SPECIAL

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance

Deliverable D6.3

Plan for community group and standardisation contribution

Document version: V1.0

SPECIAL DELIVERABLE

Name, title and organisation of the scientific representative of the project's coordinator:

Mr Philippe Rohou t: +33 4 97 15 53 06 f: +33 4 92 38 78 22 e: philippe.rohou@ercim.eu

GEIE ERCIM, 2004, route des Lucioles, Sophia Antipolis, 06410 Biot, France

Project website address: http://www.specialprivacy.eu/

Project	3
Grant Agreement number	731601
Project acronym:	SPECIAL
Project title:	Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance
Funding Scheme:	Research & Innovation Action (RIA)
Date of latest version of DoW against which the assessment will be made:	17/10/2016
Document	
Period covered:	M01-M09
Deliverable number:	D6.3
Deliverable title	Plan for community group and standardisation contribution
Contractual Date of Delivery:	30-09-2017
Actual Date of Delivery:	30-09-2017
Editor (s):	Axel Polleres (WU), Sabrina Kirrane (WU)
Author (s):	Axel Polleres (WU), Sabrina Kirrane (WU), Rigo Wenning (ERCIM)
Reviewer (s):	Martin Kurze, Benedict Whittamsmith
Participant(s):	WU, ERCIM
Work package no.:	6
Work package title:	Collaboration, Dissemination & Standardisation
Work package leader:	ULD
Distribution:	PU
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages (including cover):	60

Disclaimer

This document contains description of the SPECIAL project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the SPECIAL consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

SPECIAL has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731601.

Contents

List of Figures

1 Introduction

We do not intend to re-invent the wheel: SPECIAL aims to adapt and extend – wherever possible and appropriate – techniques and (i) *technologies developed in previous projects* and (ii) *existing standards and standardisation initiatives* in order to provide technical means to support consent, transparency and compliance obligations set forth in the General Data Protection Regulation (GDPR). The core building blocks of the SPECIAL project include: Semantic Web standards and technologies, the Big Data engine developed by the Big Data Europe project and privacy research outputs and insights from the PRIME and the PrimeLife privacy projects. Additionally, we aim to promote and disseminate our new results – again where appropriate – in terms of proposals to new standards or extensions of existing standards. That is, in order to ensure the transfer of SPECIAL research and development outcomes to relevant international research and industrial communities, the SPECIAL consortium will actively contribute to standardisation activities throughout the course of the project. As outlined in *Deliverable D6.2 Public Relations Strategy* the World Wide Web Consortium (W3C) will form the main channel for our standardisation activities. The reason for concentrating mainly on the W3C was twofold:

- The W3C is an *international* community with currently 463 members that operates worldwide (including relevant industry players and also research stakeholders); it develops *open standards* to ensure the long-term growth of the Web, where the Web and systems using or connected through Web architectures and protocols are a main focus of our project. Moreover, the W3C operates under a transparent Code of Ethics and Professional Conduct `https://www.w3.org/Consortium/cepc/`.
- The SPECIAL consortium is very well connected to the W3C already: firstly, project partner ERCIM, is the European host of the W3C, and thus in itself is a Standards Developing Organisation (SDO); and secondly, the project partners Wirtschaftsuniversität Wien (WU), Thomson Reuters (TR) and T-Labs (through Deutsche Telekom AG) are full W3C members.

SPECIAL's core standardisation objectives include: (i) the identification of vocabularies that are necessary to represent policies and personal data processing and sharing events; (ii) the development of a Transparent Linked Data Processing Platform; and (iii) the strengthening of the existing Semantic Technology stack especially in terms of unified logic, trust, security and user interfaces.

The aim of this deliverable is to present relevant standardisation activities and initiatives within the remit of SPECIAL and to outline the standardisation strategy of the project going forward. As such, this deliverable builds upon the requirements analysis in terms vocabularies that can be used to represent consent requests, policies and transparency logs and potential transparency and compliance architectures discussed in *D1.3 Policy, transparency and compliance guidelines V1* and *D1.4 Technical requirements V1*. In this report we make the following contributions:

Section 2 provides an overview of existing Semantic Web standards and discuss potential gaps based on our requirements analysis;

Section 3 highlights several specific recent and ongoing standardisation initiatives and activities;

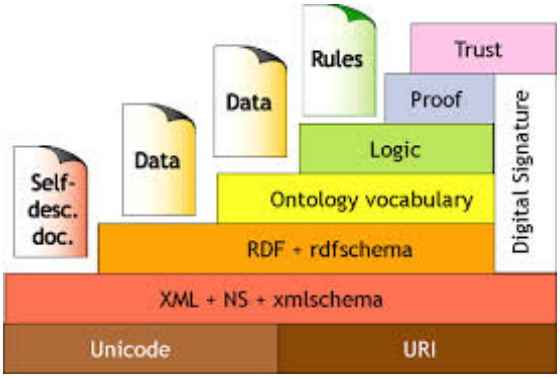

Figure 1: Semantic Web Technology Stack from Tim Berners-Lee

Section 4 discusses potential standardisation challenges and opportunities based on our initial gap analysis;

Section 5 finally presents a standardisation roadmap for the course of the SPECIAL project.

2 Survey of W3C Semantic Web Standards

The World Wide Web Consortium (W3C) is an international organisation whose members come from a variety of international organisations from both the public and the private sector. The mission of the W3C is to develop open standards to ensure the sustainability of the Web.

We start by introducing the Semantic Technology stack and by highlighting some of the layer that have been somewhat neglected to date.

Given that Semantic Web standards and technologies are one of the core building blocks of the SPECIAL project, in this section we categorise existing Semantic Web standards, and for ease of reading we recall the relevant abstracts and provide a link to the actual recommendation.

2.1 Semantic Technology Stack

In addition to the classic 'Web of documents' the W3C is helping to build a technology stack to support a 'Web of data'[1]. Over the years there has been many version of the Semantic Web technology stack such as the one presented in *Figure 1*, which was taken from a presentation at the turn of the millennium by Tim Berners-Lee[2].

When it comes to both standardisation and practical applications the Semantic Web technology stack is usually concentrated around data representation, data integration

[1] W3C Semantic Web, https://www.w3.org/standards/semanticweb/
[2] Semantic Web Technology Stack, https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

and providing semantics. While, the vertical and top layers of digital signatures (which was later extended to consider encryption), logic, proof and trust have received much less attention to date. However, the advancement of these very layers are crucial to the success of the SPECIAL project.

As such, the project presents us with a perfect use case to motivate the advancement of standardisation efforts in these underrepresented areas. More specifically, the project will be ideally placed to deliver a reference architecture that demonstrates capability, especially in terms of logic, proof, trust, and security of RDF data. However, in order to ensure generality, standardisation efforts should also be motivated by other use cases that require trust and transparency with respect to data processing and sharing.

2.2 W3C Semantic Web Standards

Semantic Web standards are at the core of the SPECIAL project and the Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance that will be developed throughout the course of the project. In this section, we highlight the existing W3C Semantic Web Standards that are referred to in the Semantic Web area of the W3C website `https://www.w3.org/standards/semanticweb/`. The standards are roughly categorised according to their function, with each section including a short note on the relevancy in terms of the SPECIAL project.

2.2.1 Machine Readable Data

The W3C standards presented below relate to the Resource Description Framework, including the data model, its semantics and the alternative serialisation mechanisms. SPECIAL will build upon Linked Data principles and the underpinning Resource Description Framework data model. Thus there are a number of recommendations that have been developed by the RDF Core Working Group that will be relevant as we progress through the SPECIAL project.

RDF Primer `https://www.w3.org/TR/2004/REC-rdf-primer-20040210/`: The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web. This Primer is designed to provide the reader with the basic knowledge required to effectively use RDF. It introduces the basic concepts of RDF and describes its XML syntax. It describes how to define RDF vocabularies using the RDF Vocabulary Description Language, and gives an overview of some deployed RDF applications. It also describes the content and purpose of other RDF specification documents.

Resource Description Framework (RDF): Concepts and Abstract Syntax `https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`: RDF Concepts and Abstract Syntax defines an abstract syntax on which RDF is based, and which serves to link its concrete syntax to its formal semantics. It also includes discussion of design goals, key concepts, datatyping, character normalisation and handling of URI references.

RDF Semantics `https://www.w3.org/TR/2004/REC-rdf-mt-20040210/`: This is a specification of a precise semantics, and corresponding complete systems of in-

ference rules, for the Resource Description Framework (RDF) and RDF Schema (RDFS).

RDF Test Cases https://www.w3.org/TR/2004/REC-rdf-testcases-20040210/: This document describes the RDF Test Cases deliverable for the RDF Core Working Group as defined in the Working Group's Charter.

RDF 1.1 Concepts and Abstract Syntax https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/: RDF 1.1 Concepts and Abstract Syntax defines an abstract syntax (a data model) which serves to link all RDF-based languages and specifications. The abstract syntax has two key data structures: RDF graphs are sets of subject-predicate-object triples, where the elements may be IRIs, blank nodes, or datatyped literals. They are used to express descriptions of resources. RDF datasets are used to organise collections of RDF graphs, and comprise a default graph and zero or more named graphs. This document also introduces key concepts and terminology, and discusses datatyping and the handling of fragment identifiers in IRIs within RDF graphs.

RDF 1.1 Semantics https://www.w3.org/TR/2014/REC-rdf11-mt-20140225/: This document describes a precise semantics for the Resource Description Framework 1.1 and RDF Schema. It defines a number of distinct entailment regimes and corresponding patterns of entailment. It is part of a suite of documents which comprise the full specification of RDF 1.1.

RDF 1.1 Turtle https://www.w3.org/TR/2014/REC-turtle-20140225/: The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web.

RDF 1.1 XML Syntax https://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/: This document defines an XML syntax for RDF called RDF/XML in terms of Namespaces in XML, the XML Information Set and XML Base. The formal grammar for the syntax is annotated with actions generating triples of the RDF graph as defined in RDF Concepts and Abstract Syntax. The triples are written using the N-Triples RDF graph serializing format which enables more precise recording of the mapping in a machine processable form. The mappings are recorded as tests cases, gathered and published in RDF Test Cases.

JSON-LD 1.0 https://www.w3.org/TR/2014/REC-json-ld-20140116/: A common JSON representation format for expressing directed graphs; mixing both Linked Data and non-Linked Data in a single JSON document. RDF 1.1 Concepts and Abstract Syntax The Resource Description Framework (RDF) is a framework for representing information in the Web.

JSON-LD 1.0 Processing Algorithms and API https://www.w3.org/TR/2014/REC-json-ld-api-20140116/: An Application Programming Interface and a set of algorithms for programmatically transforming JSON-LD documents in order to make them easier to work with in programming environments like JavaScript, Python, and Ruby.

RDF 1.1 N-Quads https://www.w3.org/TR/2014/REC-n-quads-20140225/: N-Quads is a line-based, plain text format for encoding an RDF dataset.

RDF 1.1 N-Triples https://www.w3.org/TR/2014/REC-n-triples-20140225/: N-Triples is a line-based, plain text format for encoding an RDF graph.

RDF 1.1 TriG https://www.w3.org/TR/2014/REC-trig-20140225/: This document defines a textual syntax for RDF called TriG that allows an RDF dataset to be completely written in a compact and natural text form, with abbreviations for common usage patterns and datatypes. TriG is an extension of the Turtle format.

rdf:PlainLiteral: A Datatype for RDF Plain Literals (Second Edition) https://www.w3.org/TR/2012/REC-rdf-plain-literal-20121211/: This document presents the specification of a primitive datatype for the plain literals of RDF.

2.2.2 Modeling Languages

In this section, we summarise the various recommendations that relate to W3C standard modeling languages RDF Schema and the Web Ontology Language (OWL), which are used to describe RDF data. In the context of SPECIAL the below standards are relevant in terms of modeling and reasoning over policies in the form of usage constraints, legal obligations and business rules, and also in the context of capturing and recording data processing and sharing events at different levels of granularity.

RDF Schema 1.1 https://www.w3.org/TR/2014/REC-rdf-schema-20140225/: The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. This specification describes how to use RDF to describe RDF vocabularies. This specification defines a vocabulary for this purpose and defines other built-in RDF vocabulary initially specified in the RDF Model and Syntax Specification.

OWL Web Ontology Language Overview https://www.w3.org/TR/2004/REC-owl-features-20040210/: The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full.

OWL Web Ontology Language Guide https://www.w3.org/TR/2004/REC-owl-guide-20040210/: The World Wide Web as it is currently constituted resembles a poorly mapped geography. Our insight into the documents and capabilities available are based on keyword searches, abetted by clever use of document connectivity and usage patterns. The sheer mass of this data is unmanageable without powerful tool support. In order to map this terrain more precisely, computational agents require machine-readable descriptions of the content and capabilities of Web accessible resources. These descriptions must be in addition to the human-readable versions of that information.

OWL Web Ontology Language Reference https://www.w3.org/TR/2004/REC-owl-ref-20040210/: The Web Ontology Language OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL is developed as a vocabulary extension of RDF (the Resource Description Framework) and is derived from the DAML+OIL Web Ontology Language. This document contains a structured informal description of the full set of OWL language constructs and is meant to serve as a reference for OWL users who want to construct OWL ontologies.

OWL Web Ontology Language Semantics and Abstract Syntax https://www.w3.org/TR/2004/REC-owl-semantics-20040210/.

OWL Web Ontology Language Test Case https://www.w3.org/TR/2004/REC-owl-test-20040210/: This document contains and presents test cases for the Web Ontology Language (OWL) approved by the Web Ontology Working Group. Many of the test cases illustrate the correct usage of the Web Ontology Language (OWL), and the formal meaning of its constructs. Other test cases illustrate the resolution of issues considered by the Working Group. Conformance for OWL documents and OWL document checkers is specified.

OWL Web Ontology Language Use Cases and Requirements https://www.w3.org/TR/2004/REC-webont-req-20040210/: This document specifies usage scenarios, goals and requirements for a web ontology language. An ontology formally defines a common set of terms that are used to describe and represent a domain. Ontologies can be used by automated tools to power advanced services such as more accurate web search, intelligent software agents and knowledge management.

OWL 2 Web Ontology Language Primer (Second Edition) https://www.w3.org/TR/2012/REC-owl2-primer-20121211/: This primer provides an approachable introduction to OWL 2, including orientation for those coming from other disciplines, a running example showing how OWL 2 can be used to represent first simple information and then more complex information, how OWL 2 manages ontologies, and finally the distinctions between the various sublanguages of OWL 2.

