ABSTRACT

The present deliverable provides an initial discussion of data management challenges in the context of CPaaS.io. It describes the state of the art regarding semantic vocabularies relevant to data (quality) management in general and specifically for IoT and discusses prevalent mechanisms to support data protection and privacy, putting a focus on citizen empowerment based on MyData concepts. The state of the art description provides input for answering the question on how the semantic integration layer of CPaaS.io can or should be designed and lays the grounds for developing a cross technology data management model needed in that respect.
Disclaimer

This document has been produced in the context of the CPaaS.io project which is jointly funded by the European Commission (grant agreement n° 723076) and NICT from Japan (management number 18302). All information provided in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and NICT have no liability in respect of this document, which is merely representing the view of the project consortium. This document is subject to change without notice.

Document Information

<table>
<thead>
<tr>
<th>Editors</th>
<th>Marianne Fraefel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Adrian Gschwend, Marianne Fraefel, Stephan Haller, Jan Frecé, Katsunori Shindo, Martin Strohbach</td>
</tr>
<tr>
<td>Reviewers</td>
<td>Antonio Skarmeta; Toshihiko Yamakami</td>
</tr>
<tr>
<td>Delivery Type</td>
<td>R</td>
</tr>
<tr>
<td>Dissemination Level</td>
<td>Public</td>
</tr>
<tr>
<td>Contractual Delivery Date</td>
<td>31.12.2016</td>
</tr>
<tr>
<td>Actual Delivery Date</td>
<td>27.12.2016</td>
</tr>
<tr>
<td>Keywords</td>
<td>Data Governance, Data Management, Data Quality</td>
</tr>
</tbody>
</table>
## Revision History

<table>
<thead>
<tr>
<th>Rev.</th>
<th>Date</th>
<th>Description</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>V.01</td>
<td>09.09.2016</td>
<td>ToC</td>
<td>Marianne Fraefel</td>
</tr>
<tr>
<td>V.02</td>
<td>10.10.2016</td>
<td>First draft SOTA description (preliminary)</td>
<td>Marianne Fraefel</td>
</tr>
<tr>
<td>V.03</td>
<td>18.10.2016</td>
<td>Overview Best Practices Linked and Open Data Comments Personal Data section</td>
<td>Adrian Gschwend</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Marianne Fraefel</td>
</tr>
<tr>
<td>V0.4</td>
<td>01.11.2016</td>
<td>Data on the Web best practices</td>
<td>Adrian Gschwend</td>
</tr>
<tr>
<td>V0.5</td>
<td>15.11.2016</td>
<td>Introduction Draft Best Practices OGD</td>
<td>Stephan Haller, Marianne Fraefel</td>
</tr>
<tr>
<td>V0.6</td>
<td></td>
<td>Draft Fiware / U2 Data Models Draft User Empowerment Draft IoT Data / Social Media Data Draft Best Practices OGD Draft IoT Data / Social Media Data Draft Best Practices OGD Changes Introduction</td>
<td>Adrian Gschwend, Marianne Fraefel, Katsunori Shindo, Stephan Haller</td>
</tr>
<tr>
<td>V0.7</td>
<td>30.11.2016</td>
<td>Amendments of Best Practices Revision of Drafts</td>
<td>Marianne Fraefel, Adrian Gschwend</td>
</tr>
<tr>
<td>V0.8</td>
<td>30.11.2016</td>
<td>Document completed for Feedback</td>
<td>Marianne Fraefel, Adrian Gschwend</td>
</tr>
<tr>
<td>V0.9</td>
<td>15.12.2016</td>
<td>Append Deployment Tool AGT Integration of Feedback by partners Revision of Text</td>
<td>Martin Strohbach, Marianne Fraefel, Adrian Gschwend, Jan Frecé</td>
</tr>
<tr>
<td>V0.10</td>
<td>20.12.2016</td>
<td>Append Data Protection Japan</td>
<td>Katsunori Shindo</td>
</tr>
<tr>
<td>V0.11</td>
<td>21.12.2016</td>
<td>Revision of Personal Data Section</td>
<td>Jan Frecé</td>
</tr>
<tr>
<td>V0.12</td>
<td>21.12.2016</td>
<td>Document completed for Review</td>
<td>Marianne Fraefel</td>
</tr>
<tr>
<td>V0.13</td>
<td>23.12.2016</td>
<td>Addressed reviewer comments</td>
<td>Marianne Fraefel, Jan Frecé, Adrian Gschwend</td>
</tr>
<tr>
<td>V1.0</td>
<td>23.12.2016</td>
<td>Final Document</td>
<td>Marianne Fraefel</td>
</tr>
</tbody>
</table>
# Table of Contents

1. **Introduction** .................................................................................................................................................................................7

2. **Data Management: Dimensions** .................................................................................................................................................9
   2.1 Data Management & Data Governance .................................................................................................................................9
   2.2 Data Quality ...............................................................................................................................................................................10
   2.3 Data Lifecycle ...........................................................................................................................................................................16
   2.4 Data Access ..............................................................................................................................................................................17
   2.5 Implications for CPaaS.io .......................................................................................................................................................17

3. **Data Management: Best Practices** .............................................................................................................................................18
   3.1 Content Overview .................................................................................................................................................................18
   3.2 FIWARE and u2 Data Models ..............................................................................................................................................20
   3.3 Linked Data ............................................................................................................................................................................22
   3.4 Open (Government) Data ....................................................................................................................................................37
   3.5 IoT Data ..................................................................................................................................................................................44
   3.6 Social Media Data .................................................................................................................................................................48
   3.7 Personal Data / User Empowerment ....................................................................................................................................49

4. **Conclusions & Next Steps** .........................................................................................................................................................57

5. **References** ..............................................................................................................................................................................58
List of Figures

Figure 1: The 5 + 1 V’s of Big Data ................................................................. 7
Figure 2: Data governance and related concepts ............................................. 10
Figure 3: Quality dimensions and interrelations ........................................... 12
Figure 4: Data Lifecycle .............................................................................. 16
Figure 5: Nominal Data Lifecycle ................................................................. 17
Figure 6: FIWARE-NGSI data model ............................................................. 20
Figure 7: Context Entities ......................................................................... 21
Figure 8: Examples of triples in ucode ......................................................... 22
Figure 9: Data on the Web - context ............................................................ 23
Figure 10: PROV data model .................................................................... 37

List of Tables

Table 1: Data Governance Decision Domains .............................................. 9
Table 2: Key concepts and Principles of Data Governance ........................... 10
Table 3: General Data quality Dimensions .................................................. 11
Table 4: Example for operationalizing data quality assessment .................... 12
Table 5: IoT Taxonomy .............................................................................. 13
Table 6: Categories of Social Media ............................................................. 14
Table 7: Characterisation of Social Media .................................................... 14
Table 8: Overview of data quality management methodologies ................... 15
Table 9: Common phases in data quality management methodologies .......... 15
Table 10: Linked Data best practices ............................................................ 19
Table 11: OGD best practices ...................................................................... 19
Table 12: IoT best practices ....................................................................... 20
Table 13: SMD best practices ..................................................................... 20
Table 14: Overview of (selected) Data on the Web Best Practices ............... 24
Table 15: Overview Open Data Licences (Open Knowledge International) ...... 43
Table 16: Advantages and disadvantages of centralized vs. decentralized PDS .................................................................................. 55
Table 17: Advantages and disadvantages of managed vs. unmanaged PDS ................................................................................. 56
Table 18: Advantages and disadvantages of different PDS price models ....... 56
Table 19: Advantages and disadvantages of Data vs. Analysis Access .......... 56
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMS</td>
<td>Asset Description Metadata Schema</td>
</tr>
<tr>
<td>ADMS-AP</td>
<td>ADMS Application Profile for solutions on Joinup</td>
</tr>
<tr>
<td>BP</td>
<td>Best Practice</td>
</tr>
<tr>
<td>CKAN</td>
<td>Comprehensive Knowledge Archive Network</td>
</tr>
<tr>
<td>DAMA</td>
<td>Global Data Management Community</td>
</tr>
<tr>
<td>DCAT</td>
<td>Data Catalog Vocabulary</td>
</tr>
<tr>
<td>DQ</td>
<td>Data Quality</td>
</tr>
<tr>
<td>DQV</td>
<td>Data Quality Vocabulary</td>
</tr>
<tr>
<td>DUV</td>
<td>Dataset Usage Vocabulary</td>
</tr>
<tr>
<td>DUL</td>
<td>DOLCE+DnS Ultralite ontology</td>
</tr>
<tr>
<td>DWBP</td>
<td>Data on the Web Best Practices</td>
</tr>
<tr>
<td>FIESTA</td>
<td>Federated Interoperable Semantic IoT Testbeds and Applications</td>
</tr>
<tr>
<td>FIWARE</td>
<td>Future Internet Core Platform</td>
</tr>
<tr>
<td>FOAF</td>
<td>Friend of a Friend</td>
</tr>
<tr>
<td>GDPR</td>
<td>European General Data Protection Regulation</td>
</tr>
<tr>
<td>GeoDCAT-AP</td>
<td>Geographical extension of DCAT-AP</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>ISA</td>
<td>Interoperability Solutions for European Public Administrations</td>
</tr>
<tr>
<td>LOD</td>
<td>Linked Open Data</td>
</tr>
<tr>
<td>LOV</td>
<td>Linked Open Vocabularies</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>NGSI</td>
<td>Next Generation Service Interface</td>
</tr>
<tr>
<td>OBDA</td>
<td>Ontology Based Data Access</td>
</tr>
<tr>
<td>OGD</td>
<td>Open Governmental Data</td>
</tr>
<tr>
<td>ORG</td>
<td>Organization Ontology</td>
</tr>
<tr>
<td>OWL</td>
<td>Web Ontology Language</td>
</tr>
<tr>
<td>PDS</td>
<td>Personal Data Storage</td>
</tr>
<tr>
<td>PROV-O</td>
<td>Provenance Ontology</td>
</tr>
<tr>
<td>PSI</td>
<td>Public Sector Information</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RegOrg</td>
<td>Registered Organization Vocabulary</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer</td>
</tr>
<tr>
<td>SDMX</td>
<td>Statistical Data and Metadata eXchange</td>
</tr>
<tr>
<td>SHACL</td>
<td>Shapes Constraint Language</td>
</tr>
<tr>
<td>ShEx</td>
<td>Shape Expressions</td>
</tr>
<tr>
<td>SIOC</td>
<td>Semantically-Interlinked Online Communities</td>
</tr>
<tr>
<td>SPARQL</td>
<td>SPARQL Protocol and RDF Query</td>
</tr>
<tr>
<td>SSN</td>
<td>Semantic Sensor Network</td>
</tr>
<tr>
<td>StatDCAT-AP</td>
<td>DCAT Application Profile for description of statistical datasets</td>
</tr>
<tr>
<td>URI</td>
<td>Uniform Resource Identifier</td>
</tr>
<tr>
<td>WP</td>
<td>Work Package</td>
</tr>
<tr>
<td>W3C</td>
<td>World Wide Web Consortium</td>
</tr>
</tbody>
</table>
1 Introduction

The main purpose of the CPaaS.io platform is to be the foundation for an urban data infrastructure, where data from various sources is combined, linked, aggregated, refined and presented to various applications, from end-user applications as well as specialized big data and analytics applications.

The challenges regarding data management in such a platform (cf. Barnaghi et al. 2015) are most easily explained looking at the commonly accepted 5 V’s of big data (Demchenko at al. 2013, Marr 2014):

- **Velocity**: In particular IoT data streamed in from sensors can change rapidly. Such changes – e.g., in order to immediately alert citizens about critical water levels – need to be detected quickly, in real-time\(^1\). Furthermore, old and outdated data may need to be removed from responses or only be processed in the context of trend detection and historical analysis.

- **Volume**: With hundreds of sensors providing data readings every few seconds, the volume of data in the platform can increase rapidly. Linking such data with sometimes large external data sources aggravates the issue. The data management of the CPaaS.io platform must thus be able to handle and process such data volumes while keeping acceptable response times for user applications.

- **Variety**: As already indicated, data in the platform comes in many different forms. There are different types of sources, like sensor data, open government data, other open data sets, as well as possibly also some closed data sets. There are different formats in which the data is provided or stored: raw data readings from sensors, linked data triples, and database tables. In addition, privacy sensitive data may be encrypted using fully homomorphic encryption schemes. And the storage location of the data is different – some data is stored within the platform, other data is stored on 3rd party servers and only accessible via pre-defined APIs.

\(^1\) We follow the accepted definition of real-time, meaning that a response must happen within a *guaranteed* time frame. How quickly that needs to be depends on the actual use case, but typically would be in a range from milliseconds to a few seconds.
Veracity: In particular IoT data is inherently unreliable: a sensor may be decalibrated, it may have only a limited accuracy, the data reading can be outdated etc. But also other data sets published on Open (Government) Data portals often contain errors (Neumaier et al. 2016). Services based on such data need some measures about the data quality of the individual or aggregated sources in order to assess if the data quality is good enough to be used in the specific purpose. This is very much application dependent: For an application displaying ads dynamically on a city’s digital billboard data quality is less of an issue – in the worst case, a wrong ad is displayed –, while for an application controlling a busy crossing’s traffic lights, good quality data is crucial. Data delivered through the CPaaS.io platform must thus be annotated with metadata that will enable such quality assessments. The annotations range from simple data about a sensor’s accuracy to information about provenance.

Value: The data accessible through the platform is intended to provide value, as it is seen as an enabler for both cities as well as third parties to build services on top. Providing the data as Linked Data, annotated with quality parameters, is the key in generating value-added applications.

Various authors have proposed additional V’s depending on their particular interests. In Figure 1 we have added a 6th V which is relevant to this project:

Visibility: The data delivered through the CPaaS.io platform must be visible to application services, and findable via the provided SPARQL endpoints. However, sensitive data must be protected from unauthorized access and the necessary security mechanisms have to be implemented. One key feature of CPaaS.io is the MyData approach to privacy, where a citizen can find all data related to her/him easily and has the power to decide, under what conditions and for what purposes others may get access to the data. This requires that the data is identifiable. Such identifiability actually also helps – contrary to what one may think – to increase data protection (Dungga et al. 2016).

Data quality is a crucial point for several of the aspects listed above, and it is also least understood in the context of urban data infrastructures, hence a big part in this deliverable is devoted to data quality. Secondly, the target of the CPaaS.io project is to provide data as linked data. The W3C recommendations, best practices and working group notes provide a good starting point for how this is best done. In particular, we discuss established vocabularies that allow for exposing available data as linked data. The selection of adequate vocabularies is guided by the use case trials. Finally, a third important aspect about security and data privacy is only outlined here, but it will be tackled in more depth in work package (WP) 5. The document is structured as follows:

Chapter 2 provides a short introduction to the main concepts related to data governance and data management in general and sets out in more detail, what kind of data quality issues are likely to arise with view to several data types relevant in the CPaaS.io context.

Chapter 3 details the current best practices relevant to data management in CPaaS.io structured along the different types of data already introduced in chapter 2, complementing it with requirements and approaches for handling personal data. The focus is on existing recommendations in the field of Linked and Open (Government) Data and in particular on essential annotations and ontologies that will be
relevant for managing data and data quality in CPaaS.io. For each section, we will derive the main implications with view to the CPaaS.io project.

Chapter 4 presents the overall conclusions drawn from the state of the art description and sets out how issues related to data management will be further pursued in the forthcoming deliverables of WP6 and other work packages.

## 2 Data Management: Dimensions

Before discussing best practices considered as relevant in the CPaaS.io context, we briefly introduce the concepts of data management and data governance and discuss some generic considerations on data quality with view to different types of data. The discussion of data management dimensions provides the basis for setting out the main fields of activities to be tackled in the course of the project.

### 2.1 Data Management & Data Governance

Data management puts the focus on data as an asset, which must be managed in a proper way for ensuring success (DAMA 2014). Data governance can be considered as core knowledge area of data management, related to planning and controlling the management and use of data. Generally, governance refers to "what decisions must be made to ensure effective management and use of IT" or data and to "who makes the decisions", while management "involves making and implementing decisions." (Kathri & Brown 2010: 148). Based on the data governance framework proposed by Kathri & Brown (2010), decision domains typically include data principles, data quality, metadata, data access and data lifecycle. The decision domains are interrelated and each of them raises questions related to the definition of roles (e.g. data owners, data consumers).

