HORIZON 2020 - ICT-14-2016-1

AEGIS

Advanced Big Data Value Chains for Public Safety and Personal Security

WP2 – Core Data Value Chain Transformation and Handling Methods

D2.1 – Semantic Representations and Data Policy and Business Mediator Conventions

Due date: 31.08.2017	**Delivery Date**: 04.09.2017

Author(s): Maurizio Megliola, Elisa Rossi, Cinzia Rubattino (GFT), Evmorfia Biliri, Michael Petychakis, Spiros Mouzakitis, Christos Botsikas, Giannis Tsapelas (NTUA), Yury Glikman, Andreas Schramm (Fraunhofer), Gianluigi Viscusi (EPFL), Spyridon Kousouris, Fenareti Lampathaki, Sotiris Koussouris (SUITE5)

Editor: Elisa Rossi (GFT)

Lead Beneficiary of Deliverable: GFT

Dissemination level: Public	**Nature of the Deliverable**: Report

Internal Reviewers: Christopher Tucci (EPFL), Alexander Stocker (VIF), Spiros Mouzakitis (NTUA)

EXPLANATIONS FOR FRONTPAGE

Author(s): Name(s) of the person(s) having generated the Foreground respectively having written the content of the report/document. In case the report is a summary of Foreground generated by other individuals, the latter have to be indicated by name and partner whose employees he/she is. List them alphabetically.

Editor: Only one. As formal editorial name only one main author as responsible quality manager in case of written reports: Name the person and the name of the partner whose employee the Editor is. For the avoidance of doubt, editing only does not qualify for generating Foreground; however, an individual may be an Author – if he has generated the Foreground - as well as an Editor – if he also edits the report on its own Foreground.

Lead Beneficiary of Deliverable: Only one. Identifies name of the partner that is responsible for the Deliverable according to the AEGIS DOW. The lead beneficiary partner should be listed on the frontpage as Authors and Partner. If not, that would require an explanation.

Internal Reviewers: These should be a minimum of two persons. They should not belong to the authors. They should be any employees of the remaining partners of the consortium, not directly involved in that deliverable, but should be competent in reviewing the content of the deliverable. Typically this review includes: Identifying typos, Identifying syntax & other grammatical errors, Altering content, Adding or deleting content.

AEGIS KEY FACTS

Topic:	ICT-14-2016 - Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation
Type of Action:	Innovation Action
Project start:	1 January 2017
Duration:	30 months from 01.01.2017 to 30.06.2019 (Article 3 GA)
Project Coordinator:	Fraunhofer
Consortium:	10 organizations from 8 EU member states

AEGIS PARTNERS

Fraunhofer	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.
GFT	GFT Italia SRL
KTH	Kungliga Tekniska högskolan
UBITECH	UBITECH Limited
VIF	Kompetenzzentrum - Das virtuelle Fahrzeug, Forschungsgesellschaft-GmbH
NTUA	National Technical University of Athens – NTUA
EPFL	École polytechnique fédérale de Lausanne
SUITE5	SUITE5 Limited
HYPERTECH	HYPERTECH (CHAIPERTEK) ANONYMOS VIOMICHANIKI EMPORIKI ETAIREIA PLIROFORIKIS KAI NEON TECHNOLOGION
HDIA	HDI Assicurazioni S.P.A

Disclaimer: AEGIS is a project co-funded by the European Commission under the Horizon 2020 Programme (H2020-ICT-2016) under Grant Agreement No. 732189 and is contributing to the BDV-PPP of the European Commission.

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Communities. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

EXECUTIVE SUMMARY

The scope of D2.1 is to document the preliminary efforts undertaken within the context of Tasks 2.1 and 2.2.

Towards this goal, one of the scopes of the current deliverable is to define both the design and the infrastructure for accessing and sharing linked big data vocabularies and metadata. In order to create an AEGIS data network, suitable for a wide range of stakeholders, that will include data in different formats and from different data sources, an overview of the state of the art about big data vocabularies and metadata repository is needed to point out the methods and technologies for big data and linked data management. Therefore, the requirements related to our three demonstrators of the semantic vocabularies and metadata repository provided by AEGIS will be presented, in order to choose the most suitable ones, in agreement with the user's functional requirements identified in WP3 (D3.1). In addition, we will describe each semantic vocabularies and ontologies that are going to be utilised in AEGIS, providing the semantic representations and linked data vocabularies necessary for describing the data. Their functionalities and implementation will be investigated in detail along with the integration of the LinDA Vocabularies and Metadata Repository, which was developed in the context of the EU Project LinDA.

Another aim of this deliverable is to provide the main methods and data schemas for the Data Policy and Business Mediator Frameworks. The main role of such frameworks is to provide useful tools for the stakeholders to sell and purchase data, datasets and services; these frameworks will cooperate to guarantee a secure and trusted exchange. The Data Policy Framework (DPF) will provide a set of rules and warranties about each specific dataset or service to the Business Brokerage Framework (BBF). On its side the BBF will be the tool able to create smart contracts, i.e. the blockchain technology will allow the crossing of the requests of purchasing/selling adding the rules/warranties from DPR. In D2.1, we will define a first version of the design of the core methods that will be used to power both the (DPF), as well as the (BBF). We will describe data IPR, security, trust, and quality features, highlighting their involvement in the DPF. Moreover, we will define the BBF, capitalising on the work performed in the Data Policy Framework and more specifically on the IPR annotations to be selected. Both platforms' Frameworks design and engineering are described providing detailed schemas of their related concepts.

The current deliverable comprises the preliminary list of the semantic vocabularies and metadata repository of the AEGIS platform, as well as the preliminary definition of the Data Policy and Business Mediator Frameworks. The forthcoming version (D2.3 – Update on Semantic Representation and Data Handling and Analytics Methods) of this deliverable will include the final versions of the vocabularies, metadata repository and of the frameworks, based on the feedback received by the project's demonstrators/end users.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CO	Confidential, only for members of the Consortium (including the Commission Services)
D	Deliverable
DoW	Description of Work
H2020	Horizon 2020 Programme
FLOSS	Free/Libre Open Source Software
GUI	Graphical User Interface
IPR	Intellectual Property Rights
MGT	Management
MS	Milestone
OS	Open Source
OSS	Open Source Software
O	Other
P	Prototype
PU	Public
PM	Person Month
R	Report
RTD	Research and Development
WP	Work Package
Y1	Year 1

1. INTRODUCTION

The scope of the current section is to introduce the deliverable and familiarize the user with its contents. Towards this end, this section summarises the objective of the current deliverable, its relation to the other work packages and corresponding deliverables, and analyses its structure.

1.1 OBJECTIVE OF THE DELIVERABLE

The scope of D2.1 is to document the preliminary efforts undertaken within the context of Tasks 2.1 and 2.2. Towards this end, the scope of the current deliverable is to provide the semantic representations and linked data vocabularies necessary for describing the data, as well as the main methods and data schemas for the Data Policy and Business Mediator Frameworks. Both of the aforementioned tasks will be developed until M18, hence these first considerations and decisions will constitute the basis of the future work.

The second and the third chapters of the deliverable investigate the needs and the expectations towards the AEGIS Semantic Vocabularies and Metadata Repository. Their requirements were pointed out and grouped in a set of high-level categories related both to our pilots and to possible AEGIS stakeholders' needs. Moreover, a detailed list of semantic vocabularies that covers the AEGIS requirements is reported, followed by a description of the metadata of the vocabularies and ontologies that are going to be utilised in AEGIS. The last paragraph of the Chapter 3 is dedicated to the identification and description of the functionalities of the AEGIS vocabularies and metadata repository, with a focus on the basic features that the metadata repository has to have in order to satisfy the needs of our demonstrator.

Chapter 4 instead will focus on the design of the Data Policy Framework as well as the Business Brokerage Framework, going from the state of the art to our proposal for the aforementioned AEGIS Frameworks.

1.2 INSIGHTS FROM OTHER TASKS AND DELIVERABLES

Work package 2 receives as input mainly the early reports of WP1 and WP3. Towards this end, the used datasets or measured quantities per pilot and related user stories descripted in Table 3-1 were drawn mainly from deliverable D3.1. In D3.1, we have also defined the design of the Business Brokerage Framework (BBF), giving an overview of our expectations and ideas about the BBF, defining the AEGIS Brokerage Engine as a component that will instantiate part of the methods that will be included in the AEGIS Data Policy and Business Brokerage Frameworks, which will be delivered under WP2. From D1.2, instead, comes the definition of the regulatory framework for data protection, IPR and Ethical Issues that will drive the Data Policy framework of the AEGIS platform in Chapter 4.

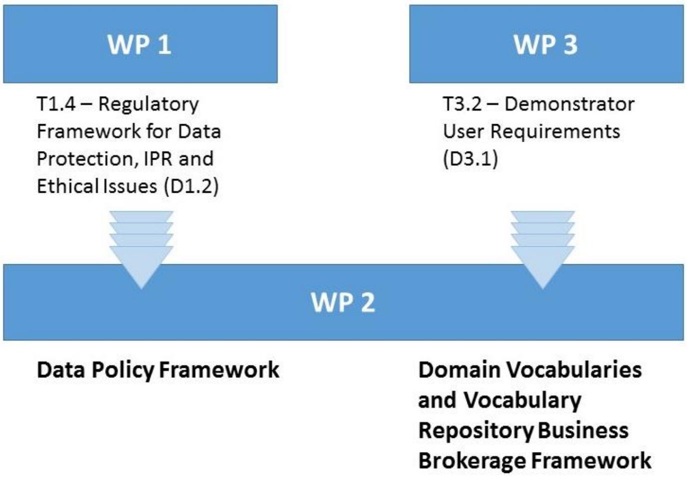

WP 1
T1.4 – Regulatory Framework for Data Protection, IPR and Ethical Issues (D1.2)
WP 3
T3.2 – Demonstrator User Requirements (D3.1)
WP 2
Data Policy Framework
Domain Vocabularies and Vocabulary Repository Business Brokerage Framework

Figure 1-1: Inputs from other Work Packages/Deliverables

The final version of the vocabularies, metadata repository, harmonisation, knowledge extraction and business intelligence will be delivered in D2.3, M18.

1.3 STRUCTURE

Deliverable 2.1 is organized in five main sections as indicated in the table of contents.

- The first section introduces the deliverable. It documents the scope of the deliverable and briefly describes how the document is structured. It also documents the relation of the current deliverable with the other deliverables, and how the knowledge produced in the other deliverables and work-packages served as input to the current deliverable.
- Following the introduction, section 2 describes the types of metadata that AEGIS has to provide, defining the standard that AEGIS has to implement, the structure and semantics of data for visualisation and analysis and the syntactical aspects of tabular data.
- In Section 3 we present the relevant for AEGIS domain ontologies and how we plan to manage them, starting from their requirements, ending with their definition and list.
- Section 4 is about the Data Policy and Business Brokerage Frameworks; it is split into two main paragraphs, one for each topic and both of them are organized following the same structure: they begin with an overview of the existing technologies/applications providing some examples of them, and finally we propose our Frameworks, describing their design and how they will be integrated in the whole AEGIS platform.
- Section 5 concludes the deliverable. It outlines the main findings of the deliverable, which will guide the future research and technological efforts of the consortium.

Deliverable D2.1 includes also an Annex, Annex 1, which shows a figure with the UML class diagram of the DCAT application profile. The figure refers to paragraph 2.1.

2. AEGIS METADATA

All data registered and/or stored in the AEGIS platform has to be properly described in metadata to enable users and tools to find, understand, and (re-)use it. We distinguish three levels of metadata:

1) "**Contextual metadata**". This is a traditional description of data as we know it from traditional Open Data portal, like the European Data Portal[1]. It includes the name of the dataset, description, information about the data publisher, publication date, original dataset publication URL, information about included in the dataset files, information about the license, dataset domain, etc.
2) "**Structural and semantic metadata**", which is essentially the types and the semantics of the data columns.
3) "**Syntactic metadata**", such as the column separation character in CSV files.

The next subsections present the levels of metadata in more detail.

2.1 CONTEXTUAL AEGIS METADATA

The contextual AEGIS metadata providing general information about data in the AEGIS platform will conform to the DCAT Application profile (DCAT-AP) version 1.1 specification[2]. DCAT-AP is a relatively new standard developed as a join initiative of:

- the Directorate-General for Communications Networks, Content & Technology: DG CONNECT;
- the Directorate-General for Informatics: DG DIGIT;
- and the Publications Office of the EU.

DCAT-AP is defined to be the standard for describing public sector datasets in Europe to enable interoperability between European public sector data portals. It is already implemented in the European Data Portal and several national European Open Data portals. It can be expected that this standard will be adopted by the most European public sector data portals in the next years. The standard is not specific for public datasets only and there are good chances that it will be adopted by the European industry as well.

AEGIS metadata will be conform to DCAT-AP specification to assure interoperability with other European Open Data portals, but will introduce an extension to address the specific of the AEGIS project, which is presented in the next section.

The DCAT-AP specification reuses classes and properties from the following namespaces defined in other specifications:

- adms: http://www.w3.org/ns/adms#
- dcat: http://www.w3.org/ns/dcat#
- dct: http://purl.org/dc/terms/
- foaf: http://xmlns.com/foaf/0.1/

[1] https://www.europeandataportal.eu/

[2] https://joinup.ec.europa.eu/asset/dcat_application_profile/description

- owl: http://www.w3.org/2002/07/owl#
- rdfs: http://www.w3.org/2000/01/rdf-schema#
- schema: http://schema.org/
- skos: http://www.w3.org/2004/02/skos/core
- spdx: http://spdx.org/rdf/terms#
- xsd: http://www.w3.org/2001/XMLSchema#
- vcard: http://www.w3.org/2006/vcard/ns#

Annex 1 presents the classes and properties of the DCAT-AP in a UML diagram.

The specification defines a number of obligatory, recommended and optional classes. The most important of them for AEGIS are:

- Catalogue - a catalogue or repository that hosts the datasets being described. In AEGIS it is the AEGIS metadata catalogue
- Dataset - a conceptual entity that represents the information published.
- Distribution - a physical embodiment of the Dataset in a particular format. In AEGIS distributions will denote data files or APIs of the registered in AEGIS datasets.

For each of the classes the DCAT-AP specification defines a rich set of attributes describing them in detail. Their description as well as description of other classes are omitted in this document and should be taken from the original DCAT-AP specification.

2.2 STRUCTURAL AND SEMANTIC METADATA

Apart from general information about data, which is important for finding the right data in the registry, it is important to understand the structure and semantics of data for its visualisation and analysis.

One of the design goals is that this part of the AEGIS vocabulary should be understandable in minutes, not days, and should be described in a single-digit number of pages or slides. With that in mind, the following decisions were taken:

- Some concepts from Frictionless Data[3] were taken and reformulated as an RDF vocabulary, slightly adapted in order to fit the requirements of AEGIS. Frictionless Data is a relatively spartanic formalism for metadata of datasets in terms of JSON data structures.
- Second, in order to understand the classes and properties introduced here, the reader should not be forced to follow references to complex vocabularies in the outside world. All RDF classes and properties here are specified from the scratch on the top of rdf(s) in a self-contained manner, and only then relations to existing concepts, if any, are established.

All RDF classes and properties defined here should also be equipped with appropriate rdfs:description properties etc.; this is not shown here.

[3] http://frictionlessdata.io

To begin with, we define two classes with the obvious subclass relation between them:

```
aegis:DataSet rdf:type rdfs:Class.

aegis:TabularDataSet rdf:type rdfs:Class.

aegis:TabularDataSet rdfs:isSubclassOf aegis:DataSet.
```

An example tabular dataset in AEGIS would then be typed as follows:

```
<my-example-dataset> rdf:type aegis:TabularDataSet.
```

A table consists of a number of rows and columns, under the homogeneity assumption that all entries of the same column possess the same type in some sense; hence it is customary, albeit imprecise, to speak of the "type of the column".

In addition, one or more (or even all) columns constitute the "primary key" of the table. This is a concept taken from the database parlance; it states that there are no two rows with coinciding values in all of the primary key columns. This is some kind of uniqueness information. When a table is to be interpreted as a value table of a function in the mathematical sense, this information describes "from where to where" the function goes. The key columns constitute the domain of the function, the remaining columns, called value columns, constitute its range. This knowledge is helpful, if not indispensable, in the interpretation and combination of tabular data in order to get new insights.

In order to reflect this, we define two properties to be used to specify table columns. We let table columns have the type "aegis:Column":

Name	**rdfs:domain**	**rdfs:range**
aegis:hasKeyColumn	aegis:TabularDataSet	aegis:Column
aegis:hasValueColumn	aegis:TabularDataSet	aegis:Column

Now, the columns themselves are described the following properties:

Name	**rdfs:domain**	**rdfs:range**
aegis:columnNumber	aegis:Column	xsd:int
aegis:columnHeader	aegis:Column	xsd:string
aegis:columnType	aegis:Column	rdfs:Class
aegis:measuredOrCountedUnit	aegis:Column	rdfs:Resource or xsd:string

Column number and column header are to be taken from the source table.

The type of a column's entries may be any "foreign" RDF class, e.g., a class from some **domain ontology** (this is known in Frictionless Data as *Rich Types*). It may also be some of the following RDF classes, which we introduce as members of the AEGIS vocabulary:

Class	**Meaning: Entries of described column contain …**
aegis:MeasureOrCount	numbers that count or measure something (as specified *via aegis:measuredOrCountedUnit*)
aegis:TimePoint	points in absolute time including date; any precision
aegis:GeoName	proper names of geographic entities
aegis:GeoCoordinates	longitude/latitude coordinates
aegis:FreeText	text in any natural language
aegis:Image	image in binary format
aegis:Video	video in binary format
aegis:Audio	audio in binary format
aegis:DatabaseKey	number or other literal, meaningful only as surrogate key of a determined table

In Section 3 we present the relevant for AEGIS domain ontologies and how we plan to manage them.

If a column type is *aegis:MeasureOrCount*, then the aegis:measuredOrCountedUnit property must also be specified for that column; it states what physical unit is measured or what kind of things are counted, e.g., by pointing into some domain ontology.

Conceptually, these classes partially overlap with the field descriptor types in Frictionless Data.

We emphasize that the part of the vocabulary described in this section refers to the *logical* structure. This means that columns whose column type is text, image, audio, or video, may well hold just *pointers* to data of the respective type, not the (voluminous) data itself.

2.3 SYNTACTIC METADATA

Syntactic metadata describes the syntactical aspects of tabular data, which is the main type of structured data in AEGIS. In AEGIS, the structured data when possible will be converted and stored in tabular format. For describing syntactical aspects, a couple of properties are defined, given in the table below (they have been adapted from Frictionless Data’s CSV Dialect Description Format). These properties are to be used at the table level; they hold for all columns simultaneously.

Property	**rdfs:domain**	**rdfs:range**	**Meaning**	**Default value**
aegis:columnDelimiter	aegis:TabularDataSet	xsd:string	column delimiter character	“,”
aegis:lineTerminator	aegis:TabularDataSet	xsd:string	line termination character	“\n”
aegis:quoteChar	aegis:TabularDataSet	xsd:string	quoting character to surround entries	“\””
aegis:doubleQuote	aegis:TabularDataSet	xsd:bool	whether two consecutive quoting characters count as a single one within an entry	true

aegis:escapeChar	aegis:TabularDataSet	xsd:string	character used to express quoting characters in strings	*‹none›*
aegis:skipInitialSpace	aegis:TabularDataSet	xsd:bool	whether white space directly after a column delimiter is ignored	true

The escape character, if specified, must be distinct from the quote character. Note that tabular data, after their ingression in an AEGIS platform, have been stripped of their headers (they have become part of the semantic metadata).

Example: The syntactical properties of a dataset of tabular data might be described as follows:

```
<my-example-dataset>
      aegis:columndelimiter “\t”;
      aegis:lineTerminator “\n”;
      aegis:escapeChar “\\”.
```

3. AEGIS DOMAIN VOCABULARIES AND VOCABULARY REPOSITORY

3.1 RELATED USER STORIES

The first step toward the creation of the AEGIS Domain Vocabularies and a repository for them is to perform an initial identification of the requirements, in terms of semantic vocabularies and metadata. Since the Public Safety and Personal Security domain is fairly wide and includes numerous diverse concepts and actors, it is consequent that the semantic vocabularies and metadata requirements will also be quite diverse, taking into account the different datasets that will be provided or used by the AEGIS pilots and therefore, covering a large range of different domains. The process of the identification of these requirements combined the recognised related datasets that are described in the collected user stories reported in D3.1, with a set of available data sources that were provided by the pilots themselves and other, external datasets, which are related to their specific activities or can be utilised in the analytical processes of the platform's services.