OWL 2 Web Ontology Language Quick Reference Guide (Second Edition) https://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/: This document provides a non-normative quick reference guide to the OWL 2 language. It also provides links to other documents, including the OWL 2 Primer for language introduction and examples, the OWL 2 Structural Specification and Functional Syntax document for more details of the functional syntax, and the OWL 2 New Features and Rationale document for new feature descriptions.

OWL 2 Web Ontology Language Document Overview (Second Edition) https://www.w3.org/TR/2012/REC-owl2-overview-20121211/: This document serves as an introduction to OWL 2 and the various other OWL 2 documents. It describes the syntaxes for OWL 2, the different kinds of semantics,

the available profiles (sub-languages), and the relationship between OWL 1 and OWL 2.

OWL 2 Web Ontology Language New Features and Rationale (Second Edition) https://www.w3.org/TR/2012/REC-owl2-new-features-20121211/: This document is a simple introduction to the new features of the OWL 2 Web Ontology Language, including an explanation of the differences between the initial version of OWL and OWL 2. The document also presents the requirements that have motivated the design of the main new features, and their rationale from a theoretical and implementation perspective.

OWL 2 Web Ontology Language Conformance (Second Edition) https://www.w3.org/TR/2012/REC-owl2-conformance-20121211/: This document describes the conditions that OWL 2 tools must satisfy in order to be conformant with the language specification. It also presents a common format for OWL 2 test cases that both illustrate the features of the language and can be used for testing conformance.

OWL 2 Web Ontology Language Mapping to RDF Graphs (Second Edition) https://www.w3.org/TR/2012/REC-owl2-mapping-to-rdf-20121211/: This document defines the mapping of OWL 2 ontologies into RDF graphs, and vice versa.

OWL 2 Web Ontology Language Profiles (Second Edition) https://www.w3.org/TR/2012/REC-owl2-profiles-20121211/: This document provides a specification of several profiles of OWL 2 which can be more simply and/or efficiently implemented. In logic, profiles are often called fragments. Most profiles are defined by placing restrictions on the structure of OWL 2 ontologies. These restrictions have been specified by modifying the productions of the functional-style syntax.

OWL 2 Web Ontology Language Direct Semantics (Second Edition) https://www.w3.org/TR/2012/REC-owl2-direct-semantics-20121211/: This document provides the direct model-theoretic semantics for OWL 2, which is compatible with the description logic SROIQ. Furthermore, this document defines the most common inference problems for OWL 2.

OWL 2 Web Ontology Language RDF-Based Semantics (Second Edition) https://www.w3.org/TR/2012/REC-owl2-rdf-based-semantics-20121211/: This document defines the RDF-compatible model-theoretic semantics of OWL 2.

OWL 2 Web Ontology Language XML Serialisation (Second Edition) https://www.w3.org/TR/2012/REC-owl2-xml-serialisation-20121211/: This document specifies an XML serialisation for OWL 2 that mirrors its structural specification. An XML schema defines this syntax and is available as a separate document, as well as being included here.

OWL 2 Structural Specification and Functional-Style Syntax (Second Edition) https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/: The OWL 2 Web

Ontology Language, informally OWL 2, is an ontology language for the Semantic Web with formally defined meaning.

2.2.3 Meta Data & Vocabularies

The W3C recommendations presented in this section include data models and vocabularies that can be used to describe meta data that is needed to classify and categorise RDF data. *SKOS Simple Knowledge Organisation System Reference* and *the Organisation Ontology* could be used to describe organisational systems and organisational structures respectively. While the *PROV-O: The PROV Ontology* and related recommendations could be used to represent provenance information relating to data processing and sharing events, usage constraints, legislative obligations and information pertaining to the transparency ledger(s). *The Protocol for Web Description Resources (POWDER)* is particularly interesting as it can be used to describe data in a manner that facilitates discovery and trust. *Data Catalog Vocabulary (DCAT)* is another recommendation that could potentially be interesting if there is a need to maintain a catalog of datasets. While the *The RDF Data Cube Vocabulary* could be used capture information pertaining to aggregations or simply statistical information in relation to compliance checking. Finally, the *Internationalisation Tag Set (ITS)* may be relevant from a localisation and an internationalisation perspective.

SKOS Simple Knowledge Organisation System Reference `https://www.w3.org/TR/2009/REC-skos-reference-20090818/`: This document defines the Simple Knowledge Organisation System (SKOS), a common data model for sharing and linking knowledge organisation systems via the Web. Many knowledge organisation systems, such as thesauri, taxonomies, classification schemes and subject heading systems, share a similar structure, and are used in similar applications. SKOS captures much of this similarity and makes it explicit, to enable data and technology sharing across diverse applications.The SKOS data model provides a standard, low-cost migration path for porting existing knowledge organisation systems to the Semantic Web. SKOS also provides a lightweight, intuitive language for developing and sharing new knowledge organisation systems. It may be used on its own, or in combination with formal knowledge representation languages such as the Web Ontology language (OWL).

The Organisation Ontology `https://www.w3.org/TR/2014/REC-vocab-org-20140116/`: This document describes a core ontology for organisational structures, aimed at supporting linked-data publishing of organisational information across a number of domains. It is designed to allow domain-specific extensions to add classification of organzations and roles, as well as extensions to support neighbouring information such as organisational activities.

PROV-O: The PROV Ontology `https://www.w3.org/TR/2013/REC-prov-o-20130430/`: This specification defines the PROV Ontology as the normative representation of the PROV Data Model using the Web Ontology Language (OWL2). This document is part of a set of specifications being created to address the issue of provenance interchange in Web applications.

PROV-DM: The PROV Data Model `https://www.w3.org/TR/2013/REC-prov-dm-20130430/`: PROV-DM is a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or thing in the world. PROV-DM is domain-agnostic, but with well-defined extensibility points allowing further domain-specific and application-specific extensions to be defined. It is accompanied by PROV-ASN, a technology-independent abstract syntax notation, which allows serialisations of PROV-DM instances to be created for human consumption, which facilitates its mapping to concrete syntax, and which is used as the basis for a formal semantics.

Constraints of the PROV Data Model `https://www.w3.org/TR/2013/REC-prov-constraints-20130430/`: PROV-DM, the PROV data model, is a data model for provenance that describes the entities, people and activities involved in producing a piece of data or thing. PROV-DM is structured in six components, dealing with: (1) entities and activities, and the time at which they were created, used, or ended; (2) agents bearing responsibility for entities that were generated and activities that happened; (3) derivations of entities from entities; (4) properties to link entities that refer to a same thing; (5) collections forming a logical structure for its members; (6) a simple annotation mechanism.

PROV-N: The Provenance Notation `https://www.w3.org/TR/2013/REC-prov-n-20130430/`: To provide examples of the PROV data model, the PROV notation (PROV-N) is introduced: aimed at human consumption, PROV-N allows serialisations of PROV instances to be created in a compact manner. PROV-N facilitates the mapping of the PROV data model to concrete syntax, and is used as the basis for a formal semantics of PROV. The purpose of this document is to define the PROV-N notation.

Protocol for Web Description Resources (POWDER): Description Resources `https://www.w3.org/TR/2009/REC-powder-dr-20090901/`: The purpose of the Protocol for Web Description Resources (POWDER) is to provide a means for individuals or organisations to describe a group of resources through the publication of machine-readable metadata, as motivated by the POWDER Use Cases [USECASES]. This document details the creation and lifecycle of Description Resources (DRs), which encapsulate such metadata. These are typically represented in a highly constrained XML dialect that is relatively human-readable. The meaning of such DRs are underpinned by formal semantics, accessible by performing a GRDDL Transform.

Protocol for Web Description Resources (POWDER): Grouping of Resources `https://www.w3.org/TR/2009/REC-powder-grouping-20090901/`: The Protocol for Web Description Resources (POWDER) facilitates the publication of descriptions of multiple resources such as all those available from a Web site. This document describes how sets of IRIs can be defined such that descriptions or other data can be applied to the resources obtained by dereferencing IRIs that are elements of the set. IRI sets are defined as XML elements with relatively loose operational semantics. This is underpinned by the formal semantics of POWDER which include a semantic extension, defined separately. A GRDDL transform

is associated with the POWDER namespace that maps the operational to the formal semantics.

Protocol for Web Description Resources (POWDER): Formal Semantics `https://www.w3.org/TR/2009/REC-powder-formal-20090901/`: This document underpins the Protocol for Web Description Resources (POWDER). It describes how the relatively simple operational format of a POWDER document can be transformed through two stages: first into a more tightly constrained XML format (POWDER-BASE), and then into an RDF/OWL encoding (POWDER-S) that may be processed by Semantic Web tools. Such processing is only possible, however, if tools implement the semantic extension defined within this document. The formal semantics of POWDER are best understood after the reader is acquainted with the Description Resources [DR] and Grouping of Resources [GROUP] documents.

Data Catalog Vocabulary (DCAT) `https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/`: DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.

The RDF Data Cube Vocabulary `https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/`: There are many situations where it would be useful to be able to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts. The Data Cube vocabulary provides a means to do this using the W3C RDF (Resource Description Framework) standard.

Internationalization Tag Set (ITS) Version 2.0 `https://www.w3.org/TR/2013/REC-its20-20131029/`: This document defines data categories and their implementation as a set of elements and attributes called the Internationalization Tag Set (ITS) 2.0. ITS 2.0 is the successor of ITS 1.0; it is designed to foster the creation of multilingual Web content, focusing on HTML5, XML based formats in general, and to leverage localisation workflows based on the XML Localisation Interchange File Format (XLIFF). In addition to HTML5 and XML, algorithms to convert ITS attributes to RDFa and NIF are provided.

CSV on the Web: A Primer `http://w3c.github.io/csvw/primer/` CSV is one of the most popular formats for publishing data on the web. It is concise, easy to understand by both humans and computers, and aligns nicely to the tabular nature of most data. But CSV is also a poor format for data. There is no mechanism within CSV to indicate the type of data in a particular column, or whether values in a particular column must be unique. It is therefore hard to validate and prone to errors such as missing values or differing data types within a column. The CSV on the Web Working Group has developed standard ways to express useful metadata about CSV files and other kinds of tabular data. There are a variety of Recommendations developed. This is why the link is to the primer that lists and explains the various standards and how they work together.

2.2.4 Query Languages

This section is dedicated to recommendations that describe SPARQL, the standard query language for RDF, including the query language itself, query protocols, entailment regimes and the various output serialisations. As SPARQL is the standard query language for RDF, in the context of SPECIAL is will be used to retrieve (and possibly update) RDF data. *SPARQL 1.1 Entailment Regimes* could potentially be used to consider implicit (i.e. inferred) data during query execution based on RDF entailment regimes. The *SPARQL 1.1 Graph Store HTTP Protocol* could potentially be used to manage a collection of RDF graphs via HTTP. While, the *SPARQL 1.1 Service Description* could be used if there is a need to describe the services offered by a particular endpoint. Finally, *SPARQL 1.1 Federated Query* would be relevant if there is a need to execute queries in a distributed manner across several SPARQL endpoints.

SPARQL Query Language for RDF `https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/`: RDF is a directed, labeled graph data format for representing information in the Web. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs.

SPARQL Query Results XML Format (Second Edition) `https://www.w3.org/TR/2013/REC-rdf-sparql-XMLres-20130321/`. RDF is a flexible, extensible way to represent information about World Wide Web resources. It is used to represent, among other things, personal information, social networks, metadata about digital artifacts like music and images, as well as provide a means of integration over disparate sources of information. A standardised query language for RDF data with multiple implementations offers developers and end users a way to write and to consume the results of queries across this wide range of information.

SPARQL Protocol for RDF `https://www.w3.org/TR/2008/REC-rdf-sparql-protocol-20080115/`: The SPARQL Protocol and RDF Query Language (SPARQL) is a query language and protocol for RDF. This document specifies the SPARQL Protocol; it uses WSDL 2.0 to describe a means for conveying SPARQL queries to an SPARQL query processing service and returning the query results to the entity that requested them. This protocol was developed by the W3C RDF Data Access Working Group (DAWG), part of the Semantic Web Activity as described in the activity statement .

SPARQL 1.1 Overview `https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/`: This document is an overview of SPARQL 1.1. It provides an introduction to a set of W3C specifications that facilitate querying and manipulating RDF graph content on the Web or in an RDF store.

SPARQL 1.1 Query Language `https://www.w3.org/TR/2013/REC-sparql11-query-20130321/`: RDF is a directed, labeled graph data format for representing information in the Web. The SPARQL specification defines the syntax and semantics of the SPARQL query language for RDF. This document describes changes that will be made to the SPARQL query language to form SPARQL 1.1 Query.

SPARQL 1.1 Query Results CSV and TSV Formats `https://www.w3.org/TR/2013/REC-sparql11-results-csv-tsv-20130321/`: The formats CSV [RFC4180] (comma separated values) and TSV [IANA-TSV] (tab separated values) provide simple, easy to process formats for the transmission of tabular data. They are supported as input datat formats to many tools, particularly spreadsheets. This document describes their use for expressing SPARQL query results from SELECT queries.

SPARQL 1.1 Query Results JSON Format `https://www.w3.org/TR/2013/REC-sparql11-results-json-20130321/`: This document describes the representation of SELECT and ASK query results using JSON.

SPARQL 1.1 Protocol `https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/`: The SPARQL Protocol and RDF Query Language (SPARQL) is a query language and protocol for RDF. This document specifies the SPARQL Protocol; it uses WSDL 2.0 to describe a means for conveying SPARQL queries to a SPARQL query processing service and returning the query results to the entity that requested them.

SPARQL 1.1 Update `https://www.w3.org/TR/2013/REC-sparql11-update-20130321/`: This document describes SPARQL-Update, an update language for RDF graphs. It uses a syntax derived from SPARQL. Update operations are performed on a collection of graphs in a Graph Store. Operations are provided to change existing RDF graphs as well as create and remove graphs in the Graph Store.

SPARQL 1.1 Graph Store HTTP Protocol `https://www.w3.org/TR/2013/REC-sparql11-http-rdf-update-20130321/`: This document describes the use of HTTP operations for the purpose of managing a collection of RDF graphs. This interface is an alternative to the SPARQL 1.1 Update protocol. Most of the operations defined here can be performed using that interface, but for some clients or servers, this interface may be easier to implement or work with. This specification may serve as a non-normative suggestion for HTTP operations on RDF graphs which are managed outside of a SPARQL 1.1 graph store.

