Data quality management

September 2019

DIGIT

Directorate-General for Informatics

ISA² Programme

ec.europa.eu/isa2

Disclaimer:

The views expressed in this study are purely those of the Author(s) and may not, in any circumstances, be interpreted as stating an official position of the European Commission.

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

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

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

The study was prepared for the European Commission by Deloitte – NRB consortium.

EUROPEAN COMMISSION

Directorate-General for Informatics (DIGIT)

Directorate D – Digital Services

Unit D2 – ISA^2 Programme

Contact: isa2@ec.europa.eu

European Commission

B-1049 Brussels

This publication has been drafted under the 2016.07 *SEMIC: Promoting semantic interoperability amongst the EU Member States*. ISA^2 is a EUR 131 million programme of the European Commission, supporting the modernisation of public administrations in Europe through the development of interoperability solutions. More than 20 solutions are already available, with more to come soon. All solutions are open source and available free of charge to any interested public administration in Europe.

Table of Contents

Table of Tables

1. INTRODUCTION

This study explores the intersection between data quality management (from a data governance point of view) and semantic interoperability: how semantic assets support and evolve data quality considerations. It describes state of the art concepts and frameworks for data quality and link those to semantic interoperability by studying how data quality can be improved.

As we are currently going through the digital age, enormous amounts of content become available each day in terms of corporate files, medical records, government documents, court hearing etc. resulting in information overload. Even though recent technological advances have promoted innovation in all kind of enterprises, organisations and IT fields, we also experience the "data crisis" of the digital age. That results in poor or low quality of data, mainly due to the inability of properly handling the massive and fast production of data. In organisations poor data quality is influenced by data transfers from legacy system, data merging processes (in company level, dataset level etc.), skill shortages, erroneous data entries. Poor data quality has multiple impacts on an organisation or enterprise including reduced productivity, customer dissatisfaction, increased operational cost, less effective decision-making, reduced ability to make and execute strategic plans, difficulties in aligning the enterprise, and so on. In order to avoid such problems and target for high data quality, an organisation should follow a data quality management strategy. The primary task in defining the appropriate strategy is the quality assessment of available data with respect to the information it conveys. Such a process is not straightforward, since there are many definitions of data quality available in literature and there have been many criteria and metrics defined in an effort to assess and measure data quality. In this study we consider a semantic approach on data quality and we use metrics to assess quality of data through the prism of semantic interoperability between organisations and public administrations.

The main objective of this document is to investigate how data quality in the context of data governance can be improved through the use of semantic methodologies. In the past decade, along with the evolution of the semantic web, several useful technologies for knowledge capturing, representation, processing and enrichment have been developed that can be used in large-scale environments for data quality improvement and management. Such technologies known as semantic web technologies, provide standards for modelling datasets, encoding general knowledge in ontologies, allowing enhancements based on automatic reasoning (improved querying, for example). Semantic web technologies may be able to shift current data quality management to the next level. Linked data has proved to be a powerful tool that gains more attention by many communities as it enables data to be interconnected by generating semantic connections among datasets and thus improving the quality of data in many ways.

The rest of this document is structured as follows: in section 2 we present the most prominent data quality dimensions that are available in bibliography. In section 3, we focus on the semantic web methodologies, technologies and open standards that can be used by public organisations to improve their data quality primarily with respect to the dimensions introduced in section 2 and additionally from a general perspective. In section 4, we focus on semantic enrichment of metadata, considering the impact of metadata in data quality and discoverability. We provide a general overview of state-of-the-art machine learning trends and how they can contribute to data quality improvement. In particular, we refer to natural language processing and human in the loop and how they can be combined to improve the quality of the data through metadata enrichment. In section 5, we conclude our study by reflecting on what was learned through this work.

Additional information on linked data and its quality dimensions is available in Annex I and a glossary of terms is available in Annex II.

2. DATA QUALITY

Data Quality Management is the way an organisation defines, manages, monitors, maintains the integrity and improves quality of data (1). It includes the processes of data acquisition, implementation of advanced data processes and effective distribution of data, and also implies the high-level management of the information. Data quality management is very important for any organisation to derive actionable and, more importantly, accurate insights from information.

Data quality is synonymous with information quality. Data quality refers to the accuracy of datasets, and their ability to analyse and create actionable insights for other users. Key elements to reaching high-quality of data are *people*,

processes and *technology*. All enterprises and organisations that deal with data need to define and follow a rigorous data quality approach to provide a solid solution to improved data quality and integrity. Such an approach should involve managing the lifecycle for data creation, transformation, and transmission in order to ensure that the resulting information meets the needs of all the data consumers within the organisation.

The first step towards data quality assessment is the identification of key metrics for measuring data quality. The definition of appropriate metrics is essential to provide the best and most solid basis for future analyses. These metrics will help to track the effectiveness of the quality improvement efforts. In literature, there have been many metrics defined in the effort to assess and measure data quality and different terms have been used to refer to the same notion such as *metrics*, *dimensions*, *criteria*. Although the term *data quality dimension* has been widely used for a number of years to describe a measure of data quality, the key data quality dimensions are not universally agreed amongst data quality experts. Depending on business expectations, there are many different criteria on how and what data dimensions should be evaluated in order to define data quality metrics. The proposed dimensions and criteria are usually developed on an ad hoc basis to solve specific problems. As proposed by DAMA UK working group in (2) in an effort to define the key data quality dimensions and provide context so there can be a common understanding between professionals, the primary dimensions for data quality assessment are:

- Completeness
- Uniqueness
- Timeliness
- Integrity or Validity
- Accuracy
- Consistency

As already mentioned, in this study, we consider data quality in relation to data governance and we mainly focus on promoting the use of semantic methodologies in order to achieve improved data quality and promote semantic interoperability. In the remaining of this section we provide the definitions of the primary data quality dimensions. In section 3.7, we illustrate how these dimensions are impacted by the semantic web methodologies presented throughout section 3.

2.1. Data quality dimensions

A Data Quality Dimension is a term used to describe a data quality measure that can relate to multiple data elements. The most recognisable and widely used dimensions are the 6 core ones as proposed by DAMA UK working group (2).

1. **Accuracy** refers to the extent to which entities and facts correctly represent the real-life phenomenon. Inaccuracy can be reflected by incorrect data values, numeric or descriptive data (gender, location, preferences etc.) or other information that is not updated. By term accuracy we usually refer to semantic accuracy. However, accuracy can be distinguished into syntactic and semantic.

    a. Syntactic Accuracy is defined as the degree to which an entity document conforms to the specification of the serialisation format and literals are accurate with respect to a set of syntactical rules.

    b. Semantic Accuracy is defined as the degree to which data values correctly represent the real-world facts. Semantic accuracy refers to accuracy of the meaning. In order to capture the semantic inaccuracies, one needs to understand whether facts precisely capture the status of the real world.

