistical process would facilitate data management and analysis, as well as data reuse. The Linked data cubes lifecycle [62] can be divided in three main stages. The first stage includes the following tasks:

- Discovery and pre-processing of input data, according to their format (e.g., xls, csv, Database tables)
- Data conversion in RDF format and reuse of controlled vocabularies
- Metadata management
- Interfaces development for data cubes publishing.

In the second stage, data cubes information is enriched by linking several related data cubes available on the Web. The activities performed in this stage allow to obtain a set of harmonised cubes also by aggregating or increasing dimensions and/or measures, starting from an initial linked data cube. The third stage relates to data cubes exploitation through data analytics and visualizations methods and tools.

Among the several vocabularies allowing the conversion of data cubes in RDF [27] graphs, the RDF data cube (QB) vocabulary [57] is one of the most used and is recommended by the World Wide Web Consortium (W3C). Concerning metadata, the Simple Knowledge Organization System (SKOS) [54] vocabulary and the Extended Knowledge Organization System (XKOS) [63] are the reference vocabularies for modelling statistical classifications. Moreover, specific vocabularies have been created for the main statistical concepts, as well as mappings of RDF format with metadata standards such as SDMX and DDI.

Nevertheless, data providers face several challenges to publish LOSD, and the lack of a common strategy to address them is one of the main factors preventing a broader LOSD exploitation. A study [64] performed for identifying the several challenges related to the conceptual description of data cubes, and their conversion in LOSD, has underlined the following open issues:

- Lack of shared solutions for standardising common modelling decisions
- Definition of standard code lists
- Improvement of LOSD tools performances.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

More in detail, the first issue concerns the design of cubes with a combination of multiple measures, or the definition of multiple units related to the same measure. A suggested approach for such kind of cubes is to publish them only if measures refer to a single observed event (e.g., sensor data observations), otherwise they should be modelled as distinct cubes. The development of further specialized vocabularies would foster the adoption of harmonised solutions and provide guidance on how to model relevant attributes for specific domains, thus increasing interoperability.

Similarly, the standardization of relevant code lists would facilitate metadata modelling, and the linkage between related information provided by several sources. Other advantages of ad-hoc vocabularies would be the description of the method used to calculate a measure, or the specification of aggregation functions that can be applied to a measure or to a dimension, such as sum or average. The provision of this additional information would enhance the definition of new measures derived by combining other measures and enable the compatibility assessment between variables. A clearer specification concerning related measures would improve the statistical analysis of a phenomenon through LOSD, reducing potential inconsistencies between the results of similar studies [65].

Concerning the enhancement of LOSD applications, in some cases, several iterations are needed to retrieve the cube observations of interest, for instance the retrieval of a subset of values used in a cube from a code list. The definition of best practices is a key element to remove the above procedural barriers to LOSD publication, and address other issues related to data and metadata management, such as naming conventions, standardization of data and metadata structures, data scaling.

2.3.5.1 Naming Conventions and IRI Normalization

Naming convention must be extended to make it possible for applications to search and find resources easily. URN, URL, URI and IRI help build a system where resources are well identified by a proper name, independent of the technical details and everlasting beyond location. URL, on the other hand guarantee resource addressability. Internationalized Resource Identifier (IRI) builds on URI widening available alphabet to international usage. Concepts of identifier and location relate to the need of uniquely identify and locate resources in the always feeble and ever changing Internet of Things. First expressed at application level protocol by a well-defined syntax [66], soon catalogues of these objects for specific area domains were created. Namespaces linked to Area Domains are wide available.

Co-financed by the Connecting Europe Facility of the European Union

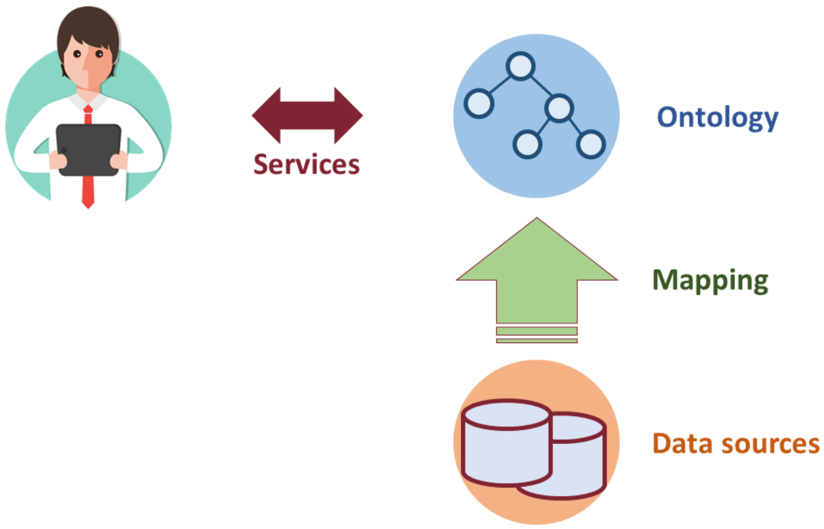
The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

2.4 LOSD Tools and Implementation

2.4.1 Monolith (Ontology Based Data Management Systems)

Ontology-Based Data Management (OBDM) is a new paradigm, rooted on the idea of using **Database Theory** and **Semantic Technologies** for data management. OBDM is characterized by the following principles:

- Let data reside where they are (no need to move data)
- Define a logic-based conceptual specification of the domain of interest (ontology)
- Map the ontology to the concrete data sources.
- Express all the services over the ontology.
- Automatically translate knowledge services to data services



MASTRO[7] is a Java tool for ontology-based data access (OBDA) developed jointly by Sapienza University of Rome and the Free University of Bozen-Bolzano. In Mastro the data source layer is represented by relational data management, the ontology is connected to data sources through a mapping establishing a semantic relation between SQL queries issued over the underlying databases and elements of the ontology. To access data, users can specify SPARQL queries over the ontology and make use of the query answering services provided by Mastro.