SPARQL 1.1 Entailment Regimes `https://www.w3.org/TR/2013/REC-sparql11-entailment-20130321/`: SPARQL is a query language and a protocol for data that is stored natively as RDF or viewed as RDF via middleware. The main mechanism for computing query results in SPARQL is subgraph matching: RDF triples in both the queried RDF data and the query pattern are interpreted as nodes and edges of directed graphs, and the resulting query graph is matched to the data graph using variables as wild cards. Various W3C standards,

including RDF and OWL, provide semantic interpretations for RDF graphs that allow additional RDF statements to be inferred from explicitly given assertions. Many applications that rely on these semantics require a query language such as SPARQL, but in order to use SPARQL, basic graph pattern matching has to be defined using semantic entailment relations instead of explicitly given graph structures. There are different possible ways of defining a basic graph pattern matching extension for an entailment relation. This document specifies one such way for a range of standard semantic web entailment relations. Such extensions of the SPARQL semantics are called entailment regimes within this document. An entailment regime defines not only which entailment relation is used, but also which queries and graphs are well-formed for the regime, how the entailment is used (since there are potentially different meaningful ways to use the same entailment relation), and what kinds of errors can arise. The entailment relations used in this document are standard entailment relations in the semantic web: RDF entailment, RDFS entailment, D-entailment, OWL Direct and RDF-Based Semantics entailment, and RIF Core entailment.

SPARQL 1.1 Service Description `https://www.w3.org/TR/2013/REC-sparql11-service-description-20130321/`: This document describes SPARQL Service Descriptions, a method for discovering and vocabulary for describing SPARQL services made available via the SPARQL Protocol. Such descriptions are intended to provide a mechanism by which a client or end user can discover information about the SPARQL implementation/service such as supported extension functions and details about the available dataset.

SPARQL 1.1 Federated Query `https://www.w3.org/TR/2013/REC-sparql11-federated-query-20130321/`: RDF is a directed, labeled graph data format for representing information in the Web. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. This specification defines the syntax and semantics of SPARQL 1.1 Federated Query extension for executing queries distributed over different SPARQL endpoints. The SERVICE keyword extends SPARQL 1.1 to support queries that merge data distributed across the Web.

2.2.5 Rule Languages & Constraints

In this section, we provide an overview of the recommendations that relate to the Rule Interchange Format (RIF) and the Shapes Constraint Language (SHACL). RIF specifies a coherent way to build more-expressive RIF dialects, using a single semantic framework. RIF could potentially be used to translate existing business rules into something that can be digested by the SPECIAL engine. While, SHACL could be used to express constraints in usage policies and legislative obligations.

RIF Framework for Logic Dialects (Second Edition) `https://www.w3.org/TR/2013/REC-rif-fld-20130205/`: This document, developed by the Rule Interchange Format (RIF) Working Group, defines a general RIF Framework for Logic Dialects (RIF-FLD). The framework describes mechanisms for specifying the syntax and semantics of logic RIF dialects through a number of generic

concepts such as signatures, symbol spaces, semantic structures, and so on. The actual dialects should specialise this framework to produce their syntaxes and semantics.

RIF Core Dialect (Second Edition) `https://www.w3.org/TR/2013/REC-rif-core-20130205/`: This document, developed by the Rule Interchange Format (RIF) Working Group, specifies RIF-Core, a common subset of RIF-BLD and RIF-PRD based on RIF-DTB 1.0. The RIF-Core presentation syntax and semantics are specified by restriction in two different ways. First, RIF-Core is specified by restricting the syntax and semantics of RIF-BLD, and second, by restricting RIF-PRD. The XML serialisation syntax of RIF-Core is specified by a mapping from the presentation syntax. A normative XML schema is also provided.

RIF Basic Logic Dialect (Second Edition) `https://www.w3.org/TR/2013/REC-rif-bld-20130205/`: This document, developed by the Rule Interchange Format (RIF) Working Group, specifies the Basic Logic Dialect, RIF-BLD, a format that allows logic rules to be exchanged between rule systems. The RIF-BLD presentation syntax and semantics are specified both directly and as specialisations of the RIF Framework for Logic Dialects, or RIF-FLD. The XML serialisation syntax of RIF-BLD is specified via a mapping from the presentation syntax. A normative XML schema is also provided.

RIF Production Rule Dialect (Second Edition) `https://www.w3.org/TR/2013/REC-rif-prd-20130205/`: This document, developed by the Rule Interchange Format (RIF) Working Group, specifies the production rule dialect of the W3C rule interchange format (RIF-PRD), a standard XML serialisation format for production rule languages.

RIF Datatypes and Built-Ins 1.0 (Second Edition) `https://www.w3.org/TR/2013/REC-rif-dtb-20130205/`: This document, developed by the Rule Interchange Format (RIF) Working Group, specifies a list of datatypes, built-in functions and built-in predicates expected to be supported by RIF dialects such as the RIF Core Dialect, the RIF Basic Logic Dialect, and the RIF Production Rules Dialect. Each dialect supporting a superset or subset of the datatypes, built-in functions and built-in predicates defined here shall specify these additions or restrictions. Some of the datatypes are adapted from [XML Schema Datatypes]. A large part of the definitions of the listed functions and operators are adapted from [XPath-Functions]. The rdf:PlainLiteral datatype as well as functions and operators associated with that datatype are adopted from [RDF-PLAINLITERAL].

RIF RDF and OWL Compatibility (Second Edition) `https://www.w3.org/TR/2013/REC-rif-rdf-owl-20130205/`: Rules interchanged using the Rule Interchange Format RIF may depend on or be used in combination with RDF data and RDF Schema or OWL ontologies. This document, developed by the Rule Interchange Format (RIF) Working Group, specifies the interoperation between RIF and the data and ontology languages RDF, RDF Schema, and OWL.

Shapes Constraint Language (SHACL) `https://www.w3.org/TR/2016/WD-shacl-20160530/`: SHACL (Shapes Constraint Language) is a language for describing and constraining the contents of RDF graphs. SHACL groups these descriptions and constraints into "shapes", which specify conditions that apply at a given RDF node. Shapes provide a high-level vocabulary to identify predicates and their associated cardinalities, datatypes and other constraints. Additional constraints can be associated with shapes using SPARQL. The vocabulary of SHACL has been designed to support similar extension languages besides SPARQL. These extension languages can also be used to define new high-level vocabulary terms. SHACL shapes can be used to communicate information about data structures associated with some process or interface, generate or validate data, or drive user interfaces. This document defines the SHACL language and its underlying semantics.

SHACL Use Cases and Requirements `https://www.w3.org/TR/2016/WD-shacl-ucr-20160122/`: To foster the development of Shapes Constraint Language (SHACL), this document includes a set of use cases and requirements that motivate a simple language and semantics for formulating structural constraints on RDF graphs. All use cases provide realistic examples describing how people may use structural constraints to validate RDF instance data. Note, that this document avoids the use of any specific vocabulary that might be introduced by the SHACL specification.

2.2.6 Transformation Languages

When it comes to translation from other data models to RDF, there are two languages that can be used to translate relational data to RDF (RDB2RDF) namely: *A Direct Mapping of Relational Data to RDF* and *R2RML: RDB to RDF Mapping Language*, and a language for extract RDF data from XML known as *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*. Considering that the relational data model is the predominant data model underpinning existing Line of Business (LOB) applications we aim to leverage existing RDB2RDF recommendations inorder to create a bridge between existing systems and the SPECIAL platform.

A Direct Mapping of Relational Data to RDF `https://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/`: The need to share data with collaborators motivates custodians and users of relational databases (RDB) to expose relational data on the Web of Data. This document defines a direct mapping from relational data to RDF. This definition provides extension points for refinements within and outside of this document.

R2RML: RDB to RDF Mapping Language `https://www.w3.org/TR/2012/REC-r2rml-20120927/`: This document describes R2RML, a language for expressing customised mappings from relational databases to RDF datasets. Such mappings provide the ability to view existing relational data in the RDF data model, expressed in a structure and target vocabulary of the mapping author's choice. R2RML mappings are themselves RDF graphs and written down in Turtle syntax. R2RML enables different types of mapping implementations. Processors

could, for example, offer a virtual SPARQL endpoint over the mapped relational data, or generate RDF dumps, or offer a Linked Data interface.

Gleaning Resource Descriptions from Dialects of Languages (GRDDL) `https://www.w3.org/TR/2007/REC-grddl-20070911/`: GRDDL is a mechanism for Gleaning Resource Descriptions from Dialects of Languages. This GRDDL specification introduces markup based on existing standards for declaring that an XML document includes data compatible with the Resource Description Framework (RDF) and for linking to algorithms, for extracting this data from the document. As languages to actually define these algorithms GRDDL typically refers to XSLT, but also XQUERY and other languages (e.g. XSPARQL) are possible. Therefore, below we also include these below in the relevant standards as well, although they are – strictly speaking – not part of the standards from the Semantic Web activity. The markup includes a namespace-qualified attribute for use in general-purpose XML documents and a profile-qualified link relationship for use in valid XHTML documents. The GRDDL mechanism also allows an XML namespace document (or XHTML profile document) to declare that every document associated with that namespace (or profile) includes gleanable data and for linking to an algorithm for gleaning the data.

GRDDL Test Cases `https://www.w3.org/TR/2007/REC-grddl-tests-20070911/`: This document describes and includes test cases for software agents that extract RDF from XML source documents by following the set of mechanisms outlined in the Gleaning Resource Description from Dialects of Language [GRDDL] specification. They demonstrate the expected behavior of a GRDDL-aware agent by specifying one (or more) RDF graph serialisations which are the GRDDL results associated with a single source document.

XSLT `https://www.w3.org/TR/1999/REC-xslt-19991116`: This specification defines the syntax and semantics of XSLT, which is a (template-based) language for transforming XML documents into other XML documents.

XQuery `https://www.w3.org/TR/2014/REC-xquery-30-20140408/`: This specification describes a query (and transformation) language called XQuery, which is designed to be broadly applicable across many types of XML data sources. The current version XQuery 3.0 is an extended version of the XQuery 1.0 Recommendation. XQuery as opposed to the template-based XSLT specification is a functional language.

XSPARQL `https://www.w3.org/Submission/2009/01/`: As the only non W3C Recommendation in this section, we mention XSPARQL as an alternative transformation languaga, that merges XQuery and SPARQL for easier, more declaratice and more concise descriptions of RDF transformations. This submission was driven by SPECIAL team members (submitted by Axel Polleres' team who is now at WU). The project and engine are still actively maintained at sourceforge under `https://sourceforge.net/projects/xsparql/`.

2.2.7 Data On the Web

Finally there are a number of specifications that specifically focus on embedding RDF in HTML documents, extending web service definitions with additional semantics and providing guidelines for publishing data on the web. In the context of SPECIAL such recommendations could be used in the context of the User Interface. Additionally *the Semantic Annotations for WSDL and XML Schema* could be used to specific the semantics of the SPECIAL web services. While, the *Linked Data Platform 1.0* could be used to access collections of RDF data (and Metadata) via HTTP.

HTML+RDFa 1.1 - Second Edition `https://www.w3.org/TR/2015/REC-html-rdfa-20150317/`: This specification defines rules and guidelines for adapting the RDFa Core 1.1 and RDFa Lite 1.1 specifications for use in HTML5 and XHTML5. The rules defined in this specification not only apply to HTML5 documents in non-XML and XML mode, but also to HTML4 and XHTML documents interpreted through the HTML5 parsing rules.

RDFa Core 1.1 - Third Edition `https://www.w3.org/TR/2015/REC-rdfa-core-20150317/`: RDFa Core is a specification for attributes to express structured data in any markup language. The embedded data already available in the markup language (e.g., XHTML) is reused by the RDFa markup, so that publishers don't need to repeat significant data in the document content.

RDFa Lite 1.1 - Second Edition `https://www.w3.org/TR/2015/REC-rdfa-lite-20150317/`: RDFa Lite is a small subset of RDFa consisting of a few attributes that may be applied to most simple to moderate structured data markup tasks. While it is not a complete solution for advanced markup tasks, it does provide a good entry point for beginners.

XHTML+RDFa 1.1 - Third Edition `https://www.w3.org/TR/2015/REC-xhtml-rdfa-20150317/`: RDFa Core 1.1 defines attributes and syntax for embedding semantic markup in Host Languages. This document defines one such Host Language. This language is a superset of XHTML 1.1, integrating the attributes as defined in RDFa Core 1.1.

Semantic Annotations for WSDL and XML Schema `https://www.w3.org/TR/2007/REC-sawsdl-20070828/`: This document defines a set of extension attributes for the Web Services Description Language and XML Schema definition language that allows description of additional semantics of WSDL components. The specification defines how semantic annotation is accomplished using references to semantic models, e.g. ontologies. Semantic Annotations for WSDL and XML Schema (SAWSDL) does not specify a language for representing the semantic models. Instead it provides mechanisms by which concepts from the semantic models, typically defined outside the WSDL document, can be referenced from within WSDL and XML Schema components using annotations.

Linked Data Platform 1.0 `https://www.w3.org/TR/2015/REC-ldp-20150226/`: A set of best practices and simple approach for a read-write Linked Data architecture, based on HTTP access to web resources that describe their state using RDF.

2.3 Other Relevant W3C notes or standard extensions

Apart from these standard, the following notes or stadard extension proposals might be relevant:

DCAT-AP `https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11` The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on W3C's Data Catalogue vocabulary (DCAT) for describing public sector datasets in Europe. Its basic use case is to enable a cross-data portal search for data sets and make public sector data better searchable across borders and sectors. This can be achieved by the exchange of descriptions of data sets among data portals.

Asset Description Metadata Schema (ADMS) `http://www.w3.org/TR/vocab-adms/` 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.

Data Quality Vocabulary `https://www.w3.org/TR/vocab-dqv/` The Data Quality Vocabulary provides a framework in which the quality of a dataset can be described, whether by the dataset publisher or by a broader community of users. It does not provide a formal, complete definition of quality, rather, it sets out a consistent means by which information can be provided such that a potential user of a dataset can make his/her own judgment about its fitness for purpose.

The Linked Open Data Repository `http://lov.okfn.org/dataset/lov/` A vocabulary in LOV gathers definitions of a set of classes and properties (together simply called terms of the vocabulary), useful to describe specific types of things, or things in a given domain or industry, or things at large but for a specific usage. This allows to find specific vocabularies that are needed for certain use cases or that can be adapted for those use cases.