The following table presents a set of indicative datasets or measured quantities that are related to each AEGIS pilot, along with the main related user stories.

Table 3-4 - Used datasets or measured quantities per pilot and related user stories

Pilot	Indicative Datasets or Measured Quantities	Main Related User Stories
Automotive	• Trip route • Road conditions • Road names • Traffic level • Road Damage • Driving behaviour • Location and map points • Road accidents • Weather conditions • Social network traffic messages • Newspaper data	VIF2, VIF3, VIF5, VIF6, VIF7, VIF8, VIF9, VIF10, VIF14, VIF15, VIF16, VIF27, VIF28, VIF29
Smart Home	• Sensor data • Indoor environmental data • Occupancy • Air quality • HVAC (Heating, Ventilation and Air Conditioning) • Lighting • Energy consumption • Location • Individual health information • Weather conditions	HYP1, HYP2, HYP3, HYP4, HYP5, HYP6, HYP7, HYP8, HYP9, HYP10, HYP11, HYP12, HYP13, HYP14, HYP15, HYP16, HYP17, HYP18, HYP31

	• Public health information • Energy prices • CO2 emissions • Crime data • Accident data • PSPS related events (from social media and RSS channels) • Social media activity • Public safety data	
Insurance	• Recording sensor data • Location • Events • Social media data • Customer information • Customer habits • Open data • IoT sensor data	HDI12, HDI34, HDI35, HDI39, HDI49, HDI58, HDI59, HDI62

Based on the information presented in the previous table, the various types of data are grouped into a set of high-level categories for semantic vocabularies that are required, as follows:

- Health
- Insurance
- Sensor
- Traffic - Road Conditions
- Driving behaviour
- Car Accidents
- Weather
- Map - Location
- Crime
- Security - Safety
- Events
- Social Media
- News
- Automotive - Transportation

Finally, a semantic representation and a more detailed definition of the available datasets needs to be provided, in order to gain further insights into the datasets of the platform. Therefore, a set of semantic vocabularies that describe datasets or data sources will be provided along with the rest of the identified vocabularies.

3.2 SEMANTIC REPRESENTATIONS AND DOMAIN VOCABULARIES

The semantic representations in AEGIS refer to all the semantic vocabularies and metadata, which are going to be gathered and analysed by the AEGIS platform. These vocabularies and metadata will be able to describe datasets from the Public Safety and Personal Security domain, along with datasets of other domains that are going to be utilised, in order to enhance the cross-sectorial analytical capabilities of the platform.

Based on the initial recognition of the different categories of datasets that are related to the AEGIS project and its vision to exploit multi-disciplinary information for Public Safety and Personal Security services, a list of semantic vocabularies is presented in the following table. The listed vocabularies are selected in such way, so as to provide a rich set of options for the process of selecting the most suitable semantic annotation of the platform's data. Furthermore, they cover a wide range of possible dataset categories, as they are already described. For each of the semantic vocabularies, a link to the respective web resource is provided, along with a short description of its content or its purpose and the relevant category.

Table 3-5: Semantic vocabularies related to AEGIS

Name	Description	Relation to AEGIS PSPS categories
DICOM - Healthcare metadata - DICOM ontology https://www.netestate.de/dicom/dicom.owl	Ontology for healthcare metadata - especially metadata found in DICOM files (Digital Imaging and Communications in Medicine, see http://dicom.nema.org/).	Health
Translational Medicine Ontology – TMO https://code.google.com/archive/p/translationalmedicineontology/	The Translational Medicine Ontology (TMO) is a high-level, patient-centric ontology that extends existing domain ontologies to integrate data across aspects of drug discovery and clinical practice. The ontology has been developed by participants in the World Wide Web Consortium's Semantic Web for Health Care and Life Sciences Interest Group.	Health
The Disease Ontology http://disease-ontology.org/	The Disease Ontology has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts through collaborative efforts of researchers at Northwestern University, Center for Genetic Medicine and the University of Maryland School of	Health

	Medicine, Institute for Genome Sciences.	
Dem@Care Lab Ontology fhttp://www.demcare.eu/ontologies/demlab.html	Dem@Care Lab Ontology for Dementia Assessment (demlab). The ontology has been developed in the framework of the Dem@Care project for representing the experimentation protocol towards diagnostic support and assessment of Dementia in a controlled environment. The aim of the protocol is to provide a brief overview of their health status of the participants during consultation (cognition, behaviours and function), and to correlate the system (sensor) data with the data collected using typical dementia care assessment tools.	Health
Ontology for Biomedical Investigation (obo) http://purl.obolibrary.org/obo/obi.owl	The Ontology for Biomedical Investigations (OBI) is build in a collaborative, international effort and will serve as a resource for annotating biomedical investigations, including the study design, protocols and instrumentation used, the data generated and the types of analysis performed on the data. This ontology arose from the Functional Genomics Investigation Ontology (FuGO) and will contain both terms that are common to all biomedical investigations, including functional genomics investigations and those that are more domain specific.	Health
Smart Home Weather (shw) http://paul.staroch.name/thesis/SmartHomeWeather.owl#	An ontology defining weather-related concepts and properties being relevant to smart home systems that provide predictive control.	Weather
Home Weather (hw) https://www.auto.tuwien.ac.at/downloads/thinkhome	Smart home ontology for weather phenomena and exterior conditions	Weather

/ontology/WeatherOntology.owl		
Food Ontology (food) http://data.lirmm.fr/ontologies/food	This ontology models the Food domain. It allows to describe ingredients and food products. Ontology used by the Open Food Facts dataset.	Health (Nutrition)
LinkedGeoData ontology (lgdo) http://linkedgeodata.org/About	LinkedGeoData ontology (lgdo) has been derived from concepts defined by Open Street Map. It uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base.	Location - Maps
Geo ontology https://www.w3.org/2003/01/geo/	This vocabulary begins an exploration of the possibilities of representing mapping/location data in RDF and it is proposed directly by W3C.	Location - Maps
Dcterms http://dublincore.org/documents/dcmi-terms/	Another popular approach to describe RDF resources (also contains location information).	Various (high-level)
Schema.org http://schema.org	Search engines including Bing, Google, Yahoo! and Yandex rely on schema.org markup to improve the display of search results, making it easier for people to find the right web pages. Schema also contains a mechanism to describe maps in the web in a linked data way. Specific classes, such as Schema.org:Map and Schema.org:Event are very popular for describing resources on the web.	Event, Location - Maps
Bbccore http://www.bbc.co.uk/ontologies/coreconcepts	This is the core model for representing things such as people, news, places, events, organisations and themes in the BBC.	Event, News
OntoFuhSen Ontology https://github.com/LiDaKrA/Ontology	OntoFuhSen vocabulary is one of the key components in a Federated Hybrid Search Engine (FuhSen) and has a	Crime

	dedicated mechanism for the crime domain.	
Italian Crime Ontology https://www.researchgate.net/publication/228971566_A_domain_ontology_Italian_crime_ontology	The purpose of using such an ontology could make it possible to achieve a homogeneous conceptual structure in the various projects in the crime domain and to add domain knowledge to the support tools.	Crime
Dbpedia http://dbpedia.org/ontology/	This ontology is generated from the manually created specifications in the DBpedia Mappings Wiki. Each release of this ontology corresponds to a new release of the DBpedia data set which contains instance data extracted from the different language versions of Wikipedia.	All (high-level)
ISO 37120 indicator URIs (iso37120). ISO 37120 – Sustainable Development and Resilience of Communities (https://www.iso.org/standard/62436.html). http://ontology.eil.utoronto.ca/ISO37120.html	Indicators for City Services and Quality of Life ontology (under TC268). This ontology defines a class for each indicator defined in the ISO 37120 standard	Crime
Generic Sensor API https://www.w3.org/TR/generic-sensor/	This specification defines a framework for exposing sensor data to the Open Web Platform in a consistent way. It does so by defining a blueprint for writing specifications of concrete sensors along with an abstract Sensor interface that can be extended to accommodate different sensor types.	Sensor
aml http://data.ifs.tuwien.ac.at/aml/ontology	This is the AutomationML ontology that represents the automation effect of a machine that has been installed a sensor and hence controlling it by itself with no additional help during the period that it is still working. This is more or less an	Sensor

	example of a working ontology with sensors.	
Ssn http://w3c.github.io/sdw/ssn/	The Semantic Sensor Network (SSN) ontology is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators.	Sensor
dady https://www.w3.org/wiki/DatasetDynamics	This is a dataset dynamic ontology that allows a course grained information of the source of data dynamics. It may also be used in the discovery of the mechanisms that change the notifications. Linked Datasets change in the course of time: resource representations and links between resources are created, updated and removed; entire graphs can change or disappear. The frequency and dimension of such changes depends on the nature of a linked data source. Sensor data are likely to change more frequently than archival data. Updates on individual resources cause minor changes when compared to a complete reorganization of a data source's infrastructure such as a change of the domain name. Anyway, in many scenarios linked data consuming applications need to deal with these kinds of changes in order to keep their local data dependencies consistent. Dataset dynamics denotes a research activity that currently investigates how to deal with that problem.	Sensor
Dtype http://www.linkedmodel.org/schema/dtype	This is a DATATYPE ontology that provides the specifications of simple data types for example the enumerations. This is extremely useful and interesting regarding sensoric data.	Sensor

Home Activity (ha) http://sensormeasurement.appspot.com/ont/home/homeActivity#	An ontology to detect activity in a smart home.	Sensor
The Machine-to-Machine Measurement (M3) Lite Ontology (m3lite) http://ontology.fiesta-iot.eu/ontologyDocs/m3-lite.owl	M3 lite taxonomy is designed for the FIESTA-IOT H2020 EU project and is based on refactoring, cleaning and simplifying the M3 ontology designed by Eurecom (Amelie Gyrard). The M3-lite is a taxonomy that enables testbeds to semantically annotate the IoT data produced by heterogeneous devices and store them in a federated datastore such as FIESTA-IoT.	Sensor
Sensor, Observation, Sample, and Actuator (SOSA) Ontology (sosa) http://www.w3.org/ns/sosa/	This ontology is based on the SSN Ontology by the W3C Semantic Sensor Networks Incubator Group (SSN-XG), together with considerations from the W3C/OGC Spatial Data on the Web Working Group.	Sensor
Security Ontology (security) http://securitytoolbox.appspot.com/securityMain	A security ontology to annotate resources with security-related information.	Security - Safety
acl https://www.w3.org/ns/auth/acl	(Basic Access Control) this an ontology that defines the authorization mechanisms and the required properties of the access class of when a resource may be written or read4. Even though this may seem indirect to PSPS, this is highly important as with anything that has to do with access management to user resources, since those are the key to user privacy.	Security - Safety
Linked Datex II (datex)	This document gives URIs to all terms used within Datex II. the Datex standard was developed for information exchange between traffic management centres,	Transportation

http://vocab.datex.org/terms#	traffic information centres and service providers in Europe.	
Road Accident Ontology https://www.w3.org/2012/06/rao.html	A Road Accident ontology and accompanying resources/tools to describe traffic or road accidents, involving people, vehicle, animal, having cause, effects, etc. with RDF/OWL.	Transportation
The Transport Disruption ontology https://transportdisruption.github.io/transportdisruption.html#classes	The Transport Disruption ontology provides a formal framework for modelling travel and transport related events that have a disruptive impact on an agent's planned travel.	Transportation
Ontology of Transportation Networks http://opensensingcity.emse.fr/scans/entity/vocabulary_8	An ontology for modelling the most important aspects of traffic networks, transportation and locomotion. The OTN ontology is more or less a direct encoding of the GDF in OWL.	Transportation
Driving Context Ontology http://vi.uni-klu.ac.at/ontology/DrivingContext.owl	An ontology for specifying the driving environment for intelligent driver assistance systems.	Transportation
TTI Core v0.03: Core Ontologies for Safe Autonomous Driving http://www.toyota-ti.ac.jp/Lab/Denshi/COIN/Ontology/TTICore-0.03/	A set of Smart Vehicle Ontologies	Transportation
RAO – Road accident Ontology https://www.w3.org/2012/06/rao.html	Road Accident ontology and accompanying resources/tools to describe traffic or road accidents, involving people, vehicle, animal, having cause, effects, etc.	Road Conditions, Accidents, Automotive
SNaP Ontologies Simple News and Press Ontologies	The news ontology is comprised of several ontologies, which describe assets (text, images, video) and the events and entities (people, places,	News

http://data.press.net/ontology/	organisations, abstract concepts etc.) that appear in news content.	
BBC Storyline Ontology http://www.bbc.co.uk/ontologies/storyline	The News Storyline Ontology is a generic model for describing and organising the stories news organisations tell. The ontology is intended to be flexible to support any given news or media publisher's approach to handling news stories.	News
rNews http://dev.iptc.org/rNews	rNews is an approved standard for using semantic markup to annotate news-specific metadata in HTML documents.	News

The following table presents a list of more general vocabularies that can be used by the AEGIS platform, so as to describe individual datasets and their quality based on multiple attributes.

Table 6-3: Semantic vocabularies about datasets and data sources

Name	Description	Link
Data Catalog Vocabulary (DCAT)	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web.	http://www.w3.org/TR/vocab-dcat/
VOAF (Vocabulary Of A Friend)	VOAF is a vocabulary specification providing elements allowing the description of vocabularies (RDFS vocabularies or OWL ontologies) used in the Linked Data Cloud. In particular, it provides properties expressing the different ways such vocabularies can rely on, extend, specify, annotate or otherwise link to each other. It relies itself on Dublin Core and voiD.	http://purl.org/vocommons/voaf

VANN	A vocabulary for annotating vocabulary descriptions.	http://purl.org/vocab/vann/
Vocabulary of Interlinked Datasets (VoID)	The Vocabulary of Interlinked Datasets (VoID) is an RDF Schema vocabulary for expressing metadata about RDF datasets. It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloguing and archiving of datasets.	http://rdfs.org/ns/void#

3.3 REQUIREMENTS OF THE VOCABULARIES AND THEIR METADATA REPOSITORY

In this section, we describe the metadata of the vocabularies and ontologies that are going to be utilised in AEGIS. Each vocabulary/ontology stored in the AEGIS Vocabulary Repository should have the following fields, in order to ensure ease of use (easy to find and use), interlinking with existing datasets or new ones, and querying the overall system for extra information.

Table 3-4: Metadata requirements for the AEGIS Vocabulary Repository

Attribute	Description
Ontology/Vocabulary Type	The ontologies/vocabularies have been categorized per type as follows: Health, Sensor, News, etc.
General Title	A title for the ontology/vocabulary.
Subtitle	A subtitle for the ontology/vocabulary.
Uniform Resource Locator (URL)	The URL for the ontology/vocabulary.
Author	The initial author of the public version of the ontology/vocabulary.
Author e-mail	The author’s email.
Maintainer / Publisher	The organization responsible for publishing the ontology/vocabulary.
License	The License of the ontology/vocabulary (e.g. Open Government License UK, Creative Commons, MIT).

Category	One or more category themes that the ontology/vocabulary belongs to.
Description	A brief description of the ontology/vocabulary category.
Created	The creation date concerning the ontology/vocabulary.
Modified	The modification date concerning the ontology/vocabulary.
Feedback	Feedback mechanisms: Request Dataset forms, Rate Datasets, View popular demands / vote best requests, and Comment.
Language Interface	The language(s) the user interface is available in.
Language Data	The language(s) the datasets themselves are available in.
Data Format	The format of the available ontology/vocabulary (Excel/ PDF/ CSV just to name a few).
Metadata	If available, the metadata standard for the data catalog of the ontology/vocabulary.
Version	The latest version of the ontology/vocabulary.
numberOfProperties	Number of properties in the ontology / vocabulary.
numberOfClasses	Number of classes in the ontology / vocabulary.
Keywords	Related keywords

As we described already, the goal of the AEGIS Vocabularies and Metadata repository is to offer the highest quality of service to the pilots and its users in general. The most important functionalities it aims to cover are the following:

- Querying
- Searching
- Management
- Importing/Exporting
- Interlinking of vocabularies
- Quality assurance

In this section, we focus on satisfying those functionalities and we make a separation between functional and non-functional characteristics. Functional are the ones that allow the repository to operate and execute the necessary mechanisms while the non-functional are those focusing mainly on the quality assurance aspects which are of uppermost importance in AEGIS.

Table 3-5: Requirements

Requirement	Categorisation
Insert new vocabularies / ontologies in the Metadata Repository	Functional
Delete vocabularies / ontologies from the Metadata Repository	Functional
Update vocabularies / ontologies in the Metadata Repository	Functional
Insert metadata about vocabularies / ontologies in the Metadata Repository	Functional
Search vocabularies / ontologies in the Metadata Repository based on different criteria and keywords	Functional
Evaluate SPARQL queries over the Metadata Repository to collect metadata about vocabularies / ontologies	Functional
Evaluate SPARQL queries over the Metadata Repository to retrieve classes and properties of vocabularies / ontologies	Functional
Search and identify PSPS datasets semantically enriched with particular vocabularies / ontologies	Functional
Ensure persistence of the Metadata Repository	Non-Functional
Ensure web-based access and availability of the Metadata Repository	Non-Functional
Compute statistics about the Repository vocabularies / ontologies	Functional
Search pilots using particular vocabularies / ontologies	Functional
Download data dumps of the Repository vocabularies / ontologies	Functional
Provide a recommendation system on top of the Metadata Repository	Functional
Search for related vocabularies / ontologies in the Metadata Repository	Functional

3.4 VOCABULARY REPOSITORY IMPLEMENTATION

The AEGIS Vocabulary Repository will be built on top of the Linda Workbench infrastructure[5]. Linda is a generic vocabulary / ontology metadata repository that allows for registering, describing, and searching vocabularies. It also supports a variety of more advanced capabilities like transformation to RDF, analytics, visualizations, and more. Currently, Linda makes available the description of more than 300 vocabularies used to describe data in the Linked Open Data cloud, which break down to thousands of classes and properties. Moreover, Linda relies on a publication pipeline where ontologies are reviewed by curators who decide if an ontology satisfies the best design practices. Additionally, the complete Linda Workbench is open source by an MIT license[6] making it easy for anyone to create and tailor the platform for different requirements.

The AEGIS Vocabulary Repository will exploit the main features of LINDA. However, extensions have been conducted to the LINDA repository to enable the satisfaction of the AEGIS needs. First, vocabularies will be described in terms of main properties, e.g. classes, predicates, hierarchies or creator, but they will be also associated with the AEGIS pilots in which vocabularies have been utilised. Second, the connectivity between ontologies is defined not only based on links but also according to the associations between the datasets and pilots where these ontologies are utilised. Third, the AEGIS Vocabulary Repository implements a notification system that keeps track of the changes of the ontologies and automatically propagates these changes to the related ontologies. In this chapter, basic features of the AEGIS Vocabulary Repository are demonstrated.

The vocabulary repository serves the purpose of presenting the final user with various ontologies, supporting the transformation of traditional data formats to Linked Data by suggesting classes and properties. The usage of the repository can take place with actions that can be grouped in the following categories:

- Navigation: Actions that let the user search for vocabularies and entities inside them, read vocabulary descriptions, download the vocabulary RDF documents in various formats and get access to vocabulary visualizations and best usage practices.

- Usage feedback: Evaluation of vocabularies, discussions and commenting, that expose the advantages and disadvantages of choosing a vocabulary’s terminology to create transformation plans and guide the user base of an enterprise to vocabularies best representing its structure, operations and needs.

- Repository enrichment: Authenticated users may create and upload new vocabularies containing ontologies that do not exist to the initial repository or are specific to the enterprise. Vocabulary owners may further update their vocabularies at any times. The repository automatically extracts metadata information contained in the vocabulary RDF document like classes and properties, as well as their relations.