Istat experimented the OBDM paradigm and Mastro system in the context of Italian Integrated System of Statistical Registers. This work focused on data quality assessment, especially in the presence of multiple and possibly mutually incoherent data sources. The outcome of this work was the paper written jointly by Istat and Sapienza titled: “On the Experimental Usage of Ontology-based Data Management for the Italian Integrated System of Statistical Registers: Quality Issues” [67]. In this

[7] http://obdasystems.com/

context Istat also experimented and adopted Eddy, a graphical Ontology owl2 compliant editor. Eddy uses a graphical language for Ontology Design called “Graphol”.
MONOLITH [68], an innovative application developed by OBDA Systems, a spinoff of Sapienza University of Rome, that combines ontology and mapping inspection, query answering. Monolith integrates MASTRO OBDM reasoner. MONOLITH is not open source and it has been purchased by Istat and used for implementing OBDM Paradigm on the Italian Integrated System of Statistical Registers.

2.4.2 Data Cube Mapping

Linked Open Data production process unfolds into several sub steps where mapping is the key link between the conceptual formal description of the knowledge object (ontology) and the actual data level, where data are stored in the form of files, tables and the like. When modelling data at the conceptual level, data structure description is separated from the semantics of area domain, meaning there are two main aspects of formal description:

1. Data Structure description
   i.e. a data mart, classification, microdata, etc.
2. Object contents description
   i.e. measure and data object, such as population, tax, revenues, territory, etc

Data descriptors are mapped to actual data sources, then triplets are produced and then stored on a triplestore, ready to be queryed

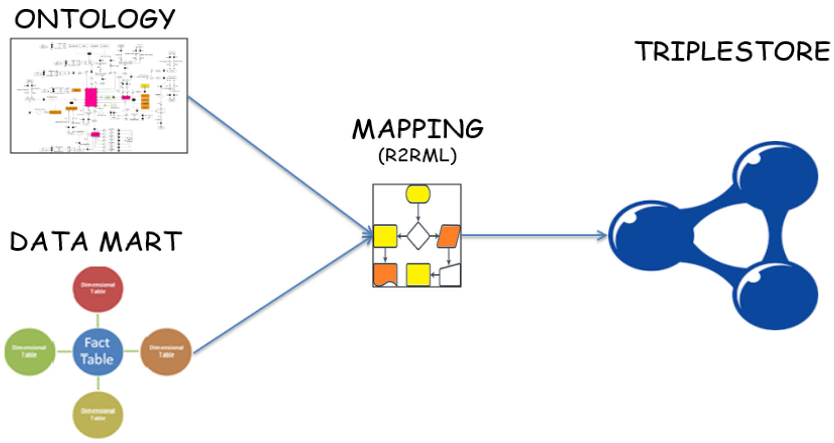


Figure 23: concepts mapping

Data Cube Metamodel describes aggregated structures. It models related concepts like Data Structure Definition (DSD), Dimension, Measure, Observations and attributes, as shown in the following Figure 24, expressed in Graphol language.

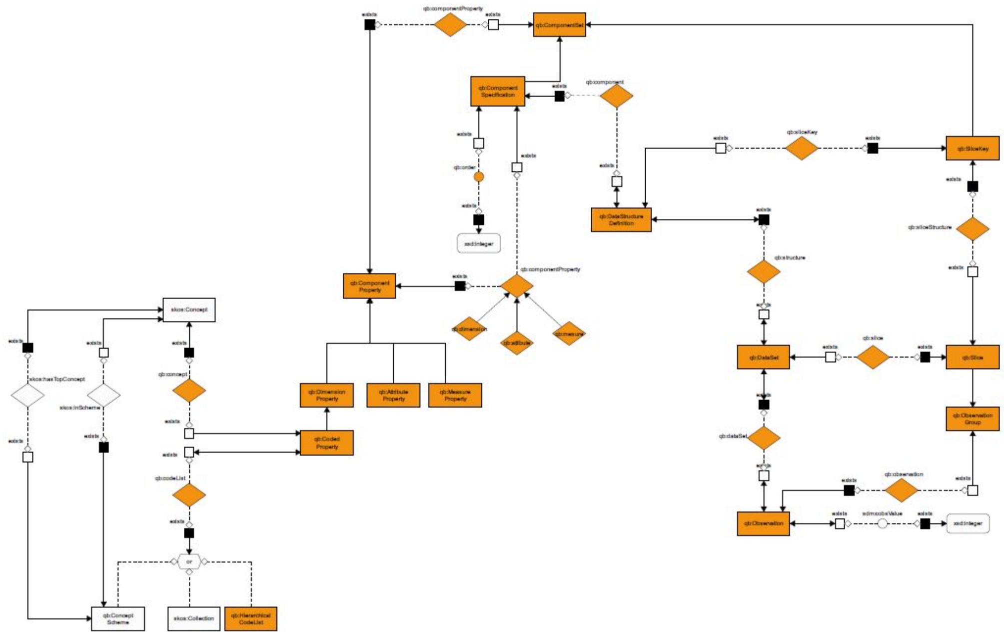


Figure 24: Graphol excerpt from qb

Each concept in the Data Cube needs to be instantiated, so that actual data needs to be modelled accordingly, and linked to its corresponding meta-model counterpart. Instantiation is shown in green in Figure 25:

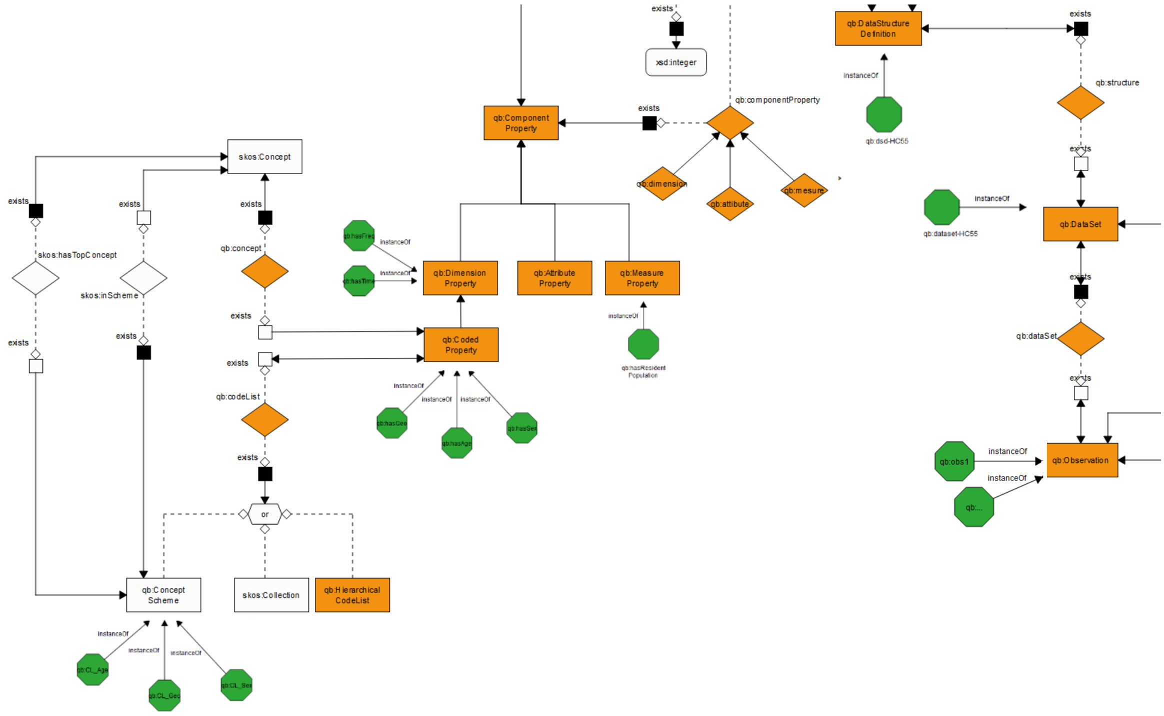


Figure 25:Graphol excerpt from qb with instances

Mapping is the act of linking the conceptual IRI descriptors to the actual data wherever is stored. Note that the same data description and instance can be mapped to several datasets sharing the same data structure and contents, even if they are stored on different systems and varying data formats. They will appear consistent to the conceptual description and fully integrated.

2.4.3 ESSNET LOSD Publishing Pipeline

A pipeline for standardized LOSD publishing was designed in the ESSNET LOSD pipeline with high flexibility and adaptability to different contexts. Pipeline process steps are standard and their implementation can vary accordingly to each publisher's needs. Standard process steps are sketched in the Archimate[8] modelling language in Figure 26:

- Data Model is designed using Domain Ontologies and Meta Ontologies.
  It uses Ontology description and produces an owl file as data object
- Mapping instantiates Knowledge Objects and links to actual data sources.
  It uses mapping rules expressed in R2RML language
- Production and management of triples in a Triplestore.
  RDF data objects are managed in this step
- Endpoint query and integration into distributed web apps.
  Data is transferred at the presentation layer by html protocol

Shared tools examples were also discussed and tested for compatibility with the pipeline. ESSNET was focused on developing the basic pipeline infrastructure to implement useful tools. Following works from the high level consortium group have explored several other aspects like metadata sharing, semantic harmonization and common data spaces.

There are two main use cases for the pipeline:

1. Data production.
   This use cases pertains to data producers like NSIs and its goal is to publish data to RDF.
2. Data consume.
   Queries use published data in some sort of third-parties applications or data analysis by final users.

Examples of consume cases are:

1. Data visualization and documentation purpose
2. Web searching and information discovery
3. Enrich application at semantic level
4. Collaborative and social computing

2.4.3.1 Process details and Application Components

Archimate schema of ESSNET LOSD publication process is sketched in Figure 26. Business layer depicting the four steps in the middle in yellow colour and application components in blue colour

[8] https://www.opengroup.org/archimate-forum/archimate-overview

below. Data Object are also included in the schema. ESSNET LOSD pipeline pertains mainly to data publication and uses appropriate Ontology vocabularies for both data structures design and metadata. In this use case, available Ontologies were used for modelling aggregated data and classification objects. The sketched process can be implemented by several components that can be substituted depending on the chosen software solutions. The figure shows one of the feasible architectures, realized with tools provided by Insight consortium[9].