Table 1: Data Governance Decision Domains (adapted from Kathri & Brown 2010)

<table>
<thead>
<tr>
<th>Decision Domains for data governance</th>
<th>Data Assets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Principles</strong></td>
<td>Data Assets</td>
</tr>
<tr>
<td>Clarifying the role of data as an asset (use of data)</td>
<td></td>
</tr>
<tr>
<td><strong>Data Quality</strong></td>
<td>Metadata</td>
</tr>
<tr>
<td>Establishing the requirements of intended use of data</td>
<td>Establishing the semantics of data</td>
</tr>
<tr>
<td><strong>Data Access</strong></td>
<td>Data Lifecycle</td>
</tr>
<tr>
<td>Specifying access requirements of data</td>
<td>Determining the definition, production, retention and retirement of data</td>
</tr>
</tbody>
</table>

The close interrelation between data governance and data quality management is also stressed by Otto (2011), who conceives data governance as organizational design task and proposes the following conceptual model for organizing data governance:
As Brous et al. point (2016) out, most “research into data governance till now has focused on structuring or organizing data governance. Evidence is scant as to which data governance processes should be implemented, what data governance should be coordinating or how data governance could be coordinated”. Based on a systematic literature review they provide a more detailed categorization of prevailing principles of data governance. These can be clustered along the following dimensions (ibid: 122):

**Table 2: Key concepts and Principles of Data Governance (adapted from Brous et al. 2016)**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Key concepts and principles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organization</strong></td>
<td>Decision-rights, balanced roles, stewardship, ownership, separation of duties and concern, improved coordination of decision making</td>
</tr>
<tr>
<td><strong>Alignment</strong></td>
<td>Meeting business needs, aligning business and IT, developing data strategy, defining data quality requirements, reducing error of use, effective policies and procedures</td>
</tr>
<tr>
<td><strong>Compliance</strong></td>
<td>Accountability, policy enforcement, due diligence, privacy, openness, security, data quality measurement</td>
</tr>
<tr>
<td><strong>Common Understanding</strong></td>
<td>Shared data commons, use of standards, metadata management, standardized data models, standardized operations, facilitates communication</td>
</tr>
</tbody>
</table>

The above mentioned models for designing data governance as well as the DAMA-DMBOK2 framework to data management (cf. DAMA 2014) presuppose the existence of established organizational structures (companies, government organizations). As Brous et al. (2016: 122) conclude, “the organization of data governance should not be seen as a “one size fits all” approach and data governance must be institutionalized through a formal organizational structure that fits with a specific organization.” In the present case we are dealing with a project organization.

### 2.2 Data Quality

Data quality is crucial for making decisions and providing services based on the results of querying various datasets and for adopting a human-centred IoT approach, which entails that applications need to be trusted and accepted (Zaveri et al. 2015, Karkouch et al. 2016, European Commission 2016)
Data quality is commonly defined as “fitness for use” from a user perspective and “free of defects” from a technical perspective. More specifically, data quality denotes the degree to which data fulfils requirements as defined by: 1) different individuals or groups of individuals, 2) standards, 3) laws and regulations, 4) business policies or 5) of data processing applications (Fürber 2016). Thus, data quality requirements depend on a given use case or application.

Within data quality research, requirements are usually clustered according to a basic classification of data quality dimensions as proposed by Wang & Strong (1996), who distinguish four main categories:

<table>
<thead>
<tr>
<th>Category</th>
<th>Description (Karchouch et al. 2016: 60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility dimensions</td>
<td>“Dimensions that describe how accessible (and in the same time secured) data are for data consumers”</td>
</tr>
<tr>
<td>Intrinsic dimensions</td>
<td>“Dimensions that describe quality that is innate in or that inherently exists within data”</td>
</tr>
<tr>
<td>Contextual dimensions</td>
<td>“Dimensions describing the quality with respect to the context of tasks using data”</td>
</tr>
<tr>
<td>Representational</td>
<td>“Dimensions describing how well data formats are representative and understandable”</td>
</tr>
</tbody>
</table>

As Zaveri et al. (2015: 2) point out, there “are already many methodologies and frameworks available for assessing data quality, all addressing different aspects of this task by proposing appropriate methodologies, measures and tools.” The choice and definition of quality dimensions within a given category as well as the metrics for measuring them may vary across existing approaches and for particular types of data (linked data, IoT data, open government data, social media data etc.) or application domains (e.g. health, energy, etc.). When considering a given set of quality dimensions for a given use case or application, it is important to also investigate the interrelations between them to adequately prioritize requirements (Zaveri et al. 2015). For structured data in particular Batini & Scannapieco (2016) point to the relationship between quality of data and quality of schemas and propose to also take schema quality dimensions into consideration with view to the life cycle of data usage. They distinguish between the following quality dimensions:

- **Accuracy**: correctness with respect to model (correct use of constructs) and with respect to requirements (correct representation in model constructs)
- **Completeness**: completeness with respect to inclusion of necessary elements in the conceptual schema and pertinence with respect to unnecessary elements
- **Redundancy**: minimality in terms of representing every part of requirements only once in the schema and normalization related to the structure of functional dependencies
- **Readability**: in terms of intuitive understanding of the represented meaning of the reality in the schema, accounting for aesthetic criteria in a diagrammatic representation as well as compactness

---

2 The term “data quality” is often used interchangeably with the term “information quality”. Although there has been no consensus about the distinction between data quality and information quality, there is a tendency to use data quality (DQ) to refer to technical issues and information quality (IQ) to refer to nontechnical issues. (Zhu et al. 2014).
2.2.1 Linked Data quality

For Linked Data, Zaveri et al. point out a set of quality dimensions that are specifically relevant (cf. dimensions marked with “∗” in Figure 3).

![Quality dimensions and interrelations (Zaveri et al. 2015: 22)](image)

Their extensive overview of existing approaches may serve as input when defining the data quality management approach with respect to assessing data quality based on defined metrics. To give one example: The quality of the interlinking dimension could be measured as follows:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Abr.</th>
<th>Metric</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interlinking</td>
<td>I1</td>
<td>detection of quality of interlinks</td>
<td>detection of (i) interlinking degree, (ii) clustering coefficient, (iii) centrality, (iv) open sameAs chains and (v) description richness through sameAs by using network measures</td>
<td>quantitative</td>
</tr>
<tr>
<td></td>
<td>I2</td>
<td>existence of links to external data providers</td>
<td>detection of the existence and usage of external URIs (e.g. using owl:sameAs links)</td>
<td>quantitative</td>
</tr>
<tr>
<td></td>
<td>I3</td>
<td>dereferenced back-links</td>
<td>detection of all local in-links or back-links: all triples from a dataset that have the resource’s URI as the object</td>
<td>quantitative</td>
</tr>
</tbody>
</table>

Rula et al. (2016) point out that by now, linked data quality assessment has only received limited attention from the linked data community. They provide an overview to semantic web standards and based on this, point out some peculiarities related to linked data quality. In particular, they point to the “schemaless” approach, which raises a need for techniques to validate RDF (Resource Description Framework) data a posteriori in order to ensure quality. Similar to Zaveri et al. (2015) they propose a method and metrics for measuring specific quality dimensions.
2.2.2 IoT Data quality

For IoT Data in particular, Karkouch et al. (2016: 60) define data quality in terms of “how suitable the gathered data (from the smart things) are for providing ubiquitous services for IoT users”. They claim that the characteristics of IoT data (volume and distribution, uncertainty and noise, smooth variations etc.) call for updated quality assessment criteria, which they discuss in more detail. As they point out, data “quality (DQ) is crucial to gain user engagement and acceptance of the IoT paradigm and services.” (Karkouch et al. 2016: 57). Based on a summary of the characteristics of IoT data, and a broad literature review they define a set of quality dimensions that are relevant to IoT. These include intrinsic and contextual quality dimensions (e.g. accuracy, confidence, completeness, data volume and timeliness) as well as accessibility (ease of access, access security) and representation dimensions (interpretability). The discussion also includes techniques for enhancing IoT data quality with focus on data cleaning, such as semantic-based integration approaches to enhance integration and interoperability of IoT data (e.g. Sensor Web Enablement Initiative by the Open Geospatial Consortium or the SSN initiative by W3C) (cf. 3.5.1).

Similarly, Qin et al. (2016: 137) point out that managing IoT data is challenging since, it is typically “produced in dynamic and volatile environments, (...) is not only extremely large in scale and volume, but also noisy and continuous”. They provide a broad review of existing techniques in the fields of data stream processing, data storing, event processing and searching of IoT. The discussion is based on an IoT taxonomy, which differentiates IoT data characteristics along the dimensions of data generation, data interoperability and data quality, which need to be addressed from a data management perspective.

Table 5: IoT Taxonomy (Qin et al. 2016):

<table>
<thead>
<tr>
<th>IoT Taxonomy</th>
<th>Data Generation</th>
<th>Data Interoperability</th>
<th>Data Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity</td>
<td>Incompleteness</td>
<td>Uncertainty</td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>Semantics</td>
<td>Redundancy</td>
<td></td>
</tr>
<tr>
<td>Dynamics</td>
<td></td>
<td>Ambiguity</td>
<td></td>
</tr>
<tr>
<td>Heterogeneity</td>
<td></td>
<td>Inconsistency</td>
<td></td>
</tr>
</tbody>
</table>

Another aspect to be addressed is the trade-off between data quality and performance in the context of real-time processing (Geisler et al. 2016).

2.2.3 Open government data quality

Open Government Data (OGD) has no value if not being used. Data quality is therefore also an issue for OGD publishers and re-users. Often, data quality is defined against the 5 star deployment scheme proposed by Berners-Lee, thus putting a focus on data formats:

* Data is available on the Web, in whatever format.
** Available as machine-readable structured data, (i.e., not a scanned image).
*** Available in a non-proprietary format, (i.e, CSV, not Microsoft Excel).
**** Published using open standards from the W3C (RDF and SPARQL).
***** All of the above and links to other Linked Open Data.”

3 https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data
As Vetrò et al. (2016) point out, data – according to this scheme – can reach the 5 star level, despite being of poor quality, e.g. regarding data accuracy. As they state, many evaluations focus on the quality of open data platforms, rather than on datasets. In their framework, they propose a set of data quality dimensions for OGD and according metrics, thereby focusing on intrinsic data quality. Several factors influencing re-usage of OGD are however not explicitly addressed, e.g. the question of licensing (cf. 3.4.5).

### 2.2.4 Social media data quality

Social media is yet another type of data and a heterogeneous source of data. Categories of social media sites are for instance: Blogs, social networking sites or Wikis (Agarwal & Yiliyasi 2010):

**Table 6: Categories of Social Media (Agarwal & Yiliyasi 2010)**

<table>
<thead>
<tr>
<th>Category</th>
<th>Social Media Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogs</td>
<td>Wordpress, Blogger, Blogcatalog, MyBlogLog</td>
</tr>
<tr>
<td>Media Sharing</td>
<td>Flickr, Photobucket, YouTube, Multiply, Justin.tv, Ustream</td>
</tr>
<tr>
<td>Micro Blogging</td>
<td>Twitter, SixApart</td>
</tr>
<tr>
<td>Social Bookmarking</td>
<td>Del.icio.us, StumbleUpon</td>
</tr>
<tr>
<td>Social Friendship Network</td>
<td>MySpace, Facebook, Friendfeed, Bebo, Orkut, LinkedIn, PatientsLikeMe, DailyStrength</td>
</tr>
<tr>
<td>Social News</td>
<td>Digg, Reddit</td>
</tr>
<tr>
<td>Wikis</td>
<td>Wikipedia, Wikiversity, Scholarpedia, Ganfyd, AskDrWiki</td>
</tr>
</tbody>
</table>

A further classification by Kaplan and Haenlein (2010) e.g. differentiates a range of social media with respect to categories such as media richness and self-disclosure displayed on the sites:

**Table 7: Characterisation of Social Media (Kaplan & Haenlein 2010)**

<table>
<thead>
<tr>
<th>Self-presentation / Self-disclosure</th>
<th>Low (e.g. Wikipedia)</th>
<th>Medium (e.g. Facebook)</th>
<th>High (e.g. Second Life)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Blogs</td>
<td>Social networking sites</td>
<td>Virtual social worlds</td>
</tr>
<tr>
<td>Low</td>
<td>Collaborative projects</td>
<td>Content communities</td>
<td>Virtual game worlds</td>
</tr>
</tbody>
</table>

Regarding data quality, the accessibility dimension provides another characteristic for differentiation. While Twitter for instance makes a lot of information available, Facebook does not (cf. Czernek 2016). Finally, as Scannapieco and Berti (2016: 440) point out, it must also be noted that social media data, “are highly unstructured and often not accompanied by metadata. This means that high percentages of these data cannot be simply used by automated processes as they are affected by high percentages of noise.” Assessing the content of social media data adds another level of complexity to data quality assessment (Chai et al. 2009): “Challenges arise in accurately evaluating the quality of content that has been created by users from different backgrounds, for different domains and consumed by users with different requirements.” As is the case for other types of data, specific data characteristics can be mapped and prioritized against established data quality dimensions (cf. Agarwal & Yiliyasi 2010).
2.2.5 Data quality management methodologies

There are many methodologies to conduct data quality management. They provide different rationales for choosing appropriate activities and techniques. A detailed analysis of methodologies is provided by Batini & Scannapieco (2016), including:

Table 8: Overview of data quality management methodologies (Batini & Scannapieco 2016)

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Extended name</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDQM</td>
<td>Total Data quality Management</td>
</tr>
<tr>
<td>DWQ</td>
<td>The Datawarehouse Quality Methodology</td>
</tr>
<tr>
<td>TIQM</td>
<td>Total Information Quality Management</td>
</tr>
<tr>
<td>AIMQ</td>
<td>A Methodology for information quality assessment</td>
</tr>
<tr>
<td>CIHI</td>
<td>Canadian Institute for Health Information Methodology</td>
</tr>
<tr>
<td>DQA</td>
<td>Data Quality Assessment</td>
</tr>
<tr>
<td>IQM</td>
<td>Information Quality Measurement</td>
</tr>
<tr>
<td>ISTAT</td>
<td>ISTAT Methodology Falorsi</td>
</tr>
<tr>
<td>AMEQ</td>
<td>Activity Based Measuring and Evaluating of Product Information Quality Methodology</td>
</tr>
<tr>
<td>COLDOQ</td>
<td>Cost Effect of Low Data Quality Methodology</td>
</tr>
<tr>
<td>DaQuinCIS</td>
<td>Data Quality in Cooperative Information Systems</td>
</tr>
<tr>
<td>QAFD</td>
<td>Methodology for the Quality Assessment of Financial Data</td>
</tr>
<tr>
<td>CDQ</td>
<td>Comprehensive Methodology for Data Quality Management</td>
</tr>
</tbody>
</table>

This list could be further complemented, e.g. with the Semantic Data Quality Management Framework (SDQM) as discussed in Fürber (2016) with a focus on data quality management activities using ontologies. As Batini & Scannapieco point out, it is most effective “to see the guidelines, phases, tasks, activities, and techniques, which together form a methodology, as a toolbox, where single pieces are to be used in connection and/or in sequence, according to circumstances and to specific characteristics of the application domain involved in the process” (2016: 402). In short, methodologies share the following basic phases for assessing and improving data quality (ibid 360f):

Table 9: Common phases in data quality management methodologies (adapted from Batini & Scannapieco 2016)

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assessment process</strong></td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>analyze information bases, databases, schemas, metadata, etc.</td>
</tr>
<tr>
<td>DQ requirement analysis</td>
<td>identify quality issues, set quality targets</td>
</tr>
<tr>
<td>Identification of critical areas</td>
<td>identify most relevant data and data flows</td>
</tr>
<tr>
<td>Process modelling</td>
<td>provide model of producing / updating information</td>
</tr>
<tr>
<td>Measurement of quality</td>
<td>select quality dimensions, define metrics</td>
</tr>
<tr>
<td><strong>Improvement process</strong></td>
<td></td>
</tr>
<tr>
<td>Evaluation of costs</td>
<td>estimate direct / indirect costs of data quality</td>
</tr>
<tr>
<td>Assignment of process responsibilities</td>
<td>identify process owners, define information production responsibilities and management activites</td>
</tr>
<tr>
<td>Assignment of responsibilities</td>
<td>identify data owners, define data management responsibilities</td>
</tr>
<tr>
<td>Identification of error causes</td>
<td>identify causes of quality problems</td>
</tr>
<tr>
<td>Selection of strategies / techniques</td>
<td>identify improvement strategies and techniques that comply with contextual knowledge, quality objectives and budget</td>
</tr>
<tr>
<td>Design of improvement solutions</td>
<td>select most effective and efficient strategies and tools for improving DQ</td>
</tr>
<tr>
<td>Process control</td>
<td>define checkpoints in production process, monitor quality during process</td>
</tr>
<tr>
<td>Process redesign</td>
<td>define process improvement actions</td>
</tr>
<tr>
<td>Improvement management</td>
<td>define new organizational rules for DQ</td>
</tr>
<tr>
<td>Improvement monitoring</td>
<td>establish monitoring activities that provide feedback on results</td>
</tr>
</tbody>
</table>
As Batini & Scannapieco (2016: 402) point out for the organizational context, it can make sense to adopt a simplified methodology in order to account for available resources (e.g. in terms of knowledge) and ensure acceptance. The constraints of a project context, need to be taken into account when defining the procedure for data quality management.