Smart Descriptions & Smarter Vocabularies (SDSVoc) `https://www.w3.org/2016/11/sdsvoc/` The Data Catalog Vocabulary, DCAT, became a W3C Recommendation in January 2014. Making use of Dublin Core wherever possible, DCAT captures many essential features of a description of a dataset: the abstract concepts of the catalog and datasets, the realizable distributions of the datasets, keywords, landing pages, links to licenses, publishers etc. But it's clear that DCAT is not a full solution. For example, it doesn't cover versioning or time and space slices; it does not relate semantically the dataset to organisations, persons, software, projects, funding; it describes datasets, not APIs and so on. Other well-established and widely used schemas for describing data include CKAN's native schema, schema.org, DDI, SDMX, CERIF, VoID, INSPIRE and the Healthcare and Life Sciences Interest Group's Dataset Description vocabulary. These provide for discovery of datasets and - in some cases - contextualization (to ascertain relevance and quality) and action (access). Of the above only CERIF provides for provenance, although the W3C Recommendation PROV is also clearly relevant here. To emphasize the variety, the UK's Digital Curation Centre - jointly with

the RDA's Metadata Standards Catalog group - manages an extensive catalog of metadata standards used in different scientific disciplines. W3C is considering further activities in this area and had a Workshop in November 2016, which WU participated in with a paper [10].

3 Other Standards and Standardisation Initiatives

In addition to existing standards there are are number of standardisaion initiatives and activities that could potentially be relevant in the course of the SPECIAL project. Within the SPECIAL consortium ERCIM, WU, TR, and T-Labs (through Deutsche Telekom) are active members of the W3C as a worldwide standardisation body around Web technologies. Through these members, we will actively liaise with active and upcoming activities within W3C, and participate in both community and working groups within W3C. For instance, TR is co-chairing the POE Working Group within W3C, a WU team member has been co-chairing the SPARQL Working Group. Additionally, we aim to monitor and engage with other standardisation organisations, groups and initiatives that are also relevant in terms of privacy, consent, policies, transparency and compliance.

Key considerations when it comes to standardisation include the relevancy of the standard for the project in general (but especially for our industrial partners), key drivers behind the standard (e.g. regulatory requirements), and uptake of the standard from an enterprise perspective. Thus far we have focused on the identification of potentially relevant standards and standardisation initiatives, as the project progresses we plan to further analyse their suitability based on key considerations such as those mentioned previously.

In the following, we will go through the relevant initiatives and recent W3C groups and describe connections or involvement where already established. Following on from this we will provide a summary of potentially relevant standards and standardisation initiatives in the wider standardisation community.

3.1 Relevant W3C Initiatives

Several related W3C groups that SPECIAL aims to liaise with were listed in [12]. Herein we amend this list and provide additional details.

W3C current or recently closed working groups:

POE `https://www.w3.org/2016/poe/wiki/Main_Page` the Permissions & Obligations working group provides as its main contribution a core information model [7] and vocabulary [6] for expressing permissions, obligations in the form of policies. Neither of which are tailored specifically for personal data handling. **SPECIAL involvement/liaison:** both WU and TR ace active members of the group, TR co-chairs the group.

PROV (the WG is closed, but we will build upon its results) defined a data model [9] and vocabulary [8] to trace and describe provenance trails of arbitry artifacts, which is relevant to document processing of personal data. **SPECIAL involvement/liaison:** the group is closed already, but the team W3C contact, Ivan Herman has formerly been affiliated with ERCIM.

Verifiable Claims `https://www.w3.org/2017/vc/charter.html` is working on a data model and syntax (i.e., we assume as well a vocabulary) to express claims that a subject makes and their verification (e.g., by means of a digital signature of an issuer) – we assume that this group is in need of privacy technologies, but also might produces relevant outcomes in terms of RDF graph signatures. etc.). **SPECIAL involvement/liaison:** Manu Sporny, one of the already confirmed supporters of our CG is an active member and editor of relevant drafts of this WG.

Dataset Exchange Working Group `https://www.w3.org/2017/dxwg/` The Dataset Exchange Working Group will revise the Data Catalog Vocabulary, DCAT, taking account of related vocabularies and the extensive work done in developing a number of its application profiles. It will also define and publish guidance on the use of application profiles when requesting and serving data on the Web. **SPECIAL involvement/liaison:** WU is not yet a member, but considering joining and has also contributed actively with a paper [10] in the recent SDVOc workshop that initiated this WG, cf. `https://www.w3.org/2016/11/sdsvoc/report`

The W3C Tracking protection working group (https://www.w3.org/2011/tracking-protection/ (DNT), could potentially provide e.g. policy use cases, provides a mechanism to embed policy preferences about tracking in HTTP header fields, but not a corresponding (RDF or similar) vocabulary. **SPECIAL involvement/liaison:** Deutsche Telekom (i.e., T-Labs) and ERCIM are active members of this group.

The W3C PING (Privacy Interest Group) is mostly working on general policy within W3C, making sure that W3C standards have a privacy section etc.. This group is mainly useful as a dissemination channel, but not in terms of producing relevant standards on its own. **SPECIAL involvement/liaison:** ERCIM is an active member of this WG.

The Spatial Data on the Web Working Group (SDWWG) `https://www.w3.org/2015/spatial/`or `http://www.opengeospatial.org/projects/groups/sdwwg` constituted as a subgroup of the OGC Geosemantics DWG in parallel with a W3C WG with the goal to clarify and formalize the standards landscape for spatial and temporal information on the web; it defined relevant time and spatial ontologies, such as the OWL Time ontology (e.g. useful for us for modeling locations, durations timestamps, retention times and context) **SPECIAL involvement/liaison:** This working group has just closed, the team W3C contact, Phil Archer has formerly been affiliated with ERCIM.

The RDF Data Shapes WG/SHACL CG a CG has recently been established as successor of the just concluded RDF Data Shapes working group (https://www.w3.org/2014/data-shapes/charter) may be relevant in the context of expressing complex policies as Linked Data, in terms of constraints over RDF graphs. **SPECIAL involvement/liaison:** WU has been an active member of the WG and remains active in the CG.

Other potentially relevant W3C community groups where we are not (yet) involved, but which we monitor:

Interledger CG `https://www.w3.org/community/interledger/` this community group discusses modularisation of ledger technology which might be relevant for the scalability of building up a transparency layer as it could be useful for interconnecting distributed/modularised ledgers.

The W3C Block chain CG `https://www.w3.org/community/blockchain/` may be useful in terms of technologies and vocabularies relating to information stored on (distributed) ledgers.

3.2 Other Standards and Initiatives

Besides the ongoing W3C initiatives there are a number of complementary efforts in other standards bodies such as ETSI, IEEE, IETF, ISO, and OASIS. Additionally there is a global consortium known as Kantara that aims to enable innovation, standardisation and best practice with respect to personal data processing.

We have contacted relevant stakeholders from these initiatives already through a pre-kickoff event before the SPECIAL project started, hosted at WU: the launch event of WU's "Privacy and Sustainable Computing Lab" on 29th and 30th of September 2016, cf. `http://www.privacylab.at/events/launch/`, hosted a dedicated session on "Standardisation Efforts to Tackle Privacy & Ethics" where we attracted speakers from several standardisation organisations (besides the W3C) from our network:

- Konstantinos Karachalios,Ph.D, Managing Director **IEEE**-Standards association;
- Kai Rannenberg, who is active in the **ISO/IEC** standardisation of IT Security and Criteria;
- Robin Wilton, Technical Outreach for Identity & Privacy @ **ISOC (Internet Society)**, who is also a member of the **Kantara initiative**'s board of trustees and active in technology forums such as the **IETF**.

3.2.1 IEEE

The Institute of Electrical and Electronics Engineers (IEEE) Standards Association focus on advancing technology through global standardisation in a manner that transforms the way people live, work and communicate. Relevant standardisation initiatives relate to ethics, privacy, security and distributed technologies such as blockchain.

IEEE P7000 Model Process for Addressing Ethical Concerns During System Design `https://standards.ieee.org/develop/project/7000.html`: The purpose of this standard is to enable the pragmatic application of a Value-Based System Design methodology in order to refine ethical system requirements in systems and software life cycles in the form of an implementable process aligning innovation management processes.

This standard is currently in the making but we shall re-assess SPECIAL's system design upon its completion with respect to the proposed process.

IEEE SA - PDP IEEE SA - PDP - Personal Data Privacy Working Group `http://standards.ieee.org/develop/wg/PDP.html`: This standard defines requirements for a systems/software engineering process for privacy oriented considerations regarding products, services, and systems utilizing employee, customer or other external user's personal data. It extends across the life cycle from policy through development, quality assurance, and value realisation. It includes a use case and data model (including metadata). It applies to organisations and projects that are developing and deploying products, systems, processes, and applications that involve personal information. By providing specific procedures, diagrams, and checklists, users of this standard will be able to perform a conformity assessment on their specific privacy practices. Privacy impact assessments (PIAs) are described as a tool for both identifying where privacy controls and measures are needed and for confirming they are in place.

IEEE P1912 WG P1912 WG - Privacy and Security Architecture for Consumer Wireless Devices Working Group `http://standards.ieee.org/develop/wg/P1912_WG.html`: This standard describes a common communication architecture for diverse wireless communication devices such as, but not limited to, devices equipped with near field communication (NFC), home area network (HAN), wireless area network (WAN) wireless personal area network (WPAN) technologies or radio frequency identification technology (RFID) considering proximity; and specifies approaches for end user security through device discovery/recognition, simplification of user authentication, tracking items/people under user control/responsibility, and supports alerting; while supporting privacy through user controlled sharing of information independent of the underlying wireless networking technology used by the devices.

IEEE SA - blockchain wg Blockchain working group `http://standards.ieee.org/develop/wg/blockchain_wg.html`: This standard provides a common framework for blockchain usage, implementation, and interaction in Internet of Things (IoT) applications. The framework addresses scalability, security and privacy challenges with regard to blockchain in IoT. Blockchain tokens, smart contracts, transaction, asset, credentialed network, permissioned IoT blockchain, and permissionless IoT blockchain are included in the framework.

3.2.2 IETF

The remit of the Internet Engineering Task Force (IETF) is to develop high quality technical documentation that shapes the way we use, manage and advance the Internet.

IETF geopriv Geographic Location/Privacy `https://datatracker.ietf.org/wg/geopriv/about/`: The IETF has recognised that many applications are emerging that require geographic and civic location information about resources and entities, and that the representation and transmission of that information has significant privacy and security implications. We have created a suite of protocols that allow such applications to represent and transmit such location objects and to allow users to express policies on how these representations are exposed and used. The IETF has also begun working on creating applications that use these

capabilities, for emergency services, general real-time communication, and other usages.

The GEOPRIV working group is chartered to continue to develop and refine representations of location in Internet protocols, and to analyze the authorisation, integrity, and privacy requirements that must be met when these representations of location are created, stored, and used. The group will create and refine mechanisms for the transmission of these representations that address the requirements that have been identified.

IETF trans Public Notary Transparency `https://datatracker.ietf.org/wg/trans/charter/`: Mitigating web site certificate mis-issuance is the initial problem of interest for this working group. Certificate Transparency (CT, RFC6962) allows such mis-issuance to be detected in interesting and useful cases, for example by an auditor acting for the web site, or one acting to check general CA behaviour. The working group will produce a standards-track version of the experimental RFC 6962 for HTTP over TLS, reflecting implementation and deployment experience since that specification was completed.

3.2.3 ISO

The International Organization for Standardization (ISO) is a global network of national standards bodies whose mission is to support the development of consensus based industry relevant standards.

ISO/IEC 29100:2011 ISO/IEC 29100:2011 Information technology – Security techniques – Privacy framework `https://www.iso.org/standard/45123.html`: ISO/IEC 29100:2011 provides a privacy framework which specifies a common privacy terminology; defines the actors and their roles in processing personally identifiable information (PII); describes privacy safeguarding considerations; and provides references to known privacy principles for information technology. ISO/IEC 29100:2011 is applicable to natural persons and organisations involved in specifying, procuring, architecting, designing, developing, testing, maintaining, administering, and operating information and communication technology systems or services where privacy controls are required for the processing of PII.

ISO/IEC 27001:2013 ISO/IEC 27001:2013 is a specification for an information security management system (ISMS). Organisations that meet the standard may be certified compliant by an independent and accredited certification body on successful completion of a formal compliance audit. Particularly it contains a classification of controls that need to be considerd for risk assessment that could be relevant for personal data handling. A recent paper by Bartolini and Muthuri [1] mentions the relation of ISO/IEC 27001:2013 concepts in the context of developing a base ontology of concepts in the GDPR.

ISO/IEC 27018:2014 ISO/IEC 27018:2014 Information technology – Security techniques – Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors `https://www.iso.org/standard/`

61498.html: ISO/IEC 27018:2014 establishes commonly accepted control objectives, controls and guidelines for implementing measures to protect Personally Identifiable Information (PII) in accordance with the privacy principles in ISO/IEC 29100 for the public cloud computing environment. In particular, ISO/IEC 27018:2014 specifies guidelines based on ISO/IEC 27002, taking into consideration the regulatory requirements for the protection of PII which might be applicable within the context of the information security risk environment(s) of a provider of public cloud services.

ISO/IEC 2382:2015 ISO/IEC 2382:2015 Information technology – Vocabulary https://www.iso.org/standard/63598.html: TBA

3.2.4 OASIS

The Organisation for the Advancement of Structured Information Standards (OASIS) is primarily made up of public and private sector technology leaders that strive towards the advancement of standards for the information society.

OASIS COEL TC The OASIS classification of Everyday living (COEL) Technical Committee OASIS COEL, https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=coel: The OASIS COEL specification provides a privacy-by-design framework for the collection and processing of behavioural data. It is uniquely suited to the transparent use of dynamic data for personalised digital services, IoT applications where devices are collecting information about identifiable individuals and the coding of behavioural data in identity solutions. The specification pseudonymises personal data at source (IDA) and maintains a separation of different data types with clearly defined roles & responsibilities (RPE) for all actors. All behavioural data are defined as event-based packets (BAP). Every packet is connected directly to an individual and can contain a summary of the consent they provided for the processing of the data. A combination of a taxonomy of all human behaviours (COEL) and the event-based protocol provide a universal template for data portability. Simple interface specifications (MMI & PQI) enforce the separation of roles and provide system-level interoperability..

OASIS PMRM TC OASIS Privacy Management Reference Model (PMRM) Technical Committee https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=pmrm: The OASIS PMRM TC works to provide a standards-based framework that will help business process engineers, IT analysts, architects, and developers implement privacy and security policies in their operations. PMRM picks up where broad privacy policies leave off. Most policies describe fair information practices and principles but offer little insight into actual implementation. PMRM provides a guideline or template for developing operational solutions to privacy issues. It also serves as an analytical tool for assessing the completeness of proposed solutions and as the basis for establishing categories and groupings of privacy management controls.