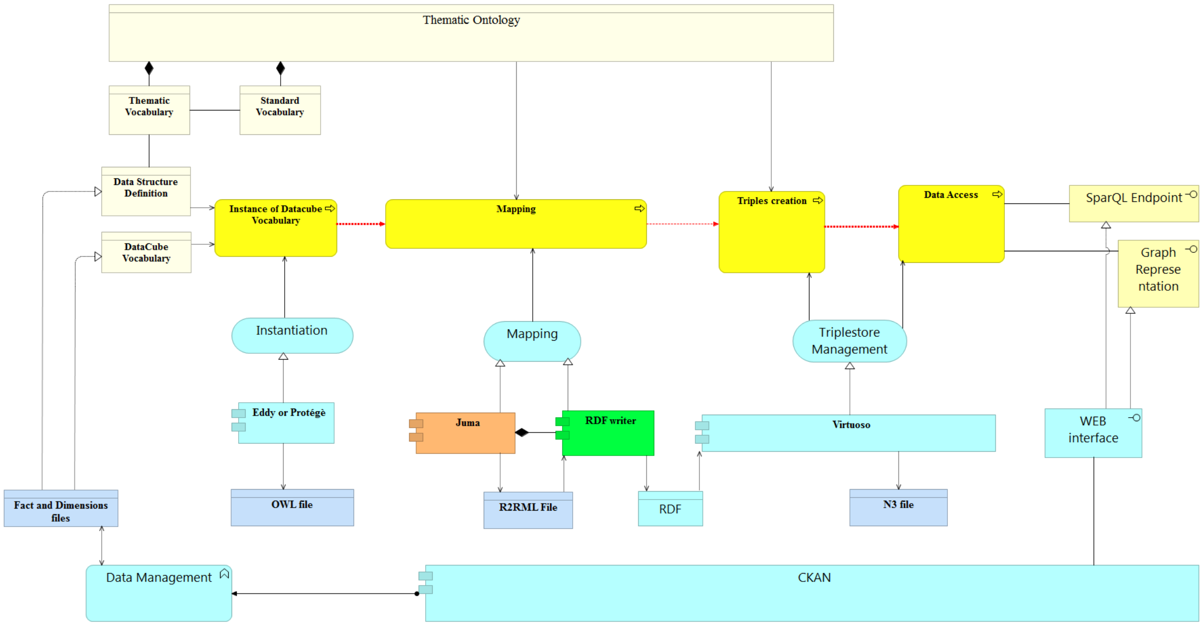


Figure 26: Process Design and Application Components

Among the tools developed by Derilinx-ADAPT-Insight consortium, there was a Graphical Mapping Editor (Figure 27) called JUMA which was much appreciated. As a matter of fact, mapping had always been a step that needed a discrete amount of work with ad hoc tools and needed to be standardized.

[9]https://losd-data.staging.derilinx.com/

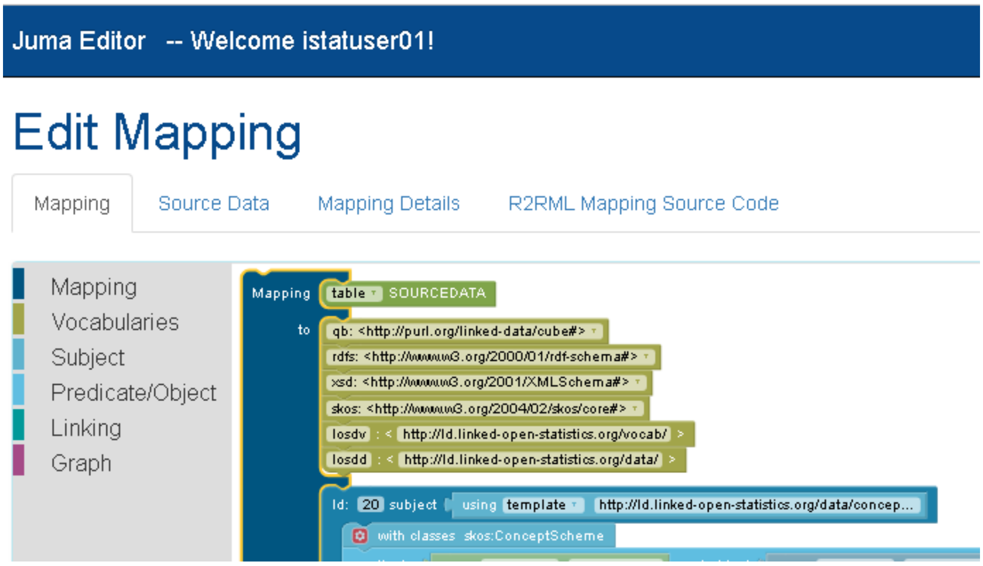


Figure 27: JUMA Graphical Editor

Juma was used to map csv data tables to several knowledge objects:

- Data Mart stored on csv file to Data Cube (qb) Vocabulary
- Classification tables to SKOS representing Data Cube dimensions

Application of several other tools to build custom pipeline were further explored during Istat LabINN (Istat Innovation Laboratory). Two implementations of the pipeline were realized:

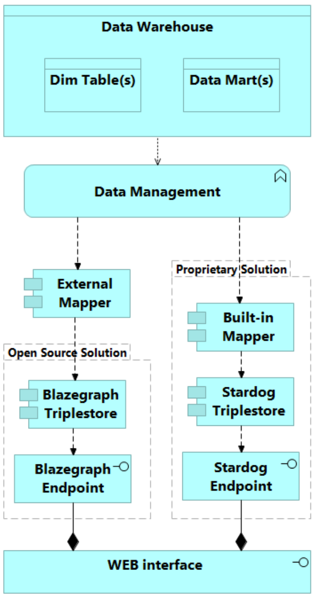


Figure 28: LOSD pipeline alternative implementations

Data sources were stored on Oracle database mimicrying Istat Data Warehouse structure and contents for this experiment. With courtesy from the producer, JUMA could be tested and adapted for the experiment too.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Freeware Blazegraph and proprietary tool Starrdog were tested as alternatives for triplestore component to build the pipeline.
Finally, these tools were used for pipeline build-up:

1. Oracle data source.
2. Stardog and Blazegraph, triple store.
3. JUMA for mapping.
4. MySQL for JUMA support.
5. J-r2rml as rdf converter.
6. Pubby for graph navigation.

Such pipeline shows how modular design can be adapted to different contexts and can be used to connect to data sources, such as databases or files in several formats. The prototype can be integrated in any existing production system and can bridge between data sources and linked open data portals. Once data have been published and made available for querying, they are ready to be consumed by final users. Data can be presented in tabular form, or can be queried directly, but LOSD technology can be used beyond the limits of human-centric dissemination process. LOSD is a machine-readable technology, meaning a computer can read, interpret, and use the data directly, making it possible to incorporate data into applications, so that LOSD acts more like an application protocol and is masked by user experience. End users applications allow to integrate published data and make it easier to visualize and analyze it. Search and navigation can also benefit from integrated tools, enriching the information discovery experience and making navigation more interesting and enjoyable for final users.

Co-financed by the Connecting Europe Facility of the European Union
The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

3 Technical barriers for LOSD and possible solutions

In this section, starting from the survey and review of technologies and initiatives presented in the previous chapters and further analysing the 2020 report "Barriers in working with Open Data" commissioned by the EU Data portal [69], some considerations related to the possible the technical obstacles in provisioning and reuse of (Linked) Open Data will be highlighted. For each of this barrier we will try to propose some suggestions in terms of best practise and technical solution that INTERSTAT is going to follow and develop to overcome them. It is important to underline that there are several stakeholders involved in the potential achievement of these objectives: these include national governments, single public institutions, but also technical developers and data managers. For this reason, in this context it is very important, in order to reach a successful adoption of any guideline or recommendation to involve all the interested entities and institution in their definitions: holding public consultations during national guideline definition could be a mean to drive ownership by giving different stakeholders an opportunity to share their expectations and influence the process at hand.