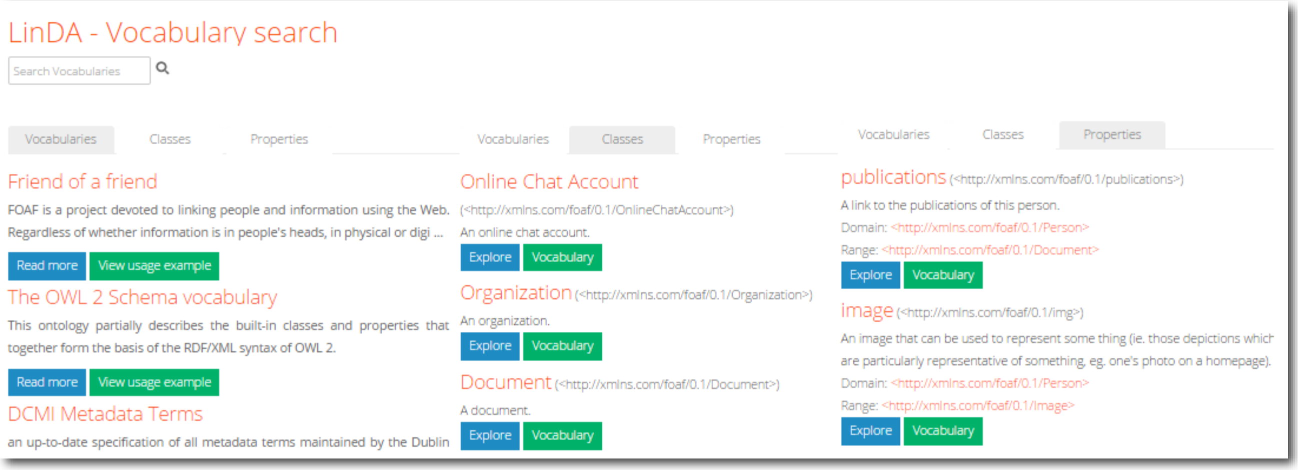
- Term suggestion: Web API methods pick the most prevalent vocabulary terms that describe real world objects and relationships.

[5] http://linda.epu.ntua.gr/

[6] https://github.com/LinDA-tools/LindaWorkbench/blob/master/LICENSE

3.4.1 Navigation

Due to the size of the vocabulary indexes, it is crucial for the usability and success of a vocabulary repository to assist term search in order for users to quickly access the intended vocabularies. When users navigate to the "Vocabularies" page, they are shown a catalogue of all repository entities, which they can select to view by vocabularies, classes, or properties:



LinDA - Vocabulary search

Friend of a friend

FOAF is a project devoted to linking people and information using the Web. Regardless of whether information is in people's heads, in physical or digi ...

The OWL 2 Schema vocabulary

This ontology partially describes the built-in classes and properties that together form the basis of the RDF/XML syntax of OWL 2.

DCMI Metadata Terms

an up-to-date specification of all metadata terms maintained by the Dublin

Online Chat Account

(<http://xmlns.com/foaf/0.1/OnlineChatAccount>)

An online chat account.

Organization (<http://xmlns.com/foaf/0.1/Organization>)

An organization.

Document (<http://xmlns.com/foaf/0.1/Document>)

A document.

publications (<http://xmlns.com/foaf/0.1/publications>)

A link to the publications of this person.

Domain: <http://xmlns.com/foaf/0.1/Person>

Range: <http://xmlns.com/foaf/0.1/Document>

image (<http://xmlns.com/foaf/0.1/img>)

An image that can be used to represent some thing (ie. those depictions which are particularly representative of something, eg. one's photo on a homepage).

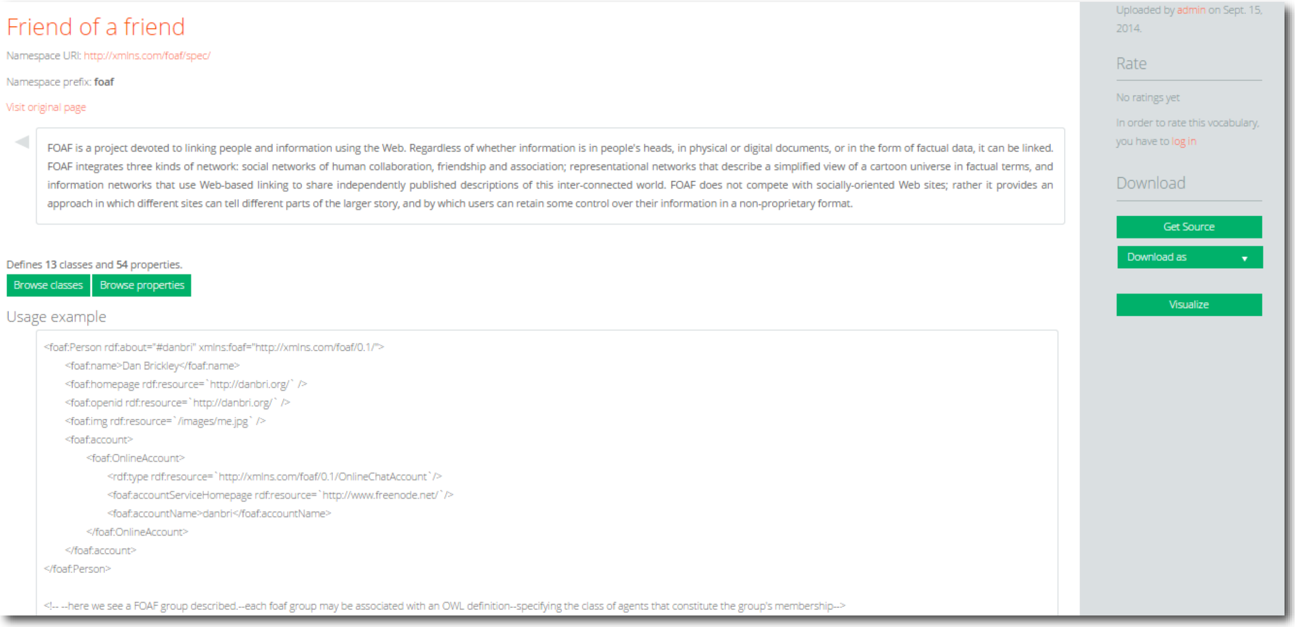
Domain: <http://xmlns.com/foaf/0.1/Person>

Range: <http://xmlns.com/foaf/0.1/Image>

Figure 3-1: The Vocabularies page (Vocabularies, Classes and Properties views)

As the vocabularies page is an object list, only a teaser of each element is shown. A teaser is composed by the name or label of the entity, a small description so that users can quickly decide if it interests them or not, and some basic links to get more detailed information about each entity.

By selecting a vocabulary, users get access to a page with more details about the selected vocabulary, which also allows them to perform actions on it, depending on their current role and permissions on the website.



Friend of a friend

Namespace URI: http://xmlns.com/foaf/spec/

Namespace prefix: **foaf**

Visit original page

FOAF is a project devoted to linking people and information using the Web. Regardless of whether information is in people's heads, in physical or digital documents, or in the form of factual data, it can be linked. FOAF integrates three kinds of network: social networks of human collaboration, friendship and association; representational networks that describe a simplified view of a cartoon universe in factual terms, and information networks that use Web-based linking to share independently published descriptions of this inter-connected world. FOAF does not compete with socially-oriented Web sites; rather it provides an approach in which different sites can tell different parts of the larger story, and by which users can retain some control over their information in a non-proprietary format.

Defines 13 classes and 54 properties.

Usage example

```
<foaf:Person rdf:about="#danbri" xmlns:foaf="http://xmlns.com/foaf/0.1/">
    <foaf:name>Dan Brickley</foaf:name>
    <foaf:homepage rdf:resource=`http://danbri.org/` />
    <foaf:openid rdf:resource=`http://danbri.org/` />
    <foaf:img rdf:resource=`/images/me.jpg` />
    <foaf:account>
        <foaf:OnlineAccount>
            <rdf:type rdf:resource=`http://xmlns.com/foaf/0.1/OnlineChatAccount`/>
            <foaf:accountServiceHomepage rdf:resource=`http://www.freenode.net/`/>
            <foaf:accountName>danbri</foaf:accountName>
        </foaf:OnlineAccount>
    </foaf:account>
</foaf:Person>

<!-- --here we see a FOAF group described.--each foaf group may be associated with an OWL definition--specifying the class of agents that constitute the group's membership-->
```

Uploaded by admin on Sept. 15, 2014.

Rate

No ratings yet

In order to rate this vocabulary, you have to log in

Download

Figure 3-2: The Friend of a friend vocabulary page

The vocabulary page contains:

- Some basic information about the vocabulary, like its namespace URI, the prefix that is commonly used for it, a link to the website where it is defined (like a W3C recommendation document or a website dedicated to the vocabulary) and a short description of its purpose and contents.
- Links to the source vocabulary document, both in its original version and in an automatically created RDF in all supported serializations (RDF/XML, n3 and NTriples), as well as a link to an automatically created vocabulary visualisation.
- Metadata about the vocabulary owner and when it was created.
- Information about classes and properties that it defines.
- Feedback controls, including rate and comment capabilities for authenticated users.
- A usage example that indicates how the major entities defined in the vocabulary are supposed to be used in order to create semantically correct RDF documents (optional).

The visualization of a vocabulary, even being limited by the number of elements that can be visualized in a web page without causing information overload, is often useful for users who want to get a quick view of the described ontology.

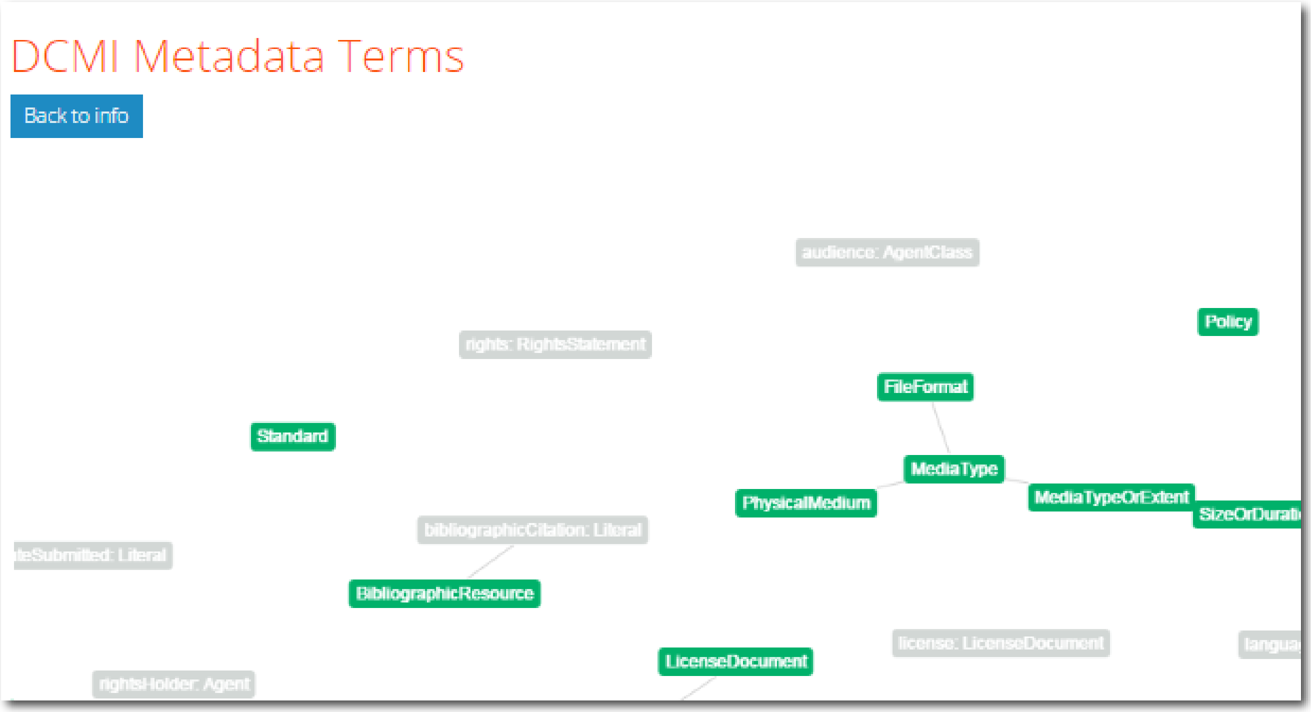
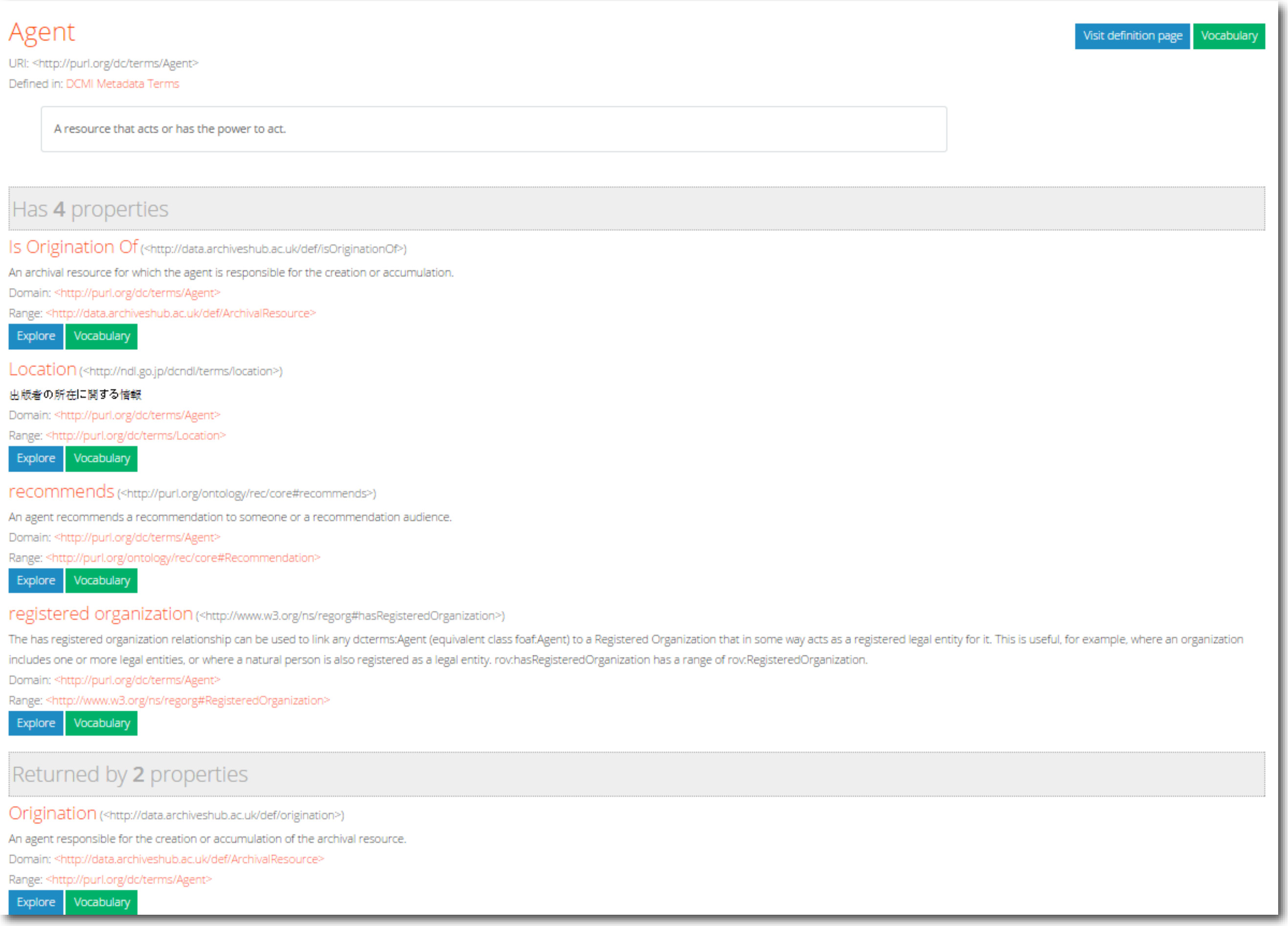


Figure 3-3: Part of the visualization of the DCMI Metadata Terms vocabulary

Users can also view the details of both classes and properties that have been extracted by the installed vocabularies.

- For both classes and vocabularies the resource URI, a humanized label, a description and a link to the vocabulary that defines them are provided.
- Classes show a list of all the classes that are the *rdfs:domain* of (properties that they *have*), as well as a list of all the classes that they are the *rdfs:range* of (properties they *return* them).
- On the other hand, *properties* present the user with the classes that are their *rdfs:domain* and *rdfs:range*.
- All elements are presented in a way that allows users to navigate between vocabularies, classes and properties seamlessly and without any interruption, something that helps them select the best entities describing real world entities and their connections.



Agent

URI: <http://purl.org/dc/terms/Agent>

Defined in: DCMI Metadata Terms

A resource that acts or has the power to act.

Has **4** properties

Is Origination Of (<http://data.archiveshub.ac.uk/def/isOriginationOf>)

An archival resource for which the agent is responsible for the creation or accumulation.

Domain: <http://purl.org/dc/terms/Agent>

Range: <http://data.archiveshub.ac.uk/def/ArchivalResource>

Location (<http://ndl.go.jp/dcndl/terms/location>)

出版者の所在に関する情報

Domain: <http://purl.org/dc/terms/Agent>

Range: <http://purl.org/dc/terms/Location>

recommends (<http://purl.org/ontology/rec/core#recommends>)

An agent recommends a recommendation to someone or a recommendation audience.

Domain: <http://purl.org/dc/terms/Agent>

Range: <http://purl.org/ontology/rec/core#Recommendation>

registered organization (<http://www.w3.org/ns/regorg#hasRegisteredOrganization>)

The has registered organization relationship can be used to link any dcterms:Agent (equivalent class foaf:Agent) to a Registered Organization that in some way acts as a registered legal entity for it. This is useful, for example, where an organization includes one or more legal entities, or where a natural person is also registered as a legal entity. rov:hasRegisteredOrganization has a range of rov:RegisteredOrganization.

Domain: <http://purl.org/dc/terms/Agent>

Range: <http://www.w3.org/ns/regorg#RegisteredOrganization>

Returned by **2** properties

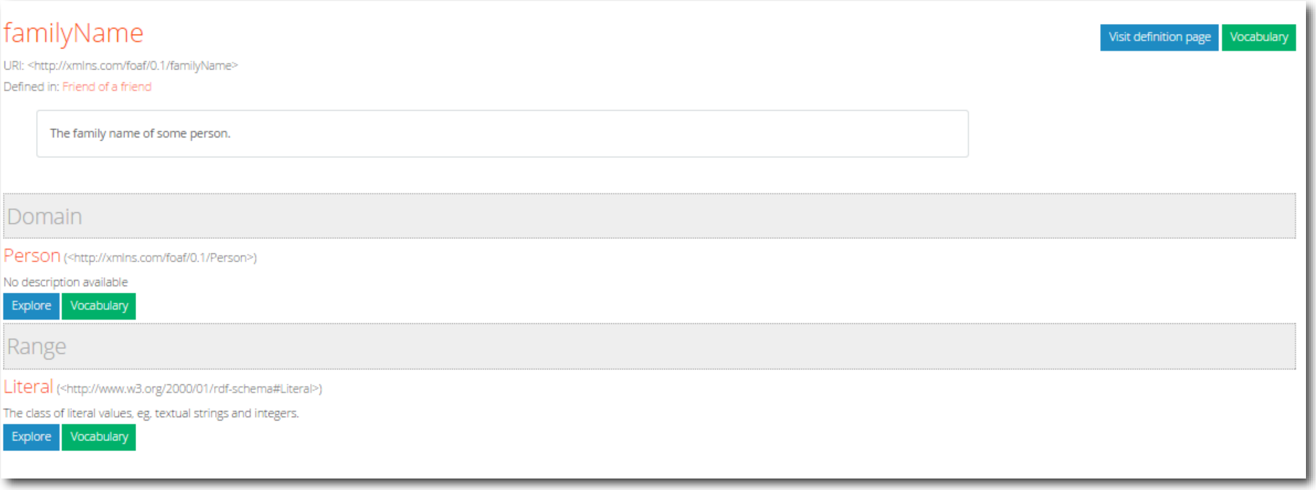
Origination (<http://data.archiveshub.ac.uk/def/origination>)

An agent responsible for the creation or accumulation of the archival resource.

Domain: <http://data.archiveshub.ac.uk/def/ArchivalResource>

Range: <http://purl.org/dc/terms/Agent>

Figure 3-4: The dcterms: Agent class page



familyName

URI: <http://xmlns.com/foaf/0.1/familyName>

Defined in: Friend of a friend

The family name of some person.

Domain

Person (<http://xmlns.com/foaf/0.1/Person>)

No description available

Range

Literal (<http://www.w3.org/2000/01/rdf-schema#Literal>)

The class of literal values, eg. textual strings and integers.