2. **Completeness** indicates if there is enough information contained in data to draw conclusions. Completeness can be measured by determining whether or not each data entry is a "full" data entry. All available data fields must be complete and sets of data records should not be missing any pertinent information.

3. **Consistency** refers to the absence of difference, when comparing two or more representations of a thing against a definition. At a practical level, it specifies that two data values pulled from separate data sets should not conflict with each other. However, consistency does not automatically imply correctness.

4. **Integrity**, also known as data validity, refers to the structural testing of data to ensure that it complies with the procedures. This means there are no unintended data errors, and it corresponds to its appropriate designation (e.g., date, month and year).

5. **Timeliness** indicates the degree to which data represent reality from the required point in time**,** corresponding to the expectation for availability and accessibility of information. In other words, it measures the time between the moment the data is expected and the moment when it is readily available for use**.** Timeliness is usually affected by the way data are collected. The more steps and intermediate systems involved in data collection; the more delay will experience in receiving and making available the needed information.

6. **Uniqueness** of data is the degree to which data is free of redundancies, in breadth, depth and scope**.** It basically indicates that there should be no data duplicates. Each data record should be unique, otherwise there is a risk of accessing outdated information.

The following table illustrates the six core data quality dimensions as defined in (2) and the relations among them.

Overview and relations of data quality dimensions

Name	Definition	Related Dimension
Accuracy	The degree to which data correctly describes the "real world" object or event being described.	Validity,
Completeness	The proportion of stored data against the potential of "100% complete"	Validity and Accuracy
Consistency	The absence of difference, when comparing two or more representations of a thing against a definition.	Validity, Accuracy and Uniqueness
Integrity/Validity	The absence of difference, when comparing two or more representations of a thing against a definition.	Accuracy, Completeness, Consistency and Uniqueness
Timeliness	Data are valid if it conforms to the syntax (format, type, range) of its definition.	Accuracy
Uniqueness	Nothing will be recorded more than once based upon how that thing is identified.	Consistency

Table 1 Data Quality Dimensions

3. SEMANTIC WEB TECHNOLOGIES FOR IMPROVED DATA QUALITY

The term Semantic Web is commonly used to refer to an extension of the World Wide Web that is constructed upon standards introduced by the World Wide Web Consortium (W3C). These standards enable the use of common data formats and provide a common framework for data sharing and reuse across applications, enterprises, organisations and communities, thus advancing the semantic web as an integrator across different content, applications and systems. Data in the semantic web should not only be stored in a machine-processable syntax, but it should also be endowed with formal semantics that clearly specify which conclusions should be drawn from the collected information. The basic structuring element of semantic web is **Linked Data**, that is data understandable by machines, which is interlinked with other data with defined relationships, can be retrieved through semantic queries. More information on linked data, their basic principles and their quality dimensions can be found in **Error! Reference source not found.**.

Semantic web technologies are technologies that implement and foster semantic web standards, enable the creation of data stores on the web, facilitate the development of vocabularies and the extraction of rules for handling data. Linked data, as the basis of semantic web, are empowered by semantic technologies such as RDF, SPARQL, OWL, and SKOS.

In this study we focus on how semantic web technologies can improve data quality metrics and contribute to enhanced data quality. In the remaining of this section we look into semantic web technologies such as knowledge representation, reasoning validation and querying. Precisely we provide a short introduction of hierarchical knowledge structures such as controlled vocabularies, thesaurus, taxonomies and ontologies, we emphasise on ontological reasoning, validation and query mechanisms and consider how they can improve data quality dimension. We also provide information about open standards like DCAT-AP that can be employed and extended by public organisations to model their data, improve their quality and facilitate interoperability.

3.1. Knowledge representation

Knowledge representation is a branch of symbolic Artificial Intelligence that refers to a machine-interpretable representation of the world. Knowledge-based systems define a computational model of some domain of interest, which can cover any part of the real world or any hypothetical system about which one desires to represent knowledge for computational purposes. "In the domain of interest, symbols serve as surrogates for real-world domain artefacts, such as physical objects, events, relationships, etc." (4). A knowledge-based system maintains a knowledge base which stores the symbols of the computational model in form of statements about the domain. By the term "conceptualisation" we refer to the mapping between the symbols used in the computer (i.e., the vocabulary) and the individuals and relations in the world. In order to share and communicate the conceptualized knowledge, it is important to use a common vocabulary and an agreed-on meaning for that vocabulary. In this study, we will discuss the hierarchical models that use formal specifications for knowledge representation such as ontologies, controlled vocabularies, taxonomies and thesaurus, with main emphasis on ontologies.

A **taxonomy** is a controlled vocabulary with a hierarchical structure. Terms within a taxonomy have relations to other terms (parent/broader term, child/narrower term). The term taxonomy tends to be used to refer to two different things: a tree-hierarchical controlled vocabulary lacking more complex relationships found in thesauri or ontologies, or any kind of controlled vocabulary.

A **thesaurus** is essentially a controlled vocabulary following a standard structure, where all terms have relationships of three kinds to each other: *hierarchical* (broader term/narrower term), *associative* (see also (7)), and *equivalent* (use/used from or see/seen from). In addition, it is common in thesauri that some or all terms have scope notes, brief explanations of how the term should be used in indexing. Term history notes may also be present.

Controlled vocabularies are the broadest category, which includes thesauri and taxonomies. Thesauri and taxonomies are specific kinds of controlled vocabularies, but not all controlled vocabularies are thesauri or taxonomies.

The term ontology comes from philosophy and is the study of what exists. In artificial intelligence an **ontology** is a specification of a conceptualisation that is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. This description defines in a formal way the components such as individuals (instances of objects), classes, attributes and relations as well as restrictions, rules and axioms and what terminology is used for them. The main characteristic of ontologies is that they provide a sharable and reusable knowledge representation of a domain and can also contribute new knowledge about it. They can be modelled with RDF and RDFs. Both rely on a data model of graph structures consisting of basic elements called triples, in the form of "subject-predicate-object", which are also used for encoding more complex data structures like lists. Another established ontology modelling language is OWL. More details about ontology-modelling languages will be provided in section 3.3.

3.2. Automated reasoning

Automated reasoning (or inferencing) lies in the intersection of artificial intelligence, theoretical computer science and philosophy and it is about deriving information that is implied by the information already present. It is closely related to knowledge representation, since the former is useless without the ability to reason with it. As a semantic web methodology, reasoning can be used to discover new relationships between data that are modelled as a set of (named)

relationships between resources. In fact, it can be considered as the set of the automatic procedures used to generate new relationships based on the data and some additional information drawn upon knowledge representation techniques, usually formed as a vocabulary (ontology) or a set of rules (8). In general, ontology-based reasoning emphasizes on classification methods, particularly on 'classes and 'subclasses' definition and how individual resources can be associated to such classes and deriving the relationships among classes and their instances. RDFS, OWL, or SKOS are the tools of choice to define ontologies apply reasoning. Rule-based reasoning, on the other hand, concentrates on the definition of a general mechanism used to discover and generate new relationships based on existing ones. Rules Interchange Format RIF (9) has been developed to cover rule-based approaches, as a mechanism to exchange rules between rule-based languages. In this study we are mainly interested in ontology-based reasoning.