3.1.1 Availability of Open Data

Several technical barriers prevent both data suppliers and data users from efficiently working with Open and Linked Data. Although a lot of data has been made available in the past years, resulting in over half a million data sets referenced at the EDP [4], the availability of Open Data persists to be a barrier.

EU countries are developing policies and datasets from multiple government departments are being opened, resulting in more and more data sets being made available. But, a lot of work remains to be done in this area, as the availability of Open Data remains to be a challenge, at least in the perception of re-users. 73% of the companies using Open Data indicate to find it difficult or very difficult to find the data they need.

Organisations indicate as a barrier the poor discoverability of the data they are looking for. Barrier which is also related to the low levels of quality in the descriptions of the datasets themselves and the plurality of platforms where the data can be found. When the descriptions are not specific enough it is hard to find the data even when it is published. When dispersed among platforms in different countries, also language barriers might play a role in the poor discoverability of Open Data. This is even more prominent when portals have only limited or very basic search functionalities.

The solution to this problem can be related to different factors, starting from the real access to data, that are officially public, but often managed by external companies that effectively control it, to the role of data managers in making data really open. From strictly technical point of view the possibility to adopt standard technologies for open data provisioning is one of the key factors to simplify and foster open data availability. But considering that a lot of open data is currently shared on legacy and not standard systems the migration to new IT infrastructures or portals can require a lot of effort and costs.

The INTERSTAT framework is going to face this problem offering technical solutions that can federate old technologies (e.g. legacy data portal) providing the access to the open data through standard metadata models (i.e. DCAT-AP) and API even they are not provided in the original portal/repository. This will allow public organisation to save cost and time avoiding to change the existing systems achieving, at the same time, the possibility to offer their open data in a standard way to be also harvested by other portals (e.g. EDP).

3.1.2 Quality of Data and metadata

The quality of Open Data appears to be even more problematic, with data being published in different structures and in different formats. At the IODC[10] 2016 in a survey, data-reusers indicated quality and the availability of Open Data as the main barriers. Figure 29 shows the diagram with the complete results: metadata, format and standardisation are the other significant barriers highlighted by the survey.

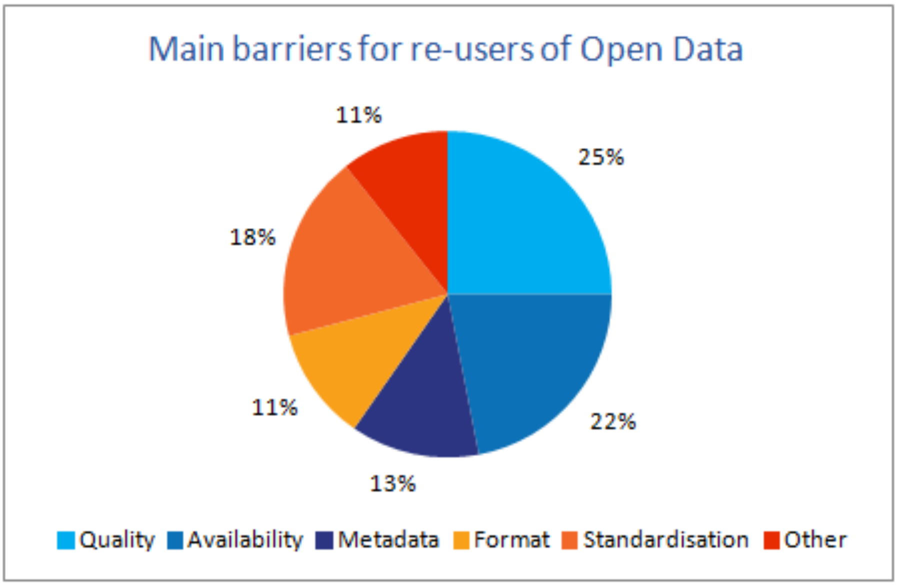


Figure 29: IODC 2016 Survey on Open Data adoption barriers. Source [69]

The quality of data and related metadata is very important, high quality and consistent (meta) data saves the re-users time fostering automated processes and creating trust. Organisations in several countries indicated that data-quality issues continue to be a barrier to the use of data accessed via

[10] International Open Data Conference

their portal, and that no easy way exist to enforce a better quality [70]. The low quality refers to both the data as well as the related metadata. An Open Data portal should not be a source of broken links, out of date and unused data nor poor meta-data, this makes it difficult for users to find the data they are looking for as their queries do not find the sets which hold this information.
Despite the fact that the number of open datasets is radically increased in the last decade, there is still a lot of low quality data (in terms of content and metadata completeness, data format ect). Furthermore, in order to achieve data quality, it is necessary to take care of the entire data lifecycle considering both the technical and organizational aspects.

3.1.2.1 Lack of DCAT-AP compliance and adherence to standards for metadata

The interoperability of Open Data portals is a challenge for organisations and a barrier for re-users of Open Data. This refers to different portals using the same standards which facilitates the sharing of data between different data catalogues. In this context the role of metadata standardisation is fundamental. As already said in section 1.3.1.1 DCAT-AP (DCAT Application Profile) is an extension of DCAT that was developed by the European Commission to improve interoperability and facilitate the availability and reuse of open data in European catalogues. Compliance with DCAT-AP (and its extensions such as GeoDCAT-AP) is increasingly supported and recognised among Member States, but there are several local agencies that are still using non-standard metadata. As we reported in section 1.2.1 EDP uses the DCAT-AP, but not all portals harvested by the EDP map their datasets in the same categories. Using common Open Data metadata standards and application profiles will help to maximise the discoverability of data across portals avoiding data publishers struggling with technical aspects.