### 2.3 Data Lifecycle

Basic steps of data quality management can be structured along the data life cycle: Its iterative phases include data acquisition, data usage and data retirement.

#### Data Lifecycle

<table>
<thead>
<tr>
<th>Data Lifecycle</th>
<th>Data Acquisition Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“relates to the problem of (1) generating new or (2) retrieving existing data and storing it onto some kind of medium (...). Hence, during data acquisition data may be filtered according to their quality or transformed to cleanse incorrect data before passing it to data usage”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Lifecycle</th>
<th>Data Usage Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“In the usage phase, data is used as an information source for humans and machines in operational or decision-making processes (...). Data may be altered, filtered, enriched or aggregated to derive additional information in this phase”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Lifecycle</th>
<th>Data Retirement Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“data retirement encompasses deleting, deactivating and archiving data. This phase is often entered when data is not used anymore or system performance slows down due to huge amounts of data to be processed”</td>
</tr>
</tbody>
</table>

Figure 4: Data Lifecycle (Redman 1996: 217, cit. in: Fürber 2016: 40)

From a generic point of view Fürber (2016: 42) identifies the data quality management capabilities that are relevant throughout the data lifecycle:

- Data quality monitoring reports (“identify instances with data quality problems”)
- Data quality assessment reports (“quality state of a data source”)
- Data cleansing functionalities (“remove data quality problems”)
- Data constraints (“data quality rules that can be automatically applied by an information system”)
- Requirements management (“manage the quality criteria”)

With view to handling data with various qualities in the big data context Wahyudi and Janssen (2016) propose a generic process pattern. Based on a discussion of data lifecycle models they identify the main tasks related to data processing from a data consumer perspective: 1) discover, 2) access, 3) exploit, 4) analyze. Their “nominal data lifecycle” gives an overview to typical functions to be performed in each phase with the management task covering all phases (ibid: 499):
2.4 Data Access

In the present context “data access” had two meanings. From a governance perspective data access focuses on an organization’s requirements and decisions on the use of their data, relates to legal compliance, privacy and security issues and materializes e.g. in data access policies (Khatri & Brown 2010). From the perspective of developing a platform that enables processing of data from various data sources, data access relates to the acquisition phase in the data lifecycle, i.e. discovering and accessing relevant data (see above, Fürber 2016, Wahyudi & Janssen 2016).

With view to data access and discoverability (see also chapter 3 below), the ontology-based data management approach is worth mentioning (cf. Lenzerini 2011). Ontology-based data access (ODBA) management builds up on work done in the semantic web community and envisages that data from multiple sources (sensors and others) are integrated by linking them to a common data model. More precisely, in “the OBDA paradigm, an ontology defines a high-level global schema of (already existing) data sources and provides a vocabulary for user queries. An OBDA system rewrites such queries and ontologies into the vocabulary of the data sources and then delegates the actual query evaluation to a suitable query answering system such as a relational database management system or a datalog engine” (Kontchakov et al. 2013: 194). In short, the idea is to provide data access of heterogeneous sources unified and coherently, while allowing data users to formulate needs at an ontological level (cf. Calbimonte 2010).

2.5 Implications for CPaaS.io

In the context of CPaaS.io we are dealing with a project organization. Data governance in the sense of establishing decision-structures relates to the interplay between the project management structure and stakeholders involved in the use cases and to making sure, that roles are adequately defined. Basic aspects of data governance and management will formally be covered by the iterations of the Data Management Plan as elaborated in WP7 (D7.2 and D7.9) based on the H2020 guidelines. With view to

---

data management in CPaaS.io, we will have to define mechanisms for handling data from multiple sources, including uncertain data across the data lifecycle. In particular, this relates to defining adequate vocabularies for data integration in CPaaS.io and for re-usage by use case applications with respect to data quality requirements. Data quality dimensions and measures will have to be defined, accounting for the different types of data used in the project.

3 Data Management: Best Practices

Data that is not discoverable has no value. Discoverability means re-using existing definitions. Once a dataset has been identified, we need to be able to judge its usefulness within the application. For that reason, we need to define best practices and common vocabularies within the CpaaS.io project. The following discussion provides a first overview to existing work that will be taken into account in the next steps of the project. We start with an introduction to the FIWARE and u2 data models in regard to exposing (part of) the data as RDF, followed by work from the Linked and Open Data communities, best practices on Open Government Data publishing and vocabularies related to IoT data and social media data. The chapter closes with considerations on approaches aimed at handling personal data.

3.1 Content Overview

Many of the presented standards in the area of Linked and Open Data are coming from the Data on the Web Best Practices Working Group\(^5\). The group describes its mission as follows: “The mission of the Data on the Web Best Practices Working Group, part of the Data Activity\(^6\), is:

1. to develop the open data ecosystem, facilitating better communication between developers and publishers;
2. to provide guidance to publishers that will improve consistency in the way data is managed, thus promoting the re-use of data;
3. to foster trust in the data among developers, whatever technology they choose to use, increasing the potential for genuine innovation.

The guidance will take two forms: a set of best practices that apply to multiple technologies, and vocabularies currently missing but that are needed to support the data ecosystem on the Web.” Multiple of these goals were reached by the Data on the Web Best Practices Working Group in 2016 and will be summarized in this document. The major outcome of the Data on the Web Best Practices Working Group is the Data on the Web Best Practices document, which is a W3C Candidate Recommendation\(^7\) at the time

\(^5\) https://www.w3.org/2013/dwbp/
\(^6\) https://www.w3.org/2013/data/
\(^7\) The W3C publication process follows these steps when advancing a technical report to Recommendation: 1) Publication of the First Public Working Draft, 2) Publication of zero or more revised Public Working Drafts, 3) Publication of a Candidate Recommendation, 4) Publication of a Proposed Recommendation, 5) Publication as a W3C Recommendation, 6) Possibly, Publication as an Edited Recommendation (https://www.w3.org/2015/Process-20150901/#rec-advance).
writing. Beside this document, a range of additional vocabularies will be discussed. The following table provides an overview:

### Table 10: Linked Data best practices

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DWBP</td>
<td>Data on the Web Best Practices</td>
<td><a href="https://www.w3.org/TR/dwbp/#intro">https://www.w3.org/TR/dwbp/#intro</a></td>
</tr>
<tr>
<td>DQV</td>
<td>Data Quality Vocabulary</td>
<td><a href="http://www.w3.org/TR/vocab-dqv/">http://www.w3.org/TR/vocab-dqv/</a></td>
</tr>
<tr>
<td>DUV</td>
<td>Dataset Usage Vocabulary</td>
<td><a href="https://www.w3.org/TR/vocab-duv/">https://www.w3.org/TR/vocab-duv/</a></td>
</tr>
<tr>
<td>-</td>
<td>Web Annotation Data Model (w.r.t. semantic annotations)</td>
<td><a href="http://www.w3.org/TR/annotation-model/">http://www.w3.org/TR/annotation-model/</a></td>
</tr>
<tr>
<td>SHACL</td>
<td>Shapes Constraint Language (w.r.t. data quality assessment)</td>
<td><a href="https://www.w3.org/TR/shacl/">https://www.w3.org/TR/shacl/</a></td>
</tr>
<tr>
<td>ShEx</td>
<td>Shape Expressions</td>
<td><a href="https://www.w3.org/2001/sw/wiki/ShEx">https://www.w3.org/2001/sw/wiki/ShEx</a></td>
</tr>
<tr>
<td>Schema.org</td>
<td>schema.org (widely adopted for basic concepts)</td>
<td>schema.org</td>
</tr>
<tr>
<td>LOV</td>
<td>Linked Open Vocabularies</td>
<td><a href="http://lov.okfn.org/dataset/lov/">http://lov.okfn.org/dataset/lov/</a></td>
</tr>
<tr>
<td>CUBE</td>
<td>RDF Data Cubes Vocabulary (w.r.t. publishing multi-dimensional data, e.g. IoT sensor information)</td>
<td><a href="http://www.w3.org/TR/vocab-data-cube/">http://www.w3.org/TR/vocab-data-cube/</a></td>
</tr>
<tr>
<td>PROV-O</td>
<td>PROV Ontology</td>
<td><a href="http://www.w3.org/TR/2013/REC-prov-o-20130430/">http://www.w3.org/TR/2013/REC-prov-o-20130430/</a></td>
</tr>
<tr>
<td>ORG</td>
<td>Organization Ontology (w.r.t. organizational structures)</td>
<td><a href="http://www.w3.org/TR/vocab-org/">http://www.w3.org/TR/vocab-org/</a></td>
</tr>
</tbody>
</table>

Best practices in the field of Open Government Data are aimed at governmental data publishers and as such provide a background to the data management issues to be expected in that respect. The Following vocabularies and application profiles will be discussed:

### Table 11: OGD best practices

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCAT</td>
<td>Data Catalog Vocabulary</td>
<td><a href="http://www.w3.org/TR/vocab-dcat/">http://www.w3.org/TR/vocab-dcat/</a></td>
</tr>
<tr>
<td>ADMS</td>
<td>Asset Description Metadata Schema</td>
<td><a href="http://www.w3.org/TR/vocab-adms/">http://www.w3.org/TR/vocab-adms/</a></td>
</tr>
<tr>
<td>-</td>
<td>Core Vocabularies</td>
<td><a href="https://joinup.ec.europa.eu/asset/core_vocabularys/asset_release/all">https://joinup.ec.europa.eu/asset/core_vocabularys/asset_release/all</a></td>
</tr>
<tr>
<td>RegOrg</td>
<td>Registered Organization Vocabulary</td>
<td><a href="http://www.w3.org/TR/vocab-regorg/">http://www.w3.org/TR/vocab-regorg/</a></td>
</tr>
</tbody>
</table>
As for IoT the following working groups and vocabularies will be discussed:

<table>
<thead>
<tr>
<th>Table 12: IoT best practices</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abbreviation</strong></td>
</tr>
<tr>
<td>SSN</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>FIESTA-IoT</td>
</tr>
</tbody>
</table>

As for social media data, the following work will briefly be discussed:

<table>
<thead>
<tr>
<th>Table 13: SMD best practices</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abbreviation</strong></td>
</tr>
<tr>
<td>SIOC</td>
</tr>
</tbody>
</table>

3.2 FIWARE and u2 Data Models

The dominant platforms in CPaaS.io are FIWARE and u2. Both platforms have their own data models.

3.2.1 FIWARE-NGSI

FIWARE is using the NGSI data model. The main elements in the NGSI data model are context entities, attributes and metadata, as shown in the figure below.

![FIWARE-NGSI data model](image)

**Figure 6: FIWARE-NGSI data model**

Entities (in NGSIv1 Context Entities) are entities that are described by Context Information. Figure 7 gives some examples on entities that can be used as Context Entities.
NGSI version 1 cannot be directly mapped to RDF within the FIWARE platform itself. In FIWARE NGSI version 2 this should change: The final specification aims at providing support for context registrations and linked data. This is done by supporting JSON-LD, which would greatly facilitate mapping FIWARE based NGSI data to RDF. However, at the time writing support for JSON-LD cannot be found in the NGSI v2 specification.

### 3.2.2 ucode

The u2 platform is using the ucode model to store data. Everything which needs to be stored gets a unique identifier attached. This is also the case for abstract concepts. The identification is a technology agnostic object identification system that provides 128-bit expandable address space for any kind of object identification. With ucode, digital information can be associated with objects and places, and the associated information can be retrieved by using the appropriate ucode. The uID center maintains a registry for ucode assignments, like the root server. Companies can apply for a ucode prefix/domain and assign their own ucodes under this domain. Based on those identifiers, one can make up a relation, informally called ucR (ucode Relation). A relation represents a triple where both the subject and the predicate are a ucode and the object is either a ucode or a literal.

---

This makes it straightforward to expose ucode as RDF, even though some of the philosophical concepts are different:

- ucR uses existing objects in the real world as its foundation. Ucodes assigned to objects and places in the real world are essential.
- Abstract concepts such as relationships are deduced from facts in the real world. Thus, ucode does not require a prepared vocabulary in advance.
- ucR is RDF but the URIs are URNs based on ucode IDs.\textsuperscript{12} This is still valid RDF but less common as Linked Data propagates the use of HTTP resolvable identifiers as URIs.

### 3.2.3 Implications for CPaaS.io

CPaaS.io needs data, which is available in the ucode and/or NGSI data model for its use cases. It is thus important that once a specific data set is found one can map and expose it as RDF easily. In FIWARE this can be done by using the upcoming NGSI v2 data model, given that it will support JSON-LD as mentioned in the specification. For ucode there are already examples available where ucode data is mapped to RDF and common vocabularies\textsuperscript{13} (Viljama et al. 2011). It would be useful to create and/or integrate tools which facilitate mapping to common RDF identifiers and vocabularies.

### 3.3 Linked Data

The CPaaS.io Platform aims at fusing data from different sources, in particular the FIWARE platform in Europe and the u2 platform in Japan. From a data consumer perspective, it should not matter where the data is stored, CPaaS.io will facilitate discovery and access to information in these data storages via generic APIs. These APIs may be based on the W3C Semantic Web and Linked Data technology stack. The RDF data model provides well-known schemas and ontologies as lingua franca, HTTP as transport

\textsuperscript{12} https://tools.ietf.org/html/rfc6588
\textsuperscript{13} See “External Specifications for Open Data Distribution Platform Systems v2.0” http://www.vled.or.jp/en/results/
layer, URIs as decentralized identifiers and multilingualism in its core. This makes it the data model of choice for bridging between data silos.

The following sections will give an overview over common schemas and ontologies in the domain of Semantic Web and Linked Data. The Linked and Open Data communities published many documents related to the objectives of this deliverable in the past months. This section will align best practices in the Linked and Open Data domain with CPaaS.io requirements. For readers that are new to Linked Data and its concepts and/or do not know a particular term this document refers to the Linked Data Glossary at W3C.

### 3.3.1 Data on the Web Best Practices (DWBP)

The Data on the Web Best Practices is a W3C Candidate Recommendation published in 2016. After roughly 10 years of Open Data movement, the document summarizes best practices and recommendations about how to publish Open Data, especially in the context of what needs to be considered to ensure that the published data is of maximum value for the public. Most of the recommendations are related to machine readability and discoverability of Open Data.

The main principles of the guide are based around the following diagram:

![Data on the Web - context](image)

**Figure 9: Data on the Web - context**

Data is published in different distributions, which are a specific physical form of a dataset. The dataset contains the metadata about the dataset itself, which is thus valid for all different distributions as well. Metadata attached to a specific distribution is only valid for that particular physical form of a dataset.

The dataset publication follows web architecture principles, the main principle of it is to publish resources with stable URIs, so that they can be referenced and make links, via URIs, between two or more resources. Finally, for promoting the interoperability among datasets it is important to adopt data vocabularies and standards.

---

14 [https://www.w3.org/TR/ld-glossary/](https://www.w3.org/TR/ld-glossary/)
3.3.1.1 The Best Practices

The following section presents a selection of the 35 Best Practices (BP) defined in the document as relevant to the CPaaS.io project. Each BP is related to one or more requirements from the Data on the Web Best Practices Use Cases & Requirements document which guided their development. Each Best Practice has at least one of these requirements as evidence of its relevance. Neither the Use Case nor the requirements are repeated in this document (please consult the original publication for details). This also applies for examples, which are provided throughout the document.

Each BP is listing the main benefits of applying it to data, based on the following list:

- **Comprehension**: humans will have a better understanding about the data structure, the data meaning, the metadata and the nature of the dataset.
- **Processability**: machines will be able to automatically process and manipulate the data within a dataset.
- **Discoverability**: machines will be able to automatically discover a dataset or data within a dataset.
- **Reuse**: the chances of dataset reuse by different groups of data consumers will increase.
- **Trust**: the confidence that consumers have in the dataset will improve.
- **Linkability**: it will be possible to create links between data resources (datasets and data items).
- **Access**: humans and machines will be able to access up to date data in a variety of forms.
- **Interoperability**: it will be easier to reach consensus among data publishers and consumers.