3.3. Knowledge representation and reasoning tools

3.3.1. Resource Description Framework RDF

The Resource Description Framework RDF (10) is a formal language for describing structured information. Based on a very simple graph-oriented data schema, it aims to enable the exchange of data on the Web between applications while preserving their original meaning and facilitating the processing and re-combination of the contained information. RDF is often viewed as the basic representation format used on the semantic web. As a simple graph-oriented data schema, every RDF graph can, in essence, be completely described by its edges and every such edge corresponds to an RDF triple "subject-predicate-object." The original graph can be split into smaller parts that can be stored one by one. Such a transformation of complex data structures into linear strings is called serialisation. There are many serialisations of RDF such as XML, Turtle, JSON-LD, N3 etc.

3.3.2. RDF Schema

RDF Schema denoted as RDFS (11) provides a data-modelling vocabulary for RDF data. RDFS is a semantic extension of RDF. It provides mechanisms for describing groups of related resources and the relationships between these resources. RDFS is an ontology language: An RDFS document is a machine-processable specification which describes knowledge about some domain of interest. Furthermore, RDFS documents have a formally defined meaning, given by the formal semantics of RDFS.

3.3.3. Web Ontology Language OWL

The W3C Web Ontology Language (OWL) (12) is a Semantic Web language, part of the W3C's semantic web technology stack, which also includes RDF, RDFS, SPARQL, etc. Its main purpose is to represent rich and complex knowledge about things, groups of things, and relations between them. OWL is a computational logic-based language that can be exploited to verify the consistency of that knowledge it expresses or to make implicit knowledge explicit. OWL ontologies can be published in the web and may refer to or be referred from other OWL ontologies.

3.3.4. Simple Knowledge Organisation System SKOS

Simple Knowledge Organisation System SKOS is an area of work developing specifications and standards to support the use of knowledge organisation systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web (13). SKOS provides a standard way to represent knowledge organisation systems using the Resource Description Framework (RDF) (14), allowing them to be passed between computer applications in an interoperable way and to be used in distributed, decentralised metadata applications, where metadata are harvested from multiple sources.

3.4. Semantic validation

Data shared between public organisations should be validated before publication. Validation rules should be defined according to the specification used for data modelling that re-uses terms from one or more vocabularies. Validation rules can be defined with, for example: Shapes Constraint Language SHACL, which is a W3C recommendation for validating RDF graphs under set of assumptions, or The Shape Expressions (ShEx) (15) language, which is uses shapes

to describe the triples involving nodes in an RDF graph. ShEx shapes can be used to communicate data structures associated with some process or interface, generate or validate data, or drive user interfaces.

3.4.1. SHACL

Shapes Constraint Language (SHACL) (16) is a World Wide Web Consortium (W3C) specification for validating graph-based data against a set of conditions. Among others, SHACL includes features to express conditions that constrain the number of values that a property may have, the type of such values, numeric ranges, string matching patterns, and logical combinations of such constraints. SHACL also includes an extension mechanism to express more complex conditions in languages such as SPARQL. A SHACL validation engine takes as input a data graph and a graph containing shapes declarations and produces a validation report that can be consumed by tools. The data and shapes graphs can be represented in any RDF serialisation format.

3.5. Query mechanisms

Up to now we have discussed semantic technologies that can be used to specify information in a machine-readable way. For example, RDF allows us to structure and relate pieces of information, and RDFS and OWL introduced further expressive means for describing more complex logical relations. But how exactly is this knowledge are accessed? Query mechanisms is the answer to this question. However, retrieving information from a knowledge base is not just a matter of query expressivity, but also addresses practical requirements such as post-processing and formatting of results. In addition, it is sometimes desirable to filter results using criteria that are not represented in the logical semantics of the underlying language. In the remaining of this section, we have a closer look at some important query languages for RDF and OWL such as SPARQL and GraphQ-LD

3.5.1. SPARQL

SPARQL is the standard language for querying the RDF data model. It is developed by the W3C Data Access Working Group and is associated with the SPARQL protocol for formulating queries across diverse data sources through the web (17). The data to be queried can be stored natively as RDF or viewed as RDF via middleware. SPARQL is based on simple queries in the form of simple graph patterns, but also provides a number of advanced capabilities such as functions for constructing advanced query patterns, for stating additional filtering conditions, and for formatting the final output, aggregation, subqueries, negation, creating values by expressions, extensible value testing, and constraining queries by source RDF graph. The results of SPARQL queries can be result sets or RDF graphs known as *answer graphs*.

3.5.2. GraphQ-LD

GraphQL is a query language that was introduced by Facebook as an alternative way of querying data through interfaces. It has been gaining increasing attention partly due to its simplicity in usage, and its large collection of supporting tools. However, since GraphQL queries represent trees and not full graphs, GraphQL is not as expressive as SPARQL, it has no notion of semantics and no notion of global identifiers. To overcome these shortcomings, GraphQL-LD (19) was introduced an extension of GraphQL with a JSON-LD context, so that it can be used to evaluate queries over RDF data. Although it is less expressive than SPARQL, it can still achieve many of the typical data retrieval tasks and it can translate GraphQL-LD queries to SPARQL algebra.

3.6. Specifications and standards for public organisations

The need for common modelling and interlinking of datasets and assets in public organisations have led many initiatives and standardisation bodies towards the creation of open standards that can be used and extended by public organisations to facilitate their needs and ease their activities.

The ISA^2 Programme supports the development of data models that enable public administrations, businesses and citizens in Europe to benefit from interoperable cross-border and cross-sector public services. In the following paragraphs we present the standards introduced by ISA^2 in the last few years, which have gained increasing interest from public bodies and organisations and have been adapted my many of them.

3.6.1. Core Vocabularies

Core Vocabularies (20) are simplified, reusable, and extensible data models that capture the fundamental characteristics of an entity, such as a person or a public organisation, in a context-neutral manner. The Core Vocabularies can become the basis for the design of context-specific data models, they can be used to annotate new and existing data models with mappings to the Core Vocabularies (21)

The **Core Vocabularies** are summarised in the following table:

Vocabulary	Description
Core Person	Captures the fundamental characteristics of a person, e.g. name, gender, date of birth, location.
Core Business	Captures the fundamental characteristics of a legal entity (e.g. its identifier, activities) which is created
Core Location	Captures the fundamental characteristics of a location, represented as an address, a geographic name or geometry.
Core Criterion and Core Evidence	Describes the principles and the means that a private entity must fulfil to become eligible or qualified to perform public services. A Criterion is a rule or a principle that is used to judge, evaluate or test something. An Evidence is a means to prove a Criterion.
Core Public Organisation:	Describes public organisations in the European Union.