One of the central components of INTERSTAT framework will be the open source platform Idra[11] that will specifically address the problem of compliance with DCAT-AP metadata providing automatic mapping functionalities between the existent/legacy metadata and DCAT-AP metamodel. Also, the support for GeoDCAT and statDCAT is planned to be introduced, during the framework development, as additional mapping capabilities for Spatial and statistical datasets.

3.1.2.2 Difficulty to maintain up-to-date metadata and data

As reported in previous sections not all the EU member states have a pre-defined approach in place to ensure that the metadata is kept up to date. Up-to-date metadata is critical for users to obtain the correct information about the data that it describes. The ability to keep the metadata up to date depends, among others, on the extent to which metadata is obtained from its source automatically. Similarly, also the update of content data itself is crucial in the open data provisioning. Deliver an open dataset without taking care of its update makes the opening data process completely useless in the long term. This still remain a problem, in particular in the context of public organisation that have no automatic data provision pipelines and manages the process completely manually.

[11] https://github.com/OPSILab/Idra

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

INTERSTAT will contribute to overcome this barrier providing tools to automatize the data and metadata updates: in particular the INTERSTAT framework will take advantage from the reuse of CEF context broker that provides functionalities of context data management with notification in real-time about data and metadata changes.

3.1.3 Formats

The provision of Open Data is still not a priority among data providers and the diversity of Open Data being available in different formats, with different licences in different languages is another limitation for re-using it. This heterogeneity of the various characteristics of Open Data limits their usability and the lack of standardization limits the opportunity for users to develop permanent solutions to reuse Open Data in their processes.

In particular, in relation to the file formats it is possible to recognised that, the machine-readability of datasets is a key factor for data reuse. Machine-readable data refers to data that can be automatically read and processed by a computer. As examples, formats that are machine-readable are CSV, XLM or XLS. Machine-readable formats can be accessed in an automated fashion. A dataset is considered as machine-readable if at least one of its distributions uses a machine-readable format. A relatively low number of machine-readable datasets might be caused by a focus on quantity by the publication team rather than on quality. At the same time, machine readable format can be an obstacle for the reuse of the dataset from the point of the final user if they are not accompanied by specific tools that offer a human readable view of the data.

Figure shows the percentage of machine-readable datasets for the top 20 catalogues mostly using machine-readable datasets.

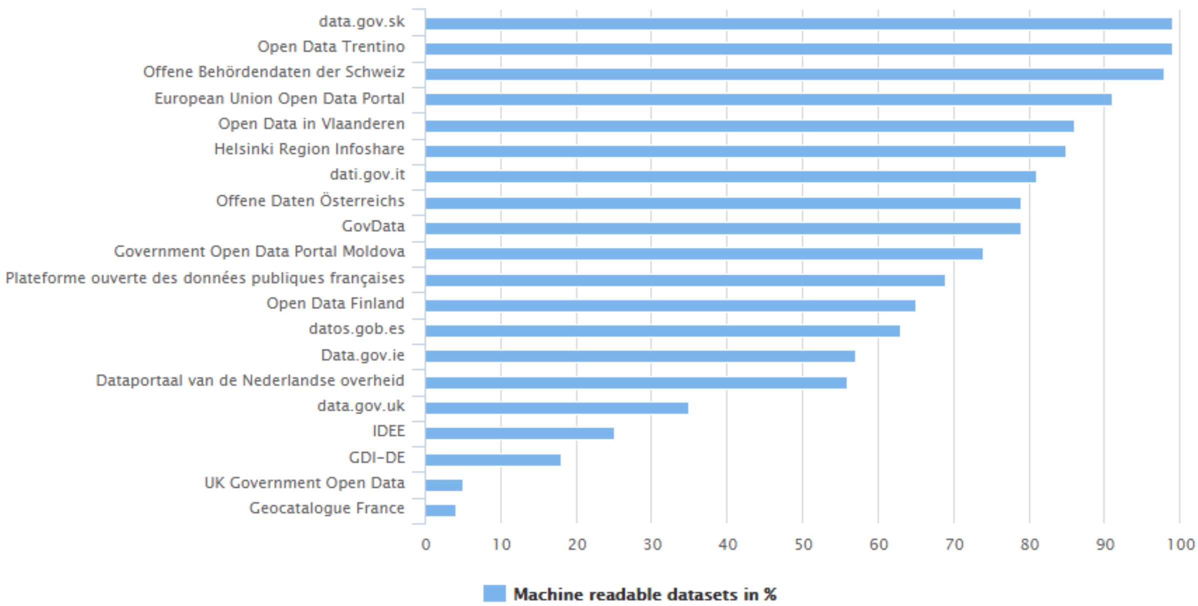


Figure 30: Top 20 catalogues mostly using common machine-readable datasets. Source: [69]

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

Heterogeneity of datasets file types and formats will be managed, by the INTERSTAT framework, through specific data connectors able to access and gather data in different formats. Specific open data provisioning pipelines developed in the project will simplify the process of providing datasets in machine-readable formats starting from legacy data.

3.1.4 Linked Open data issues

3.1.4.1 Shortage of experienced personnel in the organizations that manage and provide the data

One of the biggest problems in the era of open data is certainly the lack of specific skills regarding linked data in the personnel of organizations and public administrations that hold the data and must make them usable in linked data formats. Building knowledge in linked data is not an easy task also for technical people with several year of experience and require specific training activities and effort not always available in public organisation.
Specific tool for user friendly manipulation of LOSD will be developed in INTERSTAT: these tools will support data managers in mapping and producing linked open data.

3.1.4.2 Complexity for final user to query and visualise LOD

The complexity of LOD management is something that is not simply related to the data provider, but it is also reflected on the final user that has to access to the data. For instance, the most common language for querying linked data is SPARQL: its complexity makes it unattractive for end users who cannot understand and fully exploit the linked data potential. Also, the LOD visualisation can be very complicated for non technical-savvy users and the lack of simple human readable interfaces create a big barrier for linked data usage in the everyday life.