Figure 3-5: The foaf: familyName property page

3.4.2 Usage feedback

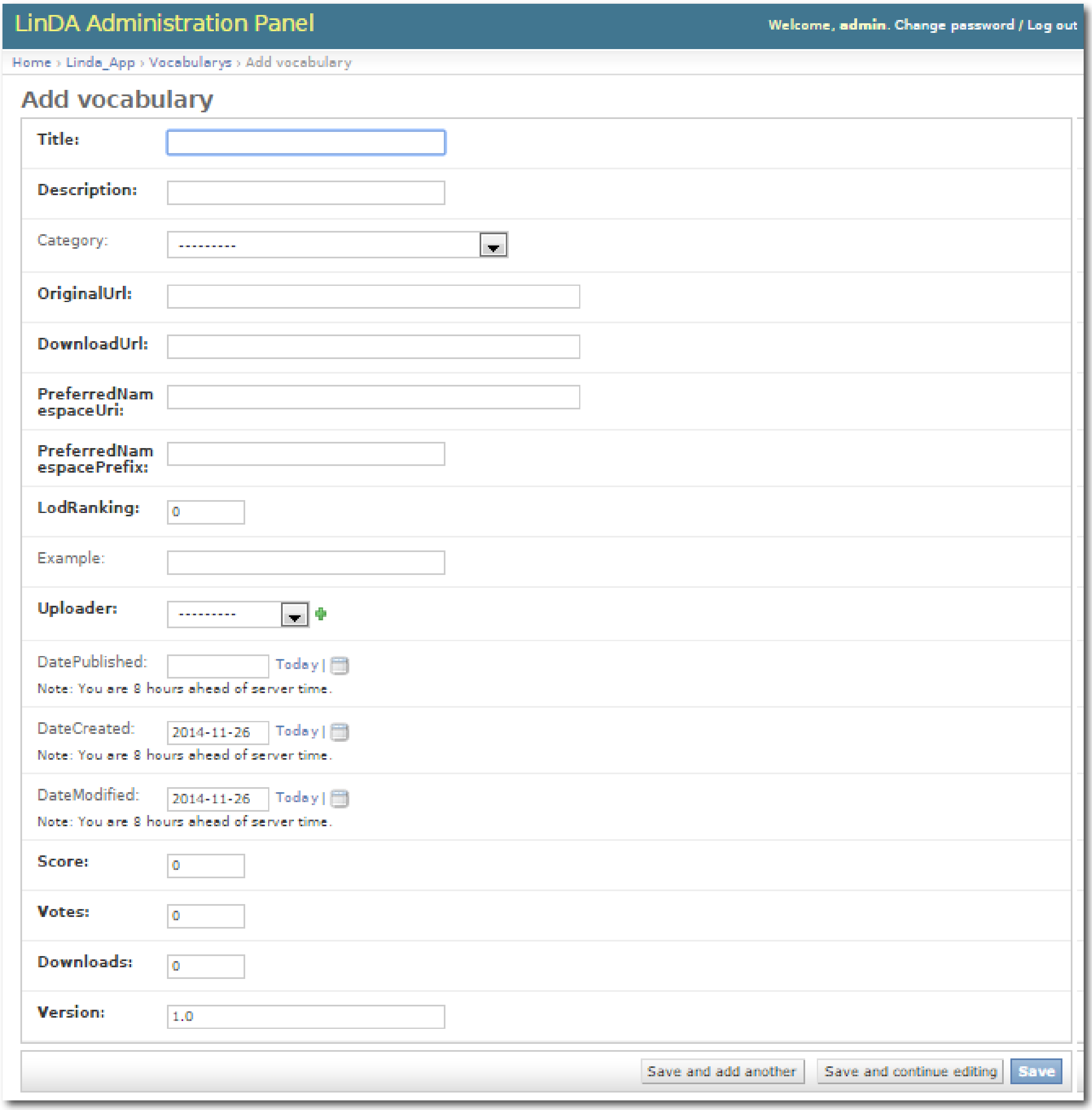
Feedback, evaluation of the presented material and community discussions are all great tools in order to promote the appropriate material according to each enterprise community needs and solve questions and problems that end users may face. In the LinDA Vocabulary and Metadata repository, two main mechanisms have been developed to let users express feedback and interact with each other:

- Vocabulary rating: By rating it, users let others know how well a particular vocabulary is suited for a specific business need. Highly rated vocabularies are more likely to contain material that can be used to describe business objects and actions well.
- Vocabulary discussions: Through commenting, statements about a vocabulary and its contents can be expressed and questions might get solved. While parsing a whole conversation is much slower than evaluating a vocabulary by its rating, it could award user with a lot of extra information about the vocabulary from other users that have used it or tried its terms out in data transformations.

3.4.3 Repository enrichment

By default the repository includes the basic set of the most common and popular terms and relations. In order to address the business needs of a specific domain, a number of actions need to be taken in order to enrich the initial repository contents with useful metadata, as well as let users add more content to the repository:

- Usage examples are a case of vocabulary metadata that was added to facilitate vocabulary usage by end users. Examples were gathered from various online sources, such as publications, standard recommendations, websites devoted to specific ontologies, presentations and online forums. They have been chosen in a way that presents the basic features of every vocabulary, giving the reader an idea of how to compose data sources based on the particular vocabulary and of the vocabulary's structure in general.
- An administration panel which lets super users create new vocabularies. After filling in the basic information required, mainly an ontology document, the repository will automatically create information about entities described in the vocabulary without user intervention. Administrators are urged not to edit existing vocabularies but to extend them using RDF constructs like *rdf:about*. Editing vocabularies will lead to inconsistencies between the local and the central repositories, and changes could be overwritten by future updates.



LinDA Administration Panel

Welcome, **admin**. Change password / Log out

Home › Linda_App › Vocabularys › Add vocabulary

Add vocabulary

Title:	
Description:	
Category:	---------
OriginalUrl:	
DownloadUrl:	
PreferredNamespaceUri:	
PreferredNamespacePrefix:	
LodRanking:	0
Example:	
Uploader:	---------
DatePublished:	Today | Note: You are 8 hours ahead of server time.
DateCreated:	2014-11-26 Today | Note: You are 8 hours ahead of server time.
DateModified:	2014-11-26 Today | Note: You are 8 hours ahead of server time.
Score:	0
Votes:	0
Downloads:	0
Version:	1.0

Figure 3-6: Creating a new vocabulary in the administration panel

4. DATA POLICY AND BUSINESS BROKERAGE FRAMEWORKS

In this assessment, we will give a definition of the Data Policy and Business Brokerage Frameworks (respectively DPF and BBF) related to their involvement in the AEGIS platform, describing the state of the art of the main features of the frameworks and the concept beyond their integration in AEGIS. The design of the Data Policy and Business Brokerage Frameworks is part of the work of WP2, T2.2.

The first overview of our expectations and ideas about the Data Policy and Business Brokerage Frameworks was given in D3.1, were we defined the AEGIS Brokerage Engine, as a component that will instantiate part of the methods that will be included in these frameworks.

The final aim of the AEGIS Brokerage Engine is to serve the AEGIS infrastructure with an endpoint that is able to create micro-contracts for artefact sharing, managing IPRs, quality and privacy issues as well. On this view, we will accurately describe each aspect of both of the DPF and BBF, in particular, we will investigate the state of the art of Data IPR, security, trust and quality features (paragraph 3.1.1), Blockchain technologies (paragraph 3.2.1.1) and Virtual currencies (paragraph 3.2.1.2).

4.1 DATA POLICY FRAMEWORK

4.1.1 Data IPR, security, trust and quality features

The current common understanding of big data can be summarized by the following definition appearing in 2013 in the first issue of *Big Data*, one of the first journals on the topic: *Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it* (Dumbill, 2013). Yet, big data hype and phenomenon followed and overlapped with the public sector interest in open government data symbolically enforced at a global level by the memoranda and directives signed by Barack Obama in the early years of his first mandate (Chignard, 2013; Obama, 2009). According to the Open Knowledge International open data are "*data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike"*(Open Knowledge International, 2017b).

Notwithstanding the clarity and appeal of the above definitions as well as the number of resources made available for improving data science skills and big or open data policies, one of the main barriers for laypersons and businesses is related to the understanding of the types of data and the capacity of current technological infrastructure and human resources to maintain, produce, and use open data. According to the Open Data Institute's data spectrum the different types of data are *closed*, *shared*, and *open data*. These types differ in terms of the following features:

- *Volume*: small, medium, and *big data* (for the latter adding their *velocity*, *variety*, *veracity*)
- *Ownership*: personal, commercial, and government data
- *Access*: internal, named, group based, public access, and open license data

Thus, it seems that while big data are worth opening in terms of access, not all the "open" data are necessarily "big" data. Furthermore, big data may be different considering their static rather than dynamic nature (*velocity* and *variety* dimensions) when considering digital data streams (DDSs) as "*dynamically evolving sources of data changing over time that have the potential to spur real-time action*"(Gabriele Piccoli & Federico Pigni, 2013).

Also, it is worth noting that one of the key issues is actually related to the accessibility of data and the *license* associated to each data, that can range from *contracts* typical of closed data to *open license* of open data or *authentication* required by shared data such as, e.g., the ones of medical research (The Open Data Institute, 2017). However, licensing is related to the requirement for open data to be *legally open*; whereas a further requirement for open data is to be *technically open* to be the data needs to be available in bulk in a machine-readable format (Open Knowledge International, 2017a), considering the different types of data (e.g., texts, statistics, images, maps, videos, sensors data, etc.) and data types available in the spectrum ranging from unstructured (not having a pre-defined data model such as e.g., textual data), structured (organized e.g. in relational databases) to semi-structured data (markup languages such as Extensible Markup Language – XML or open standards formats such as the JavaScript Object Notation - JSON).

Table 4-1: Big data features and key policy issues for the DPF and BBF

Big data features	Key policy issues			
	IPR	Security	Trust	Quality
Volume, velocity, variety, veracity		X	X	X
Ownership	X			X
Access	X	X	X	X

The issues related to the Data Spectrum are strictly connected to three key challenges for big data exploitation by laypersons as well as public as well as private organizations as well as appropriate policy design (see Table): big data Intellectual Property Rights, (IPRs), big data security, trust, and quality. In what follows we are going to discuss them at a glance.

4.1.1.1 Data IPR

Notwithstanding the generally claimed relevance of sharing and openness for creating value from by big data and open data, (Hofheinz & Osimo, 2017; OECD, 2014; The Economist, 2017), a somewhat free data economy faces two major challenges related to privacy and intellectual property, (Ekbia et al., 2015; Mattioli, 2014; Vare & Mattioli, 2014) especially when considering regulatory complexity and the fact that legal instruments vary in the different country jurisdictions and are not as robust as required by big data commercialization (Thomas & Leiponen, 2016)

Table 4-2: Intellectual property protectors and protection areas, adapted and elaborated from Vare & Mattioli (2014)

	Protection areas		
Intellectual property protector	Data ownership	Big data processing	Data sharing
Patents	Poor protection	Poor protection	Poor protection
Copyrights	Poor protection	Low protection	Poor protection

Trade Secrets	High protection (conditioned)	High protection (conditioned)	High protection (conditioned)

Considering traditional ways to protect intellectual property (Table) according to Vare and Mattioli (21014) they provide a poor protection when dealing with data or single datum, due their not being patentable or copyrightable, exception made for *trade secrets* when one can demonstrate that there (a) they have per se economic value and (b) there are reasonable efforts to keep their subject matter secret. Thus, according to Vare and Mattioli (2014), big data could fit more into a wide definition of trade secret law rather than the traditional intellectual property paradigms of patent or copyright (Mattioli, 2014), at least within the United States legal framework and commercial context. However, as pointed out by Lundqvist (2016) although the intellectual property rights concern devices, technologies, algorithms, and the infrastructure, firms holding large IP portfolios in a specific device industry as well as network, algorithm or cloud providers may try to obtain fees from access the data flowing in their systems or exclude others from accessing to it.

Taking these issues into account, in the context of big data IPR is connected to data sharing challenges, especially the *disclosure* of their provenance and pedigree, e.g. how data is initially collected and prepared (Borgman, 2012), without which data reuse and innovative applications from big data can be limited or even prevented (Mattioli, 2014). As argued by Mattioli (2014, p.547), “data devoid of context can also be devoid of meaning”. Focusing on big data practices rather than the datum itself, Mattioli (2014) proposes a hypothetical solution based on intellectual property to big data's disclosure challenges, called “dataright”. The solution is conditioned on the full and complete disclosure of data preparation practices by a data producer and provide data producers with a limited yet exclusive right in a closely-related asset-data itself, entitling them to block downstream use of data, but not reproduction or distribution.

As to these issues Lundqvist (2016, p. 1), considering Intellectual Property, Privacy Regulations and Competition Law for digitalized information, especially in the ‘Internet of Things’ domain, after an analysis of current US and European Union (EU) regulations conclude that, on the one hand, “general competition law may not be readily available for accessing generic (personal or non-personal) Data, except for the situation where the Data set is indispensable to access an industry or a relevant market”, on the other hand “sector specific regulations seem to emerge as a tool for accessing Data held by competitors and third parties.” In particular, Lundqvist (2016) points out the change in the number of institutional subject in charge of data collection from government authority or a similar public body to a scenario where private entities such as, e.g., Google, Facebook, Apple, Amazon, Microsoft, Spotify are collecting and storing big data, where most of them are personal consumer-related data. To these “digital” private players, are worth adding telecom companies and what Lundqvist (2016) calls “brick-and-mortar” firms such as, e.g., car manufacturers or refrigerator producers that start collecting data from their products as actually a by-product and stored from other connected devices, to provide new services, e.g. orienting car drivers or allowing communication between cars, but also for other business goals not strictly related to the original product.

As previously said, these data collectors having IPR portfolios for their infrastructure can prevent access to the data flows even if their IPRs don’t cover the raw data. Nevertheless, considering European Union, as noticed by Lundqvist (2016, p. 11) all industrial private collecting players should consider the General Data Protection Regulation (GDPR), being most of the data, personal data[7] and being the definition of personal data wide enough to include non-personal data (such as, e.g., meta-data) that might indirectly be in combination with other data identifying a natural person, thus becoming personal data, likewise. Indeed, according to the GDPR the "data subject" holds some rights on his/her personal data, especially with regard to portability and sharing. Nevertheless, considering the above mentioned role of trade secret as intellectual property protection and the right to portability the new EU regulation does not fully sort out the interface between them and database "sui generis protection" could be eventually applicable for holders of Data (Lundqvist, 2016, pp. 12–13), thus having a "thicket"-like regulations context similar to what Shapiro (2000) identified for patents. Furthermore, besides EU, database protection varies in the different countries jurisdiction with, e.g., the extreme of the US having no copyright protection for databases, Australian copyright law protecting them (Thomas & Leiponen, 2016, p. 83). Finally, competition law can have a key role in the EU for the exploitation of Big Data and the development of new business models, especially when considering the abuse of dominance doctrine in reference to (i) refusal to supply, (ii) exclusionary abuses and even (iii) discriminatory exclusion (Lundqvist, 2016, p. 15).

Table 4-3: Application of IPR Instruments for Semantic Metadata, adapted and elaborated from Pellegrini (2012)

	Copyright	Database Right	Unfair Practice	Patents
Documents	YES	YES	YES	NO
Base Data	NO	NO	PARTIALLY	NO
Description	YES	NO	YES	NO
Identifier	NO	YES	NO	NO
Name Space	YES	YES	YES	NO
Vocabulary	PARTIALLY	YES	YES	NO
Classification	PARTIALLY	PARTIALLY	PARTIALLY	NO
Ontology	PARTIALLY	YES	YES	PARTIALLY
Rules	PARTIALLY	YES	YES	PARTIALLY

Considering now the technical side of IPR data, especially for what concerns one of the key

[7] According to Art.4 of the EU general data protection regulation 2016/679 (GDPR) that will take effect in May 25 2018, personal data means personal data' means "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (Source: https://www.privacy-regulation.eu/en/4.htm)

features of the AEGIS approach to Big Data, namely the use of linked data, both at industrial and academic level the topic is still at a developing when not emerging status for what concerns IPRs management. As discussed by Pellegrini (2012, 2017) at industry level the few experience of media companies such as BBC Online, the New York Times, The Guardian, Reuters as well as publishing companies such as Wolters Kluwer and Reed Elsevier. As we have discussed above, also metadata can be considered personal data and when the subject is a private (or public) organization can be in principle covered by IPRs. Taking these issues into account, Table 4-3 shows the results of the analysis carried out by Pellegrini (2012) on the application of traditional IPR Instruments for semantic metadata. Furthermore, as pointed out by the study carried out by Pellegrini (2014) and Pellegrini & Ermilov (2013) on http://datahub.io in July 10, 2013 the licence scenario is heterogeneous and most of the linked data sets either do not specify licences (251 data sets or 30% of the sample) or use Creative Commons Attribution (135 data sets or the 16% of the sample), besides other licence models shown in Fehler! Verweisquelle konnte nicht gefunden werden..

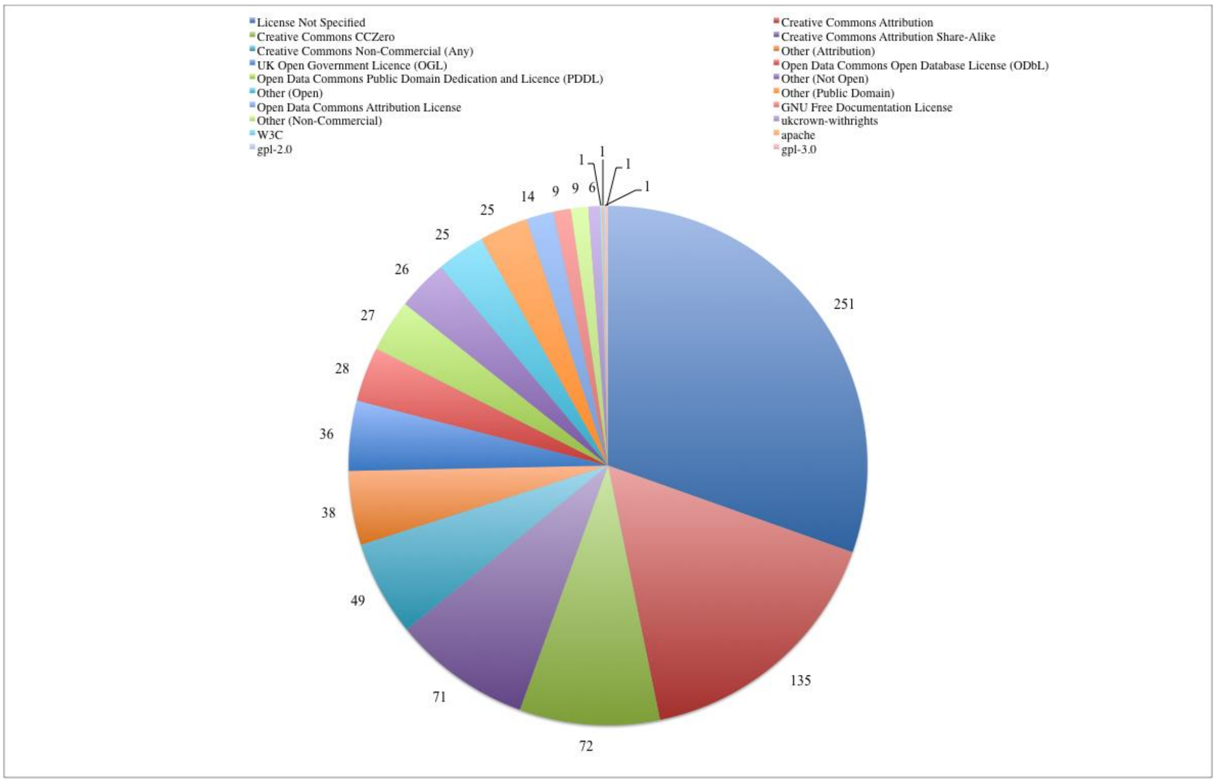


Figure 4-1: Datasets and Licenses from http://datahub.io (July 10, 2013), adapted and elaborated from Pellegrini (2014) and Pellegrini & Ermilov (2013)

The same heterogeneity has been found by the survey carried out by Jain et al. (2013) on the use of Linked Data datasets such as, e.g., DBpedia, Freebase, and Geonames for commercial purposes (the number of data sets per type of license are shown in Figure 4-).