The initial description and the “Why” paragraph of each BP is taken from the W3C document. In the line “Implementation” this document gives specific remarks on how the best practice might be implemented into the CPaaS.io platform.

<table>
<thead>
<tr>
<th>Issue / Best Practice</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td></td>
</tr>
<tr>
<td>Provide Metadata</td>
<td></td>
</tr>
<tr>
<td>Recommendation</td>
<td>Provide metadata for both human users and computer applications.</td>
</tr>
<tr>
<td>Purpose (Why)</td>
<td>“Providing metadata is a fundamental requirement when publishing data on the Web because data publishers and data consumers may be unknown to each other. Then, it is essential to provide information that helps human users and computer applications to understand the data as well as other important aspects that describes a dataset or a distribution.”</td>
</tr>
<tr>
<td>Implementation</td>
<td>This BP recommends to add metadata to the publication in both machine – and human readable form. Metadata for the machine should be provided in a specific serialization like JSON-LD or Turtle. While it is not mandatory it is recommended to use a RDF serialization for maximum interoperability and provide content-negotiation which allows servers (or clients or proxies) to decide which version of a document a user should be given</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Provide descriptive metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendation</td>
</tr>
<tr>
<td>Purpose (Why)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deliverable D6.1</th>
<th>Holistic Data Management: State of the Art</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation</td>
<td>In this BP specific fields and vocabularies are listed to provide metadata about a dataset and distribution. This includes – among others – features like title, description, data, keywords, contact points, data publisher, themes and topics. The dominant vocabulary used in this BP is Data Catalog Vocabulary (3.4.3) and Dublin Core Terms.</td>
</tr>
</tbody>
</table>

**Provide structural metadata**

| Recommendation | Provide metadata that describes the schema and internal structure of a distribution. |
| Purpose (Why)  | “Providing information about the internal structure of a distribution is essential for others wishing to explore or query the dataset. It also helps people to understand the meaning of the data.” |
| Implementation | Implementation of this BP depends on the type of data: XML provides this with XML Schema, for CSV the W3C published the Model for Tabular Data and Metadata on the Web document and for multi-dimensional data one can use Data Cubes, see RDF Data Cubes Vocabulary (3.3.7). For RDF data interpretation is in general easier as one can follow the used schema and vocabularies in the dataset itself. |

**Data Licences**

| Recommendation | Provide a link to or copy of the license agreement that controls use of the data. |
| Purpose (Why)  | “The presence of license information is essential for data consumers to assess the usability of data. User agents may use the presence/absence of license information as a trigger for inclusion or exclusion of data presented to a potential consumer.” |
| Implementation | The license should be described by re-using an existing properties like dct:license, cc:license, schema:license or xhtml:license. By pointing to a specific, stable URI of a license the machine can examine and interpret the license accordingly. |

**Data Provenance**

| Recommendation | Provide complete information about the origins of the data and any changes you have made. |
| Purpose (Why)  | “Provenance is one means by which consumers of a dataset judge its quality. Understanding its origin and history helps one determine whether to trust the data and provides important interpretive context.” |
| Implementation | This BP can be implemented by providing appropriate properties which clearly state the provenance of the data. Examples are: dct:issued, dct:publisher, dct:creator and related to that prov:actedOnBehalfOf. |

**Data Quality**

| Recommendation | Provide information about data quality and fitness for particular purposes. |
| Purpose (Why)  | “Data quality might seriously affect the suitability of data for specific applications, including applications very different from the purpose for which it was originally generated. Documenting data quality significantly eases the process of dataset selection, increasing the chances of reuse. Independently from domain-specific peculiarities, the quality of data should be documented and known quality issues should be explicitly stated in metadata.” |
| Implementation | This BP can be implemented using the Data Quality Vocabulary (3.3.1.2). |
## Data Versioning

### Provide a version indicator

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Assign and indicate a version number or date for each dataset.</th>
</tr>
</thead>
</table>

**Purpose (Why)**

“Version information makes a revision of a dataset uniquely identifiable. Uniqueness can be used by data consumers to determine whether and how data has changed over time and to determine specifically which version of a dataset they are working with. Good data versioning enables consumers to understand if a newer version of a dataset is available. Explicit versioning allows for repeatability in research, enables comparisons, and prevents confusion. Using unique version numbers that follow a standardized approach can also set consumer expectations about how the versions differ.”

**Implementation**
The approaches recommended for this BP are:

- Include a unique version number or date as part of the metadata for the dataset.
- Use a consistent numbering scheme with a meaningful approach to incrementing digits, such as SchemaVer, which is basically a generalized version of semantic versioning used for software.
- If the data is made available through an API, the URI used to request the latest version of the data should not change as the versions change, but it should be possible to request a specific version through the API.
- Use Memento [RFC7089]¹⁶, or components thereof, to express temporal versioning of a dataset and to access the version that was operational at a given datetime. The Memento protocol aligns closely with the approach for assigning URIs to versions that is used for W3C specifications, described below.

### Provide version history

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Provide a complete version history that explains the changes made in each version.</th>
</tr>
</thead>
</table>

**Purpose (Why)**

“In creating applications that use data, it can be helpful to understand the variability of that data over time. Interpreting the data is also enhanced by an understanding of its dynamics. Determining how the various versions of a dataset differ from each other is typically very laborious unless a summary of the differences is provided.”

**Implementation**

This can be implemented by using properties like dct:isVersionOf, pav:previousVersion, owl:versionInfo, pav:version and an appropriate rdfs:comment.

## Data Identifiers

### Use persistent URIs as identifiers of datasets

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Identify each dataset by a carefully chosen, persistent URI.</th>
</tr>
</thead>
</table>

**Purpose (Why)**

“Adopting a common identification system enables basic data identification and comparison processes by any stakeholder in a reliable way. They are an essential pre-condition for proper data management and reuse.”

**Implementation**

Developers may build URIs into their code and so it is important that those URIs persist and that they dereference to the same resource over time without the need for human intervention.

Implementation of this BP highly depends on the published data. Examples are given in the original W3C document and are not repeated here.

### Use persistent URIs as identifiers of datasets

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Reuse other people’s URIs as identifiers within datasets where possible.</th>
</tr>
</thead>
</table>

¹⁶ [https://www.w3.org/TR/dwbp/#bib-RFC7089](https://www.w3.org/TR/dwbp/#bib-RFC7089)
### Purpose (Why)

“The power of the Web lies in the Network effect. The first telephone only became useful when the second telephone meant there was someone to call; the third telephone made both of them more useful yet. Data becomes more valuable if it refers to other people’s data about the same thing, the same place, the same concept, the same event, the same person, and so on. That means using the same identifiers across datasets and making sure that your identifiers can be referred to by other datasets. When those identifiers are HTTP URIs, they can be looked up and more data discovered.”

### Implementation

These ideas are at the heart of the 5 Stars of Linked Data (2.2.3) where one data point links to another, and of Hypermedia where links may be to further data or to services that can act on or relate to the data in some way.

That’s the Web of Data.

This is one of the more challenging BP to implement. To make sure one provides useful links one needs to evaluate appropriate endpoints for interlinking and find a way to at least partially automate the process to make sure it is sustainable and can be re-done with the same quality level on updates. As mentioned in the W3C document this is a topic in itself and is not elaborated in details.

### Assign URIs to dataset versions and series

**Recommendation** Assign URIs to individual versions of datasets as well as to the overall series.

**Purpose (Why)** “Like documents, many datasets fall into natural series or groups. In different circumstances, it will be appropriate to refer to the current situation (the current set of bus stops, the current elected officials etc.). In others, it may be appropriate to refer to the situation as it existed at a specific time.”

**Implementation** This can be implemented by providing immutable URIs for each snapshot and point the persistent URI to the “latest version” of this document. Once there is a new version a new, again immutable snapshot is created and the persistent “latest version” URI points to this version. The dated URIs remain unchanged.

### Data Formats

#### Use machine-readable standardized data formats

**Recommendation** Make data available in a machine-readable, standardized data format that is well suited to its intended or potential use.

**Purpose (Why)** “As data becomes more ubiquitous, and datasets become larger and more complex, processing by computers becomes ever more crucial. Posting data in a format that is not machine-readable places severe limitations on the continuing usefulness of the data. Data becomes useful when it has been processed and transformed into information. Note that there is an important distinction between formats that can be read and edited by humans using a computer and formats that are machine-readable. The latter term implies that the data is readily extracted, transformed and processed by a computer.”

**Implementation** Using non-standard data formats is costly and inefficient, and the data may lose meaning as it is transformed. On the other hand, standardized data formats enable interoperability as well as future uses, such as remixing or visualization, many of which cannot be anticipated when the data is first published. It is also important to note that most machine-readable standardized formats are also locale-neutral.

This can be implemented by making data available in a machine-readable standardized data format that is easily parseable including but not limited to CSV, XML, HDF5, JSON and RDF serialization syntaxes like RDF/XML, JSON-LD, Turtle.

#### Use locale-neutral data representations

**Recommendation** Use locale-neutral data structures and values, or, where that is not possible, provide metadata about the locale used by data values.

**Purpose (Why)** “Data values that are machine-readable and not specific to any particular language or culture are more durable and less open to misinterpretation than values that use one of the
many different cultural representations. Things like dates, currencies and numbers may look similar but have different meanings in different locales. For example, the ‘date’ 4/7 can be read as 7th of April or the 4th of July depending on where the data was created. Similarly, €2,000 is either two thousand Euros or an over-precise representation of two Euros. By using a locale-neutral format, systems avoid the need to establish specific interchange rules that vary according to the language or location of the user. When the data is already in a locale-specific format, making the locale and language explicit by providing locale parameters allows users to determine how readily they can work with the data and may enable automated translation services.”

**Implementation**

There are various aspects in implementing this BP. For natural language text one should use an associated language tag to indicate the language and locale of the data. BCP47 provides the standard for language and locale identification. Date fields should be represented by using XML Schema xsd:date. In case of specific units one should assign them separately to the value, for example:

```json
{ "price": {  "value": 2000.00,  "currency": "EUR" } }
```

**Data Vocabularies**

**Reuse vocabularies, preferably standardized ones**

**Recommendation** Use terms from shared vocabularies, preferably standardized ones, to encode data and metadata.

**Purpose (Why)** “Use of vocabularies already in use by others captures and facilitates consensus in communities. It increases interoperability and reduces redundancies, thereby encouraging reuse of your own data. In particular, the use of shared vocabularies for metadata (especially structural, provenance, quality and versioning metadata) helps the comparison and automatic processing of both data and metadata. In addition, referring to codes and terms from standards helps to avoid ambiguity and clashes between similar elements or values.”

**Implementation** The complexity of implementing this BP again depends heavily on the dataset. A dataset that provides data which is often published as Open Data is easier to describe in standardized vocabularies for the simple reason that it is very likely someone else already did the work. In the past years schema.org became one of the main schemas in that regard, see Generic Schemas (3.3.5) for more information about finding appropriate schemas and vocabularies.

**Choose the right formalization level**

**Recommendation** Opt for a level of formal semantics that fits both data and the most likely applications.

**Purpose (Why)** “Formal semantics help to establish precise specifications that convey detailed meaning and using a complex vocabulary (ontology) may serve as a basis for tasks such as automated reasoning. On the other hand, such complex vocabularies require more effort to produce and understand, which could hamper their reuse, comparison and linking of datasets that use them.”

**Implementation** Choosing a very simple vocabulary is always attractive but there is a danger: the drive for simplicity might lead the publisher to omit some data that provides important information, such as the geographical location of the bus stops that would prevent showing them on a map. Therefore, a balance has to be struck, remembering that the goal is not simply to share your data, but for others to reuse it.

**Data Access**

**Provide Subsets for Large Datasets**

**Recommendation** If your dataset is large, enable users and applications to readily work with useful subsets of
| Purpose (Why) | “Large datasets can be difficult to move from place to place. It can also be inconvenient for users to store or parse a large dataset. Users should not have to download a complete dataset if they only need a subset of it. Moreover, Web applications that tap into large datasets will perform better if their developers can take advantage of “lazy loading”, working with smaller pieces of a whole and pulling in new pieces only as needed. The ability to work with subsets of the data also enables offline processing to work more efficiently. Real-time applications benefit in particular, as they can update more quickly.” |
| Implementation | Implementing an API is probably the most flexible approach for this BP. As an example, one can provide the data as RDF and select subsets via SPARQL or Linked Data fragments. |

### Provide real-time access

**Recommendation**

When data is produced in real time, make it available on the Web in real time or near real-time.

**Purpose (Why)**

“The presence of real-time data on the Web enables access to critical time sensitive data, and encourages the development of real-time Web applications. Real-time access is dependent on real-time data producers making their data readily available to the data publisher. The necessity of providing real-time access for a given application will need to be evaluated on a case by case basis considering refresh rates, latency introduced by data post processing steps, infrastructure availability, and the data needed by consumers. In addition to making data accessible, data publishers may provide additional information describing data gaps, data errors and anomalies, and publication delays.”

**Implementation**

Implementing this BP can be done by providing real-time APIs that support polling or streaming.

### Provide data up to date

**Recommendation**

Make data available in an up-to-date manner, and make the update frequency explicit.

**Purpose (Why)**

“The availability of data on the Web should closely match the data creation or collection time, perhaps after it has been processed or changed. Carefully synchronizing data publication to the update frequency encourages consumer confidence and data reuse.”

**Implementation**

To ensure fast publication of updated datasets it is important to automate the whole publication process as much as possible. Update frequency should be added to the metadata of the dataset by using the `dct:accrualPeriodicity` property.

### Data Access APIs

**Make data available through an API**

**Recommendation**

Offer an API to serve data if you have the resources to do so.

**Purpose (Why)**

“An API offers the greatest flexibility and processability for consumers of your data. It can enable real-time data usage, filtering on request, and the ability to work with the data at an atomic level. If your dataset is large, frequently updated, or highly complex, an API is likely to be the best option for publishing your data.”

**Implementation**

This BP can be implemented by supporting generic APIs like SPARQL or GraphQL. If necessary one can also provide domain specific APIs where appropriate.

**Use Web Standards as the foundation of APIs**

**Recommendation**

When designing APIs, use an architectural style that is founded on the technologies of the Web itself.

**Purpose (Why)**

“APIs that are built on Web standards leverage the strengths of the Web. For example, using HTTP verbs as methods and URIs that map directly to individual resources helps to avoid tight coupling between requests and responses, making for an API that is easy to maintain and can readily be understood and used by many developers. The statelessness of the Web can be a strength in enabling quick scaling, and using hypermedia enables rich interactions with your API.”
| Implementation | This would be provided by using SPARQL. In case of “classical” APIs the simplest form of such an implementation would be using a plain HTTP API. Better examples would be using REST and Hypermedia based APIs that provide better machine readability of the API itself. |

**Provide complete documentation for your API**

| Recommendation | Provide complete information on the Web about your API. Update documentation as you add features or make changes. |

**Purpose (Why)**

“Developers are the primary consumers of an API and the documentation is the first clue about its quality and usefulness. When API documentation is complete and easy to understand, developers are probably more willing to continue their journey to use it. Providing comprehensive documentation in one place allows developers to code efficiently. Highlighting changes enables your users to take advantage of new features and adapt their code if needed.”

| Implementation | This would be provided by using SPARQL. Hypermedia APIs are/should be self-describing as well. HTTP APIs do need specific documentation for developers. |

**Avoid Breaking Changes to Your API**

| Recommendation | Avoid changes to your API that break client code, and communicate any changes in your API to your developers when evolution happens. |

**Purpose (Why)**

“When developers implement a client for your API, they may rely on specific characteristics that you have built into it, such as the schema or the format of a response. Avoiding breaking changes in your API minimizes breakage to client code. Communicating changes when they do occur enables developers to take advantage of new features and, in the rare case of a breaking change, take action.”

| Implementation | Use Hypermedia APIs or standards like SPARQL. |

**Data Preservation**

| **Preserve identifiers**

**Recommendation** | When removing data from the Web, preserve the identifier and provide information about the archived resource. |

**Purpose (Why)**

“URI dereferencing is the primary interface to data on the Web. If dereferencing a URI leads to the infamous 404 response code (Not Found), the user will not know whether the lack of availability is permanent or temporary, planned or accidental. If the publisher, or a third party, has archived the data, that archived copy is much less likely to be found if the original URI is effectively broken.”

| Implementation | Do not delete the HTTP URI itself in the SPARQL-endpoint. Provide a way to return the correct status code instead like 410 (GONE). In case the resource is no longer available in the SPARQL endpoint but as a dump, point to the according archive. |

**Feedback**

| **Gather feedback from data consumers**

**Recommendation** | Provide a readily discoverable means for consumers to offer feedback. |

**Purpose (Why)**

“Obtaining feedback helps publishers understand the needs of their data consumers and can help them improve the quality of their published data. It also enhances trust by showing consumers that the publisher cares about addressing their needs. Specifying a clear feedback mechanism removes the barrier of having to search for a way to provide feedback.”

| Implementation | Provide ways of access like email, issue trackers or forms. |

**Make feedback available**

| **Recommendation** | Make consumer feedback about datasets and distributions publicly available. |

**Purpose (Why)**

“By sharing feedback with consumers, publishers can demonstrate to users that their concerns are being addressed, and they can avoid submission of duplicate bug reports. Sharing feedback also helps consumers understand any issues that may affect their ability to
use the data, and it can foster a sense of community among them.”