OASIS PbD-SE TC OASIS Privacy by Design Documentation for Software Engineers (PbD-SE) Technical Committee https://www.oasis-open.org/

committees/tc_home.php?wg_abbrev=pbd-se: The OASIS PbD-SE TC provides privacy governance and documentation standards for software engineers. It enables software organisations to embed privacy into the design and architecture of IT systems, without diminishing system functionality. The PbD-SE TC work follows the Seven Foundational Principles of Privacy by Design:

- Proactive not Reactive; Preventative Not Remedial
- Privacy as the Default Setting
- Privacy Embedded into Design
- Full Functionality - Positive-Sum, Not Zero-Sum
- End-to-End Security - Full Lifecycle Protection
- Visibility and Transparency - Keep It Open
- Respect for User Privacy - Keep It User-Centric

PbD-SE offers a privacy extension/complement to OMG's Unified Modeling Language (UML) and serves as a complement to OASIS' eXtensible Access Control Mark-up Language (XACML) and Privacy Management Reference Model (PMRM).

3.2.5 Kantara Initiative

The Kantara Initiative, is a global consortium that focuses on the trustworthy use of personal data in the form of innovation, standardisation and best practices.

Kantara InfoSharing WG Consent & Information Sharing Work Group https://kantarainitiative.org/groups/ciswg/:. Project VRM and other related parties wish to build a framework around which a new type of personal information can be enabled to flow, and in doing so improve the relationship between demand and supply. Our contention is that when individuals are forced to sign organisation-centric privacy policies/ terms of use then this places limitations on the information that will be shared. If such constraints were removed, and capabilities built on the side of the individual, then new, rich information will flow - including actual demand data (as opposed to derived/ predicted demand). The goal of this working group is to identify and document the use cases and scenarios that illustrate the various sub-sets of user driven information, the benefits therein, and specify the policy and technology enablers that should be put in place to enable this information to flow. Kantara Information Sharing Working Group Lexicon[3].

3.2.6 ETSI

Finally, the European Telecommunications Standards Institute (ETSI) whose mission is to produce global standards for Information and Communications Technologies, have a number of standardisation Initiatives that could potentially be relevant for special.

[3]Kantara, https://kantarainitiative.org/confluence/display/infosharing/Lexicon

ETSI TS 102 941 V1.1.1 (2012-06) Intelligent Transport Systems (ITS); Security; Trust and Privacy Management Trust and Privacy Management `https://portal.etsi.org/webapp/workprogram/Report_WorkItem.asp?WKI_ID=41193`: To update TS 102 941 to include specification for identity and key management functions, e.g. for issuing and managing long-term and short-term identities. To address updates in data structures identified in TS 103 097 and latest versions of IEEE 1609.2 .

ETSI TR 122 949 V14.0.0 (2017-03) Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Study on a generalised privacy capability (3GPP TR 22.949 version 14.0.0 Release 14)

ETSI GS INS 009 V1.1.1 (2012-09) Identity and access management for Networks and Services (INS); Security and privacy requirements for collaborative cross domain network monitoring `http://www.etsi.org/deliver/etsi_gs/INS/001_099/009/01.01.01_60/gs_INS009v010101p.pdf`: Security and privacy requirements for distributed network monitoring; identifying gaps regarding distributed processing and computation, protocols, and (anonymised) data exchange. Identify existing techniques / specifications to use and which new ones are required.

4 Potential Standardisation Opportunities

In this section, we highlight two themes arising from the initial SPECIAL requirements that have broad applicability and could benefit greatly from standardisation. Namely core vocabularies and the development of a Transparent Linked Data Processing Platform.

4.1 Core Vocabularies

In the recent MyData conference, our assumption that one of the biggest obstacles for machine readable support for privacy is lack of interoperability was confirmed, for instance in the following sessions and topics:

Interoperable (and self-souvereign) identity management (`https://mydata2017.org/session/self-sovereign-identity/`): Herein, interoperability of identity as well as the importance of decentralised governance of identity was stressed in talks such as “The Impact of Decentralised Identifiers (DIDs)” by Drummond Reed, or by “Identity as Driver Towards Interoperability” by Eugeniu Rusu.

Technical Building Blocks Workshop (`https://mydata2017.org/session/technical-building-blocks-workshop/`): This session hosted talks and discussions on interoperability of architectures, such as Geoff Revill’s talk on “Proven Interoperability Through Semantic Data Architecture”.

The Expert Workshop: Making Consent Work (`https://mydata2017.org/session/consent-workshop/`) hosted the talk “Architecting consent with open

standards" by Joss Langford, and the subsequent Open Space discussion on "Making Consent Work" also heavily stressed on the importance of open standards and interoperability for consent.

We believe that SPECIAL's choice of using URIs and Linked Data for identity and interoperability will fit in well with those discussions and initiatives, but we need to start with bottom-up development of *core vocabularies* to enable such interoperability. At the same time, we need to involve key stakeholders in order to achieve such interoperability of core vocabularies. A similar approach has already successfully been deployed, e.g., in the Open Data initiative through bottom-up core vocabularies such as DCAT (cf. Section 2) or in industry led efforts like the `schema.org` initiative for interoperable metadata for search engines in an industry context.

This is why we put the primary focus of our standardisation activities mainly on this aspect, i.e. which core vocabularies would be needed to *express personal data handling*.

In the scope of our project, we will not be able to model all possible use cases of personal data handling, but we will focus on the requirements and used cases from our partners to build such a core vocabulary.

Deriving from the MCM in Deliverable D1.3 and [11, Section 5.1.1] and the generic use case from [11, Section 2.1] we will discuss a minimal set of concepts being presented in such a core vocabulary. Along these lines, hereafter follow some basic ideas for a minimal core vocabulary of expressing consent.

The MCM contains the following core concepts as depicted in Fig. 2 which we recall and refine herein and from which we will derive requirements (denoted with the letter **R**) of concepts and properties to express in a core vocabulary around expressing and recording consent:

- "Data" describes what personal data is collected from the data subject; a core vocabulary will need classes and properties to express various kinds of personal data, such as:
  - **R1** "static" personal data, including identifiers or attributes by which an individual can be identified directly or indirectly, such as name, personal ID number, online identifiers such as nick- or username, (home or work) address, etc.
  - Various "dynamic" context-dependent personal attributes or events related to an individual, e.g.: **R2** Location of a Person, heartrate at a certain point in time, participation at an event, visiting a Website, bit also e.g. being assigned a certain IP address by a provider, etc.
- "Processing" describes operations that are performed on personal data, such as:
  - **R3** re-sharing the data with another party
  - **R4** applying algorithms to the data, including aggregation, anonymisation, as well as properties of the outcomes of the algorithm, such as anonymity metrics, as described in [11]
- **R5** "Purpose" specifies the objective of such processing, such as *why* the processing is performed, e.g.

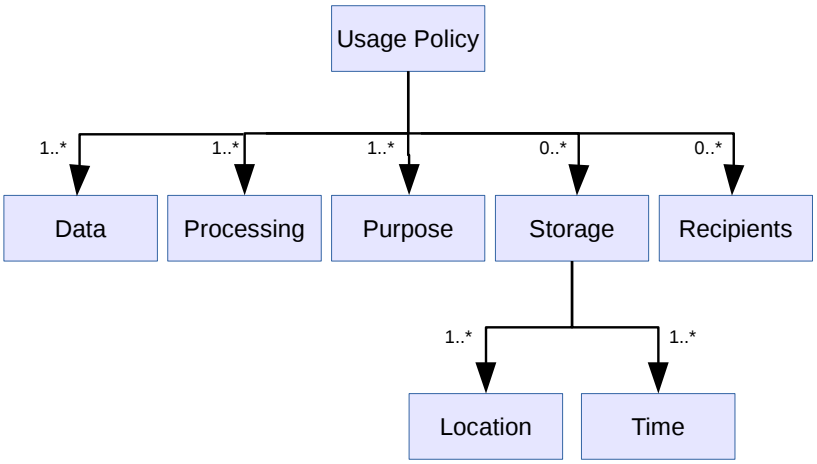

Figure 2: The minimum, core usage policy model (MCM) from [11]

  - re-sharing with a third party for marketing purposes
  - applying algorithms for improved service efficiency

- "Storage" specifies **R6** (i) where data are stored and (ii) for how long: here (i) could be on-premise with the data controller, or on a cloud service, and the duration (ii) could be a concrete time duration relative to the event data itself (e.g., "locations are stored for 24 hours", or relative to a contract between the data subject and data controller "account data is stored until the account is closed", etc.)

- **R7** "Recipients" specifies who is going to receive the results of data processing and, as a special case, whom data is shared with, that is recipients need to be categorised in at least data subjects, data processors, data controllers, but also regulators, or – as mentioned in [11], business partners, etc.

We recall that the overall goal of the MCM in D1.3 [11] was to identify concepts around provision and verifying of, as well as updating *consent*. That is, a core vocabulary will also need to **R8** express mechanism and concepts related to recording and controlling consent, including conditions defining "restriction of processing" (as per article 4 of the GDPR), which again could be tied to (possibly complex) *policies*.

Driven by the generic use case scenario from [11, Section 2.1], in the following, we will instantiate a small core vocabulary with references to and *bridging between* existing vocabularies as a starting point for discussions for a standard core vocabulary. This core is not meant to be exhaustive nor final, but – in line with the overall purpose of the present deliverable – to set out a *strategy* for standardisation. That is, we consider the following as input for discussion which we will continuously refine, until and based

on the results of the planned first SPECIAL standardisation workshop, cf. Section 5.1 below.

Finally, we stress that the approach taken here to define and grow a core vocabulary is different from abstract ontologies to model concepts of the GDPR so far, e.g. [1], or product specific non-open vocabularies for compliance checking such as under development in products like TopQuadrant's TopBraid Enterprise Data Governance (EDG).[4]

Particularly, in the following we will make reference to selected promising RDF vocabularies (indicated by their respective namespace prefixes):

- RDF/RDFS: the RDF and RDFS core vocabularies

  ```
  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
  ```

- OWL: the OWL core vocabulary.

  ```
  @prefix owl: <http://www.w3.org/2002/07/owl#>.
  ```

- FOAF & vCard: the friend-of-a-friend and vCard vocabularies have a notion of agents and persons as well as various attributes for personal data which we will make use of.

  ```
  @prefix foaf: <http://xmlns.com/foaf/0.1/>.
  @prefix vcard: <http://www.w3.org/2006/vcard/ns#>
  ```

- The DICOM ontology for healthcare metadata has attributes relevant to fitness and health, such as HeartRate.

  ```
  @prefic health: <http://purl.org/healthcarevocab/v1#> .
  ```

- schema.org offers likewise classes to describe persons, agents and a variety of other generic classses and their properties:

  ```
  @prefix schema: <http://schema.org/>.
  ```

- P3P: The P3P working group published an RDF vocabulary for expressing the concepts of P3P:

  ```
  @prefix p3p: <http://www.w3.org/2002/01/p3prdfv1#>.
  ```

- The ODRL vocabulary:

  ```
  @prefix odrl: <http://www.w3.org/ns/odrl/2/>.
  ```

- The OWL Time Vocabulary

  ```
  @prefix time: <http://www.w3.org/2006/time#>.
  ```

- The Open Time and Space Core Vocabulary Specification

  ```
  @prefix oc: <http://observedchange.com/tisc/ns>.
  ```

[4] cf. https://www.topquadrant.com/products/topbraid-enterprise-data-governance/, https://2017.semantics.cc/hodgson-ralph

- The NeoGeo vocabulary for describing topological relations between features.

  ```
  @prefix spatial: <http://geovocab.org/spatial#>.
  ```

- The GeoSPARQL vocabulary for representing geospatial data in RDF,

  ```
  @prefix geosparql: <http://www.opengis.net/ont/geosparql#>.
  ```

  The WGS84 Geo Positioning vocabulary for representing latitude, longitude and altitude information in the WGS84 geodetic reference datum.

  ```
  @prefix geo: <https://www.w3.org/2003/01/geo/wgs84_pos#>.
  ```

- The PROV vocabulary,

  ```
  @prefix prov: <http://www.w3.org/ns/prov#> .
  ```

In order to discuss define a core vocabulary we will work through the requirements **R1**-**R8** not in chronological order, but building up in an intuitive order, defining base an (extensible) set of base concepts and properties under the namespace prefix `sp:`[5]

```
@prefix sp: <http://www.special-privacy.eu/ns/> .
```

as well as mapping those to existing ontologies and vocabularies:

4.1.1 R7 "Recipients" and agents

This covers classes properties to describe *agents* involved in Personal Data Sharing and Processing as defined in Article 4 of the GDPR, such as "subject", "controller", "processor". We assume in the context of SPECIAL, that all these are – according to the Linked Data principles – identified/identifiable by URIs.

```
sp:Subject rdfs:subClassOf foaf:Person; schema:Person;
  rdfs:comment "natural person as per Art. 4 (1) of the GDPR".
sp:Controller rdfs:subClassOf foaf:Agent;
   rdfs:comment "controller as defined by Art. 4 (7) of the GDPR".
sp:Processor rdfs:subClassOf foaf:Agent;
  rdfs:comment "processor as defined by Art. 4 (8) of the GDPR".
sp:Recipient rdfs:subClassOf foaf:Agent;
  rdfs:comment "recipient as defined by Art. 4 (9) of the GDPR".
sp:ThirdParty rdfs:subClassOf foaf:Agent;
  rdfs:comment "third party as defined by Art. 4 (10) of the GDPR".
```

Here, we understand `foaf:Person`s to subsume natural persons, whereas we understand `foaf:Agent`s to subsume "natural or legal person, public authority, agency or another body" as per Art. 4 of the GDPR.

[5] Note that for the moment this namespace URL is not yet registered nor dereferenceable and – depending on the outcome of further standardisation discussions – may be subject to change.