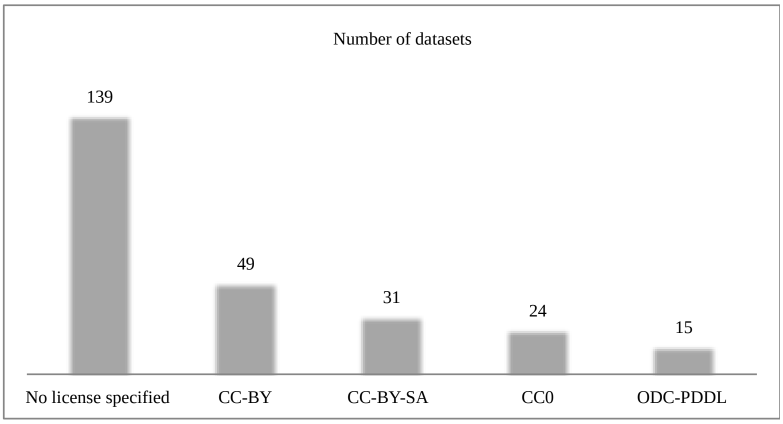


Figure 4-2: Type of license per linked data data sets, adapted and elaborated from Jain et al. (2013). [8]

Considering now available languages and approaches for describing IPRs and delivery/publishing data, the main effort has been on digital rights management (DRM), especially focusing on open and linked data, where the main languages are the Open Digital Rights Language (ODRL), Creative Commons Rights Expression Language (CCREL), and Open Data Commons (Pellegrini & Ermilov, 2013). Nevertheless, the research on IPR of big data as well as open linked data may benefit from and reuse/adapt the languages, models, and frameworks conceptualized and formalized in the (web) service licensing literature such as the one proposed by Gangadharan & D'Andrea (2011a, 2011b), using ODRL as a "rights expression language for describing machine interpretable licenses for services" (2011b, p. 50). Furthermore, another worth considering topic for languages, models, and frameworks to be reused/adapted/developed is the one of value-based (web) service contract specification (Liu et al., 2015) matchmaking (Comerio, 2013; Comerio, Paoli, Palmonari, & Panziera, 2014).

These issues are also connected to what (2016, p. 2) identify as the fourth phase in the late twentieth and early twenty-first century of the relationship between law and technology "*involving a new approach to regulation, the code-ification of law, which entails an increasing reliance on code not only to enforce legal rules, but also to draft and elaborate these rules*." (Filippi & Hassan, 2016, p. 2), where blockchain plays a key role. as mentioned in the AEGIS deliverable D1.2 ("The AEGIS Methodology and High Level Usage Scenarios", p. 99), AEGIS use of Blockchain technology for IPR and data sharing agreements through semi-automatic negotiation of micro-contract will be based on the predefined data handling policies, schemes and annotations defined in the DPF. They will ensure IPRs on data artefacts and data usage in relation to the data to be contributed to the platform. A Blockchain-based Intellectual Property (IP) Model is expected to allow all users to have clear insights and access to all copyright information on the dataset and on any dataset element and, at the same time, to be able to bring an easier payout system to IP owners and licensors of data.

Furthermore, as also discussed in the AEGIS deliverable D1.2 (p. 99), among the state of the art proposals at academic (J. Kishigami, S. Fujimura, H. Watanabe, A. Nakadaira, & A. Akutsu, 2015; S. Fujimura et al., 2015), the Ascribe "Ownership Layer" (McConaghy & Holtzman, 2015) is worth considering. This solution provides a powerful tool allowing proof-of existence

[8] Legenda: CC = Creative Commons; CC-BY = Creative Commons Attribution alone; CC-BY-SA = Creative Commons Attribution + ShareAlike; CC0= Creative Commons Freeing content globally without restrictions; ODC-PDDL= Open Data Commos Public Domain Dedication and License (PDDL)

on Blockchain for IP and innovation (in AEGIS, for dataset or data element) and makes easier the whole process of licensing and copyright transfer. Ascribe tackles the compelling need for a workable solution to the ownership and attribution issues by ensuring "ownership processing", that makes ownership actions of digital property universally accessible. Ascribe's approach is twofold, being based both on a registry with easy and secure legals and on visibility of data on usage / provenance of the content. It has two components, respectively ensuring IPR transparency and management, based on an ownership registry for easy secure disposition of rights. There is an ownership registry with easy and secure legals, which formalize (via a creator and consumer-friendly Terms Of Service) existing copyright rights on digital objects traditionally difficult to be leveraged, whilst the bitcoin-inspired blockchain serves for securely recording ownership transactions. In the registry it is possible to register a work, transfer ownership, grant licenses, loans and rentals. The registry also provides the time-stamping evidence of ownership actions through bitcoin-inspired blockchain. Ascribe enables to record intellectual properties on the Bitcoin-inspired blockchain, which is used as a distributed database to store the registry records (that track the history of ownership, the so-called "provenance"). It, thanks to the combination with cryptography, is able to make the registry global, robust, and impairment-resistant, whilst shielding the parties' personal identity (thanks to cryptography again). Ascribe has been proven in several domains and is being used both by individual creators and by institutions (e.g. marketplaces, libraries, archives, museums, galleries) and organizations, including new startups.

4.1.1.2 Security

Big data raise critical questions concerning their benefits to public good rather than the economic interests of private organizations, especially for what concerns privacy breaches and security of personal data (Boyd & Crawford, 2012). Furthermore, as pointed out by Bertino (2015, p. 760) "*many relevant applications of big data are in security, including cyber security, homeland protection, and healthcare, and in many such applications personal identifiable information may be required by the involved parties, such as law enforcement agencies.*" Accordingly, big data has to strictly satisfy the requirements identified by Bertino and Sandhu (2005, p. 2) data security solutions:

- *secrecy or confidentiality* of data against unauthorized disclosure,
- *integrity* as "prevention of unauthorized and improper data modification", and
- *availability* as "the prevention and recovery from hardware and software errors and from malicious data access denials making the database system unavailable."

Taking the above issues into account, from a functional point of view, Murthy et al. (2014, p. 29) identify four key security and privacy challenges for big data (*infrastructure security*, *data privacy*, *data management*, and *integrity and reactive security*), further refined by Ye et al. (2016, p. 269) as follows:

- *Infrastructure security* (Hadoop Security, Cloud Security, DoS Attack) related to the *variety* and *velocity* of big data;
- *Data privacy* (Encryption, Data Anonymization, Access Control) related to the *volume* and *value* of big data;
- *Data management* (key Management, Data Provenance, Monitoring and Auditing) related to the *volume*, *variety*, *velocity*, and *veracity* of big data.

Furthermore, due to the increasing relevance and availability of location and trajectory data, Ye et al. (2016, pp. 270–271) point out the relevance of privacy-preserving trajectory publishing techniques in big data, arguing the need for tailored privacy-preserving methods and techniques besides anonimization, such as, e.g., *generalization and suppression*, *perturbation*, and *differential privacy*. Considering now *privacy-preserving data publishing* (PPDP) Fung et al. (2010) present a survey, which focuses on attack (*record linkage*, *attribute linkage*, *table linkage*, *probabilistic attack*) and privacy models (k-Anonymity, MultiR k-Anonymity, l-Diversity, Confidence Bounding, (α, k)-Anonymity, (X, Y)-Privacy, (k, e)-Anonymity, (ε, m)-Anonymity, Personalized Privacy, t-Closeness, δ-Presence, (c, t)-Isolation, ε-Differential Privacy, (d, γ)-Privacy, Distributional Privacy), anonymization operations (*Generalization and Suppression*, *Anatomization and Permutation*, *perturbation*) , information metrics, and anonymization algorithms. As for record linkage, a research stream has emerged focused on Privacy-Preserving Record Linkage (PPRL) for big data, whose challenges (*scalability*, *linkage quality*, and *privacy*) and techniques have been analyzed and discussed by Vatsalan et al. (2017). Furthermore, due to the relevance of the above mentioned differential privacy techniques, Zhu et al. (2017) presents a survey of the research on *differentially private data publishing (DPDP)* and *differentially private data analysis (DPDA)*, identifying a set of challenges for the two of them such as *query number* (DPDP), *accuracy*, and *computational efficiency* (DPDP-DPDA).

Besides the identification of challenges, methods and techniques through surveys, at the state of the art, also frameworks and methodologies are being proposed for the different steps of the big data value chain, such as the one presented by Alouneh et al. (2016) focused on protecting big data during their analysis through an early classification of data before their being moved, copied or processed; the classification then activates security procedures according to their actual criticality level. Furthermore, considering the storage of big data, especially the adoption of cloud solution and the issue of keeping the integrity and security of the outsourced data, Sookhak et al. (2017) proposes a remote data auditing (RDA) technique based on algebraic signature properties for a cloud storage system with minimum computational and communication costs and ii) a data structure-Divide and Conquer Table (DCT) aimed at supporting dynamic data operations (e.g. append, insert, modify, and delete). Also, as for differential privacy it is worth mentioning the GUPT platform proposed by Mohan et al. (2012), which implement a model of data sensitivity that i) degrades privacy of data over time, thus enabling efficient allocation of different yet constant levels of privacy for different user applications, and ii) introduces techniques for improving the accuracy of output. Moreover, considering the relevance of graph data to the AEGIS project, it is also worth mentioning the algorithms proposed by Karwa et al. (2014, p. 22:1) aiming at releasing "*useful statistics about graph data while providing rigorous privacy guarantees*", which works on datasets of e.g. social ties or email communication, satisfying the *edge differential privacy*, and whose output approximates answers to *subgraph counting queries*.

Another facet of big data security is related to the balance between protection of digital contents through digital rights management (DRM) and their actual use (Gaber, 2013; Ku & Chi, 2004; S. Lee, H. Park, & J. Kim, 2010). As to these issues, Lee et al. (2010) have proposed a secure DRM interoperability scheme for minimizing disclosure of the security properties of DRM providers while preserving their profits through a designated proxy re-encryption scheme, also allowing the providers to manage and trace their digital contents. Another interesting proposal related to the DRM topic is the one by Win et al. (2012), introducing a privacy preserving content distribution mechanism for DRM without relying on the trusted third party assumption, by using primitives such as blind decryption and one way hash chain.

Finally, considering the increasing availability of sensors data enabled by the Internet of Things (IoT), the security issues are at the state of the art also connected to data quality due the heterogeneity and number of data sources and connected devices (Sicari, Rizzardi, Miorandi, Cappiello, & Coen-Porisini, 2016). To face these issues, Sicari et al. (2016; 2016) have proposed a lightweight and cross-domain prototype of a distributed architecture for IoT with minimum data caching functionality and in-memory data processing.

4.1.1.3 Trust

Trust has been discussed and investigated in multiple fields and received attention in areas connected to big data, such as, e.g., data sharing in smart cities (Cao et al., 2016), data reuse (Yoon, 2017), social networks (Sherchan, Nepal, & Paris, 2013), and cloud computing (Corradini, De Angelis, Ippoliti, & Marcantoni, 2015; Monir, AbdelAziz, AbdelHamid, & EI-Horbaty, 2015). Also, since the early investigation on the role of trust with regard to noncoercive/persuasive power rather than coercive power in Electronic Data Interchange (EDI) adoption by Hart & Saunders (1997), trust has been a key topic in the information systems (IS) research area. In particular, McKnight et al. (2011) have pointed out the difference between *trust in people* and *trust in technology*, showing common features of risk and uncertainty in their contextual condition, but difference in the object of dependence (moral and volitional agents vs. artifacts generally lacking volition and moral agency) as well as the perceptual (actually, not objective in nature) expectations of the users, which refer to different attributes for people and technology, such as, respectively, *competence vs. functionality*, *benevolence vs. helpfulness*, *predictability/Integrity vs. reliability* (Mcknight et al., 2011, p. 12:4-5). Consequently, trust can be generally seen from a *system* perspective or from a *user* perspective, the latter coming from sociology (Song, 2017, p. 4); also, among its properties are worth noting its being *context specific*, *dynamic*, *propagative*, *non-transitive*, *composable*, *subjective*, *asymmetric*, *self-reinforcing*, event sensitive (Antoniou et al., 2007, p. 47:8-10).

Taking these issues into account, trustworthiness of data is the second big challenge related to the effective and value added use of big data, where the semantics of the application domain is one of key complexity factors (Bertino, 2015, pp. 758–759). As for trustworthiness of big data, Bertino (2015, p. 759) has identified the following key research directions summarizing the current challenges for the topic: *data correlation techniques*, *high assurance and efficient provenance*, *source correlation techniques*.

Table 4-4: Provenance frameworks with their storage model, query support and level of provenance integration, adapted from Zafar et al. (2017, p. 55)

Provenance Framework	Storage Model	Query Support	Provenance Integration
PASS	Berkeley DB	Query tool	OS Level
LinFS	RDBMS	SQL	
ES3	XML database	XPATH, Xquery	
PreServ	File System + Berkeley DB	Java API+Xquery	Process Level

Karma2	XML database	XPATH, Xquery	Workflow + Process Level
Taverna	Relational RDF Store	SPARQL	Workflow Level
Swift	RDBMS	SQL	
Pegasus	RDF files + RDBMS	SPARQL + SQL	
VisTrails	RDBMS + XML	Visual QBE	
Kepler	File System	API	
REDUX	RDBMS	SQL	

Furthermore, considering provenance as a meta-data that describe the history of data and processes, Zafar et al. (2017, p. 50) have focused on secure provenance schemes, surveying available frameworks (see Table) for the provenance lifecycle (provenance collection, provenance storage, provenance query and analysis), thus identifying a set of secure provenance requirements (confidentiality, privacy, integrity, availability, unforgeability, non-repudiation, chronology) and proposing a taxonomy of secure provenance scheme, which considers (Zafar et al., 2017, p. 56):

- *Implementation primitives* (Watermarking, Signatures, Hashes or Checksum, Encryption)
- *Trusted Platform* (trusted software and trusted hardware)
- *Scope* (identification of information leakage, Identification of guilty party) and
- *Access control* (traditional ACL, Provenance-based)

Moreover, at the state of the art, also the difference between trust and distrust in terms of their antecedents (reputation, environmental scanning, and defensive posture) for the related dispositions and beliefs have been empirically studied, e.g., by Simon (2016) on consumers using credit card in technology-driven transactions. As for these issues, focusing on big data, open data and user-generated data, Kostkova et al. (2016) question the challenges and opportunities they bring to healthcare, especially with regard to responsibility, accountability, and public policies that both protect personal information and enable the use of data.

Considering now technology and computational solutions to the challenge of trust of big data, Yin et al. (2017) has proposed a recommendation algorithm focusing on social trust between users, thus, designing a collaborative filtering recommendation algorithm (CFRAT) and a hybrid recommendation algorithm based on the trust and similarity (HRAT) to analyze the impact of trust in recommendation grounded on big data. As for these issues, the use of big data associated to services or provided through them may benefit from the output of the research carried out in the web services area, such as, e.g., the trust rating method proposed by Yamasaki (2011) for information providers, exploiting the structure of human relationships and meta-level communication protocol over the social web service. Visualization of big data is another issue encompassing trust at user level, where big data exploration and analysis often require sampling or Approximate Query Processing (AQP) for fast answers to exploratory needs with a consequence trade off in terms of quality and trust, thus requiring solutions as *Pangloss*, an optimistic visualization tool based on AQP proposed by Moritz et al. (2017)**.**

As for web data, Liow and Lee (2016) presents a *data certification scheme* for data providers employing a common language to describe data attributes (e.g. accuracy, quality), the support from the providers as well as the legal restrictions on licensing and rights. Also, Sacco et al. (2011) have proposed an access control framework for structured data based on semantic web meta-formats made up of i) a light-weight vocabulary, named *Privacy Preference Ontology (PPO)*, for defining fine-grained privacy preferences restricting access to information represented as Linked Data (Owen Sacco & Passant, 2011), and ii) a privacy preference manager, named *MyPrivacyManager*, for the definition by users of privacy preferences based on the vocabulary. Moreover, the issue of trusted and secure access of linked data has been faced by Sayah et al. (2016) by focusing on selective disclosure and proposing a data-annotation approach to enforce access control policies for modular, fine-grained, and efficient selective disclosure on top of RDF data. Furthermore, focusing on querying the trustworthiness of information resulting from the combination of different RDF data sources, Hartig (2009) has proposed, a trust-aware extension to SPARQL, named tSPARQL. Moving from academic research to industry solutions for trusted access control of linked data, Lázaro and Carnero (2013) have discussed extensions to traditional role-based multi-domain access control approaches suitable to improve mobile collaboration among companies in logistic, manufacturing, and e-Commerce. Finally, trust is a key component to enable new business models from data, as also pointed out by Minzheong (2017) who first identify three contradictory trust stances (*internal optimization vs. external interaction*, *control vs. orchestration of personal data*, and *end-user vs. ecosystem value*) and trust stages (*source stage: trust data collection*; *process stage: trust value evaluation*; *result stage: trust value dissemination*) in order to suggest three strategic directions for trust based business models: *trust management* (on the 'Source' stage), *orchestrated data sharing* (on the 'Process' stage) and *authorization management* (on the 'Result' stage).

4.1.1.4 Quality

The growing interest and use of big data and analytics by organizations has moved the research and practice of data quality from a primary focus on content to "usage and context", as argued by Shankaranarayanan and and Blake (2017). Nowadays, a key issue for big data is the fact pointed out by Batini et al. (2015) that data and information quality can be considered "in the wild", thus, including not only traditional database systems but also social networks, sensors data, as well as open data and linked open data (LOD), among others. Taking these issues into account, at the state of the art Batini & Scannapieco (2016, pp. 99–110) and Batini et al. (2015) have identified five cluster of dimensions relevant to big data as well as to open data[9]:

- *Accuracy* (including *syntactic accuracy*, *semantic accuracy*, *currency*, *timeliness, reliability, precision*)
- *Completeness* (including *schema completeness*, *property completeness*, *linkability completeness, relevancy*)
- *Consistency* (including *logical consistency, domain consistency, format consistency, topological consistency, numerical consistency* and *temporal consistency*)
- *Redundancy* (including *conciseness, spatial redundancy, temporal redundancy*)
- *Readability* (including the *understandability* dimension)
- *Accessibility* (including *licensing, availability, and interoperability*)

[9] For a detailed discussion of the mentioned clusters and dimensions we refer the reader to Batini & Scannapieco (2016, pp. 99–110) and Batini et al. (2015).

- *Trustworthiness* (including dimensions as *believability*, *verifiability, reputation*)

Furthermore, from the interdisciplinary areas of space information science Liu et al.(2016) consider mobile data, social media data, volunteering data, and searching engine data and the three relevant paths for the use of big data in space information science (yet we argue suitable to be generalized to other research subject and areas and to the steps of the AEGIS value chain), i.e. *data collection*, *data processing* and *data analysis* to survey and assess state of the art big data studies and identify the following data quality and usage problems: *authoritativeness*, *information incompleteness and noise*, *representativeness*, *consistency* and *reliability*, and *ethical problems*. In general, the big data four "Vs' (*volume*, *velocity*, *variety*, and *veracity*) represent the factors that challenge traditional data quality practice and research, this latter still not strongly tackling the topic on all its facets, exception made for an interest and focus on volume and variety (Shankaranarayanan & Blake, 2017, p. 9:20-21).

As for these issues, considering big data initiatives in financial institutions such as, e.g. banks or insurance (the latter relevant to AEGIS platform), Haryadi et al. (2016) has identified eleven dimensions relevant to the sector (most of them already present in the above-mentioned classification: *accuracy*, *believability*, *relevancy*, *currency*, *completeness*, *comprehensiveness*, *consistency*, *uniqueness*, *timeliness*, *validity*, and *traceability*. Moreover, an interesting finding by the analyses of Haryadi et al. (2016) on a sample of European financial institutions is their prevalent use of internal (structured) data instead of external unstructured or open data. Taking these issues into account Haryadi et al. (2016) identify three supporting aspects of big data quality (*discovery*, *accessibility*, and *operationality*) and ten antecedents which impact big data quality and value creation: the big data four "Vs'"; the credibility of source and content; tools, techniques, and technology; authenticity of data collecting method; clarity of data use policy; analytics skills and multidisciplinary team; centralized data repository; agile capability; compelling business case (Haryadi et al., 2016, p. 121).