**Implementation**
Point to the feedback via Data Usage Vocabulary. Ideally as raw/linked data as well.

## Data Enrichment

### Enrich data by generating new data

**Recommendation**
Enrich your data by generating new data when doing so will enhance its value.

**Purpose (Why)**
“Enrichment can greatly enhance processability, particularly for unstructured data. Under some circumstances, missing values can be filled in, and new attributes and measures can be added from the existing raw data. Datasets can also be enriched by gathering additional results in the same fashion as the original data, or by combining the original data with other datasets. Publishing more complete datasets can enhance trust, if done properly and ethically. Deriving additional values that are of general utility saves users time and encourages more kinds of reuse. There are many intelligent techniques that can be used to enrich data, making the dataset an even more valuable asset.”

**Implementation**
Interlinking data sets is a topic on its own. Evaluate best practices in this domain, for example based on machine learning.

## Republication

### Provide Feedback to the Original Publisher

**Recommendation**
Let the original publisher know when you are reusing their data. If you find an error or have suggestions or compliments, let them know.

**Purpose (Why)**
“Publishers generally want to know whether the data they publish has been useful. Moreover, they may be required to report usage statistics in order to allocate resources to data publishing activities. Reporting your usage helps them justify putting effort toward data releases. Providing feedback repays the publishers for their efforts by directly helping them to improve their dataset for future users.”

**Implementation**
Provide metadata about the publisher in the application.

### Follow Licensing Terms

**Recommendation**
Find and follow the licensing requirements from the original publisher of the dataset.

**Purpose (Why)**
“Licensing provides a legal framework for using someone else’s work. By adhering to the original publisher’s requirements, you keep the relationship between yourself and the publisher friendly. You don’t need to worry about legal action from the original publisher if you are following their wishes. Understanding the initial license will help you determine what license to select for your reuse.”

**Implementation**
Read the original license and see if it can be applied for the particular use case. If there is no license given, contact the data owner and ask.

### Cite the Original Publication

**Recommendation**
Acknowledge the source of your data in metadata. If you provide a user interface, include the citation visibly in the interface.

**Purpose (Why)**
“Data is only useful when it is trustworthy. Identifying the source is a major indicator of trustworthiness in two ways: first, the user can judge the trustworthiness of the data from the reputation of the source, and second, citing the source suggests that you yourself are trustworthy as a republisher. In addition to informing the end user, citing helps publishers by crediting their work. Publishers who make data available on the Web deserve acknowledgment and are more likely to continue to share data if they find they are credited. Citation also maintains provenance and helps still others to work with the data.”

**Implementation**
Provide this in your own metadata. Provide a human readable form as well for users of the application.
3.3.1.2 Data Quality Vocabulary (DQV)

The Data Quality Vocabulary (DQV) is a direct result of the Data on the Web Best Practices Working group. It provides a framework to describe the quality of a data set. Judgement of data quality might depend heavily on the use case, the vocabulary thus does not provide a formal definition of what data quality means. Instead it provides a way to make it easier to publish, exchange and consume quality metadata.

The Data Quality Vocabulary acts as an extension to the DCAT vocabulary presented in chapter 3.4.3. It is focusing on metadata around data quality itself, update frequency, whether it accepts user corrections, if it provides persistence commitments etc. Providing information with this vocabulary will foster trust in the data amongst developers.

The two main concepts of Data Quality Vocabulary are (taken from the specification):

- A Quality Dimension (dqv:Dimension) is a quality-related characteristic of a dataset relevant to the consumer (e.g. the availability of a dataset).
- A Quality Metric (dqv:Metric) gives a procedure for measuring a data quality dimension, which is abstract, by observing a concrete quality indicator. There are usually multiple metrics per dimension; e.g. availability can be indicated by the accessibility of a SPARQL endpoint, or that of an RDF dump. The value of a metric can be numeric (e.g. for the metric “human-readable labelling of classes, properties and entities”, the percentage of entities having an rdfs:label or rdfs:comment) or boolean (e.g. whether or not a SPARQL endpoint is accessible).

This is a very flexible model, however it requires that data owners define and publish appropriate Quality Dimensions and Quality Metrics on their own or re-use existing definitions elsewhere.

3.3.1.3 Dataset Usage Vocabulary

One of the main goals to provide data to the public as Open Data is to foster re-use of said data sets. Providing information about usage of these data sets is of interest for both the data owner and potential users of the data. The owner gets more justification to invest time and money into publishing its data as open data and potential consumers can see applications of the data in real world examples. This aspect is often ignored so neither of them knows who is using which data in what way and application.

The Dataset Usage Vocabulary provides three different models: Citation, usage and feedback.

**Citation Model:** The citation model allows to cite data on the Web. It does that by mapping some of the commonly used RDF classes and properties to bibliographic fields like author, title, year and publisher. Data owners are encouraged to provide the necessary RDF classes and properties to enable proper citation. It is also possible to annotate a dataset with a memorandum or publication that gives more insights about a dataset.

---

17 [http://www.w3.org/TR/vocab-dqv/#intro](http://www.w3.org/TR/vocab-dqv/#intro)
18 [https://www.w3.org/TR/vocab-duv/](https://www.w3.org/TR/vocab-duv/)
**Usage Model:** The usage model allows to track usage of data. The two main classes used for this are `duv:Usage` and `duv:UsageTool`. The tool would be a description of an application while the usage points from the dataset to the tool.

**Feedback Model:** The feedback model is used to rate data and share feedback about it. It provides a way to describe such feedback in a similar way used on online shopping platforms among the Internet. It consists of different classes related to rating, among others `duv:UserFeedback`, `dqv:UserQualityFeedback` and `duv:RatingFeedback`.

### 3.3.2 Best Practices for Publishing Linked Data

In 2014, W3C published a compilation of Best Practices for publishing data as Linked Data.\(^{19}\) The document is partially overlapping with the Data on the Web Best Practices but has a clear focus on Linked Data publishing. It also mentioned in the Data on the Web Best Practices document. Due to the Linked Data nature, it is focusing on URI design principles, URI persistence, standard vocabularies and Linked Data specific questions like looking up a specific URI.

### 3.3.3 Web Annotation Data Model

Annotating is the act of creating associations between distinct pieces of information. Simple examples include a comment or tag on a single web page or image, or a blog post about a news article. In more complex scenarios one might for example annotate parts of a text with additional, semantic information.

The Web Annotation Data Model specification\(^{20}\) describes a structured model and format to enable annotations to be shared and reused across different hardware and software platforms. Common use cases can be modeled in a manner that is simple and convenient, while at the same time enabling more complex requirements, including linking arbitrary content to a particular data point or to segments of timed multimedia resources. The Web Annotation Data Model supersedes the Open Annotation Data Model which should no longer be used.

An annotation typically consists of a body and a target. The body describes the annotation itself and is related to the target. The relation depends on the annotation itself but the body is most frequently somehow “about” the target.

It is not uncommon to use the Web Annotation Data Model as a base for more specific annotation models. An example of such an implementation is the Fusepool Annotation Model\(^{21}\) which is built on top of Open Annotation but introduces some shortcuts and simplifications for easier querying of the data via SPARQL.

---

\(^{19}\) https://www.w3.org/TR/ld-bp/

\(^{20}\) http://www.w3.org/TR/annotation-model/

3.3.4 Shapes Constraint Language (SHACL) and Shape Expressions (ShEx)

RDF is a graph data model. By design it is possible to express any relationship between a subject and an object. In real world applications, it is often necessary to define structural constraints and validate RDF instance data against those. There are two competing standards available for this task: Shapes Constraint Language (SHACL)\(^{22}\) and Shape Expressions (ShEx).

ShEx and SHACL differ on various points. As Labra Gayo et al. (2016) put it, ShEx is more Grammar oriented while SHACL is more constraint oriented. Both ShEx and SHACL behave similarly with simple examples. The RDF syntax of ShEx is simpler to write and easier to read than SHACL, which is in that regard very graph oriented. ShEx’s focus is more on validation results while SHACL’s focus is more on validation errors. ShEx defines an abstract syntax that could be implemented in different serialization formats that are not necessarily RDF. SHACL is defined as an RDF vocabulary and can be serialized accordingly. In more complicated examples there are differences between the two standards.

ShEx is older than SHACL and was proposed at the W3C RDF Validation workshop in 2013. As a result of this workshop the W3C Data Shapes Working Group\(^{23}\) was created and the output of this working group is SHACL. However, there is still no stable standard available at the time writing so, unlike ShEx, the form of SHACL might still change and stable implementations cannot be expected yet.

For pure validation, ShEx seems to be more straightforward. For other use cases SHACL might be the better choice as the schema can be used for related use cases like creating forms and/or user interfaces dynamically. Some tools implement support for either ShEx, SHACL or both. One open source RDF validator implementing both ShEx and SHACL is available at Github\(^{24}\).

3.3.5 Generic Schemas

In 2011 several search engine giants launched schema.org (Guha et al. 2015), an initiative to "create and support a common set of schemas for structured data markup on web pages".\(^{25}\) The Semantic Web community first criticized this effort but the involved parties started talking with each other and later several people from the Semantic Web community started providing a "proper" RDF mapping\(^{26}\). Meanwhile schema.org seems to use "a simple RDF-like graph data model" and exposes its schema as embedded RDF. However, currently there is no content negotiation\(^{27}\) in place and the only language available for description of classes and labels is English.

Over the past years schema.org had a huge impact; many sites started to include structured information within their websites and the support of first RDFa\(^{28}\) and later JSON-LD\(^{29}\) made people use Semantic Web

\(^{22}\)https://www.w3.org/TR/shacl/
\(^{23}\)https://www.w3.org/2014/data-shapes/wiki/Main_Page
\(^{24}\)https://github.com/labra/rdfshape
\(^{25}\)Homepage at schema.org, source code at Github: https://github.com/schemaorg/schemaorg
\(^{26}\)See according Github issue: https://github.com/schemaorg/schemaorg/issues/317
\(^{27}\)A good introduction to content negotiation can be found at Mozilla Developer Network: https://developer.mozilla.org/en-US/docs/Web/HTTP/Content_negotiation
\(^{28}\)RDF in Attributes, see RDFa Primer: http://www.w3.org/TR/xhtml-rdfa-primer
\(^{29}\)JSON for Linking Data (JSON-LD) is a JSON based RDF serialization. See json-ld.org.
technologies without being really aware of it. This increases visibility and perception of the Semantic Web as a whole.

Using schema.org within CPaaS.io as one of the main schemas makes sense for various reasons:

- Its popularity for search engine optimization (SEO) makes it well known. Most web- and application developers probably heard of it already or even use it.
- Due to its SEO origin, schema.org provides many classes and properties related to the CPaaS.io use cases.
- There are mappings available for other popular schemas\(^{30,31}\).
- While the descriptions are only available in English, they are pretty understandable and well maintained. This is not always the case in the schema world.

At the time writing, there is also an effort underway to provide schema.org definitions for IoT related things. This will be covered in more detail in chapter 3.5.3.

### 3.3.6 Discovery Services: LOV

Sometimes none of the well-known schemas provides the appropriate attribute or class needed in the specific use case. In this situation schema search engines can help. At the time writing the best option available is Linked Open Vocabularies or in short LOV.\(^{32}\) It provides an extensive and well maintained catalogue of vocabularies that can be searched. In case one needs a new attribute or class, LOV should be the first place to look for it.

### 3.3.7 RDF Data Cubes Vocabulary

The RDF Data Cube Vocabulary\(^{33}\) is the vocabulary of choice when it comes to publish multi-dimensional data in RDF. It was first published in 2012 and has since then found many applications in the real world. The model is compatible with the widely-used cube model that underlies SDMX\(^ {34}\), an ISO standard used by many international organizations for publishing statistical data.

The base of the RDF Data Cubes Vocabulary is a DataSet. It represents a collection of statistical data that corresponds to a defined structure. The data itself can be roughly described as belonging to one of the following kinds (taken from the W3C specification):

- Observations: This is the actual data, the measured values. In a statistical table, the observations would be the values in the table cells.
- Organizational structure: To locate an observation within the hypercube, one has at least to know the value of each dimension at which the observation is located, so these values must be specified for each observation. Datasets can have additional organizational structure in the form of slices.

\(^{30}\) [http://dcmi.github.io/schema.org/mappings.html](http://dcmi.github.io/schema.org/mappings.html)
\(^{31}\) [https://github.com/mhausenblas/schema-org-rdf/tree/master/mappings](https://github.com/mhausenblas/schema-org-rdf/tree/master/mappings)
\(^{32}\) [http://lov.okfn.org/dataset/lov/](http://lov.okfn.org/dataset/lov/)
\(^{33}\) [http://www.w3.org/TR/vocab-data-cube/](http://www.w3.org/TR/vocab-data-cube/)
\(^{34}\) [https://sdmx.org/](https://sdmx.org/)
• Structural metadata: Having located an observation, we need certain metadata in order to be able to interpret it. What is the unit of measurement? Is it a normal value or a series break? Is the value measured or estimated? These metadata are provided as attributes and can be attached to individual observations, or to higher levels.

• Reference metadata: This is metadata that describes the dataset as a whole, such as categorization of the dataset, its publisher, and a SPARQL endpoint where it can be accessed.

The RDF Data Cube vocabulary provides all the necessary meta-information to create generic tools on top of RDF data cubes. This makes it a very powerful vocabulary for any kind of data that might be represented in multi-dimensional tables. For CPaaS.io this is relevant in regards to series of data coming from IoT devices.

3.3.8 Organization Ontology

The Organizational Ontology (ORG)35 allows representation of organizational structures as Linked Data. This includes the organization itself with associated metadata like classification & related structures, organizational units, sites (as in location) and reporting structures. It is also possible to model the history of an organization as these might change over time.

3.3.9 PROV-O

Provenance plays an important role when it comes to re-using published information. Provenance is related to trust and only if one understands the provenance one is able to judge if the data set can be trusted or not. Providing appropriate information about provenance can be done with the PROV Data Model. The goal of PROV is to enable the wide publication and interchange of provenance on the Web and other information systems.

The Provenance Incubator Group recommends that a provenance framework should support36:

1. the core concepts of identifying an object, attributing the object to person or entity, and representing processing steps;
2. accessing provenance-related information expressed in other standards;
3. accessing provenance;
4. the provenance of provenance;
5. reproducibility;
6. versioning;
7. representing procedures;
8. and representing derivation.

PROV is addressing various audiences, in the CPaaS.io context we focus on the RDF representation of PROV, better known as the PROV Ontology or PROV-O37. The following diagram provides a high level overview of the structure of PROV records.

---

35 https://www.w3.org/TR/vocab-org/
36 https://www.w3.org/TR/prov-overview/
37 http://www.w3.org/TR/prov-o/
In the context of publishing data via the CPaaS.io platform, Entities could be data sets available in those platforms. This dataset was created in an Activity. And this Activity was executed by an Agent. A complete introduction to PROV can be found in the PROV Model Primer. \(^{38}\)

### 3.3.10 Implications for CPaaS.io

Implementing relevant Data on the Web Best Practices is fundamental for discovery of data sets within the CPaaS.io platform. The presented vocabularies are relevant in the context of CPaaS.io and its use cases.

### 3.4 Open (Government) Data

Several use case trials in CPaaS.io involve the re-usage of Open Government Data. Many governmental and other organizations worldwide are publishing open data. The term “open government data” refers to “content that is published on the public Web by government authorities in a variety of non-proprietary formats”.\(^{39}\) For readers that are new to Open Data and its principles this document refers to the Open Data Handbook\(^{40}\) at Open Knowledge International. Best practices related to Open Government Data as described subsequently provide a basis for assessing the data to be retrieved from open government portals and to be processed in CPaaS.io. Several Best Practices from the linked open community also apply to Government Linked Data and will not be repeated here; W3C provides an overview to interrelations in that respect.\(^{41}\)

---

\(^{38}\) [http://www.w3.org/TR/prov-primer/](http://www.w3.org/TR/prov-primer/)

\(^{39}\) [http://www.w3.org/TR/ld-glossary/#open-government-data.](http://www.w3.org/TR/ld-glossary/#open-government-data.)