Table 2 ISA² Core Vocabularies

Public organisations and administrations can use and extend the Core Vocabularies in the following contexts:

- **Information exchange between systems**: The Core Vocabularies can become the basis of a context-specific data model used to exchange data among existing information systems.
- **Data integration**: The Core Vocabularies can be used to integrate data that comes from disparate data sources.
- **Data publishing**: The Core Vocabularies can be used as the foundation of a common export format for data in base registries like cadasters, business registers and service portals.
- **Development of new systems**: The Core Vocabularies can be used as a default starting point for designing the conceptual and logical data models in newly developed information systems.

3.6.2. DCAT-AP

The Data Catalogue Vocabulary (DCAT) developed by W3C is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web (22). Using DCAT to describe datasets in data catalogues, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogues. It further enables decentralized publishing of catalogues and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.

The DCAT Application Profile for Data Portals in Europe (DCAT-AP) (23) is a specification based on the DCAT focusing on public sector data and datasets. It facilitates better organisation of public sector data and aims at improved discoverability of datasets. Its basic use case is to enable a cross-data portal search for data sets and make public sector data better searchable across borders and sectors. DCAT-AP provides a common specification for describing public sector datasets in Europe to enable the exchange of descriptions of datasets among data portals. DCAT-AP allows:

- **Data catalogues** to describe their dataset collections using a standardised description, while keeping their own system for documenting and storing them.
- **Content aggregators**, such as the European Data Portal (24) to aggregate such descriptions into a single point of access.

- **Data consumers** to more easily find datasets from a single point of access.

DCAT and DCAT-AP have inspired the creation of other application profiles for geographical GeoDCAT-AP (25) and statistical StatDCAT-AP (26) as illustrated in the following table.

DCAT Profiles	Description
DCAT	Facilitates interoperability between data catalogues published on the Web.
DCAT-AP	Describes geospatial datasets, dataset series and services.
GeoDCAT-AP	Describes geospatial datasets, dataset series, and services. It provides an RDF syntax binding for the union of metadata elements defined in the core profile of ISO 19115:2003 and those defined in the framework of the INSPIRE Directive.
StatDCAT-AP	Provides a dissemination vocabulary for statistical open data, defining a number of additions to the DCAT-AP model that can be used to describe statistical datasets.

Table 3 DCAT and DCAT profiles

3.6.3. ADMS

The Asset Description Metadata Schema (ADMS) (27) is a simple specification used to describe reusable solutions, such as data models and specifications, reference data and open source software, enabling organisation to document them and helping everyone to search and discover them. Specifically, ADMS targets different groups of professional within organisations and allows:

- Solution providers, such as standardisation organisations and public administrations, to describe their interoperability solutions using the standardised descriptive metadata terms of ADMS, while keeping their own system for documenting and storing them;
- Content aggregators to aggregate such descriptions into a single point of access;
- ICT developers to more easily explore, find, identify, select and obtain interoperability solutions from a single point of access.

3.7. How semantic web technologies improve data quality

Public organisations like base registers, tax authorities etc. hold assets of high value that present a reuse potential both in national and in cross-national level. Providing these data in the form of semantic assets such as ontologies, schemata, domain models, controlled lists, catalogues, taxonomies and reference datasets and making them available through the use of semantic technologies in machine-readable formats will definitely increase its reuse and discoverability and promote interoperability across organisations.

Public organisations and administrations are strongly encouraged to employ knowledge representation methodologies to model their data, adapt widely-accepted, reusable and extendible vocabularies and specifications such as DCAT-AP, ADMS and Core Vocabularies to better describe, organise their assets and make them reusable and interoperable in semantic level, use validation and querying mechanisms to ensure the integrity of their data, reason upon their ontological models and link them with external resources so as to extract new domain knowledge and eventually enhance their data quality and facilitate interoperability.

- **Semantic accuracy** is promoted since ontologies express relationships and enable linking of multiple concepts to other concepts in a variety of ways.
- The interconnectedness and interoperability of ontological data models facilitates data **uniqueness.** The usage of controlled vocabularies as alphabetically ordered list of concepts (terms), explicitly enumerated and

provided with unambiguous, non-redundant definitions resolves terminology inconsistencies and data redundancies, in breadth, depth and scope and promote uniqueness of data.

- **Completeness** of data is improved since ontological reasoning can be used to make implicit knowledge explicit and resolve missing data issues.

- By ensuring a common understanding of information and by making explicit domain assumptions, ontologies and ontological reasoning improve data **consistency**. Also, by improving metadata and provenance, and by allowing organisations to make better sense of their data, ontologies enhance data quality, in terms **completeness** and **consistency.**

- Data **consistency**, **accuracy** and **completeness** are improved through reasoning. Inference improves the quality of data integration by discovering new relationships (completeness), automatically analysing the content of the data, and managing knowledge in general. Inference based techniques are also capable of discovering and then resolving possible inconsistencies (consistency) in the (integrated) data and can resolve missing data values (completeness).

- The extensibility and extendibility of ontologies enables new relationships and concept matching to be easily added to existing ontologies. This implies that ontological data models evolve with the growth of data without impacting dependent processes and systems if something goes wrong or needs to be changed. This way ontologies impact the **timeliness** dimension of data quality.

- Data **validity** and **syntactic accuracy** is promoted through semantic validation. An organisation can use SHACL specification for representation and validation of its governance data, to ensure that data are valid and conform to the syntax (format, type, range) of its definition, and achieve improved data quality and management.

Semantic technology	Tool	Data Quality Dimension
Ontologies and Reasoning	RDF, RDFS, OWL, SKOS	Accuracy, consistency, uniqueness, timeliness
Validation	SHACL, ShEX	Validity, syntactic accuracy
Querying	SPARQL, GRAPHQL-LD	Consistency,

Table 4 Impact of semantic technologies on data quality dimensions

Besides data quality dimensions, there are other aspects related to data quality that are impacted by the use of semantic web technologies.

- **Representation, integrations and analytics**. Ontologies provide the means to represent any data formats, including unstructured, semi-structured or structured data, enabling smoother data integration, easier concept and text mining, and data-driven analytics.

- **Access and discoverability**. Ontological data models can address the challenges of accessing and querying data in large organisations. They also improve efficiency and quality of data with a view of its subsequent search and/or analysis.