Besides the effort by practitioners and academics on identifying key factors and dimensions of big data quality, at the state of the art frameworks and models are being proposed, likewise. Jorge et al. (2016) have proposed a model ("*3As Data Quality-in-Use model*") for assessing the business value of data in big data initiatives by their context of use. The model is centered on the *adequacy* of data as ‘‘the state or ability of data of being good enough to fulfill the goals and purposes of the analysis’’(Merino et al., 2016, p. 126); accordingly, the model considers three core data quality characteristics: *contextual adequacy*, *operational adequacy* and *temporal adequacy*. Looking now at a specific set of data as the one produced by sensors, cameras, car devices, which can be considered part of the variety facet of big data, Piccoli and Pigni (2013, p. 54) have distinguished them as digital data streams due to their dynamically evolving nature changing over time. As for these issues, Geisler et al. (2016) have proposed an ontology-based data quality framework for relational Data Stream Management Systems (DSMS) and a data quality management methodology for data streams, evaluated in domains such as transportation systems and health monitoring. Yet, integration of the different types of data is a core issue for (big data) data quality and among the framework proposed by academics, it is worth mentioning *QDflows*, an ontology-based system presented by Abdellaoui et al.

(2017) aimed at designing quality-aware data flows. Finally, among the different subjects of what we could define the "galaxy" of big data rather than a simple data spectrum (see above the introduction to this Section), linked (open) data, one of the key topic of AEGIS, have received a specific attention from both academic and practitioners with a consequent production of framework and quality assessment tools surveyed by Zaveri et al. (2016). Besides them, it is worth mentioning here the proposal by Debattista et al. (2016) of a conceptual methodology for assessing Linked Datasets, and a framework for Linked Data Quality Assessment, named Luzzu that are aimed at having output which are suitable for machine consumption.

4.1.2 AEGIS DPF

The AEGIS Data Policy Framework (DPF) is responsible for realizing the underlying AEGIS premise that any AEGIS asset, including datasets, data-as-a-service, data micro-services, algorithms and intelligence reports, can be provided to the PSPS data value stakeholders on demand at any time through concrete licenses / policies that are encapsulated into smart contracts in the AEGIS Business Brokerage Framework (BBF) (as described in section 4.2.2).

Taking into account state-of-the art approaches on data IPR, trust and quality as defined in sections 4.1.1.1, 4.1.1.3 and 4.1.1.4, the AEGIS DPF aims at facilitating the PSPS stakeholders to answer questions like: Are we allowed to use a specific dataset, algorithm or data-as-a-service? Do the actual qualities of the data asset meet the denoted qualities in the agreement between the corresponding data asset provider and consumer? Are we allowed to republish a derivative intelligence report built based on a data asset or a collection of data assets?

However, such questions typically require extensible models that are able to capture contractual terms for data contracts, and their representation in a form to be reasoned by automatic techniques is not always possible. Moreover, certain properties of data assets, such as quality and IPR, increase the complexity in the definition of the AEGIS DPF to be used for agreeing on and monitoring data contracts at runtime. It is thus noted that the present draft version of the AEGIS DPF will only provide the foundations of the data asset policies that will be implemented in the AEGIS platform and will be further expanded and complemented with concrete examples in its final iteration.

As depicted in the following figure, a Data Asset complies with a specific Data Asset Policy that governs every Data Asset contract / Transaction among a Data Asset Provider and Consumer. A Data Asset Policy in AEGIS thus aims at:

- Defining the detailed terms according to which a data asset can be used, on the basis that any use outside the policy terms would constitute an infringement.
- Specifying the expected data asset quality, as well as the delivery and payment terms.
- Clarifying the liability of data asset providers and consumers (e.g. in case of failure of the provided data asset).

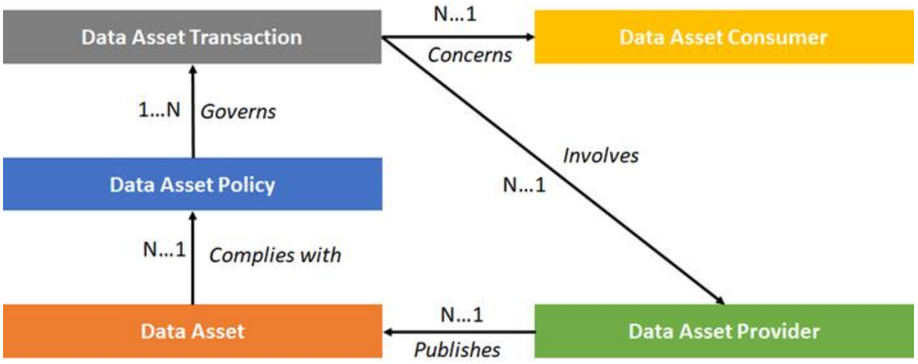


Figure 4-3: High-level Data Asset Policy Concept in AEGIS

In addition to the core metadata defined in section 2 of the present deliverable, different concerns and metadata concur to define the AEGIS DPF including:

- **Data Assets Rights (DAR)** encapsulating the rights that the data asset provider authorises the consumer to exercise for the specific data asset in order to clarify and assure the corresponding intellectual property rights. In accordance with the Creative Commons Rights Expression Language (CC REL)[10] and considering the approaches mentioned in section 4.1.1.1, the set of common data right terms for data assets offered by the AEGIS platform are classified in the following categories:
    - **Permissions** including actions on the data asset that may or may not be allowed or desired, i.e.: Distribution (restricted or unrestricted publication and distribution of a data asset); Reproduction (from a given data asset, temporary or permanent reproductions can be created by any means and in any form, in whole or in part, including of any derivative data assets or as a part of collective data assets); Derivative Works (creation and distribution of any update, adaptation, or any other alteration of a data asset or of a substantial part of the data asset that constitutes a derivative data asset); Sharing (that permits Open-Public-Group based-Named-Internal access[11] to a data asset).
    - **Requirements** including actions that may or may not be requested of the data asset consumer, i.e.: Notice (copyright and license notices to be kept intact); Attribution (credit to be given to copyright holder and/or provider); Share Alike (derivative works to be licensed under the same terms or compatible terms as the original work); Source Code (to be provided when exercising some rights granted by the license); Copyleft (derivative and combined works must be licensed under specified terms, similar to those on the original work); Lesser Copyleft (derivative works must be licensed under specified terms, with at least the same conditions as the original work; combinations with the work may be licensed under different terms).

[10] https://creativecommons.org/ns

[11] https://theodi.org/data-spectrum

  - **Prohibitions** including actions a data asset consumer may be asked not to do, i.e.: Commercial Use (exercising rights permitting or forbidding use of a data asset for commercial purposes).

- **Quality of Data Assets (QoDA)**, a complex concept that, depending on the data asset type (i.e. big/small/open datasets, data-as-a-service, data micro-services, algorithms, intelligence reports) consists of the following facets in AEGIS in alignment with the quality dimensions analyzed in sections 4.1.1.4 and 4.1.1.3:
  - **Accuracy** as a measure of correctness and precision (e.g. whether the dataset is error-free or the performance of an algorithm in terms of results is satisfactory).
  - **Completeness** defining the degree to which a data asset is sufficient in depth, breadth and scope.
  - **Consistency** by ensuring internal validity, i.e. two or more values do not conflict with each other.
  - **Credibility** as the degree to which a data asset is considered as trustworthy, traceable and reliable (e.g. through provenance, through the reputation of the data asset provider, by publishing the identity of the provider).
  - **Timeliness** as a measure of how sufficiently up-to-date a data asset (e.g. a dataset or data-as-a-service) is for a certain task, representing the timespan that such a data asset remains valid.

- **Pricing Model** that, considering the aspects described in sections 4.1.1.1, consists of:
  - **Price Scheme** including transaction, PAYG (pay-as-you-go) and subscription schemes. In detail, the transaction model allows data asset providers to charge for each single use of a data asset. The PAYG model is applicable in the case of data-as-a-service (provided through APIs) and allows charging the data asset consumers every time they call the provided APIs to retrieve data. The subscription model allows consumers to purchase data assets for a fixed period (e.g., a week, a month, or a year) and only pay once for this period with or without maximum limitations for how frequent they access a data asset.
  - **Cost** reflecting the exact amount to be paid for a certain period of time for use and/or offline retention.
  - **Coverage** referring to the geographic coverage of the data asset in question (e.g. full Europe coverage, specific countries or regions packages, specific areas packages).
  - **Exclusivity of use** that defines whether the data asset consumer requires exclusive use and the corresponding data asset becomes unavailable in the AEGIS platform (as long as the relevant data contract is active).
  - **Duration of use**, the time period for which the data consumer has paid for use of the data asset in case of a subscription scheme.
  - **Duration of offline retention**, the time period for which the data consumer is allowed to have offline / local access to the data asset.
  - **Maximum Use**, i.e. number of calls for assets use per day in case of a PAYG or subscription scheme.

- **Policy Terms** consisting of more detailed terms regarding a data asset’s evolution, support, indemnification, and limitation of liability and taking into account the approached analysed in sections 4.1.1.2 and 4.1.1.3. In this version of the AEGIS DPF, the following policy terms are defined:

- **Liability** defining the data liability disclaimer and conditions.
- **Privacy Compliance** to indicate whether and how the privacy aspects of a data asset have been appropriately handled through anonymization, fabrication, synthetisation, etc. depending on the level the data asset belongs to, e.g. Level 0 – Open data assets without any privacy aspects, Level 1 – Data assets with small privacy concerns, Level 2 – Data assets with significant privacy concerns and Level 3 – Data assets with severe privacy concerns.
- **Online Availability Guarantees** describing the expected Quality of Service in case of data-as-a-service assets.
- **Versioning & updates** whether the data asset consumer has access to updates and latest versions of the data asset.
- **Applicable Law** including the regulatory framework of the country that is responsible for settlement of any disputes.

As depicted in the figure below, the AEGIS DPF concerns accompany either each data asset or each transaction in the AEGIS BBF and are applicable in a different way depending on the nature of the data asset in question (i.e. dataset, data-as-a-service, data micro-service, algorithm, intelligence report). In the case of the Data Assets Rights and the Quality of Data Assets concerns, each data asset contains the corresponding meta-data while the meta-information regarding the Pricing Model and the Policy Terms appears in each transaction of the data asset.

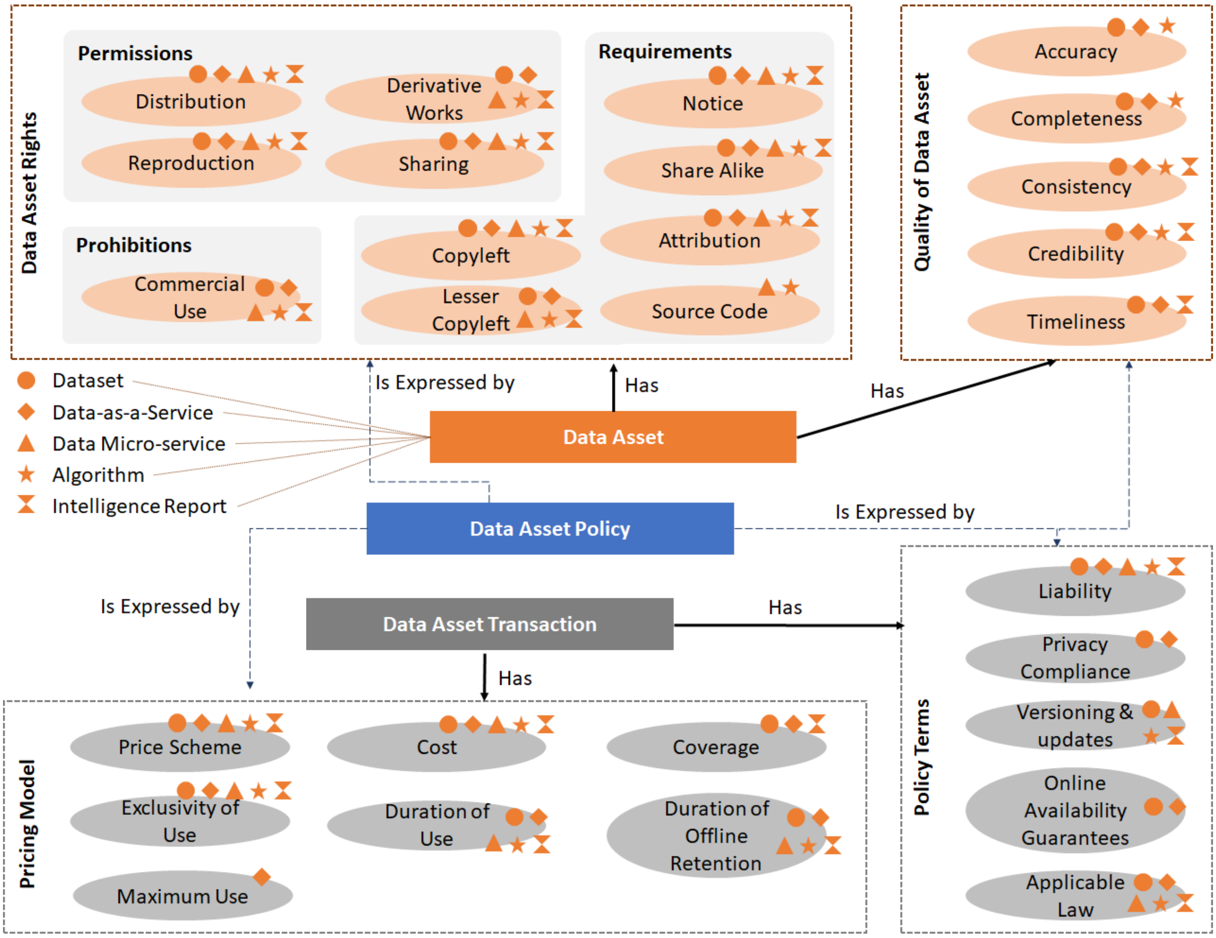


Figure 4-4: Data Asset Policy Concerns in AEGIS

4.2 BUSINESS BROKERAGE FRAMEWORK

In this assessment we will give a definition of Business Brokerage Framework (BBF) related to its involvement in the AEGIS platform, describing the state of the art of the main features of the BBF, the blockchain technology (section 3.2.1.1) and the virtual currency approach (section 3.2.1.2). In section 3.2.2 we will describe the BBF that we would like to develop with and within AEGIS.

The AEGIS BBF is an endpoint of the AEGIS infrastructure that is able to create micro-contracts for data sharing. Such service eases and encourages the data sharing offering a set of contracts templates covering a wide range of sharing possibilities, i.e. considering many features as determinants for the micro-contract creation (data type, data weight, access/visibility rules, national legislation). Since the data sharing is a key point of the AEGIS platform, and the actual data sharing percentage from the results of the D1.1 questionnaire seems to be very low, the BBF is, in our opinion, one of the most meaningful and innovative AEGIS aspects.

4.2.1 BBF Technologies

Hereinafter the BBF Technologies will be investigated in order to give an overview of the state of the art; in particular, we will focus on the Blockchain technology (paragraph 3.2.1.1) and on the Virtual currencies (paragraph 3.2.1.2).

4.2.1.1 Block chain technologies

The Blockchain technology is a relatively new concept that is generating significant interest across a wide range of industries. The reason for the interest in Blockchain is its central attributes that provide security, anonymity and data integrity, creating a decentralized environment where no third party is in control of the transactions and data. As the field of applications for blockchains grows, industry leaders are customizing and tailoring the technology to fit very particular uses[12]. The main applications that involve blockchain technologies are:

- Financial services: allowing for the entire financial services industry to dramatically optimize business processes by sharing the data in an efficient, secure, and transparent manner;
- Identity: tracking and managing digital identities both secure and efficient, resulting in seamless sign-ons and reduced fraud;
- Internet-of-Things: settling scalability, privacy, and reliability;
- Money: through cryptographic digital currencies allows for a new system of robust, transparent, and efficient monetary management; this is probably the most common application of blockchain since principal theories of blockchain architectures used today were first outlined and defined in the original Bitcoin white paper written and published by Satoshi Nakamoto in 2008;
- Real Estate: allowing for a significant gain in efficiency in how records are stored and recorded.

Start at the back, to have a better understanding on blockchain technology, we would like to provide a definition of what 'Blockchain' means[13].

Blockchain definition:

A blockchain is a distributed database solution (also named as 'ledger'), that maintains a continuously growing list of data records that are confirmed by the nodes participating in it. Data are recorded in a public ledger, including information of every transaction ever completed. Each block is then 'chained' to the next block, using a cryptographic signature. This allows block chains to be used like a ledger, which can be shared and accessed by anyone with the appropriate permissions and which do not require any third party organization in the middle. The information about every transaction ever completed in Blockchain is shared and available to all nodes. In addition, the nodes in Blockchain are all anonymous, which makes it more secure for other nodes to confirm the transactions. As aforementioned, Bitcoin was the first application that introduced Blockchain technology. Bitcoin created a decentralized environment for cryptocurrency, where the participants can buy and exchange goods with

[12] H. Vranken, Sustainability of bitcoin and blockchains (2017) http://dx.doi.org/10.1016/j.cosust.2017.04.011

[13] http://www.blockchaintechnologies.com

digital money (we will describe in deep both Bitcoin and other virtual currencies in paragraph 3.2.1.2).

Blockchain technology is at the basis of the development of some prototype applications such as IoT, smart contracts, smart property, digital content distribution, Botnet, and P2P broadcast protocols. This shows that Blockchain technology is not limited to applications in cryptocurrencies even if cryptocurrencies are the widely investigated application nowadays. In fact, the idea of a public ledger and a decentralized environment can be applied to various other fields in different industries, which can possibly be even more interesting than cryptocurrencies. However, we also found a set of different applications developed for the Bitcoin environment (i.e. BitConeView, BitIodine), rather than using Blockchain technology in some other environment. These types of applications, helping users to analyse the Bitcoin network and studying how Bitcoin transactions are completed, with a visual presentation can help to understand the essence of Blockchain, and how a decentralized transaction environment actually works. Analysis applications can also help to identify frauds and possible security issues by following the flows of transactions. Another major direction for applications is security with applications such as CoinParty and CoinShuffle that help the Bitcoin network to become more secure, by adding an extra layer of privacy for the users. In the future, increased sizes and user bases in various Blockchains will trigger the need to conduct more research on the challenges and limitations in topics related to scalability. In addition, the security and privacy of Blockchain will be always a first priority topic for research.

Blockchain 2.0

Within this category, the many ideas that firms are developing to utilize the benefits of Blockchain outside of financial services are listed. Since many are only nominally connected to Bitcoin protocol, the term Blockchain 2.0 is used to group these ideas together. IBM has introduced a protocol for smart contracts that is based on the underlying Blockchain technology. IBM is also trying to get the same technology to work with currencies besides Bitcoin. Another area that stands to benefit from Blockchains is the auditing profession. Using a Blockchain the accounting entries between two trading partners can easily be compared while maintaining data privacy. This solution could significantly reduce the reliance on auditors for testing financial transactions.

Blockchains are being examined as a means for handling loyalty-points programs. Others are examining Blockchains as an effective way to validate information about luxury goods. Similarly, vendors of tickets to events are looking at using Blockchains to help prevent fraud. The healthcare sector will be a big user of Blockchains. Storing patient data securely and accurately is a major concern of all health care providers. It is strongly possible that the public sector will become a large user of Blockchains. Several municipalities are looking at Blockchains for recoding property transactions. Other municipalities are examining using Blockchains for tamper-proof voting records and vehicle registries.

Other forms of distributed ledger consensus are:

- **Ethereum** (www.ethereum.org): it is a decentralized platform that runs smart contracts through applications that run exactly as programmed without any possibility of downtime, censorship, fraud or third party interference. These apps run on a custom built blockchain, an enormously powerful shared global infrastructure that can move value around and represent the ownership of property. This enables developers to create markets, store registries of debts or promises, move funds in accordance with instructions given long in the past (like a will or a futures contract) and many other things that have not been invented yet, all without a middleman or counterparty risk;

- **Ripple** (ripple.com): it is probably the most developed financial service using a Blockchain. It offers a means to make simpler and faster cross-border payments using a distributed approach to the global network. Ripple connects banks, payment providers, digital asset exchanges and corporates via RippleNet to provide one frictionless experience to send money globally;

- **Hyperledger** (www.hyperledger.org): Hyperledger is an open source collaborative effort created to advance cross-industry blockchain technologies. It is a global collaboration, hosted by The Linux Foundation, including leaders in finance, banking, Internet of Things, supply chains, manufacturing and Technology. Hyperledger incubates and promotes a range of business blockchain technologies, including distributed ledger frameworks, smart contract engines, client libraries, graphical interfaces, utility libraries and sample applications. The Hyperledger umbrella strategy encourages the re-use of common building blocks and enables rapid innovation of DLT components;

- **MultiChain**: it is an open platform for blockchain applications. MultiChain helps organizations to build and deploy blockchain applications with speed; it aims to remove perceived problems associated with bitcoin by limiting the visibility of the ledger to certain participants, allowing institutions to set controls on transactions permitted and by forgoing distributed mining. On the subject of privacy, Multichain allows users to set a list of permitted users that can act as nodes that refer information on the network and 'miners' that verify transactions, including a method by which nodes can verify whether other nodes have been approved;

- **Eris**: Eris Industries packages blockchain and smart contract concepts to make them more usable and apply them to the customer project. The best projects for this sort of technology are ideas that need a form of decentralized trust and security.