\(^{41}\) [http://www.w3.org/standards/techs/gld#w3c_all](http://www.w3.org/standards/techs/gld#w3c_all)
3.4.1 OGD Portals

In order to share data, e.g. based on the Directive on the Re-use of Public Sector Information (PSI) in Europe\(^{42}\), establishing an Open Government Portal for data sharing is currently considered as best practice. 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."\(^{43}\) The Share-PSI 2.0 Thematic Network at W3C\(^{44}\) promotes data portals as a solution to meet the challenge that government bodies are heterogeneous, that their information often lacks interoperability and that their datasets have different formats. A portal may fulfil the following functions:

- Making datasets available, catalogued and searchable
- Promoting and facilitating the provision of metadata
- Making applications that re-use the data visible
- Acting as a community hub

Data portals can be implemented by developing from scratch software, by using commercial solutions or open source packages. As the Share-PSI 2.0 Thematic Network best practice states, the best known example of open source software is CKAN by the Open Knowledge Foundation.\(^{45}\) A well-known commercial software is Socrata (mainly used in the US).

In collaboration with the W3C Data on the Web Best Practices Working Group, the network further provides a list of best practices for governmental open data publishers. The list complements the technical best practices by W3C Working Group (e.g. metadata, APIs, cf. 3.3.1) with non-technical best practices (e.g. policies, charging, ecosystem). Similarly, the Open Data Goldbook\(^{46}\) for data managers and data holders provided by the European Commission provides best practices for publishing open government data and in that respect refers to the already mentioned work by W3C.

A compilation of open data portals suggests that more than 2600 (governmental and other) open data portals have been set up worldwide, without providing further information on the software used to this end.\(^{47}\) Though slightly outdated, according information on European data portals can be retrieved from the OpenDataMonitor, revealing that CKAN is widely adopted.\(^{48}\) The Global Open Data Index\(^{49}\) assesses the state of open data release in currently 122 countries thereby differentiating between a range of relevant data domains and the technical and legal openness of the published data. The assessment shows that available datasets do not necessarily meet quality expectations, e.g. in terms of machine-readability.

---


\(^{43}\) https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11

\(^{44}\) http://www.w3.org/2013/share-psi/

\(^{45}\) http://ckan.org/.

\(^{46}\) https://www.europeandataportal.eu/en/providing-data/goldbook

\(^{47}\) http://opendatainception.io/


\(^{49}\) http://index.okfn.org/.
As ISA states, “opening-up data, e.g. in Open Data portals, often happens in an adhoc manner, and in many cases thousands of datasets is published without adhering to commonly-agreed data and metadata standards and without reusing common identifiers. Hence, a fragmented data-scape is created, where finding, reusing, integrating and making sense of data from different sources is a real challenge.” In order to meet this challenge, a linked data approach is promoted.50

3.4.1.1 European OGD portals

The European Union offers two portals as single entry points for accessing open data in Europe.

- European Union Open Data Portal: offers data from the institutions and further bodies of the European Union (EU)51
- European Data Portal: harvests metadata of PSI from public data portals in Europe52

The European Union Open Data Portal makes the catalogue metadata available as Linked Data, already publishes datasets as Linked Data and will increase the amount of Linked Data. The metadata stored in the portal can be queried by using a SPARQL endpoint.53 The used metadata vocabulary (based on DCAT and DCT and compatible with ADMS) is provided as worksheet specification and as ontology.

The European Data Portal also enables SPARQL queries on linked data stored in the dataset repository and supports conventional searches with filter mechanisms.54 Furthermore the portal provides a Metadata Quality Dashboard for assessing three criteria: accessibility, machine-readability55 and compliance with DCAT-AP.56 A quick check reveals that accessibility clearly can be improved, that the ratio of machine readable data is still rather low and that datasets are not fully compliant with DCAT-AP.

The EU-funded project Open Data Monitor (ODM)57 provides similar information on Open Data quality to be encountered in European OGD portals. The project has however come to an end and the information does not seem to be updated any more.

An additional study published by the European Data Portal on the open data portal maturity in Europe reveals that the majority of European open data portals provide an API (24 of 29 countries, 86%) and that efforts to provide machine-readable data are on the way. Data quality is however still a major technical barrier for reaching full open data maturity. Another barrier relevant for re-users of OGD relates to the legal issue of licenses.58

52 https://www.europeandataportal.eu/
55 Cf. e.g. ODM on Github for a list of machine-readable formats: https://github.com/opendatamonitor/odm.restapi/blob/master/odmapi/def_formatLists.py#L5-L39.
56 https://www.europeandataportal.eu/mqa-service/en
57 http://opendatamonitor.eu/
3.4.1.2 Japanese OGD portals

The Japanese government is promoting the Open Data initiative, in which the government widely discloses public data in machine-readable formats and allows secondary use of the public data for profit-making or other purposes. This initiative has the goal of improving people’s lives and stimulating corporate activities, thereby contributing to social and economic development of Japan.

The website DATA.GO.JP\(^{59}\) serves as a catalogue of this Open Data and is intended to provide information about public data available for secondary use. The website provides extensive search capabilities and a CKAN API. The website provides its metadata of this catalogue by JSON format and RDF format.

3.4.2 CKAN

CKAN is a data management system aimed at data publishers wanting to make their data open and available. It provides tools to facilitate this publishing step and helps finding and using data. As mentioned, CKAN is used in many regions, especially on Open Government Data platforms. In the context of CPaaS.io use cases, CKAN is used in related European and Japanese regions.

CKAN and other portal frameworks consist of “a content management system, some query and search features, as well as Representational State Transfer Application Programming Interface (RESTful API) to allow agents to interact with the platform and automatically retrieve the metadata and data from these portals. The metadata usually can be retrieved in a structured format via the API (e.g., as JavaScript Object Notation (JSON) data)” (Neumaier et al. 2016). It must be noted however, that data quality completely depends on the data provider. There is no additional work done on the data sets except adding some meta information. The data that gets pushed into the system is the data, which is made available to the user. Quality issues not only relate to datasets and resources, but also to metadata provided through the portal (missing or incorrect metadata) (ibid).

3.4.3 Data Catalog Vocabulary (DCAT)

The Data Catalog Vocabulary is a W3C Recommendation published in 2014.\(^{60}\) DCAT is an RDF vocabulary for describing datasets in data catalogs, e.g. government data catalogs. Using DCAT enables:

- Increasing discoverability and consumption of metadata across catalogs
- Decentralized publishing of catalogs and federated dataset search
- Interoperability between data catalogs on the web

DCAT makes use of terms from other vocabularies (e.g. Dublin Core) and defines a minimal set of classes and properties of its own. Also, DCAT acknowledges the broad range of possible format of the datasets described in a catalog and refers to complementary vocabularies for providing format-specific information.

DCAT defines three main classes and one additional class:

---

\(^{59}\) http://www.data.go.jp/, also available in English: http://www.data.go.jp/?lang=english

\(^{60}\) http://www.w3.org/TR/vocab-dcat/
• dcat:Catalog: represents the catalog
• dcat:Dataset: represents a dataset\(^{61}\) in a catalog.
• dcat:Distribution: represents an accessible form of a dataset (e.g. downloadable file, RSS feed, web service)
• dcat:CatalogRecord: describes a dataset entry in the catalog

3.4.3.1 DCAT Application Profile for data portals in Europe (DCAT-AP)

The DCAT Application Profile for data portals in Europe (DCAT-AP) is based on the Data Catalogue vocabulary (DCAT) and provides a specification for describing public sector datasets in Europe. DCAT-AP (V1.1) was prepared for the ISA Programme of the European Commission in 2015\(^{62}\) and has been implemented as the common vocabulary in the European Open Data Portal for harvesting datasets from 67 data portals of 34 countries.\(^{63}\) Thus, DCAT-AP is used to enable cross-data portal and cross-lingual search for data sets.

The 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.”\(^{64}\)

The classes and properties defined in DCAT-AP are focusing on data to be exchanged: “there are no requirements for communicating systems to implement specific technical environments. The only requirement is that the systems can export and import data in RDF in conformance with this Application Profile.”\(^{65}\)

3.4.3.2 StatDCAT-AP

StatDCAT-AP\(^{66}\) is an extension of DCAT-AP, which enables search for statistical data sets across portals. A first Draft has been prepared for the ISA Programme of the European Commission in 2016.\(^{67}\) StatDCAT-AP adds metadata elements from established standards in statistics and provides recommendation for the use of controlled vocabularies in the domain (cf. Dekkers et al. 2016).

3.4.3.3 GeoDCAT-AP

GeoDCAT\(^{68}\)-AP is an extension of DCAT-AP, which enables search for geospatial data sets across portals. An updated version of the initial specification has been prepared for the ISA Programme of the European Commission in 2016.\(^{69}\)

---

\(^{61}\) In DCAT the term “dataset” is defined as a “collection of data, published or curated by a single agent, and available for access or download in one or more formats”.


\(^{63}\) [https://joinup.ec.europa.eu/asset/dcat_application_profile/description](https://joinup.ec.europa.eu/asset/dcat_application_profile/description)

\(^{64}\) [https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11](https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11)

\(^{65}\) [https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11](https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11)

\(^{66}\) [https://joinup.ec.europa.eu/node/150075](https://joinup.ec.europa.eu/node/150075)

\(^{67}\) [https://joinup.ec.europa.eu/node/154143/](https://joinup.ec.europa.eu/node/154143/)
Commission in 2016.\(^{69}\) It provides a basis for describing geospatial data, data series and services and a RDF syntax (cf. Perego et al. 2016).

### 3.4.3.4 Asset Description Metadata Schema (ADMS)

The Asset Description Metadata Schema (ADMS) has the status of a W3C Working Group Note, published by the Government Linked Data Working Group in 2013.\(^{70}\)

ADMS is “a profile of DCAT, used to describe semantic assets (or just ‘Assets’), defined as highly reusable metadata (e.g. xml schemata, generic data models) and reference data (e.g. code lists, taxonomies, dictionaries, vocabularies) that are used for eGovernment system development.”\(^{71}\)

ADMS has been revised under the ISA Programme of the European Commission in 2016.\(^{72}\) A new version of the ADMS Application Profile (ADMS-AP) is meant to replace ADMS with the goal of integrating further interoperability solutions.

### 3.4.4 Core Vocabularies

As part of the ISA programme, the European Commission has published a set of Core Vocabularies for describing fundamental characteristics of entities that are relevant in the area of government metadata management.\(^{73}\) Core Vocabularies can be used by public administrations in the context of:

- Development of new systems: starting point for designing data models
- Information exchange between systems: basis of context-specific data models for data exchange
- Data integration: integrate data from disparate data sources / create data mash-ups.
- Open data publishing: foundation of a common export format for data in base registries

The following core vocabularies are available:\(^{74}\)

- Person vocabulary: characteristics of a person (e.g. name, gender, date of birth, location).
- Business vocabulary (Registered Organisation vocabulary, cf. 3.4.4.1): characteristics of a legal entity (e.g. identifier, activities), created through a formal registration process (governmental register)
- Location vocabulary: characteristics of a location (represented as address, geographic name or geometry).
- Public Service vocabulary: characteristics of a service offered by public administration.

ISA furthermore provides a Core Data Model Mapping Directory (collection of mappings between Core Vocabularies and related Core Data Models)\(^{75}\) and a handbook on the topic (European Union 2015).

---

\(^{70}\) http://www.w3.org/TR/vocab-adms/
\(^{71}\) http://www.w3.org/standards/techs/gld#w3c_all
\(^{72}\) https://joinup.ec.europa.eu/asset/adms_revision/description
\(^{74}\) https://joinup.ec.europa.eu/asset/core_vocabularies/asset_release/all
\(^{75}\) http://mapping.semic.eu/
3.4.4.1 Registered Organization Vocabulary

The ISA Core Business Vocabulary has been renamed into Registered Organization Vocabulary (RegOrg) in 2013 and published as Working Group Note by W3C.\(^{76}\) It is a profile of the Organization Ontology for describing organizations that have gained legal entity status through a formal registration process. This is in general useful for publishing information about government related structures and/or content of public registries, such as company registries.

The vocabulary itself is straightforward, its key class is `rov:RegisteredOrganization`. Attached to this class are some basic properties like its legal name and alternative names. With public registries, organizations often get a unique identifier assigned. This can be expressed in various ways using this vocabulary. It also provides additional metadata like the type of organization, its status or things like its activity, which in itself should be recorded using a controlled vocabulary.

3.4.5 Open Data Licences

In the context of Open (Government) Data, licensing is an important issue. Generally, the term “licence” is defined as “a legal document giving permission to use information” and it is considered as “a mechanism that gives people and organisations permission to re-use information and other material that is protected by copyright or database right. A licence should also provide clarity as to what users and re-users are permitted to do and whether there are any restrictions on the extent of that permission” (UK Government Licensing Framework, 2014, cit. in Khayyat & Bannister, 2015: 233). Organizations such as Open Knowledge International promote the adoption of Open Licences when publishing Open data. Such licences should conform to the “Open Definition”, in terms of being: reusable (not specific to an organization/jurisdiction), compatible (with one or more predefined licenses) and current (considered as best practice, e.g. in a given domain). Recommended licenses are:\(^{77}\)

**Table 15: Overview Open Data Licences (Open Knowledge International)\(^{78}\)**

<table>
<thead>
<tr>
<th>License</th>
<th>Domain</th>
<th>BY</th>
<th>SA</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creative Commons CCZero (CC0)</td>
<td>Content, Data</td>
<td>N</td>
<td>N</td>
<td>Dedicate to the Public Domain (all rights waived)</td>
</tr>
<tr>
<td>Open Data Commons Public Domain Dedication and Licence (PDDL)</td>
<td>Data</td>
<td>N</td>
<td>N</td>
<td>Dedicate to the Public Domain (all rights waived)</td>
</tr>
<tr>
<td>Creative Commons Attribution 4.0 (CC-BY-4.0)</td>
<td>Content, Data</td>
<td>Y</td>
<td>N</td>
<td>Attribution for data(bases)</td>
</tr>
<tr>
<td>Open Data Commons Attribution License (ODC-BY)</td>
<td>Data</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)</td>
<td>Content, Data</td>
<td>Y</td>
<td>Y</td>
<td>Attribution-ShareAlike for data(bases)</td>
</tr>
<tr>
<td>Open Data Commons Open Database License (ODbL)</td>
<td>Data</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

BY: = requires attribution  
SA: = requires share-alike

\(^{76}\) [http://www.w3.org/TR/vocab-regorg/](http://www.w3.org/TR/vocab-regorg/)  
\(^{77}\) [http://opendefinition.org/licenses/](http://opendefinition.org/licenses/)  
\(^{78}\) [http://opendefinition.org/licenses/](http://opendefinition.org/licenses/)
Open Knowledge International also lists a set of further licences that may not meet all criteria and provides an FAQ for license-related questions, another categorization with more information on the licences is also provided in (Khayyat & Bannister, 2015: 252). From the perspective of re-users of Open Government Data, licencing poses some challenges, since published data may be owned and licensed by third parties. As Krötsch & Speiser (2011: 354) point out, "checking license compliance is a tedious manual task (...)" and therefore a "natural goal (...) is to accurately model usage policies as part of the data so as to enable their easy exchange and automated processing". They propose a semantic framework for modelling self-referential policies, such as Creative Commons Share-Alike licence. When data from different sources are combined, the handling of licences gets quite complex. This may even be the case if data is coming from the same licence "family". E.g. within Creative Commons, one licence may trump the other and in some cases this may pose limitations on data use: "For example of data from a BY-NC-SA (attribution, no commercial use, share alike) are mixed with data from a BY-SA set. NC is inherited, i.e. all downstream uses must be non-commercial, but BY-SA requires that there be no restriction on subsequent use. (Khayyat & Bannister, 2015: 246)

### 3.4.6 Implications for CPaaS.io

CPaaS.io will use OGD from (probably) CKAN based portals. Despite existing best practices for Open Data publishers, high data quality cannot be presupposed. The Metadata Quality Dashboard provided by the European Open Data Portal can be used for an overall assessment of the quality of European portals. Data management in the project will have to identify and address potential data quality related to both, data resources and metadata problems with respect to the use cases. Furthermore, the mechanisms for ensuring that data as needed by use case applications is accessible in a in a sustainable manner will have to be defined. Taking on a generic point of view, the handling of licensing when using and distributing data via CPaaS.io will have to be analyzed in more detail.