Specific tool for user friendly visualisation of LOSD will be developed in INTERSTAT: these tools will support non technical end users querying and exploring linked open data with a simple approach.

3.1.4.3 Data Harmonisation and Federation

Many existing data in the public administration are kept in simple structured formats (e.g. CSV) that could be converted into Linked Data; in order to avoid that this activity remain a mere exercise in style, the real critical factor is the ability to identify the relationships existing among different datasets and incorporate them in the transformation into LD.
Semantic harmonization become a central topic for the dataset: linkability can’t be addressed efficiently when data aren’t harmonized nor standardized. Perform data linking of multiple datasets without a clue about each dataset semantics can drive to meaningless results. The best solution to overcome this issue is to have an early dataset design and enforcement of semantic harmonization to

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

guarantee federation and linkability capabilities among different data providers. In addition, it has been demonstrated by different studies that there is a lack of shared solutions for standardising common modelling decisions and different data providers proceed in LOD publication using different strategies.

INTERSTAT will dedicate a large part of the effort to address the issues related to data harmonisation and federation: this will be mainly faced from two different points of view: (1) as mentioned in the section above the open platform Idra will be in charge of federating and harmonise heterogenous open data from what concern the data access API and the metadata compliance with DCAT-AP. (2) identification of common ontologies for aggregated data is another key activity that will be performed in INTERSTAT with the final scope of achieve data harmonisation in the field of open statistical data.

3.1.5 Access to data via API

A standard access to the open dataset via Application Programming Interface (API) is a key requirement to enable interoperability with external systems and applications that want to provide services on top of it. The different European and national meta-catalogues initiatives are in the direction of providing a unique and standard access to datasets harvested by heterogenous open data portals of different public organisations: this effort is anyway not enough to include still high quantity open data that cannot be harvested with the traditional techniques. Moreover, the availability of API is also more relevant when we deal with dynamic data whose nature makes it most useful when, it is intended to be reused, without delay, as soon as it is collected. An example would be a weather system, which automatically retrieves real-time data from weather stations to continually improve and update its forecasts, or a public transport application that not only informs passengers of schedules, but also of location as well as the estimated time of arrival. The API is the most widely used technology for real-time data delivery and a key topic of the new open data policies of many countries.

Standard API provisioning is one of the ultimate objectives of the INTERSTAT technical framework: all the data collected and harmonised by the framework will be accessible through a set of different APIs in order to be accessible by various systems and stakeholders: it will be offered a support for the SDMX API, CKAN API, NGSI-LD that will cover the needs of different usage scenario of the Linked Open Statistical Data.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

References

[1] E. D. Portal, “Open Data Maturity Report,” 12 2020. [Online]. Available: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n6_2020.pdf.

[2] EU, “DCAT Application Profile for data portals in Europe,” [Online]. Available: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe.

[3] EU, [Online]. Available: https://www.europeandataportal.eu/.

[4] EU, [Online]. Available: https://data.europa.eu/euodp/en/home.

[5] CKAN, “CKAN Open Data Platform,” [Online]. Available: https://ckan.org/.

[6] DRUPAL, “Drupal CMS,” [Online]. Available: https://www.drupal.org/.

[7] EU, “EU Login,” [Online]. Available: https://webgate.ec.europa.eu/fpfis/wikis/display/NAITDOC/EU+Login+-+European+Commission+Authentication+Service.

[8] EU, “EDP Architecture,” [Online]. Available: https://www.europeandataportal.eu/sites/default/files/edp_factsheet_portal_architecture_online.pdf.

[9] Apache Foundation, “Lucene Solr,” [Online]. Available: https://lucene.apache.org/solr/.

[10] E. D. Portal, “ Data Supplier Guidelines,” [Online]. Available: https://www.europeandataportal.eu/sites/default/files/edp_s1_gdl_data-supplier-guidelines_v9.pdf.

[11] EU, “Decision 2011-883,” [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF.

[12] “DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL,” [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32007L0002&rid=1.

[13] “INSPIRE Metadata Implementing Rules,” [Online]. Available: https://inspire.ec.europa.eu/documents/inspire-metadata-implementing-rules-technical-guidelines-based-en-iso-19115-and-en-iso-1.

[14] Tyler tech, “SOCRATA,” [Online]. Available: https://www.tylertech.com/products/socrata.

[15] SOCRATA, “COMMUNITY,” [Online]. Available: https://open-source.socrata.com/.

[16] T. Berners-Lee, “Linked Data,” 2006. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html.

[17] L. D. M. I. G. H. &. S. E. Ibáñez, “An assessment of adoption and quality of linked data in European open government data.,” in *International Semantic Web Conference* , 2019.

[18] “OMA-TS-NGSI-Context-Management-V1-0-20120529-A: "NGSI Context Management".,” [Online]. Available: http://www.openmobilealliance.org/release/NGSI/V1_0-20120529-A/OMA-TS-NGSI_Context_Management-V1_0-20120529-A.pdf.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

[19] "FIWARE NGSI REST binding version 2," [Online]. Available: http://fiware.github.io/specifications/ngsiv2/stable/.

[20] ETSI, "Context Information Management (CIM)," 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.03.01_60/gs_CIM009v010301p.pdf.

[21] EU, [Online]. Available: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home.

[22] W3C, "XML," [Online]. Available: https://www.w3.org/TR/xml/.

[23] Adobe, "PDF-A," [Online]. Available: https://acrobat.adobe.com/it/it/acrobat/about-adobe-pdf.html.

[24] "ISO 19005-4:2020," [Online]. Available: https://www.iso.org/standard/71832.html.

[25] IETF, "CSV," [Online]. Available: https://tools.ietf.org/html/rfc4180.

[26] OASIS, "Open Document," [Online]. Available: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office.

[27] W3C, "RDF," [Online]. Available: https://www.w3.org/RDF/.

[28] W3C, "RDF-Schema," [Online]. Available: https://www.w3.org/TR/rdf-schema/.

[29] OpenStreetMaps, "Shapefiles," [Online]. Available: https://wiki.openstreetmap.org/wiki/Shapefiles.