Blockchains can be classified as public blockchains, private blockchains or consortium blockchains. Bitcoin is an example of a public blockchain, in which all records are visible to the public and everyone can take part in the consensus process. A private blockchain is fully controlled by one organization, with a closed group of known participants, which implies a centralized rather than a decentralized network. A consortium blockchain is partially decentralized, where transactions are validated by a selected set of nodes. Private and consortium blockchains may permission other users to read records in the blockchain. Public blockchains rely on a consensus protocol such as proof-of-work, which ensures that transactions cannot be tampered as long as no single miner controls more than 50% of the network's hash power. Transactions in private or consortium blockchains are editable as long as the major participants have reached an agreement, and hence a strong consensus protocol such as proof-of-work is not required. This reduces security, but improves efficiency and latency.

Focusing on AEGIS, the blockchain technology will be adopted to design and develop the BBF, in particular the micro-contracts by which the AEGIS users could sell or purchase datasets and in general, services.

The AEGIS micro-contracts will be smart contracts as almost like a blockchain-based vending machine. These contracts, in fact, are made with a set of rules (the blockchain verifies the execution of performance related to the laws) evaluated by an automated system (validation step) that implements terms of multiparty agreements. In other words one side chooses to perform an action (puts in coins) and the machine verifies that performance and responds (dispenses item and change) providing a cryptographic mechanisms for integrity. Before blockchain technology, this type of smart contract was impossible because parties to an agreement of this sort, would maintain separate databases. With a shared database running a blockchain protocol, the smart contracts auto-execute, and all parties validate the outcome instantaneously and without need for a third-party intermediary.

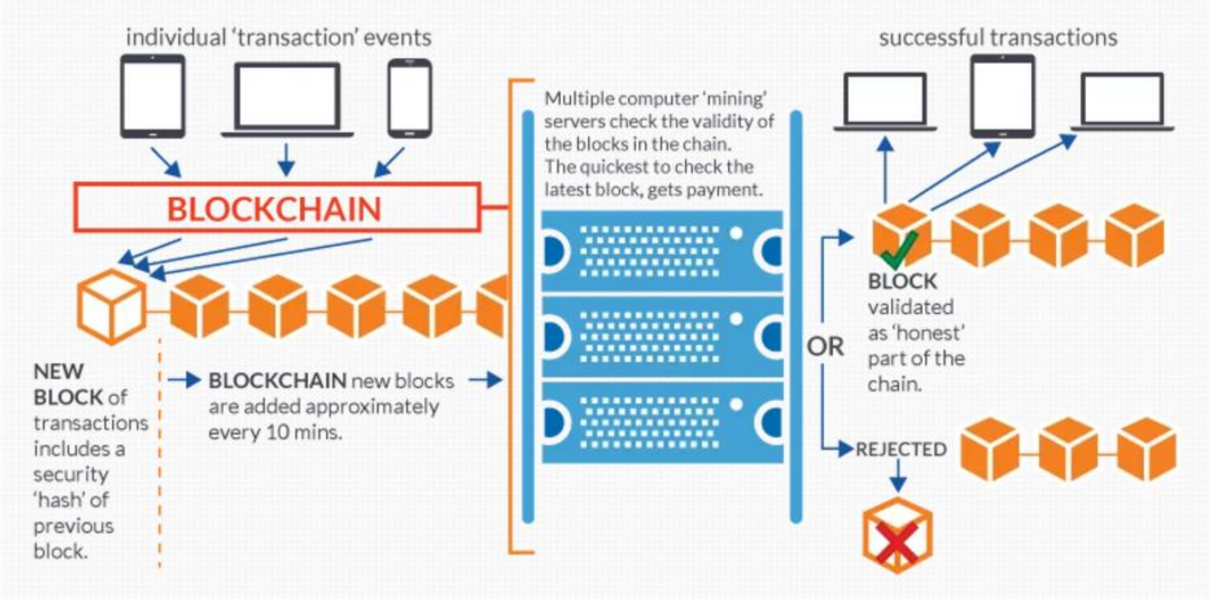

individual 'transaction' events
successful transactions
BLOCKCHAIN
Multiple computer 'mining' servers check the validity of the blocks in the chain. The quickest to check the latest block, gets payment.
NEW BLOCK of transactions includes a security 'hash' of previous block.
BLOCKCHAIN new blocks are added approximately every 10 mins.
OR
BLOCK validated as 'honest' part of the chain.
REJECTED

Figure 4-5: Description of how Blockchain work (http://blockchain.open.ac.uk/)

As the figure above shows (Figure 4-5), one of the key point of blockchain technologies is their ability to provide a secure source of truth, that could be applied to smart contracts, with automated approvals, calculations, and other transacting activities that are prone to lag and error.

The following table (Table 4-5) resumes the blockchain-based smart contracts benefits.

Table 4-5: Blockchain-based smart contracts benefits

Blockchain-based smart contract benefits	Description
Speed and real-time updates	Using software code to automate tasks, the speed of a wide variety of business processes increases.
Accuracy	The risk of manual error decreases.

Lower execution risk	The decentralized process of execution virtually eliminates the risk of manipulation, non-performance, or errors, since the network rather than an individual party manage execution automatically.
Limited failure risk	Since the blockchain technology uses a peer-to-peer network, if there is a failure in any node, the other nodes will continue to operate, maintaining the system's availability.
Fewer intermediaries	The reliance on third-party intermediaries that provide “trust” services such as escrow between counterparties could be reduced or eliminated
Lower cost	New processes enabled by smart contracts requiring less human intervention and fewer intermediaries will therefore reduce costs.
New business or operational models	Because smart contracts provide a low-cost way of ensuring that the transactions are reliably performed as agreed upon, they will enable new kinds of businesses, from peer-to-peer renewable energy trading to automated access to vehicles and storage units.
Auditability and trust	All transactions on the Blockchain are visible to all its participants, with the corresponding increase in auditability and trust. In the meanwhile, changes to the Blockchain are extremely difficult and in the very rare case such a change occurred, it would be visible to the other users.

4.2.1.2 Virtual currency approach

The overall AEGIS brokerage service will simulate a virtual currency approach using cutting edge blockchain technology, in order to validate transactions and showcase how the platform can include monetisation services that may be put in place after the end of the project for compensating data providers.

Currency transactions between persons or companies often use a centralized transaction system and all data and information are controlled and managed by a third party organization, rather than the two principal entities involved in the transaction. Making a digital payment or currency transfer requires a bank or credit card provider as a middleman to complete the transaction. In addition, a transaction causes a fee from a bank or a credit card company. The goal of Blockchain technology is to create a decentralized environment where no third party is in control of the transactions and data. The principles of the blockchain technology, widely described in paragraph 3.2.1.1 are the knowledge at the basis of the following.

From a technology perspective, existing monetary systems require paper-based cash or utilizing a private third party service (e.g. Visa, American Express) to send money at distances. From an economic perspective, holders of government issued currencies (e.g. United States Dollar, European Euro) are required to trust centralized authorities that overall monetary valuations will remain stable and that online transfers or holdings cannot be seized.

Many researchers consider cryptocurrencies as the next evolution of money; in their opinion, as many things in our world transition to becoming digital, so will our money. We can distinguish two basic kinds of cyber or digital currencies. Both are virtual currencies but serve different purposes. One is a pure virtual currency normally restricted to controlled environments such as inside of a social network or an online game. This type of digital currency is subject to a centralized authority that controls the supply of the digital currency. It can still be used to purchase items, but normally only within the confines of the centralized authority. Amazon Coins is an example of this digital currency. This type of digital currency does not use a Blockchain system since the validation is derived from the issuing entity.

The advent of cryptographic digital money has leapfrogged over this archaic system by using blockchain technologies to create a new truly person-to-person (Peer-to-peer) environment of money transfer. There is no need for a centralized party to control a cryptocurrency, nor is there any type of restrictions or rules of usage. Cryptocurrencies (also called a crypto-asset or crypto money) provide anybody with an internet connection, with global, nearly-instant, and frictionless money. This is possible by using advanced encryption and blockchain technologies to provide a robust and secure network of money management.

Cryptocurrencies use various timestamping schemes to avoid the need for a trusted third party to verify the transactions added to the blockchain ledger. Bitcoin, the most popular cryptocurrency, uses a Proof-of-work scheme, which is also known as “Mining”. Other cryptocurrencies achieve the same result with alternative approaches that are often labelled Consensus Protocols or Consensus Platforms.

Bitcoin is the first cryptocurrency to prove successfully the viability of a cryptographic-backed public money supply that is open to anyone. From a market capitalization point of view and public adoption point of view, bitcoin is currently the most popular cryptocurrency. However, there are close to 1,000 different types of cryptocurrencies currently available on coin market cap, the most popular place to discover and track cryptocurrency prices. Among the many choices available, different cryptocurrencies provide different benefits over others. Some cryptocurrencies such as Litecoin provide faster confirmation times than bitcoin. Newer cryptocurrencies such as Ether refer to themselves as crypto assets and use their native token Ether to power a decentralized virtual machine that can execute peer-to-peer smart contracts. It is not possible to define “the best cryptocurrency”, as it depends on what the user intend to use it for.

The three main cryptocurrencies that may prove a relevant inspiration for the simulated virtual currency approach of the AEGIS BBF are **Bitcoin**, **Litecoin** and **Primecoin**.

<u>Bitcoin</u>

An unknown programmer, or a group of programmers, under the name Satoshi Nakamoto, invented Bitcoin and released it as open-source software in 2009. As aforementioned (see paragraph 3.2.1.1) Bitcoin is probably the most common application of blockchain since principal theories of blockchain architectures used today were first outlined and defined in the original bitcoin white paper written and published by Satoshi Nakamoto in 2008. Bitcoin is still using the original Blockchain to record transactions. Currently, it seems that Bitcoin has by far the largest market share, used to purchase a myriad of goods and services with choices of these products and services ever expanding. Therefore, it is highly possible that Bitcoin is important as one of the future research topics, and it will attract industry and academia to conduct more research from both business and technical perspectives. Since the Blockchain technology related to Bitcoin is relatively recent[14], there are some technical challenges and limitations still open:

Throughput: the potential throughput of issues in the Bitcoin network is currently maximized to 7tps (transactions per second). Other transaction processing networks are VISA (2,000tps) and Twitter (5,000tps). When the frequency of transactions in Blockchain increases to similar levels, the throughput of the Blockchain network needs to be improved.

Latency: to create sufficient security for a Bitcoin transaction block, it takes currently roughly 10 minutes to complete one transaction. To achieve efficiency in security, more time has to be spent on a block, because it has to outweigh the cost of double spending attacks. Double-spending is the result of successful spending of money more than once. Bitcoin protects against double spending by verifying each transaction added to the block chain, to ensure that the inputs for the transaction have not been spent previously. This makes latency a big issue in Blockchain currently. Making a block and confirming the transaction should happen in seconds, while maintaining security. To complete a transaction e.g. in VISA takes only a few seconds, which is a huge advantage compared to Blockchain.

Size and bandwidth: the size of a Blockchain in the Bitcoin network is over 50,000MB (February 2016). When the throughput increases to the levels of VISA, Blockchain could grow 214PB in each year. The Bitcoin community assumes that the size of one block is 1MB, and a block is created every ten minutes. Therefore, there is a limitation in the number of transactions that can be handled (on average 500 transaction in one block). If the Blockchain needs to control more transactions, the size and bandwidth issues have to be solved.

Security: the current Blockchain has a possibility of a 51% attack. In a 51% attack a single entity would have full control of the majority of the network's mining hash-rate and would be able to manipulate Blockchain. To overcome this issue, more research on security is necessary.

Wasted resources: mining Bitcoin wastes huge amounts of energy ($15million/day). The waste in Bitcoin is caused by the Proof-of-Work effort. There are some alternatives in industry fields, such as proof-of-stake. With Proof-of-Work, the probability of mining a block depends on the work done by the miner. However, in Proof-of-Stake, the resource that is compared is the amount of Bitcoin a miner holds. For example, someone holding 1% of the Bitcoin can mine 1% of the "Proof-of-Stake blocks". The issue with wasted resources needs to be solved to have more efficient mining in Blockchain.

Usability: the Bitcoin API for developing services is difficult to use. There is a need to develop a more developer-friendly API for Blockchain. This could resemble REST APIs.

Versioning, hard forks, multiple chains: a small chain that consists of a small number of nodes has a higher possibility of a 51% attack. Another issue emerges when chains are split for administrative or versioning purposes. Overall, Blockchain as a technology has the potential to change the way how transactions are conducted in everyday life. Anonymity, data integrity and security attributes set a lot of interesting challenges and questions that need to be solved and assessed with high quality research. Scalability is also an issue that needs to be solved for future needs.

[14] Before 2013 there were not significantly publications about Blockchain [Jesse Yli-Huumo et al., *Where Is Current Research on Blockchain Technology?—A Systematic Review,* PLoS ONE11(10):e0163477.doi:10.1371/journal.pone.0163477]

Litecoin

Litecoin is probably the second most widely known cryptocurrency. While Litecoin is based on the same protocol as Bitcoin, its method of validation is designed to be much easier to use since it uses a simpler algorithm designed for Linux backup systems. This simplified approach enhances the user's ability to maintain the validation process. It also generates a richer reward in a shorter period, while requiring the use of simpler resources. Most importantly, it is able to generate blocks at a faster rate than the system behind Bitcoin. The simpler and quicker method for generating a Blockchain may make their system more valuable to others seeking to apply Blockchains.

Primecoin

Primecoin also has a different protocol for proving the validity of transactions. Primecoin validates its transactions by finding long chains of prime numbers, known as Cunningham chains. It also has the advantage of generating blocks faster than Bitcoin. Both alternative cryptocurrencies may have Blockchain protocols that could potentially benefit firms trying to use Blockchain for their endeavors more than the Bitcoin protocols.

Future research development

Because the size of the current Blockchain applications is quite small, the focus of the researchers is not too high as demonstrates the low number of high quality publications in journal level publication channels. From now on, if Blockchain solutions are used by tens of millions of people and the number of transactions is drastically multiplied, more research on e.g. size, latency and bandwidth, and wasted resources needs to be carried out to guarantee scalability. Another research gap is the lack of research on usability especially from the perspective of developers. For instance, the problem of using Bitcoin API has not been tackled yet. This requires to be studied and improved in the future. This could lead to more applications and solutions to the Bitcoin environment. The third research gap is that the majority of current research is carried out in the Bitcoin environment, rather than in other Blockchain environments. Research on e.g. smart contracts requires to be carried out to increase knowledge outside cryptocurrencies. Despite the fact that Blockchain was first introduced in the cryptocurrency environment, the same idea can be used in many other environments. As a consequence, it is essential to conduct research on using Blockchain in other environments, because it can unveil and produce better models and possibilities for carrying out transactions in different industries.

4.2.2 AEGIS BBF

In AEGIS, as sharing, utilisation, and exploitation of data-related assets is amongst the core aims that would lead to novel business opportunities and would strengthen the data value chain concept of the project, a lightweight Business Brokerage Framework (BBF) has been designed in order to formally dictate transaction terms and oversee the smooth and rightful execution of them. In general, the AEGIS BBF is thought of acting as a supervisor method in all asset operations performed over the platform, in case a user selects an asset that is not owned by him originally (e.g., has not been uploaded by himself onto the platform).

In principle, data sharing and exchange is being performed without a supervisor, as long as there is no monetary transaction and as long as both parties (the "seller" and the "customer") respect certain rules, as those implied by the licenses of the assets they will exchange (for example the license might prohibit the "customer" to extend an asset, or it might oblige the "seller" to provide certain guarantees regarding the asset's functionalities) as defined with the help of the AEGIS DPF. However, this practise is solely based on the good intentions and will of both parties and does not generally meet the criteria of building a "trusted" environment of asset exchange. In this respect, the AEGIS BBF comes as a methodology used to strengthen trust between parties over the whole AEGIS value chain, and create an undisputable ledger of transactions that is really essential when talking about a platform to become a prominent place for designing and deploying business-ready services in the public safety and personal security domain.

The Business Brokerage Framework (BBF) that is proposed concerns different assets that can be uploaded and then shared over the AEGIS platform, with the most popular of those being:

- Data Artefacts. Those are datasets or other data carrying structures (thus also data APIs) which are used to port data into the platform to be used by the different users.
- AEGIS Third Party Micro-Services. These refer to micro-services (such as data cleansing methods, etc.) which are uploaded (or constructed) on the AEGIS platform by other users, and are offered to the community to facilitate the needs of different target groups that work with data and are in need of specific tools and methods that are able to manage these data.
- Data Analysis Algorithms. Algorithms used for data analysis, that have been specifically designed for certain purposes, or that are based on default algorithms, specifically fine-tuned or/and trained to match the needs of the domain.
- Analytics Outputs and Visualisation. Outputs of analyses that are already developed by other users and that can be re-used by other AEGIS users.
- Combinations of the above.

As identified in the previous sections (under 4.2.1), blockchain offers a very modern method for implementing business brokerage engines, especially when transactions are amongst different peers and when contracts can be to an extent be signed and executed in an electronic manner. As such, the AEGIS BBF is envisioned to be supported by a blockchain implementation that will allow the different users to perform transactions (whether these include a type of payment or not) and all comply to the same rules.

What is of importance for the implementation of the proposed solution is the definition of various nodes within the AEGIS ecosystem that will play the role of the miners and that will hold the distributed ledger of the transactions. As AEGIS will be a central infrastructure, and most users will access it through its web interface, it becomes obvious that users themselves cannot play the role of the "miners". This would be possible if as "users" we consider machines that interact via APIs with the AEGIS platform, so nodes could be placed in such machines, providing them also certain benefits, such as lower fees. However, as this scenario is not evident at this point, the consortium will need to create a distributed blockchain infrastructure to initially support the operation of the platform, and gradually insert new, user-hosted nodes into this.

The AEGIS BBF is also closely related with the AEGIS DPF discussed previously in section 4.1.2, as the DPF acts as a reference point for the BBF, towards identifying important assets' characteristics, which are essential for a transaction (such as the originators, the licenses, the pricing schemes, etc.). In line with this, the BBF should be implemented as a layer that oversees all transactions performed over the platform, and is also closely related with all I/O interfaces of the platform (for example with the methods to select datasets, newly third party uploaded algorithms, etc.).

The following figure provides an illustrative presentation of the BBF framework that can be applied to the AEGIS platform and be implemented with the help of blockchain technology.