- **Granularity.** Ontological data models and hierarchical knowledge representation systems (e.g. SKOS representation) can achieve different granularities of data (i.e. level of detail).

- **Interlinking and navigation.** One of the main features of ontologies is that, by having the essential relationships between concepts built into them and by enabling automated reasoning about details that they 'work and reason' with concepts and relationships in ways that are close to the way humans perceive interlinked concepts. In addition to the reasoning feature, they provide a more coherent and easy navigation from one concept to another within the ontology structure.

- **Relevancy**. Efficient query mechanisms can improve the relevancy of the data. Relevancy refers to the provision of information which is in accordance with the task at hand and important to the users' query. Relevancy is highly context dependent and is highly recommended in information systems dealing with big flow of information since the process of retrieving the relevant information becomes complicated.

4. SEMANTIC ENRICHMENT FOR IMPROVED DATA QUALITY

Up to now we have focused on the main semantic web technologies that can be used to promote data quality within public organisations and administrations. The basic idea is to use semantic web technologies to model and link available data, inference, validate and query them, thus improving their accuracy and consistency, contributing to its completeness, ensuring their uniqueness and validity.

Apart from semantic web technologies, there are other cutting-edge methodologies and tools from various areas of artificial intelligence (such as machine learning, natural language processing etc.) that can be employed by public organisations together with semantic web technologies to improve and promote data quality and at the same time ensure interoperability.

4.1. Semantic enrichment

Interoperability in organisations that handle administrative data is essential to improve data quality. One of the main barriers of interoperability is data discoverability. Data discovery starts with the metadata since metadata describes and provides useful information about dataset entities and assets. Locating the available datasets is often difficult or even impossible due to poor or missing metadata. The existence of good quality metadata facilitates easy access, discoverability, comprehension and preservation over time. Public organisations can enrich their content with an array of metadata with the aim of ensuring that content is distributed broadly, adaptable for multiple purposes, and rendered interoperable with other relevant content.

Generally, a semantic enrichment process aims at improving (meta)data about an entity or asset by adding new statements about it. The term enrichment is usually used to refer to the methodology followed or its result, which is the new (meta)data obtained at the end of the process. A semantic enrichments process can be manual, semi-automatic (combination of manual and automatic) or automatic (e.g. by means of information extraction) (30).

The notion of manual enrichment is not new, since humans have been performing this task during the last decades. Currently manual enrichment has evolved into a process where domain experts perform the enrichment task by adding tags and linking objects to available (linked)data resources. Within the scope of manual enrichment, crowdsourcing efforts seek to empower more people to create new (meta)data about objects. Furthermore, there are initiatives that seek to guide crowdsourcing efforts into creating finer-grained metadata, shifting the focus from basic tags to a richer form by linking to contextual resources.

Although manual enrichment processes are precise and accurate when performed by experts and qualified annotators, they are still limited in their coverage. This is due to the fact that the number of items that need to be enriched in datasets appears too high compared to the available human resources. Even though crowdsourcing efforts aim to provide a solution, such efforts are often limited to quite specific datasets or collections, and with a specific objective in mind (e.g., focusing on specific data fields). Automatic enrichment attempts to solve the problem of limited coverage of manual enrichment processes, by employing various computational methodologies of information extraction to automatically enrich available metadata and extract new one.

The main components involved in semantic enrichment (30) are the following:

- **Source:** the source objects, that is assets or dataset items, whose set of (meta)data is being enriched or extended

- **Target:** the set of resources used to enrich the source (meta)data, i.e. the values that will appear at the end of the process as in the enriched metadata set for the source. Targets can be of different types, from simple uncontrolled strings to resources published as linked data.

- **Rules:** they specify how the enrichment process between the source and target should be performed. In automatic enrichment, rules are of the form of instructions to create links based on matches between the various string representations attached to the resources in the source and the target.

The results of the semantic enrichment can be:

- Simple tags: non-semantic strings are attached to the object in order to describe it. However, the exact relation between the object and the string remains unknown.

- Typed links: links of a certain type between the source object and other resources (e.g. linking the resource representing an entity with a resource representing a concept, or a label for such a concept). In the RDF model, these enrichments are a set of RDF statements.

Depending on the type of link, the enrichment result can be an equivalence or other semantic relationship (*broader/narrow*er), or any domain relationship (*dc:subject, dc:*relation). The linked resources can be of same type (e.g. two objects, places, concept etc.) or they can be of different types (e.g. an object is linked to a conceptual subject). For example, (30):

  - A semantic equivalence relationship (using *owl:sameAs* or skos:exactMatch) between resources of same type, is called **co-referencing.**

  - **A** semantic relationship between resources of the same type from two different Knowledge Organisation Systems (KOS) is called **alignment** (or matching).

  - **A** typed relationship between resources of different type is called **contextualisation**.

Finally, the semantic enrichment process may be divided into the following tasks (31):

- **Analysis:** Analysing the available (meta)data in the original source descriptions, selecting potential target(s) and creating rules to enrich the source with the target. Deciding about the enrichment methodologies can also be part of this analysis activity, since it can have impact on the choice of targets and the definition of rules.

- **Linking:** (automatically) applying the rules to connect the source resources to the target ones.

- **Augmentation:** adding more values from the target data to the original asset (meta)data, after the basic enrichment data has been produced. For instance, when an object is enriched with a SKOS concept It could include data about broader or narrower concepts.

4.2. Machine learning for semantic enrichment

The rise of artificial intelligence, machine learning and neural networks in combination with semantic web technologies provide the means to automate enrichment processes, at least partially. Current developments in the aforementioned fields has demonstrated impressive outcomes in a variety of domains, in which intelligent models perform at least as well as and sometimes even better than humans.

Machine learning systems present the ability to automatically learn and improve from experience without using explicit instructions, by only relying on patterns and inference instead. Thus, such systems are perfect candidates for performing automated semantic enrichment tasks. While machine learning systems continuously learn from raw, unstructured data, and extract knowledge, advanced AI systems are able to combine this newly extracted form of knowledge with existing knowledge from ontologies and structured data in order to produce at the same time new knowledge, answers and explanations, leading to richer, more complete and up to date metadata. Therefore, recent trends in machine learning are in the direction of coupling knowledge completion, approximate inferencing and automatic reasoning with data-

driven statistical and neural network-based approaches. The principled combination of knowledge representation, reasoning and learning (32) can provide new powerful ways of semantic enrichment that will exploit structured and unstructured data sources to extract new knowledge about existing data in terms of enriched and new metadata.

Very often the original (source) metadata or the content itself (text documents, audio files, images, maps, etc.) contain mentions of the concepts, places and other contextual resources that are in the target datasets. This allows the utilisation of machine learning methodologies and tools that can exploit such traces to create semantic links between source and target resources. In other cases, the actual data assets and/or their metadata can be processed by deep learning systems in order to extract new and even structured information that can serve as new metadata.