4.1.2 R1 Static Personal "Data" and and attributes

Herein, we consider relatively static (of course also these attributes can change over time, but we assume them to be relatively relatively stable over time, that is, *typically* to remain constant or change only rarely (e.g. once or twice) for the duration of a contract between a data subject and a data controller/processor) attributes and classes. As a starting point, we take the ones listed above, namely, name, personal ID number, online identifiers such as nick- or username, (home or work) address, but also, email-address or date of birth:

name: attributes denoting personal names are present in the FOAF and vCard Ontologies.

```
sp:name rdfs:subPropertyOf foaf:name, vcard:fn.
sp:givenName rdfs:subPropertyOf foaf:givenName, vcard:given-name.
sp:familyName rdfs:subPropertyOf foaf:familyName, vcard:family-name.
```

We note that the vCard ontology supports more complex structures to model names, which we leave out in this initial version of a core vocabulary.

personal ID number: in general, we recommend to use/mint URIs of a particular, dedicated namespace for personal IDs issues by a certain organisation, e.g. a personal ID number, such as a national ID or a customer number/ID with a certain organisation. E.g. company "ExampleCompany" who registered/owns the domain `http://example.org/` could in the most straightforward manner simply mint a IRI scheme as follows to assign a URI to its customer with the ID "08154711"

```
@prefix exID: <http://example.org/customer-ID/> .
exID:08154711 a sp:Subject.
```

However, it is not straightforward to semantically identify this URI as an ID issued by a specific company. To this end, alternatively, IDs can be mapped to RDF as a literals that adheres to some lexical space and is *issued* by some organisation. This could be modelled using for instance using the vocabulary currently being defined by the Verifiable Claims WG [2]. The Verifiable Claims WG has a respective example for a modeling a passport number as JSON-LD (using the namespaces as per JSON-LD context `https://w3id.org/identity/v1`) in its first public working draft defining dedicated properties/attributes for "issuer" , "issued" date/time, as well as date/time when and id or claim "expires", cf. [2, Example 9]. We shall follow the modeling of IDs and claims and integrate them in our model, if necessary in future versions of this document and have invited stakeholders of the Verifiable Claims working group explicitly to our planned first workshop.

online identifiers such as nick- or username: could – again as a starting point – be modeled using FOAF, which offers explicit attributes for online accounts and user names, for instance:

```
sp:onlineUserName rdfs:subPropertyOf foaf:accountName .
sp:nickname rdfs:subPropertyOf foaf:nickname .
```

(home or work) address: both vCard and schema.org offer attributes and concepts for modelling addresses. We exemplify a core vocabulary herein borrowing from schema.org's PostalAddress:

```
sp:HomeAddress rdfs:subClassOf schema:PostalAddress.
sp:WorkAddress rdfs:subClassOf schema:PostalAddress.
sp:address rdfs:subPropertyOf schema:address .
sp:addressCountry rdfs:subPropertyOf schema:addressCountry.
sp:addressLocality rdfs:subPropertyOf schema:addressLocality.
sp:addressRegion rdfs:subPropertyOf schema:addressRegion.
sp:streetAddress rdfs:subPropertyOf schema:streetAddress.
sp:addressCountry rdfs:subPropertyOf schema:streetAddress.
sp:postalCode rdfs:subPropertyOf schema:postalCode.
```

The use of an own namespace, instead of just reusing e.g. schema.org or vCard properties has the advantage of being able to declare e.g. special categories of personal data, for instance in our case for the personal address, we could declare the value of a certain attribute to belong to a specific category according to P3P, such as the physical address as "Physical contact data" according to P3P:[6]

```
sp:address rdfs:range p3p:physical .
```

An example is given following the one at `http://schema.org/PostalAddress` as follows:

```
:me sp:address [
    a sp:HomeAddress ;
    sp:addressLocality "Mexico Beach" ;
    sp:addressRegion "FL" ;
    sp:addressCountry "US" ;
    sp:streetAddress "3102 Highway 98"
 ] .
```

The level of granularity of modelling addresses and regions, and/or whether – instead of strings –referring to ontologies for countries and regions is another open point of potential discussion in developing a respective core vocabulary.

4.1.3 R2 Dynamic Personal "Data" and and attributes

These are more dynamic personal data attributes typically associated with a duration that can be modelled in several ways, as such statements need some kind of reification; for a recent article outlining different possibilities of RDF reification such as RDF reification, n-ary relations, named graphs and singleton properties, cf. [5]. Herein, we will use singleton properties, that is, specific per-triple properties that can attach metadata, such as duration. For expressing duration, we recommend the use of the OWL Time ontology.

```
sp:locatedAt rdfs:subPropertyOf oc:locatedAt .

@prefix now: <http://example.org/2017-09-26T13:20:00Z/>
```

[6] cf. `https://www.w3.org/TR/P3P/#Categories`

```
now:locatedAt rdfs:subPropertyOf sp:locatedAt .
now:heartrate rdfs:subPropertyOf health:HeartRate.

:me now:locatedAt [
  geo:lat 48.21667 ;
  geo:long 16.4
] .

now:locatedAt
  time:hasBeginning [
    a time:Instant ;
    :inXSDDateTimeStamp "2017-09-26T13:17:03Z"^^xsd:dateTimeStamp
  ] ;
  time:hasEnd [
    a time:Instant ;
    :inXSDDateTimeStamp "2017-09-26T13:18:14Z"^^xsd:dateTimeStamp
  ] .
```

The schema of using time-stamped singleton properties can be used for any dynamic personal data attributes to express transaction time, whereas the time interval is expressed by an interval in the OWL time ontology.

4.1.4 R3 & R4 Expressing processing of data

When certain data is being re-shared we will again need some form of reification to express this. As we assume data is being shared in the form of an RDF graph, it makes sense to consider the shared data in a *named graph*, thus one option to express re-sharing is to refer to this named graph. We will use TRIG syntax [3] for the following example.

Likewise, we can define attributes to define what has happened with a certain such named graph using the PROV vocabulary to express re-sharing or other processing steps performed with it.

For instance, let us assume dataset `ex:mydata` has been anonymised by algorithm `ex:alg1` under the responsibility of data controller `ex:c1`, and the resulting dataset has been shared with company `ex:c2`. We will now exemplify how this can be expressed using PROV and ORDL:

```
ex:mydata { ex:me sp:name "John Doe" . ... }

ex:anonmydata prov:wasGeneratedBy
   [ a prov:Activity;
     odrl:action sp:anonymize;
     prov:startedBy ex:alg1;
     prov:endedBy ex:alg1;
     prov:wasAssociatedWith ex:c1 ;
     prov:startedAtTime "2017-09-26T13:17:03Z"^^xsd:dateTimeStamp ;
     prov:endedAtTime "2017-09-26T13:18:13Z"^^xsd:dateTimeStamp ;
     prov:used ex:mydata
   ] .

ex:anonmydata1 prov:wasGeneratedBy
   [ a prov:Activity;
```

```
    odrl:action odrl:share;
    prov:wasAssociatedWith ex:c1, ex:c2;
    prov:qualifiedAssociation [
     a prov:Association;
     prov:agent ex:c1;
     prov:hadRole sp:Controller ],
    [
     a prov:Association;
     prov:agent ex:c2;
     prov:hadRole sp:Recipient ];
    prov:atTime "2017-09-26T14:01:00Z"^^xsd:dateTimeStamp ;
    prov:used ex:anonmydata .
]
```

From the viewpoint of proposing a best practice for vocabulary use for modeling personal data processing, the illustrative example here highlights several open questions (which should be discussed in a respective standardisation activity started by SPECIAL) about:

- PROV has its own attributes for time intervals, whereas (as we have seen above) OWL Time has another way of defining intervals, which should probably be unified.
- The ODRL action for sharing (`odrl:share` is used here as an attribute of a PROV activity to indicate data sharing; there might be other, more preferable ways to combine this.
- we will need to define in a core vocabulary a number of specific of `odrl:action`s for typical categories of processing; e.g., in the example above we have introduced an odrl:action to describe the process of anonymisation. The level of granularity for such actions is to be determined based on use cases and (interoperability) requirements.
- PROV cannot express the "direction" of an activity associated with two parties directly, in this case the controller and the recipient, which we modelled herein with two roles in a `prov:qualifiedAssociation`.

4.1.5 Modeling R5 Purpose of & R6 Storage Restrictions related to Consent

A minimal vocabulary should be able to express at least that: **Consent** given by a **data Subject** at a certain point in time (**transaction time** which again can be modeled by singleton properties) permits (i.e., is a subclass of a **odrl:Permission** in ORDL) certain processing **odrl:Action** for a specified **purpose** on personal data for a specified **duration** (**validityTime/-Duration**) under certain **restrictions/conditions** (also expressible in ODRL).

We leave the exact details of modeling consent in RDF to further discussion, we remark that while specifically ORDL covers modeling **actions**, it does not yet cover **purpose** or **storage restrictions** explicitly, which will need an extension or specific new profile of the current ODRL vocabulary. As the POE/ODRL working group is still active we will strive to resolve this as part of the planned standardisation activity started by SPECIAL.

4.2 Transparent Linked Data Processing Platform

In this section, we discuss what a Transparent Linked Data Processing Platform might look like. The primary goal is to investigate where we can leverage existing standards and what are the potential avenues for further standardisation. The goal of the section is not to define concrete standardisation requirements but rather to provide the basic information necessary to initiate a discussion with various stakeholder groups. Although SPECIAL focuses on transparency and compliance with respect to personal data processing and sharing from a platform perspective it would be wise to broaden the discussion to transparent and compliant data processing in general, thus ensuring a greater appeal to various stakeholder groups. For example, other use cases in this context could include public procurement, media, and political transparency to name but a few.

4.2.1 Core high level requirements

Based on the detailed requirements presented in *D1.3 Policy, transparency and compliance guidelines V1*, it is possible to derive four elementary building blocks that are necessary for transparent data processing. In this section we provide a high level summary of said building blocks:

- *Policies* include legislative obligations, business rules and usage constraints that govern data processing and sharing. Functional requirements include the ability to create, read, updated and delete policies (CRUD), and also the ability to specify logical formalisms and inference rules that can be used to simplify policy specification and compliance checking.
- *Events* relate to all data processing and sharing events performed by data controllers, processors and third parties. Here again, functional requirements include the ability to create, read, updated and delete events (CRUD). However, as highlighted in *D1.3*, novel update and delete functions may be needed in order to ensure the consistency of the event log.
- *Transparency* refers to the functionality that is necessary to demonstrate how data [will be/is] processed and with whom it [will be/is] shared.
- *Compliance checking* is necessary in order to verify if data processing and sharing practices conform to relevant regulatory, business and usage policies.

4.2.2 An n-tier architecture

In *D1.4 Technical requirements V1* we presented a basic initial architecture for the SPECIAL platform. In this section, we broaden the discussion to consider the different architectural layers of a transparent Linked Data platform. An n-tier architecture is particularly relevant in an industry context as it provides support for many highly desirable features such as: scalability, availability, flexibility and ease of integration. In this section, we discuss the relevant architectural layers in the context of existing Semantic Web standards. The aim is not to discuss the merits of one choice over the other but rather to highlight what standards are available, with a view to discussing potential gaps in terms of standardisation that need to be addressed.

Data Sources The proposed Transparent Linked Data Processing platform will build upon Linked Data principles and the underpinning RDF data model. Motivations for this choice include the fact that RDF is particularity suitable for data integration (especially in terms of downstream processing), it enables reuse of vocabularies, and it leverages the existing web infrastructure. From a standardisation perspective there are a number of existing W3C recommendations concerning the RDF model, its semantics and the various serialisations (a summary of the relevant standards can be found in *Section 2.2.1*).

Resource Access Layer When it comes querying and manipulating RDF data, the standard query and update language is SPARQL. However, the suite of standards also includes the *SPARQL 1.1 Entailment Regimes* recommendation which describes inference at the data access layer (i.e. query layer). A summary of relevant W3C recommendations is provided in *Section 2.2.4*. Other relevant standards include: transformation languages such as *A Direct Mapping of Relational Data to RDF* and *R2RML: RDB to RDF Mapping Language* (cf. *Section 2.2.6*).

Service Interface Layer In order to provide transparency with respect to data processing and sharing and to verify compliance with relevant policies, it is necessary to hook into existing organisational infrastructure. Here there are a number of possible alternatives ranging from minimal to full system integration. Consequently, there are a number of possible intersection points from a data, business logic and user interface perspective. Here, it would be wise to assume that parts of the system may reside inside the company while others may be hosted by external cloud providers. In terms of existing standards, besides the SPARQL query language itself there are a number of protocols for accessing RDF data (and Metadata) via HTTP, and also for federating queries over a variety of data sources such as: *Linked Data Platform 1.0* (*Section 2.2.7*) and, *SPARQL 1.1 Protocol* and *SPARQL 1.1 Graph Store HTTP Protocol* (*Section 2.2.4*).

Business Logic Layer In terms of SPECIAL the business logic is primarily concerned with obtaining consent for personal data processing and sharing, providing transparency concerning the use of personal data in existing systems, and verifying compliance with usage policies and regularity constraints. Here we need the ability to specify rules based on deontic concepts (i.e. positive and negative permissions and obligations). Here initiatives such as the *Permission and Obligation Expression WG* who are working towards standardising a core model and vocabulary that could be used to represent policies would be very relevant. From a rules perspective existing standardisation efforts have focused on rule interchange in the form of *Rule Interchange Format (RIF)* and constraint specification in the form of the *Shapes Constraint Language (SHACL)*. Both of which are described in *Section 2.2.5*. Additional initiatives (outside of the usual standardisation groups) include the *SPARQL Inferencing Notation (SPIN)* that focus on adding rules and constraints to the SPARQL query language.

Presentation layer One of the primary requirements of the User Interface (UI) is that users should be able to access the information obtained from multiple systems via a single interface. From a presentation perspective, existing standardisation efforts

have primarily focused on embedding RDF in HTML documents (cf. *Section 2.2.7*) as opposed to making it easy to integrate data from multiple sources.

Discover layer Although the discovery layer may not be strictly necessary it would be useful for service discover in an IoT setting or for maintaining a list of template models and components that can plug into a mashup infrastructure. Primary standardisation efforts with respect to discovery include the *Data Catalog Vocabulary (DCAT)* and the *Protocol for Web Description Resources (POWDER)* that allow catalogs of resources to be marked up in a manner than enables discovery (cf. *Section 2.2.3*).

Clients By adopting standard web technologies it should be possible to design for a wide range of devices such as laptops, tablets, and mobile phones. However, typical questions with respect to suitability of thin or thick clients for various tasks, and enabling intelligent web clients would need further consideration.

4.2.3 Potential standardisation opportunities

Although each physical layer serves a particular purpose, there are three major components from a logical perspective, namely: (i) the presentation component (composed of the presentation layer and the relevant business logic), (ii) the data component (made up of the data and data access mechanisms), and (iii) the service component (which is the glue that binds the data and presentation components together, but also serves as an interface to existing company systems. At the current point in time we see the presentation and the services components as the most pressing in terms of standardisation need, however there are also a number of gaps with respect to the data component. The results of our preliminary analysis are presented below:

Data Component One could argue that the data component is the most mature in terms of standardisation, however there are a number of limitations from a security perspective. The current semantic we stack does not differentiate between open (publicly accessible) and closed (access controlled) data sources.