### 3.5 IoT Data

Many CPaaS.io use cases are relying on IoT data. Both the FIWARE and u2 implementation platforms provide data collected from IoT devices, some of the data can be directly compared and used once the semantic barriers between the two platforms are removed. To interface to these data sets one will need a common understanding of the underlying semantics. This can be solved by creating and using appropriate vocabularies. The IoT topic is also highly discussed in several communities, among others the one from schema.org and the Web of Things Interest Group at W3C.

FIWARE and u2 store their data in their specific internal format. For FIWARE this is the NGSI data model and for u2 the ucode model. Details for these data models can be found in the subchapter on FIWARE and u2 Data Models above (3.2). Using these data sets is fine as long as data sets do not have to provide interoperability.

From the Web of Things Interest Group website:

---

79 [http://opendatacommons.org/faq/licenses/](http://opendatacommons.org/faq/licenses/)
81 Related Github issue proposing the group: [https://github.com/schemaorg/schemaorg/issues/1272](https://github.com/schemaorg/schemaorg/issues/1272)
82 [https://www.w3.org/WoT/IG/](https://www.w3.org/WoT/IG/)
The Internet of Things (IoT) suffers from a lack of interoperability across platforms. As a result, developers are faced with data silos, high costs and limited market potential. This can be likened to the situation before the Internet when there were competing non-interoperable networking technologies. The Internet makes it easy to develop networked applications independently of those technologies. W3C is seeking to do the same for the Internet of Things.

To achieve this goal, we need platform independent APIs for application developers, and a means for different platforms to discover how to inter-operate with one another. The approach we are taking is based upon rich metadata that describes the data and interaction models exposed to applications, and the communications and security requirements for platforms to communicate effectively. A further aspect is the need to enable platforms to share the same meaning when they exchange data. We are therefore seeking to enable expression of the semantics of things and the domain constraints associated with them, building upon W3C’s extensive work on RDF and Linked Data.  

This completely applies to the CPaaS.io platform. Note that both the schema.org group and the W3C Interest Group are heavily work in progress, at the time writing neither of the two groups have evolved to a stage, in which publishing of a stable standard can be immediately expected (e.g. formation of a W3C Working Group).

3.5.1 SSN Ontology

Semantic Sensor Network (SSN) Ontology seems to be one of the oldest efforts to bring IoT devices in contact with Semantic Web and Linked Data technologies. It was initially developed by the Semantic Sensor Network Incubator Group within the period of 2009 to 2011. Unfortunately, some of the namespaces used in the specification are no longer resolving. For that and other reasons the Spatial Data on the Web Working Group took over responsibility for SSN and re-published a revised version as a Public Working Draft in 2016. At the time writing this document is also heavily work in progress.

The goal of SSN is to provide an ontology for describing sensors and the observations they make of the physical world. SSN is published in a modular architecture that supports the judicious use of "just enough" ontology for diverse applications and is based on OWL-2 DL. The SSN ontology can be seen as a general purpose vocabulary because it is application domain independent and provides concepts about sensors and their observations (Khan et al. 2015).

SSN gets criticized by practitioners for its complexity, which makes it hard for newcomers to use. SSN dates from a time where vocabularies were built in very formal ways and took heavy use of features from knowledge representation languages like Web Ontology Language (OWL). This is in stark contrast to

---

https://www.w3.org/WoT/IG/
https://www.w3.org/TR/vocab-ssn/
https://www.w3.org/2005/Incubator/ssn/charter
https://www.w3.org/2015/spatial/wiki/Main_Page
https://de.wikipedia.org/wiki/Web_Ontology_Language
lightweight and community-based efforts like schema.org, which seem to get much more traction within the past years.

As an effort to facilitate this, the new draft of the SSN ontology can now be used independently of DOLCE ultralite ontology (DUL). DUL is partially responsible for the complexity of the ontology. The changes are not complete yet; it remains to be seen if the simplification of SSN will result in additional utilization. Vocabularies like FIESTA-IoT are mapping to SSN where possible.

### 3.5.2 WoT Current Practices and Thing Description

The Web of Things (WoT) Interest Group is collecting concepts and technologies to enable discovery and interoperability of Internet of Things (IoT) services. As mentioned above the group is work in progress, one of the main outcomes so far is the Thing Descriptions and Interactions. Charpenay et al. (2016: 6) state:

> “Alongside the implementation work done within the interest group, two main concepts have emerged. First, WoT Things should describe themselves and be able to exchange their description with other agents. We call such a description a Thing Description (TD). Second, WoT Things should expose to the Web a set of Interactions, which corresponds to their interface to the physical world”.

This leads to two main concepts:

- **Thing Description**: Semantic resource formally describing a unique WoT Thing that a software agent can interact with. Examples of WoT Things include building rooms, manufactured products, mechanical systems but also digital control devices, i.e. any real-world entity without a priori restriction.

- **Interaction**: Web resource of arbitrary content format acting as a digital proxy for any real-world entity that is not already digital information. Such entities can be physical quantities like temperature or pressure, natural phenomena like raise of temperature or object motion, arbitrary states like on/off, etc.

In the current specification, there are concepts for Thing, Property, Action and Event. The last three are part of interaction. Some aspects of their definition is left open on purpose. This allows alignment with existing IoT vocabularies.

There is also a W3C Web of Things Community Group. W3C is in the stage of forming a Working Group, the Community Group is likely to draft specifications, based on requirements from both Groups.

### 3.5.3 IoT + Schema.org

In Summer 2016 the schema.org community started to discuss support of Internet of Things-related applications of schema.org. In a more detailed proposal the schema.org team explains its motivation in more detail under the label "IoT + schema.org".

---

89 [https://www.w3.org/community/wot/](https://www.w3.org/community/wot/)
“Adopting and improving schema.org’s vocabularies for IoT can reduce needless duplication and focus attention on innovation rather than replication. Schema.org’s approach is open, organic and community based, and is supported by a technical design that includes several extensibility mechanisms designed to support decentralization without encouraging fragmentation.”

The big difference in this effort is the organic and community based approach. schema.org introduced this and is very successful in getting traction and applications, unlike many other, older and more “elaborate” vocabularies out there which are often created by small focus groups and/or companies within a specific time frame.

Beside core schema.org terms for IoT (e.g. hypothetical addition of a basic type like “Device”, or improvements around measurements, actions and events) it will support a so-called “hosted extension” for IoT, which is a feature of schema.org when specific extensions can be hosted under a schema.org namespace. Existing examples are auto.schema.org or health-lifesci.schema.org.

The IoT + schema.org effort is currently split in three phases: Part of phase 1 is setting up topical discussions and collection of existing works. Phase 2 is exploration of existing or news proposed schemas as extensions to schema.org, phase 3 is modeling and prototyping explorations and dialogue with relevant standards efforts such as W3C WoT. At the time writing phase 2 is active, phase 3 is primarily planned during 2017.

It is currently difficult to predict what will happen within this effort. Based on the success of schema.org one can assume that once there is a first version released (currently targeted at schema.org release v3.2) vendors will pick it up fast and integrate it into their IoT devices.

### 3.5.4 FIESTA-IoT Ontology

The FIESTA-IoT\(^\text{92}\) is framed within the EU H2020’s FIESTA-IoT project that aims to seamlessly support the federation of testbeds through the usage of semantic-based technologies. FIESTA-IoT is re-using some concepts from various related ontologies like Semantic Sensor Network (SSN), M3-lite (a lite version of M3 ontology), WGS84, IoT-lite, Time, and DUL ontology. Detailed arguments for the development of FIESTA-IoT and an analysis of the current state of existing ontologies is provided in a paper (Agarwal et al. 2016). The paper goes into great details about other related ontologies not mentioned in this deliverable, although latest developments about SSN, schema.org and WoT Current Practices are not mentioned.

FIESTA-IoT can be seen as an up to date application of some earlier concepts and ontologies. It tries not to create new concepts but borrowing from existing ones and by addressing interoperability issues and various aspects of IoT device observations.

\(^{90}\) [https://github.com/schemaorg/schemaorg/issues/1272](https://github.com/schemaorg/schemaorg/issues/1272)

\(^{91}\) [http://iot.webschemas.org/docs/iot-gettingstarted.html](http://iot.webschemas.org/docs/iot-gettingstarted.html)

3.5.5 Preliminary work by project partners

Regarding IoT, the project can make use of the work done by AGT International that has developed a (proprietary) deployment ontology for describing physical IoT deployments. The deployment ontology builds on the SSN ontology and covers concepts to manage IoT events, devices and people and groups of people and their relations in the context of a deployment. The ontology provides concepts for describing events including their location, duration and venues in which IoT devices are being deployed. The ontology provides concepts for both stationary and mobile deployments. Stationary deployments are associated with floor plans on which devices can be placed and associated with additional metadata such as observation zones and the duration of a deployment. AGT uses mobile associations to model deployments of mobile devices such as smart phones with mobile entities such as users. Here the ontology reuses and extends some concepts of the FOAF ontology.

The Deployment Tool, a tool to manage the deployment knowledge base developed in Task 6.5, fully builds on the deployment ontology. AGT has already provided an initial version of the deployment tool to partners in the project and modelled metadata for the Color Run event as part of the Event Management Use Case. The data generated by this tool is used for metadata enrichment of raw data and analytic results as described in deliverable D2.1.

The ontology also provides some annotation properties that allow for controlling how a UI such as the deployment tool displays ontology classes and instances.

3.5.6 Implications for CPaaS.io

It is at the time writing hard to say which of the presented, non-exhaustive list of vocabularies will win. Currently there is a choice between abstract concepts (SSN), very concrete implementations (FIESTA-IoT) with many classes and attributes or early efforts (Thing Description and IoT + schema.org). Regarding the CPaaS.io use cases, going into the direction of FIESTA-IoT is the most reasonable choice currently. This decision may be revised depending on the adoption of the mentioned vocabularies and how alignment goes with other candidates.

In its initial incarnation, the CPaaS.io platform will provide data in either the NGSI or ucode data model. The semantic layer will be implemented step by step, starting with metadata about the respective data sets. In further steps interoperability with RDF vocabularies will be added to the CPaaS.io platform, driven by use cases. The project seeks to adopt a linked data approach.

3.6 Social Media Data

Representing social media data can be done with multiple vocabularies. The oldest one is FOAF, also known as “Friend of a Friend”.\(^93\) It got extended by the SIOC (Semantically-Interlinked Online Communities) Core Ontology\(^94\) which modelled additional elements needed for online publications like blogs, discussion forums, web sites etc. Recently many of the elements introduced by FOAF and/or SIOC

\(^93\) http://xmlns.com/foaf/spec/
\(^94\) http://rdfs.org/sioc/ns#
can be found within the schema.org namespace as well. One of the main purposes of schema.org is search engine optimization or SEO. It makes sense to enrich web pages with more specific descriptions about its content. Thus many of the things FOAF and SIOC defined can now be found in schema.org. The FOAF specification is still actively maintained and defines mappings to schema.org classes via owl:equivalent properties for FOAF Person (Person), Image (ImageObject) and Document (CreativeWork). For SIOC there are some manually maintained mappings available, but they might be outdated and/or incomplete.

### 3.6.1 Implications for CPaaS.io

Social media data can be considered a potential source of data to be used in the platform. The inclusion of social media data could enhance the usefulness of some use cases, in particular event management targeting a more interesting experience for event participants, like for example the management of the Color Run. Hence we address this topic here, even though at the time of writing it is not fully clear yet if and how social media is going to be used.

### 3.7 Personal Data / User Empowerment

Privacy and data protection are of particular relevance to CPaaS.io. Several use case trials in CPaaS.io involve end-users and entail the collection and use of personal data, e.g. from the IoT (cf. deliverables D1.1 and D2.1). Privacy protection is also an issue regarding the re-usage of open (government) data. The focus of open data inherently is on non-personal data (i.e. not containing information about individuals). Open data publishers therefore take precautions to render data non-personal by anonymising them. Unless the information of the data is reduced below utility, anonymization is however reversible (e.g. with the use of auxiliary data) (Simperl et al. 2016: 12). One of the approaches to deal with concerns associated with data protection risks of publishing (non-personal) open data is to engage in a stakeholder dialogue with data subjects and consumers (Simperl et al. 2016: 18). This is one way of adopting a user-centric approach in addition to the indispensable technical approaches for ensuring the privacy of personal data. There are numerous factors influencing technical and policy choices within CPaaS.io and the following chapters are going to illustrate the most relevant ones for the decisions to be made within CPaaS.io regarding policies and technical solutions in WP5. The focus is on current considerations and initiatives that are relevant to enhancing user empowerment.

#### 3.7.1 Data Protection

##### 3.7.1.1 European General Data Protection Regulation (GDPR)

Some years ago, the European Commission has put forward a data protection reform with the aim of strengthening citizen’s rights, while fostering the development the digital market in Europe through

---

creating trust. The new Directive\textsuperscript{96} has entered into force in 2016 and will have to be transposed to national law by 2018.\textsuperscript{97} Essential principles in the EU data protection rules include.\textsuperscript{98}

Related to data subjects

- Transparency: Facilitation of exercising data subjects rights through easy accessible and easy to understand information and communication on the processing of personal data (e.g. information on the controller, the purposes of processing, the recipients of the data, period of data storage, etc.)
- Right of access and right to data portability: Information on (category of) personal data and it’s processing and access to a copy of the data. Transmittance of personal data to other data controllers.
- Right to rectification and erasure (“right to be forgotten”): Correction of inaccurate personal data, completion of incomplete data. Erasure of personal data (e.g. based on lack of ground for processing, consent withdrawal etc.)

Related to data controllers and processors:

- Data protection by design and by default: Implementation of appropriate technical and organizational measures for data protection and security (e.g. data minimization, pseudonymisation, encryption). Adoption of privacy-friendly default settings

While there is no point in replicating existing and forthcoming data protection regulations in detail here, their purpose is worth stressing again. The stronger data protection rules are meant to increase trust in online services and fostering their uptake. Data protection regulations are laying the grounds for expected (and legally enforced) “best practice” in service provision related to data privacy and data security.

3.7.1.2 European Commission Staff Working Document on IoT

The European Commission has adopted a set of policy measures with the aim of advancing the Internet of Things in Europe and associated economic and societal innovation. This includes the publication of the supporting Staff Working Document on “Advancing the Internet of Things in Europe”.\textsuperscript{99} Envisaged actions are grounded on three pillars:

- A single market for the IoT: Focus on Interoperability
- A thriving IoT ecosystem: Focus on open platforms
- A human-centred IoT: Focus on empowering people regarding data protection and security

CPaaS.io addresses the target of interoperability by integrating data flows from numerous different sources. For the data consumer the fact that data originated from technically differing sensors or other data sources does not only remain invisible but above all irrelevant. The data consumer can access data

\textsuperscript{97} http://ec.europa.eu/justice/data-protection/reform/index_en.htm
\textsuperscript{99} http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15276
reports or can trigger own queries on the data provided by CPaaS.io as long as it coheres to the open SPARQL standard supported by a wide variety of tools.

CPaaS.io is striving for wide applicability and accessibility of the concept. Hence, it is built with a strong focus on open and royalty-free standards and licenses. Using open standards like the FIWARE APIs, the standards declared by the Kantara initiative, the RDF and SPARQL standards among other standards defined by W3C keeps the concept open not only for further instances beyond this project, but also open for improvement and development at a later point in time.

In line with the goals of the GDPR, adopting a human-centred IoT approach means strengthening trust, security, personal data protection and privacy, while taking into account the needs of the digital IoT industry (European Commission 2016: 27). Data protection rules apply whenever data that relates to identified or identifiable natural persons are processed. The goal of data protection is not to deprive the digital industry from answers but rather to distinguish between trends on a multitude of levels, e.g. societal, local, cultural, generational, linguistic, etc. and information relating or relatable to a single person. It is important for the data subject to understand that even if personal information is shared through CPaaS.io or any other platform operating under the GDPR, this information is only published in heavily aggregated form not allowing for conclusions with regard to individual data subjects.

Despite all these precautions, the data subject remains the sole sovereign regarding its personal data. Consequently, s/he is the last authority when it comes to his/her data and therefore has to be able to exercise total control over his/her data. In order to enable the data subject to exercise such control, CPaaS.io’s approach is twofold:

One the one hand, simply accessible visualizations of the data currently stored and being (potentially) published is provided in combination with a graphical user interface supporting the data subject in the formulation of the access conditions for this data. Without any programming skills being necessary, data access conditions can be defined with a few mouse clicks.