By utilizing machine learning systems to process data assets, enrich semantically their metadata and extract new ones, organisations content owners and distributors can improve the quality of their data in terms of accuracy, completeness and increase the discoverability of their assets.

4.2.1. Natural language processing and deep learning

Natural language processing (NLP) entails the application of algorithms to identify and extract the natural language rules such that the unstructured language data is converted into a form that is machine understandable. The basic goal of NLP is to process the unstructured text and to produce a representation of its meaning.

Since textual metadata are the main descriptive metadata of the items (assets, records etc.) of an organisation, they can undergo NLP and machine learning based information extraction in order to associate unstructured text with a structured representation of its meaning. Various tasks of information extraction: Named entity recognition, named entity linking, temporal information extraction, relation extraction, knowledge base construction and reasoning. Named entity recognition and disambiguation (NERD) is an important task in the pipeline of information extraction from textual metadata, in order to identify in them occurrences of *named entities* i.e. predefined categories such as person names, organisations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. and associate them to linked data resources. Natural language processing can be combined with dictionary lookup methodologies and dictionary generation techniques, in order to compile a fast to search dictionary of resources from a selected subset of the available predefined thesauri, scan the textual metadata to extract occurrences of the dictionary terms and improve dictionary accuracy.

Recently, deep learning methods have been very successful in the area of natural language processing (33), achieving very high performance across many different tasks. Deep learning methods employ multiple processing layers to learn hierarchical representations of data and have produced state-of-the-art results in many domains. There is a large variety of underlying tasks and machine learning models powering NLP applications. Recently, deep learning approaches have achieved very high performance across many different NLP tasks. Convolutional Neural Networks, Recurrent neural networks, attention mechanisms, representation learning are some of the main deep learning methodologies used to assist nature language processing.

4.2.2. Human in the loop and active learning

Along with the evolution of artificial intelligence, crowdsourcing has matured and evolved into a more pragmatic approach, where the crowd produces metadata for datasets that serve as trainers for intelligent systems. This new form of crowdsourcing has created a variety of new opportunities for improving upon methods of semantic annotation, thus creating intriguing new opportunities for data-driven machine learning. Crowdsourcing leverages more arbitrary crowd-based human computation to supplement automated machine learning tasks.

The concept of Human in the loop leverages both human and machine intelligence to create machine learning models, where humans are directly involved in training, tuning and testing data for a particular machine learning algorithm. There is a special category of such algorithms, referred to as active learning, where the learning process is assisted by humans in cases where the system's confidence is below an accepted threshold.

The practice of human in the loop in combination with active learning algorithms serves as a powerful tool for semantic annotation in the context of metadata enrichment (34). The benefit of such an approach is that machine intelligent, accuracy and precision are combined with human intelligence that usually entails expert knowledge to derive high quality metadata. Depending on the application and the dataset that need to be enriched in terms of metadata, active learning approaches can be adjusted.

4.3. Prerequisites for semantic enrichment

Semantic enrichment can be applied to structured, semi-structured and even unstructured data. Depending on the nature of data different semantic enrichment approaches can be followed with different workflows and varying requirements in resources and investment (35).

Structured data are catalogues of all kinds, databases, curated datasets, metadata repositories, name authorities, Knowledge Organisations Systems. They are usually stored in databases; they follow an explicit data model and all key/value pairs have identifiers and clear relations. Semantic enrichment of structured data is normally applied to components in metadata records where data values are available in a controlled/normalized form, (e.g. entities for place, agent, concept, and time period). The *source* and *target* components involved in the process can be metadata descriptions of any standard, KOS vocabularies and other contextual resources (e.g., GeoNames, Wikidata, DBpedia, etc.), or information resources (e.g., Wikipedia entries, biographies, geo-maps) where the focused subjects are the entities in metadata descriptions or KOS vocabularies. The target and source position and the directions of linking can be switched depending on the enrichment needs. Metadata *alignment, co-referencing* and c*ontextualisation* can be easily performed on structured data as described in section 4. There are multiple cases from the cultural domain (35) proving that semantic enrichment has been successfully applied to enhance the quality of structured data with significant impact, thus encouraging the application of semantic enrichment other domains such as governance.

Semi-structured data are data with unstructured sections within metadata descriptions, or unstructured parts of otherwise structured datasets. Semi-structured data enrichment can be powered by multiple taxonomies and domain ontologies, and benefit from machine learning and other artificial intelligence technologies, such as NLP as described in sections 4.2 and 4.2.1. Different taxonomies can be employed to classify the extracted entities and different knowledge bases can be utilized to disambiguate them. A simple workflow of such type can be seen as recognizing named entities mentioned in text, assigning them as pre-defined types, and linking them with their matching entities in a knowledge base.

Unstructured data are documents, texts and all kinds of media data. These kinds of data present great diversity in type, nature, and quality, and are the most challenging to process. For unstructured data, the enrichment process identifies relations between concepts in documents and associates the unstructured data with a context that is further linked to the structured knowledge of a domain. The process relies both on human and machine actions. Crowdsourcing is coupled with advanced AI systems to extract knowledge from raw data and combine it with existing knowledge from ontologies and structured data in order to produce new knowledge, and deliver richer, more complete and up to date metadata. Ontology-based approaches combined with machine learning methods provide mechanisms for new knowledge extraction and to this end, Linked Data contributes in making semantic enrichment possible and effective, utilizing the web to connect related data that wasn't previously linked.

Semi-structured and unstructured data require more complicated semantic enrichment workflows (e.g. model developing, batch processing, validating, disseminating, etc.), might need significant additional investments, resources as well as human effort and computational power.

Another key issue in the application of semantic enrichment is the proper evaluation and selection of enrichment targets, the set of resources used to enrich the source data, as this plays a key role to the quality of enrichments and consequently to the quality of the enhanced data (30), (31). The selection should be made based on good knowledge of the source data usually obtained after analysis, the identified quality issues that need to fix through the enrichment process, the evaluation of available target datasets with respect to availability, access, granularity, coverage, quality, connectivity etc. Depending on the form and the type of source dataset to be enriched, there are case where there is no available target set. In such cases the target set should be constructed.

Finally, the results of the semantic enrichment process should provide appropriate representations of enrichments in terms of the format and standard they follow. It is also important to publish metadata about the enrichment, e.g., provenance information about how the enrichment was provided), confidence on its correctness. (30). Such information enables organisations to monitor the quality of the (enriched) data and utilize only data with the required characteristics (including performance against quality indicators) for particular purposes.