[30] W3C, "OWL," [Online]. Available: https://www.w3.org/OWL/.

[31] IETF, "JSON," [Online]. Available: https://tools.ietf.org/html/std90.

[32] Wikipedia, "GeoJSON," [Online]. Available: https://en.wikipedia.org/wiki/GeoJSON#:~:text=GeoJSON%20is%20an%20open%20standard,based%20on%20the%20JSON%20format..

[33] IETF, "GeoJSON," [Online]. Available: https://tools.ietf.org/html/rfc7946.

[34] EU, "DIGICOM," [Online]. Available: https://ec.europa.eu/eurostat/web/ess/digicom.

[35] EU, "ESS Vision 2020," [Online]. Available: https://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020.

[36] W3C, "Data Quality Vocabulary," [Online]. Available: https://www.w3.org/TR/vocab-dqv/.

[37] "RAMON Metadata Server," [Online]. Available: https://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC.

[38] S. Capadisli, "Statistical Linked Dataspaces," [Online]. Available: https://csarven.ca/statistical-linked-dataspaces.

[39] EU, "Data Quality Reporting," [Online]. Available: https://ec.europa.eu/eurostat/en/web/quality/quality-reporting .

[40] EU, "European Interoperability Framework," [Online]. Available: https://ec.europa.eu/isa2/eif_en.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

[41] EU, "COM(2017) 134 final," [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:2c2f2554-0faf-11e7-8a35-01aa75ed71a1.0017.02/DOC_1&format=PDF.

[42] EU, "INSPIRE Directive," [Online]. Available: https://inspire.ec.europa.eu/inspire-directive/2.

[43] EU, "eIDAS Regulation EU N°910/2014," [Online]. Available: https://ec.europa.eu/futurium/en/content/eidas-regulation-regulation-eu-ndeg9102014.

[44] EU, "European eGovernment Action Plan 2016-2020," [Online]. Available: https://ec.europa.eu/digital-single-market/en/european-egovernment-action-plan-2016-2020.

[45] EU, "The single digital gateway," [Online]. Available: https://ec.europa.eu/growth/single-market/single-digital-gateway_en.

[46] EU, "EIF toolbox Recommendations," [Online]. Available: https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/solution/eif-toolbox/recommendations.

[47] EU, "European Interoperability Framework - Implementation Strategy," [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:2c2f2554-0faf-11e7-8a35-01aa75ed71a1.0017.02/DOC_2&format=PDF.

[48] EU, "European Data Governance," [Online]. Available: https://ec.europa.eu/digital-single-market/en/european-data-governance.

[49] W3C, "Linked Data Cookbook," [Online]. Available: https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook.

[50] r. R. T. M. H. Thomas Steine, "How Google is using Linked Data Today and Vision For Tomorrow," [Online]. Available: https://research.google/pubs/pub37430/.

[51] "Ontopia Framework," [Online]. Available: https://github.com/italia/daf-ontologie-vocabolari-controllati.

[52] DCMI, " Dublin Core™ Metadata Initiative," [Online]. Available: https://dublincore.org/.

[53] L. M. Dan Brickley, "FOAF Vocabulary Specification," [Online]. Available: http://xmlns.com/foaf/spec/.

[54] W3C, "SKOS Simple Knowledge Organization System," [Online]. Available: https://www.w3.org/TR/skos-primer/.

[55] ISTAT, "GENERIC GSIM-SUM ONTOLOGY," [Online]. Available: https://www.istat.it/en/archivio/217565.

[56] W3C, "PROV-O: The PROV Ontology," [Online]. Available: https://www.w3.org/TR/prov-o/.

[57] W3C, "The RDF Data Cube Vocabulary," [Online]. Available: https://www.w3.org/TR/vocab-data-cube/.

[58] SDMX, [Online]. Available: https://sdmx.org/.

[59] EU, "Eurostat's SDMX InfoSpace," [Online]. Available: https://ec.europa.eu/eurostat/web/sdmx-infospace/welcome.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union

[60] SDMX, “Framework for SDMX Technical Standards,” [Online]. Available: http://docs.sdmx.org/en/i1-doc/Sections/Section1/SDMX_2-1_SECTION_1_Framework.html.

[61] S. D. F. A. P. M. S. Raffaella Maria Aracri, “Official Statistics meets the Semantic Web:How SDMX and RDF can live together,” [Online]. Available: https://ec.europa.eu/eurostat/cros/system/files/Aracri_etal%20SDMX-RDF-NTTS_2015_unblinded.pdf.

[62] T. E. K. T. Kalampokis Evangelos, “ICT Tools for Creating, Expanding, and Exploiting Statistical Linked Open Data,” vol. Statistical Journal of the IAOS, 2016.

[63] XKOS, “An SKOS extension for representing statistical classifications,” [Online]. Available: https://ddialliance.org/Specification/XKOS/1.2/OWL/xkos.html.

[64] D. Z. K. T. Evangelos Kalampokis, “On modeling linked open statistical data,” vol. Journal of Web Semantics, 2018.

[65] E. T. A. K. K. T. E. Kalampokis, “Open Statistics: The Rise of a New Era for Open Data?,” 2016.

[66] EU, “URIs in the context of the Europeana Data Model,” [Online]. Available: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/FAQs/URIs%20in%20EDM_pro.pdf.

[67] e. a. Aracri, “On the experimental usage of ontology-based data management for the italian integrated system of statistical registers: quality issues,” 2018.

[68] e. a. Lepore Lorenzo, “Monolith: An OBDM and knowledge graph management platform,” 2019.

[69] EU, “Barriers in working with Open Data,” 2020. [Online]. Available: https://www.europeandataportal.eu/sites/default/files/edp_analytical_report_n5_-_barriers_in_open_data.pdf.

[70] EU, “Recomendation on Open Data Portal,” [Online]. Available: https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

Co-financed by the Connecting Europe Facility of the European Union

The contents of this publication are the sole responsibility of INTERSTAT consortium and do not necessarily reflect the opinion of the European Union