An average data subject, however, will broadly speaking need more support than just a visualization of the currently stored data to make up his/her mind whether sharing the data is a good idea or not, and if yes, under which conditions the data should be shared. There are several conceivable aspects to this problem. As the data-subject is supposed to act in a self-determined way according to the GDPR, information is the key to success when it comes to understanding the own data vault and technological support when it comes to the definition of the sharing conditions. Information concerning data is a very broad term and many of the meta-data are not very informative to the untrained data subject. Consequently, the meta-data of personal data has to be either better prepared or simply better understood to make the approach productive and therefore attractive. There are several roles conceivable to mend this problem:

- **The Data Subject**: The simplest possibility is to pass the responsibility on to the data subject and demand its sufficient training. This rather normative approach, however, does not support a broad acceptance of this new technology and is therefore not very likely to support its dissemination.

- **The Data Creator**: Another way is addressing the data creators and determining which meta-information for the data subject has to be attached to each data set by the data creator. An
obvious advantage of this approach is the deep knowledge a data creator has regarding the data s/he creates as long as the data creator is human. As soon as a device acts as a data creator, at least some of the meta-information attached to the data sets will have to be previously prepared for the device by a human being. The disadvantage of this approach is the concentration of power at the data creator. Putting the data creator in charge of the decision which information is attached to a data set and which is not, creates opportunities for manipulation. Therefore, some kind of control mechanism has to be put in place, when putting the data creators in charge of the meta-data.

- **CPaaS.io**: The third possibility besides the data subject or the data creator is the system handling the data itself. This approach eliminates mistakes made by data subjects due to lacking training or data creators due to lacking loyalty by handing the decision over to a neutral IT system. As a system capable to reliably augment data sets with fitting meta-data, however, is unavailable, the CPaaS.io operator would have to do this manually or simplify the situation by only accepting a very narrow range of previously standardized data types.

- **Supported Data-Subject**: A combination of the first and third approach is conceivable as well, resulting in a system supporting the data subject with additional information, assumptions, statistical probabilities or links to additional information and guidebooks.

Important questions to reflect on are (European Commission 2016: 28).

- How to ensure end-user understanding of the role, functioning and possible impacts of IoT services?
- What precautions to take for making valuable information accessible on legitimate grounds (medical information)
- How to facilitate user control over their data and over data sharing without assuming technological knowledge

The European General Data Protection Regulation (GDPR) promotes several techniques to protect personal data:

- Anonymization: “removing personally identifiable information where it is not needed”
- Pseudonymisation: “replacing personally identifiable material with artificial identifiers”
- Encryption: “encoding messages so only those authorised can read it” (European Commission 2016: 29).

Anonymisation and pseudonymisation have both been used for decades. While pseudonymisation inherently bears a higher risk than anonymisation, as there is still the possibility to revert the process and re-identify the data set, it also offers the possibility to learn more after the data analysis by addressing a few interesting cases from the data set directly. Anonymisation of larger data sets must however nowadays be considered a thing of the past due to the immeasurable volume of data available to interrelate the anonymized data set with and the new approaches and tools built in the last few years exactly to cope with large volumes of versatile data (Narayanan & Welten 2014 and 2015; PCAST 2014).

As for users’ trust, the Commission services list a range of challenges to be addressed, e.g. how to ensure “a context-based security and privacy, which reflects different levels of importance (e.g. emergency crisis,
home automation)” (European Commission 2016: 29f.); an issue that is also relevant in the CPaaS.io context.

### 3.7.1.3 Inquests of Personal Data Protection in Japanese Government

In 2003, Japanese government enacted the Act of the Protection of Personal Information. The purpose of this Act is to protect the rights and interests of individuals, while taking consideration of the usefulness of personal information. In view of a remarkable increase in the utilization of personal information due to development of the advanced information and communications society, the focus is on clarifying the responsibilities of the state and local governments, etc. with laying down basic principles, establishment of a basic policy by the government and the matters to serve as a basis for other measures on the protection of personal information, and by prescribing the duties to be observed by entities handling personal information, etc., regarding the proper handling of personal information.

In 2013, The IT general strategy headquarters of the Japanese government established a study group on personal data. This study group aims to promote the revitalization of existing industries, to maximize the power of the private sector by utilizing personal data, and to create new business and new services, while preserving protection of personal information and privacy. This study group focuses on procedures for clarifying the purpose of protecting personal data, clarifying the range of protected personal data, and appropriately using and distributing personal data that takes privacy into account. Regarding procedures, we are considering procedures to be taken by business operators in expanding the purpose of use and exceptional provision (opt-out) of principal consent principle in third party offering.

### 3.7.2 User Centricity

Data Security and data protection have been important topics for years. The new focus coming up with legislations, like the ones from Europe and Japan mentioned in the chapters above, is shifting from purely protecting the data itself to perceiving the data subject and his/her personal data as one unit entitled to protection. Protecting the data subject in this context means protecting the data subject’s right for self-determination. Consequently, the data subject must be protected, or in other words empowered to execute, when s/he decides to allow or disallow access to personal data (Bus & Nguyen 2013).

#### 3.7.2.1 MyData Approach

A MyData approach focuses on personal data management and processing and therefore revolves around the data subject and its decisions and needs. It empowers the data subject to exert his/her right to decide on who gets access to their personal data and what this data is used for. The idea bases on the individual’s role as a gatekeeper, deciding who gets access to which part of his/her personal data by giving or withdrawing consent.

---

100 [http://www.japaneselawtranslation.go.jp/law/detail_main?id=130](http://www.japaneselawtranslation.go.jp/law/detail_main?id=130)
As the term “MyData” has more and more been coined by the specific approach of the MyData Alliance\(^{101}\) the term will not be used in a general sense anymore in order to prevent misunderstandings. The MyData Alliance is a network running their version of personal data storages (Mortimer 2011) quite similar to the CPaaS.io but also with a few distinct differences. The MyData Alliance agrees in the goals to empower the data subject to decide the fate of his/her data. The way to reach this goal, however, differs from CPaaS.io to the MyData Alliance in significant ways. The MyData Alliance aims to prevent a provider lock-in by defining the storage standards in a way allowing the data subject to change his person data storage provider, should s/he feel the need. Since CPaaS.io has been built with a smart city context in mind, with centralized authorities or event coordinators acting as a trusted third party, switching the provider is not a priority, while making the system interoperable for different cities or events still is.

While monetizing personal data can be within CPaaS.io’s scope, its primary focus is on providing data concerning a city and its inhabitants, where and when it is needed, and protect the data subject’s privacy while doing so (see use case descriptions in D2.1). Furthermore, the MyData Alliance aims to store a broad variety of data types if not all personal data, while CPaaS.io is focusing on providing a city with relevant information on a super-individual scale. This lowers the complexity and therefore security of the CPaaS.io approach on the one hand. On the other hand, it means that public and private data have to be combined to get a complete picture, while protecting the private data part from illicit access. While the MyData Alliance has opted to allow users to sell their data, the CPaaS.io approach choses to grant access to data analysis. SPARQL queries can be made upon public and anonymised private data, as long as the query itself does not isolate single or small groups of individuals and private data is only reported back in an aggregated state.

Due to these differences, the generic term of Personal Data Store will be used to describe the idea of private citizens being able and allowed to manage their personal data and access to it by granting or withdrawing their consent, in one way or another.

### 3.7.2.2 Consent & Information Sharing Work Group (CISWG) (Kantara Initiative)

There is another aspect worth highlighting in this section. It is the need for providing a mechanism that allows citizens to give explicit consent on some actions like for instance revealing private information, or enrolling a service. The Consent & Information Sharing Work Group (CISWG) thanks to Kantara Initiative\(^{102}\) is aimed at gathering information of the different use cases and scenarios that illustrate the exchange of such information so as to specify the policy and technology enablers that should be selected to enable this information to flow.

For CPaaS.io, User Managed Access (UMA) is at the core of consent and information sharing concept, allowing an Internet user to control a unified control point for authorizing who and what can get access to their digital resources, no matter where the resource is located. UMA reduces the flow of data and information by not transferring all the attributes found associated with an account or a resource, but rather allows for questions to be asked in order to test their suitability for receiving authorization.

\(^{101}\) [https://mydatafi.wordpress.com/](https://mydatafi.wordpress.com/)

\(^{102}\) [https://kantarainitiative.org/groups/ciswg/](https://kantarainitiative.org/groups/ciswg/)
Therefore e.g., with UMA the day of birth is not needed to check for age-based access. Rather the question “Is the attribute ‘age’ of this identity 21 or higher?” is asked and either confirmed or denied. In this way, no information is unnecessarily revealed and nevertheless rules can be enforced.

3.7.3 Aspects of generic Personal Data Storage Models

Personal Data Storage (PDS) Models generally offer their customers or users an opportunity to store some or all of their personal data in one, secure place (Kirkham et al. 2011). There is a wide selection of conceivable PDS models (Kirkham et al. 2013), so the focus of this chapter is to discuss the main modelling characteristics and finally point out their implication for CPaaS.io. As a baseline, this subsection is to be read against the most simple but still widespread manner of storing personal data: saving the data on a local data carrier (e.g. HDD/SDD, CD/DVD, USB drive, etc.) and storing the carrier. Solutions like cloud file storages are considered equivalent, as the data is copied to another data carrier in order to preserve the possibility to restore the data it contains.

3.7.3.1 Centralized vs. decentralized PDS

Most PDS models operating these days have a centralized setup. Users are given the possibility to store all or some of their personal data with a PDS provider. In this setup, the provider becomes a trusted third party, as the data subjects have to trust that the PDS provider will not abuse access to their personal data.

Alternatively, there is the possibility of a decentralized PDS, where the user data is stored in any number of decentralized data stores, preferably where they have been created. The PDS platform addresses these decentralized data stores when data or information concerning data is needed. The addressing of the decentralized data stores can be realized in different ways, allowing for more or fewer layers of anonymity.

Table 16: Advantages and disadvantages of centralized vs. decentralized PDS

<table>
<thead>
<tr>
<th></th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralized PDS</td>
<td>• Lower operational efforts</td>
<td>• PDS provider is additional trusted third party</td>
</tr>
<tr>
<td>(e.g. hubofallthings.com)</td>
<td>• Security measures must be applied once</td>
<td>• High-value target for attacks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Larger impact in case of successful attack</td>
</tr>
<tr>
<td>Decentralized PDS</td>
<td>• Lower risk of total data theft or loss due to decentralized setup</td>
<td>• Fragmented security due to decentralized setup</td>
</tr>
<tr>
<td>(e.g. bitdust.io)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.7.3.2 Managed vs. unmanaged PDS

A crucial aspect of a PDS model to decide is whether the model shall be managed or not. Unmanaged PDS do not offer any templates, forms, tables, etc. to fill the personal data into, but rather only disk space for the user files to be saved on, while managed PDS target specific kind of personal data such as health data, geo location data, consumer data, identities, etc.
### Table 17: Advantages and disadvantages of managed vs. unmanaged PDS

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Managed PDS (e.g. midata.coop)</td>
<td>• Easier data submission and management due to templates, forms, help texts, etc.</td>
</tr>
<tr>
<td></td>
<td>• Statistics or conditional exports possible</td>
</tr>
<tr>
<td>Unmanaged PDS (e.g. dropbox.com)</td>
<td>• In combination with a centralized PDS, the PDS provider must have read-access to data.</td>
</tr>
<tr>
<td></td>
<td>• Encryption can be applied prior to uploading data to the PDS to prevent access by PDS provider</td>
</tr>
<tr>
<td></td>
<td>• All kind of personal data can be stored.</td>
</tr>
</tbody>
</table>

### 3.7.3.3 PDS Price model

Providing a PDS service causes effort and hence produces costs. Consequently, all PDS providers who are not run by a governmental or charity organization have to come up with a price model or rely on the data subject to self-host a solution and cover the costs. Apart from differences in price heights, these are two main approaches what to use as a mean to counter costs.

The well-known model is the classic financial model, where the PDS user pays either a volume-dependant or fixed sum per month to cover the PDS provider’s costs. Whether the PDS user uses the PDS to monetize the personal data is an individual choice and can be done to counterbalance the PDS costs. Alternatively, PDS providers can be paid by granting them access to personal data and consenting to the PDS provider to sell data or access to data for his own profit. Obviously, there are numerous different possibilities how to define which data can be shared for which price or cannot be shared at all.

### Table 18: Advantages and disadvantages of different PDS price models

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Model / Self-Hosting (e.g. storj.io)</td>
<td>• Personal data remains untouched</td>
</tr>
<tr>
<td></td>
<td>• Higher financial costs</td>
</tr>
<tr>
<td>Data-based Model (e.g. datacoup.com)</td>
<td>• No or lower financial fees</td>
</tr>
<tr>
<td></td>
<td>• Control over personal data lost to a certain degree</td>
</tr>
</tbody>
</table>

### 3.7.3.4 Data vs. Analysis Access

Two elements are essential to a data analysis: the data and the analysis algorithm. While the classic way is moving the data to the analysis algorithm in order to generate insight, it works just as well the other way around. In the former case, access is granted to data, while in the latter case access is granted to an analysis result regarding data.

### Table 19: Advantages and disadvantages of Data vs. Analysis Access

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Access (e.g. midata.coop)</td>
<td>• After data has been handed out, there is no way of guaranteeing that it will only be used for the intended purpose.</td>
</tr>
<tr>
<td></td>
<td>• Data leaves its silo and is thus potentially exposed</td>
</tr>
<tr>
<td></td>
<td>• Data consumer is responsible for a correct analysis</td>
</tr>
<tr>
<td>Analysis Access (e.g. openpds.media.mit.edu)</td>
<td>• An analytics provider has to be found and compensated</td>
</tr>
<tr>
<td></td>
<td>• The steps to facilitate an analysis access are more complicated and thus costly.</td>
</tr>
</tbody>
</table>
3.7.4 Implications for CPaaS.io

For CPaaS.io the introduction of the European General Data Protection Regulation sets a framework to operate in from a European perspective and is a relevant factor for many design decisions regarding the handling of personal data within CPaaS.io. The requirement for data transparency for the data subjects means that all data across a smart city has to be classified according to its privacy. Data provided by sources considered to publish open information only, like open governmental data or open data provided by scientific institutions, are trusted to provide publishable material only and are not re-checked for their privacy status. All remaining data, such as data incoming from IoT sensors, video feeds, GPS trackers, etc. are in a first step considered private, before finding the evidence for it to be publishable information. While CPaaS.io does indeed handle personal data, it is not striving to be a fully-fledged all-rounder PDS. CPaaS.io will only record, manage and analyze personal data in the context of smart city related tasks. CPaaS.io will use open data responsibly and as a collector of personal data in the context of use cases will ensure compliance with data protection regulation and a relation to the smart city context. As with all PDS providers CPaaS.io will offer its users an overview of all saved personal data and the option to report wrong data for it to be corrected or correct it themselves, if legally possible. Unwarranted or simply unwanted data can be entirely removed by the users, if legally possible.

4 Conclusions & Next Steps

The purpose of this document was to describe the state of the art regarding data quality annotations in general and specifically for IoT data, data linking and semantic annotations and identification of relevant privacy mechanisms. The state of the art description lays the grounds for answering the question on how the semantic integration layer of CPaaS.io can or should be designed.

This semantic integration layer can be implemented in different phases and levels. Base functionality of the layer will expose existing data sets in a way that they are discoverable. For that it will expose metadata using best practices and vocabularies discussed in this document. For data sets with a relevance for CPaaS.io use cases it is important to expose as much data as possible as Linked Data. For that modules will be implemented that allow access to FIWARE or u2 data as RDF. Other relevant data sets can be integrated via the Semantic Web technology stack, examples are RDF datasets published in CKAN or specific SPARQL endpoints. RDF acts as the generic data model of CPaaS.io and the described vocabularies provide a base for developing the platform.

A SPARQL endpoint for the CPaaS.io platform will be implemented in D6.2. Based on the state of the art and aligned to the use cases, more detailed concepts and models for managing data in the CPaaS.io project will be developed in D6.3.

The non-functional and functional requirements identified in this deliverable will be represented in a structured manner in D3.2.

The design decisions related to the management of personal data will be addressed in more details in WP5.
5 References


Kontchakov, Roman; Rodríguez-Muro, Mariano; Zakharyaschev, Michael (2013): Ontology-Based Data Access with Databases: A Short Course. In: Rudolph, S. et al. (eds.) Reasoning Web. Semantic