Although there is a W3C WG note, entitled *LDP Access Control Usecases and Requirements for Access Control for the Linked Data Platform* `https://www.w3.org/TR/ldp-acr/`, that setsforth a charter for developing a standard for HTTP-based access control, to date there hasn't been any concrete activities. Likewise, the combination of encryption and digital signatures and web semantic web technologies has not been taken up to date.

Service Component When it comes to accessing RDF data (and Metadata) via HTTP there are a number of different possibilities. For example the following protocols can be used to read and write to RDF documents and RDF data stores:

- SPARQL 1.1 Protocol `https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/`:
- Linked Data Platform 1.0 `https://www.w3.org/TR/ldp/`:
- SPARQL 1.1 Graph Store HTTP Protocol `https://www.w3.org/TR/sparql11-http-rdf-update/`:

The differences and similarities between the alternative approaches is presented in an article entitled: *Linked Data Platform (LDP) vs SPARQL Graph Store HTTP Protocol (GSP)* `https://www.w3.org/2012/ldp/wiki/Linked_Data_Platform_(LDP)_vs_SPARQL_Graph_Store_HTTP_Protocol_(GSP)`.

However, in certain contexts it may be desirable or even necessary to employ a middleware component such as a messaging queue. Message queuing services, are commonly used for exchanging messages between devices and systems that are not always connected to each other. Relevant standardisation efforts in this context include the *Linked Data Notifications* `https://www.w3.org/TR/ldn/` protocol.

The primary question is how and where should we hook into existing systems and what protocols are necessary to support transparent data processing.

Presentation Component The term *Mashup* is commonly used to denote an application that uses content from several different sources and displays this content in a single UI. From a transparency perspective, mashups could potentially be used to support integration of data coming from several different sources (most likely under the control of different controllers/processors). Here standardisation could come in the form of a general plug and play framework for mashups, including standard templates and interfaces especially in terms of generic policies, events, transparency and compliance controls that could be adopted by various companies to ensure that data subjects can manage all of their data via a generic UI. Key considerations include:

- Identifier consolidation, which is necessary to integrate related data from multiple sources;
- Schema mapping in order to uncover semantic equivalence between systems and domains;
- Data transformation into a common data model that supports integration, reuse and analytics; and
- Simplification of presentation and maintenance using standard RDF and OWL based reasoning.

5 SPECIAL Standardisation Roadmap

The SPECIAL consortium aims to run a number of standardisation workshops that will be used to bring together a diverse array of external stakeholders such as citizens, public administrators, companies, as well as representatives from other projects such ICT-14 and the CSA in ICT-18. The initial workshop has the primary goal to initiate the launch a working group (WG) or community group (CG) within the World Wide Web Consortium (W3C) that will not only serve to inform the project but may also serve as a precursor to the eventual standardisation of the SPECIAL project activities.

5.1 Workshops

We plan one or two workshops, ideally under the umbrella of W3C, with the goal of a resulting CG or WG charter, first one in the first quarter of 2018, the 2nd workshop

earliest towards the end of 2018.

The W3C Process document[7] sets out expectations for workshops and symposia under the governance of W3C. The goal of a Workshop is usually either to convene experts and other interested parties for an exchange of ideas about a technology or policy, or to address the pressing concerns of W3C Members. As organisers of such a workshop we plan to solicit position papers for the workshop program and to use those papers to choose attendees and/or presenters.

To this end, we have prepared briefing materials for approval by the W3C Management which set out the scope of the workshop and why it is timely and beneficial to W3C's mission. We will, as a next step, consult with W3C who would go on the workshop's program committee agree on the candidates for chairs. We will pro-actively collaborate with W3C staff and the chairs and program committee on the final call for papers, reviewing the submissions, and preparing the agenda.

5.1.1 1st Workshop: Linked Data Vocabularies for Transparency and Privacy controls

The call for participation will have to be disseminated no later than six weeks before the workshop's scheduled start date, which is why depending on W3C's approval, we might still need to adapt the planned workshop dates accordingly, compared to the dates named in Appendix A . It may in fact make sense to give a longer gap than that to give potential participants more time to prepare their position papers and to make their travel arrangements. As the planned venue for the first workshop is WU, invitation letters for visa application purposes will be handled by the team of WU.

The first workshop shall focus on core vocabularies for interoperability of privacy, under the base assumption of SPECIAL that Linked Data can provide a basic interoperability layer. Therefore the workshop shall gather interested parties in the following topics to discuss standardisation potential for such core vocabularies:

- Core Vocabularies for privacy and transparency in linked data
  - Identity management vocabularies
  - Categories of sensitive and personal data
  - Modeling personal data usage, processing, sharing, and tracking
  - Modeling/Interlinking aspects of privacy and provenance
  - Modeling consent
- Vocabularies to model privacy policies, regulations, and involved (business) processes Modeling permissions, obligations, and their scope in dealing with personal data (particularly, profiles and extensions of ORDL and POE)
- Reasoning about formally declared privacy policies, in order to detect policy violations, breach and enforce policies
- Finally, how to link and liaise with related efforts such as W3C's Social Web WG, Verifiable Claims WG, ODRL/POE WG, Credentials CG, and PROV WG or other

[7] https://www.w3.org/2017/Process-20170301/#GAEvents

(non-W3C) efforts (e.g. OASIS XDI, OASIS COEL, Kantarainitiative's CISWG) under one hood through joint/linked vocabularies and vocabulary profiles?

In order to satisfy the W3C's process requirements outlined above, we have drafted a call for the first workshop, which you find in Appendix A of this deliverable. Appendix **??** (consortium-confidential) contains a list of potential stakeholders interested in such workshop (currently 15 from 12 different organisations, 11 of which are W3C members, excl. WU and ERCIM) and Appendix B contains an invitation letter which we sent to these stakeholders (and around 20 more who did not yet respond) in order to gather support for the workshop. These stakeholders include for now mostly persons involved in specific, relevant technical initiatives or projects; apart from that, we shall further reach out to lead contacts in other standardisation consortia mentioned in Section 3.2, namely:

- Konstantinos Karachalios,Ph.D, Managing Director **IEEE**-Standards association
- Kai Rannenberg, who is active in the **ISO/IEC** standardisation of IT Security and Criteria
- Robin Wilton, Technical Outreach for Identity & Privacy @ **ISOC (Internet Society)**, active in **IETF**, and member of the **Kantara initiative**'s board of trustees

In addition to these stakeholders, as soon as we have W3C's management approval, we explicitly plan to contact the following ICT-14 and ICT-18 projects to solicit their submissions in the form of position papers describing their use cases relating to personal data handling:

- **ICT-18**: e-Sides - Ethical and Societal Implications of Data Sciences: Start date: 2017-01-01, End date: 2019-12-31. `http://cordis.europa.eu/project/rcn/206175_en.html`
- **ICT-18**: SODA - Scalable Oblivious Data Analytics, Start date: 2017-01-01, End date: 2019-12-31. `http://cordis.europa.eu/project/rcn/205932_en.html`
- **ICT-18**: MH-MD - My Health - My Data: Start date: 2016-11-01, End date: 2019-10-31 `http://cordis.europa.eu/project/rcn/206202_en.html`
- **ICT-14**: SLIPO - Scalable Linking and Integration of Big POI data, Start date: 2017-01-01, End date: 2019-12-31. `http://cordis.europa.eu/project/rcn/206003_en.html`
- **ICT-14**: AEGIS - Advanced Big Data Value Chain for Public Safety and Personal Security, Start date: 2017-01-01, End date: 2019-06-30. `http://cordis.europa.eu/project/rcn/206179_en.html`
- **ICT-14**: EW-Shopp - EW-Shopp - Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey, Start date: 2017-01-01, End date: 2019-12-31. `http://cordis.europa.eu/project/rcn/207028_en.html`

- **ICT-14**: euBusinessGraph - Enabling the European Business Graph for Innovative Data Products and Services, Start date: 2017-01-01, End date: 2019-06-30. `http://cordis.europa.eu/project/rcn/206353_en.html`
- **ICT-14**: Data Pitch - Accelerating data to market, Start date: 2017-01-01, End date: 2019-12-31. `http://cordis.europa.eu/project/rcn/206193_en.html`
- **ICT-14**: FashionBrain - Understanding Europe's Fashion Data Universe, Start date: 2017-01-01, End date: 2019-12-31. `http://cordis.europa.eu/project/rcn/206358_en.html`
- **ICT-14**: QROWD - QROWD - Because Big Data Integration is Humanly Possible, Start date: 2016-12-01, End date: 2019-11-30. `http://cordis.europa.eu/project/rcn/206181_en.html`
- **ICT-14**: BigDataOcean - BigDataOcean - Exploiting Ocean's of Data for Maritime Applications, Start date: 2017-01-01, End date: 2019-06-30. `http://cordis.europa.eu/project/rcn/205983_en.html`

The planned timeline for the planned workshop and follow-up standardisation activities is as follows:

- 15 October 2017: approval by W3C management and immediate publication/dissemination of call for papers.
- 08 December 2017: deadline for submission of expressions of interest and position statements
- 22 December 2017: acceptance notification and registration instructions sent
- 12 January 2018: program announced
- 19 January 2018: deadline for registration
- 24-25 January 2018: Workshop

These milestones reflect the current assumptions and might change and be adapted depending mainly on the approval date by W3C's management as part of the workshop report (second quarter of 2018). In the unlikely case, we do NOT get approval by W3C's management, we will resort to found a CG immediately (see section5.2 and announce/repurpose the planned workshop at WU as a CG kickoff workshop.

5.1.2 Further Workshops

As a second step, after the initial goal of (1) achieving a common understanding of base vocabularies necessary to express and interchange information related to privacy, consent and personal data handling (which is essentially covering the points described in Section 4.1 of the present deliverable, we will explore standardisation (2) along the different routes related towards the final goal of a Transparent Linked Data Processing Platform as outlined in Section 4.2 as well as (3) revisiting and extending the base Semantic Web and Linked Data standards for 2.1.

As for a timeline, we will reconsider what is possible within the scope of SPECIAL after the first workshop, but in principle it makes sense to re-consider a workshop around (2) between the release of *D3.2 Policy & events release [M16]* and the delivery of *D6.4 Market analysis and plan for exploitation [M18]* in order to have results of a respective workshop considering the standardisation potential on an architectural/platform level as input for D6.4 (and also in order to get relevant different market players involved, if needed.

As for (3), we constantly assess our project results for relevant input to standards and will propose extensions of existing standards (such as e.g. encrypted and compressed RDF serialisations such as HDT-crypt [4] as we need them, and consider submitting those potentially useful input for standards as *W3C member submissions*: W3C's Member Submission process allows Members to propose technology or other ideas for consideration by the Team. After review, the Team MAY publish the material at the W3C Web site.

5.2 W3C Community Group (CG)

W3C Community and Business Groups are aimed to provide an open forum, where (Web) developers and other stakeholders develop specifications, hold discussions, develop architecture proposals and test suites, and connect with W3C. These groups are proposed and run by the community itself and serve therefore as an ideal vehicle for building a community around technologies to enable policy-compliant personal data processing and transparency. We plan to choose the model of a community group, for its low-entry-barrier: CGs do not require fees and are open for both W3C members and non-members to join.

The goal of the outcome of the first workshop is the foundation of a community group around standard vocabularies and linked-data based architectures to scalably maintain and interchange personal data processing and transparency.

Such a Community Group (CG) within W3C shall serve as a basis for the coordination of concrete standardisation activities and as a framework to cooperate with other stakeholders and initiatives that are not SPECIAL Partners as outlined above. Specifically, we consider the level of agreement and standard proposals not yet at a level to propose a concrete standard or vocabulary for W3C's standard recommendation track, so we will strive to have a charter draft[8] and scope of the WG fixed as a result of the discussion at and prior to (based on the submitted and accepted papers) the 1st SPECIAL standardisation workshop, to be both inclusive for a wide community and in line with SPECIAL's project goals.

Finally, SPECIAL plans to offer its expertise and services as a "think tank" to organisations with relevance to the project's aims including the European Commission, national regulators, standardisation bodies or data protection authorities, particularly coordinated through the planned W3C CG. That is, we will provide SPECIAL's results through the CG, to obtain and also re-include feedback, thus creating a virtuous circle with our stakeholders.

[8]The charter draft to be followed is available at `http://w3c.github.io/cg-charter/CGCharter.html`

6 Conclusions

Existing Semantic Web standards are at the core of the SPECIAL project and the Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance that will be developed throughout the course of the project. In this deliverable we provide a survey of existing W3C Semantic Web Standards, point to several potentially interesting standardisation initiatives (both in the W3C and by other standardisation bodies), and highlight potential standardisation opportunities within the remit of the SPECIAL project.

Our initial standardisation goal is to identify the base vocabularies necessary for expressing consent requests, usage policies, and data processing and sharing events that are necessary for demonstrating transparency with respect to personal data processing and facilitate automatic compliance checking.

Following on from this, we aim to further investigate the components that are necessary to develop a Transparent Linked Data Processing Platform, and the standardisation potential from a data, service and presentation perspective.

An additional potential output of the SPECIAL project is a reference architecture, which further advances the the logic, proof, trust, and security layers of the Semantic Web Technology Stack. Considering that we aim to build on existing W3C standards and several partners are already part of existing W3C standaridation activities, our standardisation roadmap includes the organisation of dedicated workshops under the W3C umbrella and the formation of a W3C community group that will serve as a basis for the coordination of concrete standardisation activities and as a framework to cooperate with other stakeholders and initiatives.

Acknowledgements

We thank Simon Steyskal (WU) for valuable comments. We also thank the presenters at the MyData 2017 conference named in Section 4.1 for valuable discussions and inputs. Liewise, we would like to thank Konstantinos Karachalios, Kai Rannenberg, and Robin Wilton for valuable discussions and insights regarding standards around privacy and the relevance of privacy within standardisation organisations, which they shared with us during the launch event of WU's "Privacy and Sustainable Computing Lab"[9] in Sept. 2016, as well as all the members of the lab for discussions. Finally, we thank all the other stakeholders that already replied to our request for support for the planned 1st SPECIAL standardisation workshop, cf. Appendices A,B, and **??**.

[9] http://www.privacylab.at

Bibliography

[1] Bartolini, C. and Muthuri, R. (2015). Reconciling data protection rights and obligations: An ontology of the forthcoming eu regulation. In *Workshop on Language and Semantic Technology for Legal Domain (LST4LD)*, Hissar, Bulgaria.