5. CONCLUSIONS

This study explored the intersection between the data quality management (from a data governance point of view) and semantic interoperability: how semantic assets can support and evolve data quality considerations. It described fundamental concepts for data quality and linked those to semantic interoperability studying how, data quality can be improved by adapting a semantic approach for data representation and organisation. At first level, the data quality metrics introduced by DAMA UK WG were demonstrated, which are utilized to evaluate the Data Quality. Then, the main focus of the study was on the trends of semantic web technologies and how they can impact data quality dimensions, improve data quality and promote interoperability between public organisations. In particular, among the key elements to reaching high data quality, i.e. *people*, *processes* and *technology*, we focused on technology. We illustrated that by employing knowledge representation technologies (ontologies, thesauri, vocabularies, open standards) and mechanisms to model and organise governance data, public organisations can improve their quality of data and achieve interoperability. The use of ontologies enables automated reasoning, which can infer new relationships and properties and thus contribute in data accuracy and completeness. Semantic web query languages can be used to enhance the relevance of data. RDF validation mechanisms (i.e. SHACL) can improve the integrity and semantic accuracy of data. Finally, we presented how semantic enrichment of metadata can be reached. Machine learning techniques like natural language processing combined with deep learning can be used to systematically enhance the quality of governance data (structured, semi-structured and unstructured) and in combination with human in the loop methodologies can improve data discoverability and accessibility and provide data of high accuracy and completeness. Semantic enrichment can be applied to structured, semi-structured and unstructured data. Depending on the nature of data, different semantic enrichment approaches can be followed with different workflows and varying requirements in resources and investment.

Annex I. QUALITY IN LINKED DATA

As part of the semantic web, **linked data** is built upon semantic web technologies. Linked data enables us to relate data by generating semantic connections among datasets and thus improve the quality of data in many ways. The basic principles for publishing and interlinking structured data on the web (34) are:

1. **Use URIs as names for things.** The use of URIs is encouraged to identify things. As in the web of documents, in linked data, a URI is used to identify a document describing an entity.

2. **Use HTTP URIs so those names can be dereferenced.** The use of the identification mechanism (URIs) is advocated through specific protocols such as the application-level protocol HTTP, to achieve interoperability between independent information systems

3. **Provide useful information by using the standards (RDF, SPARQL) upon dereferencing of those URIs.** It is assumed that each URI identifying an entity is dereferenceable

4. **Include links using externally dereferenceable URIs to discover more things**. Linked data distributed across the web apply a standard mechanism for specifying the connections between real-world objects

Linked data connects entities and the RDF links enable the process of discovering, accessing, and integrating data in a straightforward way.

Quality Dimensions in Linked Data

The above principles measure how much a dataset conforms to the linked data principles (36). In general, measuring the quality means evaluating a set of dimensions that capture specific aspects of data quality. Linked data quality dimensions definition poses a number of unique challenges. These are:

- Linked data refers to a web-scale knowledge base consisting of interlinked published data from a multitude of autonomous information providers. The quality of provided information may depend on the intention of the data provider. Among other issues, linked data providers may publish datasets with incomplete or inaccurate metadata that influence the quality of the datasets themselves.

- The increasing diffusion of the linked data paradigm allows consumers to fully exploit vast amount of data that were not available in the past. Intuitively, as the size of data increases, it becomes more and more difficult to assess the quality of data.

- Datasets in linked data may often be used by third-party applications in ways not expected by the original creators of the dataset.

- Linked data provides data integration through interlinking data between heterogeneous data sources. The quality of integrated data is related to the quality of original data sources, which is not straightforward to be modelled.

- Relevant linked data can be considered as a dynamic environment where information can change rapidly and cannot be assumed to be static (velocity of data). Changes in linked data sources should reflect changes in the real world; otherwise, data can soon become outdated. Out-of-date information can reflect data inaccuracy problems and can deliver invalid information.

Quality of linked data includes a number of novel aspects, such as coherence via links to external datasets, data representation quality, or consistency with regard to implicit information. There have been efforts that evaluate how the state of the art in data quality research fits the characteristics of linked data, and how semantic technologies and tools(like SPARQL and SPIN (37)) can be utilized to identify data quality problems in linked data automatically (38).

The quality dimensions of linked data are further complicated by the fact that both the Closed World Assumption (CWA) and the Open World Assumption (OWA) can hold. While CWA is the usual assumption to hold, the interconnected nature

of linked data makes OWA the natural assumption, which has an impact on the difficulty of defining and evaluating the compliance between data and schema. A relation between two instances can hold even if the schema does not model such relation between the concepts the instances belong to; conversely, we cannot conclude that a relation between two concepts of different schemas does not hold because it is not represented in the data instances. Usually in the literature on linked data, the CWA is implicitly assumed to hold for the definition and assessment of quality dimensions, such as completeness and consistency.

Additionally, considering that in a linked data schemaless approach is often followed (that means that RDF data are fist published and subsequently and optionally the schema is specified), in order to ensure the quality of RDF data in terms of accuracy and consistency, there have been tools and techniques developed for a-posteriori validation of RDF data.

Annex II. GLOSSARY

Term/Acronym	Description
ADMS	Asset Description Metadata Schema, a simple specification used to describe interoperability solutions helping everyone to search and discover them
Controlled vocabularies	Organised arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.
Core Vocabularies	Core Vocabularies are simplified, reusable, and extensible data models that capture the fundamental characteristics of an entity, such as a person or a public organisation, in a context-neutral manner.
Data Quality	The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use. (3)
Data Quality Management	Data quality management is a set of practices that aim at maintaining a high quality of information.
DCAT	Data Catalogue Vocabulary DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web
DCAT-AP	DCAT Application Profile for Data Portals in Europe (DCAT-AP) is a specification based on the Data Catalogue Vocabulary (DCAT) developed by W3C.
Inferencing	A rule or process that derives a new fact from a given set of facts. There are three main methods: deduction, abduction, and induction. Examples of these styles of inference can be seen in theorem proving, expert systems, and machine learning, respectively.
Interoperability	According to the ISA Decision, interoperability means the ability of disparate and diverse organisations to interact towards mutually beneficial and agreed common goals, involving the sharing of information and knowledge between the organisations, through the business
ISA²	The ISA² Programme supports the development of digital solutions that enable public administrations, businesses and citizens in Europe to benefit from interoperable cross-border and cross-sector public service
ISO	International Standardisation Organisation. Independent, non-governmental international organisation with a membership of 164 national standards bodies.
Linked Data	The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF provides the foundation for publishing and linking your data. Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files.
Metadata	Data about data – that is, data describing the structure, content or use of some other data.
OASIS	Non-profit consortium that drives the development, convergence and adoption of open standards for the global information society.
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
Semantic Interoperability	Semantic interoperability enables organisations to process information from external sources in a meaningful manner. It ensures that the precise meaning of exchanged information is understood and preserved throughout exchanges between parties.
Semantics	The term semantics (from Greek σημαντικός "significant") is used in many different contexts (like logic, linguistics, or programming languages) Probably the most appropriate corresponding English term is "meaning." (5)
SHACL	Shapes Constraint Language
ShEx	The Shape Expressions Language
SKOS	Simple Knowledge Organisation System
SPARQL	SPARQL Query Language for RDF
Structured data	Data that has enforced composition to specified datatypes and relationships and is managed by technology that allows for querying and reporting.

Taxonomy	A controlled vocabulary with a hierarchical structure. Terms within a taxonomy have relations to other terms (parent/broader term, child/narrower term)
Thesaurus	A controlled vocabulary where all terms have relationships of three kinds to each other: *hierarchical* (broader term/narrower term), *associative* and *equivalent* (use/used from or see/seen from).
Turtle	Terse RDF Triple Language (Turtle) is a syntax and file format for expressing data in the RDF data model.
Unstructured data	Computerised information which does not have a data structure that is easily readable by a machine, including audio, video and unstructured text such as the body of a word-processed document – effectively this is the same as multimedia data.
URI	Uniform Resource Identifier
Vocabularies	At times it may be important or valuable to organise data. Using OWL (to build vocabularies, or "ontologies") and SKOS (for designing knowledge organisation systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data
W3C	The World Wide Web Consortium (W3C) is an international community that develops open standards to ensure the long-term growth of the Web.
XML	eXtensible Markup Language., a mark-up language designed to store and transport data.

Table 5 Glossary

Annex III. REFERENCES

(1) *The DAMA Guide to the Data Management Body of Kno*wledge (DAMA-DMBOK by DAMA International, Publisher: Technics Publications, *2009*

(2) *The* six primary dimensions for data quality assessment, DAMA Working GROUP, white paper, 2013

(3) DAMA UK, working group, https://www.damauk.org/

(4) J.F. Sowa. *Knowledge Representation*. Brooks Cole Publishing, Pacific Grove, CA, USA, 2000

(5) P. Hitzler, M. Krtzsch, S. Rudolph, *Foundations of Semantic Web Technol*ogies, Chapman & Hall/CRC, 2009

(6) About Taxonomies & Controlled Vocabularies, http://www.taxonomies-sig.org/about.htm

(7) http://www.taxonomies-sig.org/about.htm

(8) What is inference, at: https://www.w3.org/standards/semanticweb/inference

(9) Rule Interchange Format, https://en.wikipedia.org/wiki/Rule_Interchange_Format

(10) Resource Description Framework, https://www.w3.org/RDF/

(11) RDF Schema Semantic Web Standards, https://www.w3.org/2001/sw/wiki/RDFS

(12) OWL Semantic Web Standards, https://www.w3.org/OWL/

(13) SKOS, Simple Knowledge Organisation System, https://www.w3.org/2004/02/skos/

(14) SKOS, Simple Knowledge Organisation System Reference, https://www.w3.org/TR/skos-reference/

(15) Shape Expressions Language 2.1 , http://shex.io/shex-semantics/

(16) Shape Constraint Language, SHACL, https://www.w3.org/TR/shacl/#shacl-rdfs

(17) SPARQL https://www.w3.org/TR/rdf-sparql-query/

(18) SPARQL Protocol for RDF, https://www.w3.org/TR/rdf-sparql-protocol/

(19) R. Taelman, M. Vander Sande, and R. Verborgh, *GraphQL-LD: Linked Data Querying with GraphQL*, available at : https://ruben.verborgh.org/publications/taelman_iswc_demo_2018/

(20) Core Vocabularies, https://ec.europa.eu/isa2/solutions/core-vocabularies_en

(21) e-Government Core Vocabularies handbook, available at: https://ec.europa.eu/isa2/sites/isa/files/e-government_core_vocabularies_handbook.pdf

(22) Data Catalog Vocabulary, https://www.w3.org/TR/vocab-dcat/

(23) DCAT Application Profile, https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en

(24) European Data Portal, https://www.europeandataportal.eu/en/homepage

(25) GeoDCAT-AP, https://joinup.ec.europa.eu/release/geodcat-ap/v101

(26) StatDCAT-AP, https://joinup.ec.europa.eu/release/statdcat-ap-v100

(27) Asset Description Metadata Schema ADMS, https://ec.europa.eu/isa2/solutions/asset-description-metadata-schema-adms_en

(28) N. Loutas et al. Realising a Federation of Repositories of Reusable Metadata. International Conference on Dublin Core and Metadata Applications, [S.l.], p. 156-161, sep. 2013. ISSN 1939-1366. Available at: http://dcpapers.dublincore.org/pubs/article/view/3689/1912 .

(29) Joinup, https://ec.europa.eu/isa2/solutions/joinup_en

(30) *Report on the enrichment and the evaluation*, Europeana, 2015, https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf

(31) *Multilingual and Semantic Enrichment Strategy*, Europeana, 2014: http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/MultilingualSemanticEnrichment//Multilingual%20Semantic%20Enrichment%20report.pdf

(32) L. Serafini and A. S. d'Avila Garcez, *Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge,* ArXiv, 2016, available at*: https://arxiv.org/abs/1606.04422*

(33) T. Young, D. Hazarika, S. Poria and E. Cambria, *Recent Trends in Deep Learning Based Natural Language Processing* [Review Article], in IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55-75, Aug. 2018.

(34) A. Chortaras et al., 2018, *WITH: Human-Computer Collaboration for Data Annotation and Enrichment*. In Proceedings of the The Web Conference 2018 (WWW '18), pp. 1117-1125., available at: http://linkeddata.org/docs/ijswis-special-issue

(35) M. L. Zeng, *Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article*. El profesional de la información, v. 28, n. 1, e280103, 2019, https://doi.org/10.3145/epi.2019.ene.03

(36) A. Rula, A. Maurino, C. Batini, *Data Quality Issues in Linked Data*, Chapter from Book: Information Quality in Healthcare (pp.87-112), 2016

(37) SPIN SPARQL Inferencing Notation https://spinrdf.org/

(38) Fürber C., Hepp M. (2010) Using SPARQL and SPIN for Data Quality Management on the Semantic Web. In: Abramowicz W., Tolksdorf R. (eds) Business Information Systems. BIS 2010. Lecture Notes in Business Information Processing, vol 47. Springer, Berlin, Heidelberg

.

An action supported by ISA²

ISA² is a EUR 131 million programme of the European Commission which develops digital solutions that enable interoperable cross-border and cross-sector public services, for the benefit of public administrations, businesses and citizens across the EU. ISA² supports a wide range of activities and solutions, among which is the Semantic Interoperability Community (SEMIC) action.

ISA² solutions can be used free of charge and are open source when related to IT.

More on the programme

ec.europa.eu/isa2

Contact ISA²

isa2@ec.europa.eu

Follow us

@EU_ISA2
@Joinup_eu

isa² programme