[2] Burnett, D. C., Sporny, M., Longley, D., and Kellogg, G. (2017). Verifiable claims data model and representations.

[3] Carothers, G., Seaborne, A., Bizer, C., and Cyganiak, R. (2014). RDF 1.1 TriG: RDF Dataset Language.

[4] Fernandez, J. D., Kirrane, S., Polleres, A., and Steyskal, S. (2017). HDTcrypt: Efficient compression and encryption of RDF datasets. *Semantic Web Journal*. under submission, available at `http://www.semantic-web-journal.net/content/hdt-crypt-efficient-compression-and-encryption-rdf-datasets`.

[5] Giménez-García, J. M., Zimmermann, A., and Maret, P. (2017). Ndfluents: An ontology for annotated statements with inference preservation. In *14th European Semantic W Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*, pages 638–654.

[6] Iannella, R., Steidl, M., McRoberts, M., Myles, S., Birmingham, J., and RodrÃŋguez-Doncel, V. (2017). Odrl vocabulary & expression.

[7] Iannella, R. and Villata, S. (2017). Odrl information model.

[8] Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). Prov-o: The prov ontology.

[9] Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., and Tilmes, C. (2013). Prov-dm: The prov data model.

[10] Neumaier, S., Umbrich, J., and Polleres, A. (2016). Challenges of mapping current CKAN metadata to DCAT. In *W3C Workshop on Data and Services Integration*, Amsterdam, the Netherlands.

[11] Piero Bonatti, Sabrina Kirrane, R. W. (2017). Deliverable D1.3: Policy, transparency and compliance guidelines v1.

[12] Schlehahn, E., Rohou, P., and Bos, B. (2017). Deliverable D6.2: Public relations strategy.

Appendix A

Draft Call for 1st SPECIAL standardisation workshop

H2020-ICT-2016-2017
Project No. 731601

Linked Data Vocabularies for Transparency and Privacy controls

A W3C Workshop on Linked Data and Privacy

24–25 January 2018, WU Wien, Vienna, Austria, Europe

Important Dates
(tentative)

08 December 2017:
Deadline for submission of expressions of interest and position statements
22 December 2017:
Acceptance notification and registration instructions sent
12 January 2018:
Program announced
19 January 2018:
Deadline for registration
24-25 January 2018, 8am:
Workshop

Expressions of interest and position statements

Fill out the **expression of interest form** or submit a position statement

Host

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS W3C gratefully acknowledges WU Vienna for hosting this workshop.

Sponsors

The Workshop is supported by the European Commission's H2020 Programme's Special project

Become a sponsor!

Sponsorship package info

With the new GDPR coming into force in May 2018, more data controls in data exchanges are needed. One solution is to use linked data. While there are frameworks for provenance and even rights management, the semantics of many other aspects of privacy are not formalized yet in an interoperable, interlinkable manner amenable, in the spirit of linked data.

Want to attend? Have something insightful to share?

We plan to hold the workshop on two sites in Europe (Vienna) at Vienna University of Economics and Business (WU Wien) and in the US (**tentative: Stanford**) in parallel with video connection . We will have a limited number of possible attendees at the workshop. We want to fill the room with deep knowledge about privacy and Linked Data on the Web, in order to determine what necessary vocabularies need to be agreed upon and put in place to protect privacy and enable protected interchange of transparency records between data processors, controllers and data subjects about personal data, handling in compliance with data protection regulations and personal policies. We won't just be listening to presentations, but will be actively participating in topic breakouts and working discussions. Our goal is to arrive at interchangeable and open solution proposals to manage and protect privacy online in a manner auditable, controllable and transparent to the data subject. The first step in this direction is the definition of agreed upon vocabularies to represent personal data, processing of personal data and related policies and regulations.

If you want to participate, please fill out the **expression of interest form** or submit a position statement.

Please note, **expressions of interest and position statements are not presentation proposals**. This is a workshop, not a conference, and any presentations will be short, with topics suggested by expressions of interest and decided by the chairs and program committee. Our goal is to actively discuss topics, not to watch presentations.

Attendees are encouraged to read all accepted expressions of interest prior to the workshop, to facilitate informed discussion.

Attendance is free for all invited participants, and open to the public (space allowed), whether or not W3C members.

Unfortunately, the workshop budget does not allow us to provide travel or lodging expenses to attendees.

Workshop topics

Possible topics include, but are not limited to the following:

- Core Vocabularies for privacy and transparency in linked data
  - Identity management vocabularies
  - Categories of sensitive and personal data
  - Modeling personal data usage, processing, sharing, and tracking
  - Modeling/Interlinking aspects of privacy and provenance
  - Modeling consent
- Vocabularies to model privacy policies, regulations, and involved (business) processes

Modeling permissions, obligations, and their scope in dealing with personal data (particularly, profiles and extensions of ORDL and POE)
- Reasoning about formally declared privacy policies, in order to detect policy violations, breach and enforce policies
- Finally, how to **link** and liaise with related efforts such as W3C's Social Web WG, Verifiable Claims WG, ODRL/POE WG, Credentials CG, and PROV WG or other (non-W3C) efforts (e.g. OASIS XDI, OASIS COEL, Kantarainitiative's CISWG) under one hood through joint/linked vocabularies and vocabulary profiles?

Out of Scope

- Solutions for pure access control for linked data are not in scope, as we ar particularly interested in aspects of privacy and personal data protection policies.

Position statements

An author of a position statement accepted is not required to attend (you can fill out the expression of interest form instead), but it does help set the topic discussions and to establish a particular point of view. If you wish, you can send us a position statement at <team-privacyws-submit@w3.org> (mailinglist needs to be established still), by 08 December 2017 (tentative). Our program committee will review the expressions of interest, and select the most relevant topics and perspectives.

A good position statement should be a few paragraphs (between 500 and 1000 words) and should include:

- Your background in Privacy, Linked Data and Web technologies
- What topic you would like to lead discussion on, including concrete ideas on how this topic relates directly to the Web or the Web of data
- Links to related supporting resources, activities and working groups
- Any other topics you think the workshop should cover in order to be effective
- A focus on technical issues, not process or platform preference. We plan to talk about the **what**, not the **how**.

Position statements must be in English, and HTML or plain-text format; images should be included inline in HTML using base64-encoded data URIs. You may include multiple topics, but we ask that each person submit only a single coherent position statement. All suitable submitted expressions of interest will be published and linked to from this workshop page.

Who Should Attend

Attendance is open to all, and our aim is to get a diversity of attendees from a variety of industries and communities, including:

- People dealing with Linked Big Data
- Privacy advocates
- Data protection authorities
- Security and privacy researchers
- Developers of Privacy Enhancing Technologies

Standardization Counter-arguments

There are a lot of voices and conflicting opinions in the privacy communities. Are you skeptical that standardization should be discussed at all? Are the same technologies that are critizized for enabling DRM actually useful/usable to protect and enforce privacy? We also welcome expressions of interest on issues that pose challenges to standardization, helpful to frame workshop topics and serve as a reality check. Please label these submissions **"Standards Con"** to distinguish them.

Event Archive Policy: Video and Transcripts

For posterity and for those unable to attend this workshop, we may be recording video and/or audio of the event, and will provide live notes (minuted in IRC) of the presentations and group discussion. Participants will be asked to sign a media waiver.

Goals

The goal of the workshop is exploratory. One of the primary outcomes is to bring different voices and perspectives together.

While we hope to identify opportunities and possible timelines for standardization, *we do not anticipate that W3C will form a Working Group as a direct result of this workshop*. Instead, if we do identify areas that need Web standardization, our aim would be to incubate and refine these ideas, to make sure that the right steps are taken at the right time for the key stakeholders involved.

What is W3C?

W3C is a voluntary standards consortium that convenes companies and communites to help structure productive discussions around existing and emerging technologies, and offers a Royalty-Free patent framework for Web Recommendations. We focus primarily on client-side (browser) technologies, and also have a mature history of vocabulary (or "ontology") development. W3C develops work based on the priorities of our members and our community.

Logistics

W3C's Workshop on is located at Vienna Unniversity of Economics and Business (WU Wien) near Prater in Vienna, Austria.

Organizations interested in becoming sponsors are encouraged to contact the organizers.

Venue

Vienna Unniversity of Economics and Business (WU Wien)

Welthandelsplatz 1

2. Bezirk

1020k, Vienna

Austria

Wednesday–Thursday, 24–25 January 2018

8:00–17:00

Hotels nearby include: To be added

Social Media and Remote Participation

Tweets and other social messages are encouraged to use the hashtag #specialprivacy. Please be respectful and accurate when quoting others.

We may have a live video stream… details will follow.

Program

Program Committee

Chairs

- Sabrina Kirrane, WU Wien
- Rigo Wenning, W3C/ERCIM

Committee

The committee (to be confirmed) is listed on a separate page.

Participants, Position Statements, and Expressions of Interest

You can read all the current expressions of interest (alternate view).

Schedule

The workshop will focus around several topics identified by the expressions of interest. Each topic will be introduced by one or more related lightning talks, and will be explored more in-depth by discussion breakouts, conlcuded with joint summaries of the breakouts. The goal of the discussion is not be to resolve the technical issues of the topic, but to determine its relevance and priority to standardization.

This schedule may change based on discussions with the program commitee.

Day 1: 24 January

08:00–08:30	Registration
08:30–08:45	Opening remarks by the chairs
08:45–09:00	Keynote:
09:00–09:45	Introductions of participants, annd general Lightning talks
09:45–10:00	Break
10:00–10:30	**Keynote:**
10:30–11:00	Lightning Talks on **Identity**, including reputation, personal data, KYC (4–5 talks)
11:00–12:15	Exercise: **Identity** (3 phases: breakout, report out, group discussion)
12:15–13:30	**Lunch**, Birds of a Feather Topic tables
13:30–14:00	Lightning Talks on **Provenance**, including consent, licensing of IP, assets, and services (4–5 talks)
14:00–15:00	Exercise: **Provenance** (3 phases: breakout, report out, group discussion)
15:00–15:30	Break (includes self-organizing evening plans)
15:30–16:00	Lightning Talks and open group discussion
17:00–	Closing statements

17:15	
18:30–24:00	Self-organized evening plans

Day 2: 25 January

08:00–08:30	Registration
09:00–09:45	Lightning Talks on **Data Protection** and **The Kitchen Sink**, including all other topics
09:45–10:00	Break
10:00–12:00	bla
12:00–12:15	Exercise: Passions & Commitments
12:15–13:30	**Lunch**, Discussions of Passions & Commitments
13:30–13:45	Personal Exercise: “Propose what would you like to see W3C or this community work on together”
13:45–14:00	Exercise: Dot Voting on proposals
14:00–14:45	Facilitated Discussion: Emerging Priorities for W3C & Community
14:45–15:00	Break
15:00–15:30	Close: Recap by Axel Polleres Actions Items, & Commitments
15:30–17:30	Demos and Open Discussion

Speakers

Speaker Bio

Host

Workshop Host(s)

WU Wien

Sponsors

Please help us finding sponsors!

Becoming a Sponsor

For details on the available sponsorship opportunities for this workshop, see our Sponsorship Packages.

Becoming a Sponsor

W3C Workshops, meetups, and other events bring you into direct contact with leading Web technology experts: representatives from industry, research, government, and the developer community.

Whether your interests are focused on a particular topic being discussed by a Working Group, or you wish to reach a diverse international audience setting W3C's strategic direction, sponsorship helps your organization reach W3C's engaged participants.

Sponsorships offset a portion of our meeting costs, so W3C welcomes multiple sponsors for each event. All proposals for sponsorship are subject to W3C approval.

If you're interested in being a sponsor of the W3C Linked Data and Privacy Workshop, please contact Bernard Gidon, EMEA Business Development, at <bgidon@w3.org> or +33.

For additional information, please visit the W3C sponsorship program.

Questions? Contact Rigo Wenning <rigo@w3.org>.

Appendix B

Letter sent to Stakeholders

From: **Axel Polleres** axel.polleres@wu.ac.at
Subject: Invitation to support our planned W3C workshop on Linked data Vocabularies for Transparency and Privacy controls
Date: 28 September 2017 at 22:28
To: XYZ
Cc: Sabrina Kirrane sabrina.kirrane@wu.ac.at, Rigo Wenning rigo@w3.org

Dear XYZ,

Due to the increasing importance of interoparability of systems enabling privacy and transparency about personal data handling, we are planning a W3C workshop around potentional starting points for standardization in terms of privcay interoperability vocabularies and protocols.

Among other international movements, on a European level the new EU General Data Protection Regulation (GDPR) will come into force on the 25th of May next year. With the regulation comes the need for additional controls, concerning data processing and exchange, as well as standards and best practices around privacy online.

In order to address these topics, we propose to organise a workshop that aims to bring together relevant stakeholders, under the umbrella of W3C.

The workshop which will be organised by the SPECIAL EU-funded Horizon2020 project [1], will ideally be held early next year (tentative date so far 24-25 of January 2018).

The idea is to organise a joint "parallel" workshop between Europe (physically located at Vienna University of Economics and Business in Vienna [2] and the US (location to be determined) connected by video conference, on potential standardisation around privacy, consent, transparency and non-tracking.

Please find a draft call attached. For announcement, dissemination and to foster discussion between the various stakeholders, we seek the assistance of the W3C, with the desired outcome of the foundation of a respective community and/or working group.

We kindly ask you for your support for this important workshop, and (if applicable) in your role as employee of a W3C member organisation to get timely approval for the workshop by the W3C, due to the growing importance of the topic.

In particular, we would kindly ask you to reply to us in the following points:

[ ] I support the planned workshop as member of the W3C (that is, the AC Representative of my organisation will officially support the workshop).

[ ] I can promote the workshop through another venue, apart from W3C, please provide details:

[ ] I plan to actively participate and submit a contribution to the workshop.

Thanks a lot and we look forward to working with you towards a dynamic and productive workshop,

Axel Polleres, Vienna University of Economics and Business (WU Wien)
Sabrina Kirrane, Vienna University of Economics and Business (WU Wien)
Rigo Wenning, W3C/ERCIM

1. SPECIAL (Scalable Policy-awarE linked data arChitecture for prIvacy, trAnsparency and compLiance), cf. https://www.specialprivacy.eu/
2. http://www.wu.ac.at/

--
Prof. Dr. Axel Polleres
Institute for Information Business, WU Vienna
url: http://www.polleres.net/ twitter: @AxelPolleres

W3C Workshop
on Link...acy.pdf