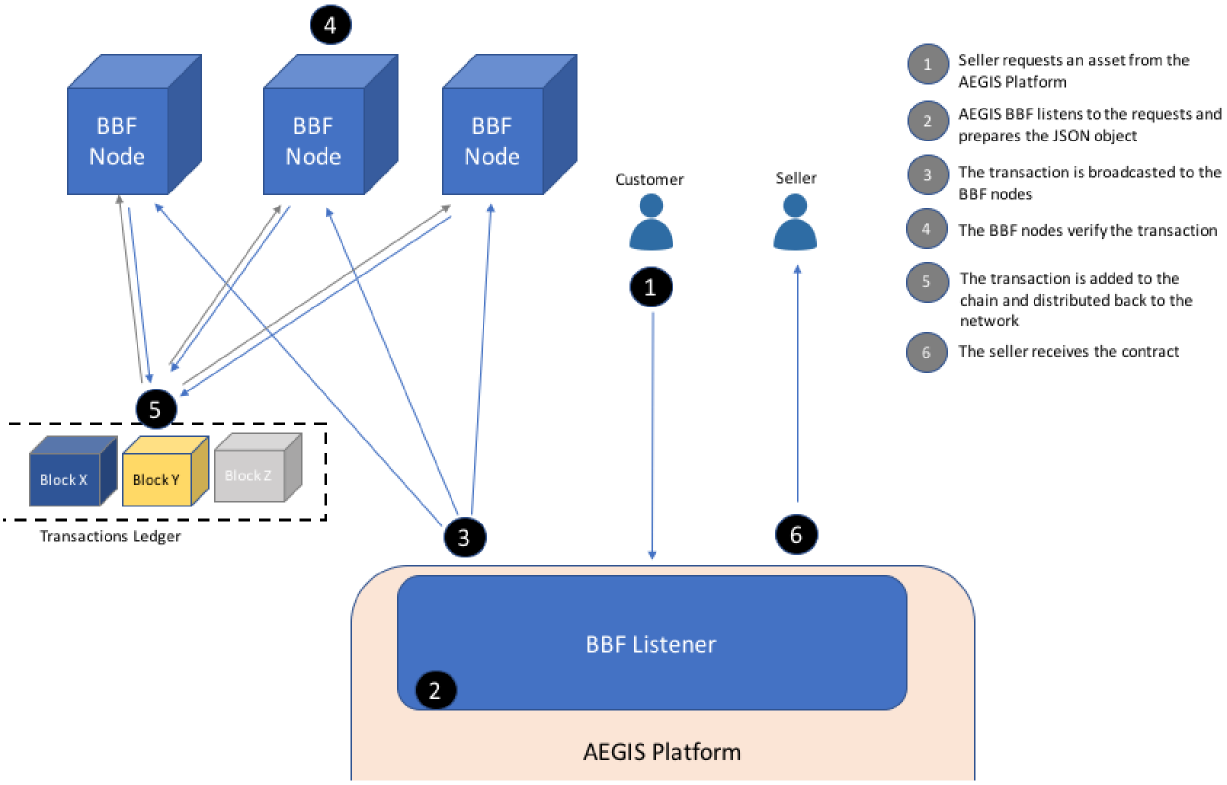

4
BBF Node
BBF Node
BBF Node
Customer
Seller
1
5
Block X
Block Y
Block Z
Transactions Ledger
3
6
BBF Listener
2
AEGIS Platform
1 Seller requests an asset from the AEGIS Platform
2 AEGIS BBF listens to the requests and prepares the JSON object
3 The transaction is broadcasted to the BBF nodes
4 The BBF nodes verify the transaction
5 The transaction is added to the chain and distributed back to the network
6 The seller receives the contract

Figure 4-6: AEGIS BBF Concept

Taking into account the underlying approach of the blockchain protocol (described under section 4.2.1) and the data trust and security aspects (as elaborated in sections 4.1.1.2 and 4.1.1.3), the AEGIS BBF contains a small set of rules that are in place to better describe each type of transaction and offer a mutual understanding about the code of conduct of each transaction amongst the involved parties. Those rules are the following:

#1 Each transaction is unique and can be identified as unique
#2 Each transaction should have one, and only one seller
#3 Each transaction should have one, and only one customer
#4 Each transaction refers to one, and only one asset
#5 Each transaction can be automatically, or manually executed, based on the asset under transaction
#6 Each transaction is accompanied by the date and time that it has been executed. In case of automatic transaction, time is specified by the system based on the system request. In case of manual transactions, the transaction's data and time is considered as the time

of acceptance of the transaction by a “seller”

#7 Each transaction is accompanied by a contract document specifying certain aspects for the transaction

#8 Each transaction is public

#9 The specifics of each transaction contract may not be disclosed publicly

#10 No other parties gain any benefit from a transaction that they are not part of

The rules identified above are mostly “guidelines” and suggest what users should expect from the BBF that is present over the platform. With regards to the benefits for the “seller”, these can be described when publicly setting an asset, resulting into the creation of automatically executed transactions (in case licensing and pricing is straightforward), or in manual transaction (where there has to be an exchange of documents between both parties until they finally agree).

Each transaction to be performed over the platform can be described with the following JSON object in the BBF.

```
{"AEGISBBFTransaction": {

  "transactionid": idoofTransaction,

  "datetime": "TimetampofTransaction",

 "executiontype": "Automatic/Manual",

 "assetconcerned": "AssetURI",

  "parties": {

          {"seller": PersistentUserIDofSeller},

         {"customer": PersistentUserIDofCustomer}
  }

  "contract": "ContractTemplateURI"
}}
```

Figure 4-7: AEGIS BBF JSON Example

The main elements as expressed in rules #1-#6 are present in the JSON object and are the ones that seal a transaction as valid.

With regards to rule #6, it is noted that the automatic execution of micro-contracts, as those envisioned in AEGIS, can be supported by the provision of contract template documents. As such, the JSON object above contains the “contract” field that is used to point to a template document of that kind. Such document templates should be present in the platform to facilitate rapid execution of baseline transactions, while more complicated transactions could be supported by specific contracts that may be uploaded to the platform by the “seller” in each case. As such, each transaction will be accompanied by such a template document, which will be filled in with the related data for each transaction in alignment with the AEGIS DPF and that would specify the obligations of each party.

As identified above, the AEGIS BBF comes as a lightweight solution that is able to oversee transactions and log them in order to enhance trust to the platform’s operations and more importantly between the different users of the platform. However, certain limitations apply and are not tackled by the project at this stage, as these consider aspects that have to be dealt with at the pilot operation phase, in conjunction with the DPF. Such issues have to do with IPRs and automatic compatibility checking of different licenses, data transportation and usage rights and obligations in the different member states.

Another consideration for the platform in the post-project era is whether it will intervene to change rule #10, in terms of the platform gaining a small percentage commission on each monetary transaction, to keep up with running expenses and invoke a business model for the platform itself. Such a consideration will be discussed over the course of the project and decided in the exploitation activities in WP7.

5. Conclusion

The objective of this deliverable is to constitute the backbone of the **AEGIS approach** towards **semantic vocabularies and metadata repository** along with the preliminary definition of the design of the core AEGIS methods to be applied for the **Data Policy and the Business Brokerage Frameworks**.

AEGIS will adopt the DCAT Application profile (DCAT-AP) version 1.1 specification as the basis for AEGIS datasets metadata. In perspective, this will increase interoperability with European open data portals. The DCAT-AP specification cannot cover all AEGIS needs. Therefore the document proposes some additional vocabularies for describing structural, semantic and syntactic metada as well as domain vocabularies.

The creation of the AEGIS domain vocabularies and their own repository started from the identification of the requirements, in terms of semantic vocabularies and metadata, leading to a well-defined list of semantic vocabularies able to describe datasets or data sources concerning different categories to exploit multi-disciplinary information for Public Safety and Personal Security services. Concerning the AEGIS Vocabulary Repository, the main requirements are the ease of use (easy to find and use), the ability of interlinking with existing datasets or new ones, and querying the overall system for extra information. We will exploit the Linda Workbench infrastructure, enhancing its features in order to join the AEGIS needs, for instance a notification system for ontologies changing tracking or the possibility for the user to enrich the initial repository contents with useful metadata, as well as allowing users to add more contents to the repository.

The scope of the Data Policy and Business Brokerage Frameworks is to provide to the AEGIS stakeholders a mean for secure data sharing through licenses/policies that are encapsulated into smart contracts, ensuring data quality and within the respect of the legislation. The DPF, in particular, will be responsible of the definition of the terms according to which a data asset can be used, specifying data quality, clarifying the liability of data asset providers and consumers and providing delivery and payment terms. Both these frameworks defined in our first concept have to work in a semi-automatic way, monitoring data contracts at runtime.

In the next steps, the actual concepts of DPF and BBF will be implemented in the AEGIS platform and will be further expanded and complemented with concrete examples in its final iteration. The data-value chain of the project will be defined creating a “trusted” environment of asset exchange.

6. REFERENCES

Abdellaoui, S., Nader, F., & Chalal, R. (2017). QDflows: A System Driven by Knowledge Bases for Designing Quality-Aware Data Flows. *J. Data and Information Quality, 8*(3–4), 14:1--14:39. http://doi.org/10.1145/3064173

Alouneh, S., Hababeh, I., Al-Hawari, F., & Alajrami, T. (2016). Innovative methodology for elevating big data analysis and security. *2016 2nd International Conference on Open Source Software Computing (OSSCOM)*. http://doi.org/10.1109/OSSCOM.2016.7863685

Antoniou, A., Korakas, C., Manolopoulos, C., Panagiotaki, A., Sofotassios, D., Spirakis, P., & Stamatiou, Y. C. (2007). Trust-Centered Approach for Building E-Voting Systems. (M. A. Wimmer, J. Scholl, & A. Gronlund, Eds.)*6th International Conference, EGOV 2007*. Regensburg, Germany: Springer Berlin/Heidelberg. http://doi.org/10.1007/978-3-540-74444-3

Batini, C., Palmonari, M., & Viscusi, G. (2014). Opening the Closed World: A Survey of Information Quality Research in the Wild. In L. Floridi & P. Illari (Eds.), *The Philosophy of Information Quality SE - 4* (Vol. 358, pp. 43–73). Springer International Publishing. http://doi.org/10.1007/978-3-319-07121-3_4

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. *Journal of Database Management, 26(1),* 60–82. http://doi.org/10.4018/JDM.2015010103

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer.

Berners-Lee, T. (2006). Design issues: Linked data.

Bertino, E. (2015). Big Data - Security and Privacy. In *2015 IEEE International Congress on Big Data* (pp. 757–761). http://doi.org/10.1109/BigDataCongress.2015.126

Bertino, E., & Sandhu, R. (2005, March). Database security-concepts, approaches, and challenges. *IEEE Transactions on Dependable and Secure Computing*. http://doi.org/10.1109/TDSC.2005.9

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63(6),* 1059–1078. http://doi.org/10.1002/asi.22634

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Inform Commun Soc, 15*. http://doi.org/10.1080/1369118X.2012.678878

Cao, Q. H., Khan, I., Farahbakhsh, R., Madhusudan, G., Lee, G. M., & Crespi, N. (2016). A trust model for data sharing in smart cities. 2016 *IEEE International Conference on Communications (ICC)*. http://doi.org/10.1109/ICC.2016.7510834

Chignard, S. (2013). A brief history of Open Data. Retrieved March 17, 2017, from http://parisinnovationreview.com/2013/03/29/brief-history-open-data/

Corradini, F., De Angelis, F., Ippoliti, F., & Marcantoni, F. (2015). A Survey of Trust Management Models for Cloud Computing. In *CLOSER* (pp. 155–162).

Debattista, J., Auer, Sö., & Lange, C. (2016). Luzzu&Mdash;A Methodology and Framework for Linked Data Quality Assessment. *J. Data and Information Quality, 8(1),* 4:1--4:32. http://doi.org/10.1145/2992786

Dumbill, E. (2013). Making Sense of Big Data (Editorial). *Big Data*, 1(1), 1–2. http://doi.org/10.1089/big.2012.1503

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., … Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523–1545. http://doi.org/10.1002/asi.23294

Fung, B., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR), 42*(4), 14.

Gaber, T. (2013). Digital rights management: Open issues to support E-commerce. *E-Marketing in Developed and Developing Countries, IGI Global*, 69–87.

Gabriele Piccoli, & Federico Pigni. (2013). Harvesting External Data: The Potential of Digital Data Streams. *MIS Quarterly Executive*, 12(1), 143–154.

Geisler, S., Quix, C., Weber, S., & Jarke, M. (2016). Ontology-Based Data Quality Management for Data Streams. *J. Data and Information Quality,* 7(4), 18:1--18:34. http://doi.org/10.1145/2968332

H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, & C. Cheng. (2016). A survey of security and privacy in big data (pp. 268–272). http://doi.org/10.1109/ISCIT.2016.7751634

Hart, P., & Saunders, C. (1997). Power and Trust: Critical Factors in the Adoption and Use of Electronic Data Interchange. *Organization Science, 8*(1), 23–42. Retrieved from http://www.jstor.org/stable/2635226

Hartig, O. (2009). Querying Trust in RDF Data with tSPARQL BT - The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009 Heraklion, Crete, Greece, May 31–June 4, 2009 Proceedings. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, … E. Simperl (Eds.), (pp. 5–20). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-02121-3_5

Haryadi, A. F., Hulstijn, J., Wahyudi, A., Voort, H. van der, & Janssen, M. (2016). Antecedents of big data quality: An empirical examination in financial service organizations. *2016 IEEE International Conference on Big Data* (Big Data). http://doi.org/10.1109/BigData.2016.7840595

Hofheinz, P., & Osimo, D. (2017). *Making Europe a Data Economy: A New Framework for Free Movement of Data in the Digital Age.*

Karwa, V., Raskhodnikova, S., Smith, A., & Yaroslavtsev, G. (2014). Private Analysis of Graph Structure. *ACM Trans. Database Syst., 39*(3), 22:1--22:33. http://doi.org/10.1145/2611523

Kostkova, P., Brewer, H., de Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., … Tooke, J. (2016). Who Owns the Data? Open Data for Healthcare. *Frontiers in Public Health,* 4, 7. http://doi.org/10.3389/fpubh.2016.00007

Ku, W., & Chi, C.-H. (2004). Survey on the Technological Aspects of Digital Rights Management BT - Information Security: 7th International Conference, ISC 2004, Palo Alto, CA, USA, September 27-29, 2004. Proceedings. In K. Zhang & Y. Zheng (Eds.), (pp. 391–403). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-30144-8_33

L. L. Win, T. Thomas, & S. Emmanuel. (2012). Privacy Enabled Digital Rights Management Without Trusted Third Party Assumption. *IEEE Transactions on Multimedia, 14*(3), 546–554. http://doi.org/10.1109/TMM.2012.2189983

Liow, M. L. F., & Lee, H. Y. (2016). Instilling Trust and Confidence in a Data Driven Economy - Preliminary Design of a Data Certification Scheme. *2016 International Conference on Cloud Computing Research and Innovations (ICCCRI)*. http://doi.org/10.1109/ICCCRI.2016.31

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 134–142. http://doi.org/http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

Lundqvist, B. (2016). *Big Data, Open Data, Privacy Regulations, Intellectual Property and Competition Law in an Internet of Things World- The Issue of Access* (Stockholm Faculty of Law Research Paper Series No. 1).

Mattioli, M. (2014). Disclosing Big Data. *Minn. L. Rev.*, *99*, 535.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Trans. Manage. Inf. Syst.*, *2*(2), 12:1--12:25. http://doi.org/10.1145/1985347.1985353

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A Data Quality in Use model for Big Data. *Future Generation Computer Systems*, 63, 123–130. http://doi.org/http://dx.doi.org/10.1016/j.future.2015.11.024

Mohan, P., Thakurta, A., Shi, E., Song, D., & Culler, D. (2012). GUPT: Privacy Preserving Data Analysis Made Easy. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 349–360). New York, NY, USA: ACM. http://doi.org/10.1145/2213836.2213876

Monir, M. B., AbdelAziz, M. H., AbdelHamid, A. A., & EI-Horbaty, E. S. M. (2015). Trust management in cloud computing: A survey. *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*. http://doi.org/10.1109/IntelCIS.2015.7397227

Moritz, D., Fisher, D., Ding, B., & Wang, C. (2017). Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2904–2915). New York, NY, USA: ACM. http://doi.org/10.1145/3025453.3025456

Murthy, P., Bharadwaj, A., Subrahmanyam, P. A., Roy, A., & Rajan, S. (2014). *Big Data Taxonomy. Big Data Working Group*. Cloud Security Alliance (CSA).

Obama, B. (2009). Transparency and open government. Memorandum for the heads of executive departments and agencies. Retrieved April 7, 2016, from https://www.whitehouse.gov/open/documents/open-government-directive

OECD. (2014). *Data-driven Innovation for Growth and Well-being - INTERIM SYNTHESIS REPORT October 2014.*

Open Knowledge International. (2017a). How to Open up Data. Retrieved March 21, 2017, from http://opendatahandbook.org/guide/en/how-to-open-up-data/

Open Knowledge International. (2017b). What is Open Data? Retrieved March 20, 2017, from http://opendatahandbook.org/guide/en/what-is-open-data/

Ortiz, P., Lázaro, O., Uriarte, M., & Carnerero, M. (2013). Enhanced multi-domain access control for secure mobile collaboration through Linked Data cloud in manufacturing. *2013 IEEE 14th International*

Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). http://doi.org/10.1109/WoWMoM.2013.6583372

Pellegrini, T. (2012). Integrating Linked Data into the Content Value Chain: A Review of News-related Standards, Methodologies and Licensing Requirements. In *Proceedings of the 8th International Conference on Semantic Systems* (pp. 94–102). New York, NY, USA: ACM. http://doi.org/10.1145/2362499.2362513

S. Lee, H. Park, & J. Kim. (2010). A secure and mutual-profitable DRM interoperability scheme (pp. 75–80). http://doi.org/10.1109/ISCC.2010.5546755

Sacco, O., & Passant, A. (2011). A privacy preference ontology (PPO) for linked data. In *LDOW2011 March 29, 2011, Hyderabad, India*.

Sacco, O., Passant, A., & Decker, S. (2011). An Access Control Framework for the Web of Data. *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*. http://doi.org/10.1109/TrustCom.2011.59

Sayah, T., Coquery, E., Thion, R., & Hacid, M.-S. (2016). Access Control Enforcement for Selective Disclosure of Linked Data BT - Security and Trust Management: 12th International Workshop, STM 2016, Heraklion, Crete, Greece, September 26-27, 2016, Proceedings. In G. Barthe, E. Markatos, & P. Samarati (Eds.), (pp. 47–63). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-46598-2_4

Shankaranarayanan, G., & Blake, R. (2017). From Content to Context: The Evolution and Growth of Data Quality Research. *J. Data and Information Quality*, 8(2), 9:1--9:28. http://doi.org/10.1145/2996198

Shapiro, C. (2000). *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting. Innovation Policy and the Economy* (Vol. 1). http://doi.org/10.2139/ssrn.273550

Sherchan, W., Nepal, S., & Paris, C. (2013). A Survey of Trust in Social Networks. *ACM Comput. Surv.*, *45*(4), 47:1--47:33. http://doi.org/10.1145/2501654.2501661

Simon, S. J. (2016). Trust and distrust as distinct constructs: Evidence from data theft environments. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. http://doi.org/10.1109/ISI.2016.7745461

Sicari, S., Cappiello, C., De Pellegrini, F., Miorandi, D., & Coen-Porisini, A. (2016). A security-and quality-aware system architecture for Internet of Things. I*nformation Systems Frontiers*, 18(4), 665–677. http://doi.org/10.1007/s10796-014-9538-x

Sicari, S., Rizzardi, A., Miorandi, D., Cappiello, C., & Coen-Porisini, A. (2016). A secure and quality-aware prototypical architecture for the Internet of Things. *Information Systems*, 58, 43–55. http://doi.org/http://dx.doi.org/10.1016/j.is.2016.02.003

Song, M. (2017). Trust-based Business Model in Trust Economy: External Interaction, Data Orchestration and Ecosystem Value. *Proceedings of The 9th MAC 2017*, 192.

Sookhak, M., Gani, A., Khan, M. K., & Buyya, R. (2017). Dynamic remote data auditing for securing big data storage in cloud computing. *Information Sciences*, 380, 101–116. http://doi.org/http://dx.doi.org/10.1016/j.ins.2015.09.004

The Economist. (2017). Data is giving rise to a new economy. Retrieved July 18, 2017, from https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy

The Open Data Institute. (2017). The Data Spectrum helps you understand the language of data. Retrieved March 20, 2017, from https://theodi.org/data-spectrum

Thomas, L. D. W., & Leiponen, A. (2016). Big data commercialization. *IEEE Engineering Management Review, 44*(2), 74–90. http://doi.org/10.1109/EMR.2016.2568798

Vare, T., & Mattioli, M. (2014). Big Business, Big Government and Big Legal Questions. Retrieved July 18, 2017, from http://www.managingip.com/IssueArticle/3382483/Archive/Big-business-big-government-and-big-legal-questions.html

Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges BT - Handbook of Big Data Technologies. In A. Y. Zomaya & S. Sakr (Eds.), (pp. 851–895). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-49340-4_25

Yamasaki, S. (2011). A Trust Rating Method for Information Providers over the Social Web Service: A Pragmatic Protocol for Trust among Information Explorers and Information Providers. *2011 IEEE/IPSJ International Symposium on Applications and the Internet*. http://doi.org/10.1109/SAINT.2011.110

Yin, C., Wang, J., & Park, J. H. (2017). An improved recommendation algorithm for big data cloud service based on the trust in sociology. *Neurocomputing*, *256*, 49–55. http://doi.org/http://dx.doi.org/10.1016/j.neucom.2016.07.079

Yli-Huumo J, Ko D, Choi S, Park S, Smolander K (2016) Where Is Current Research on Blockchain Technology?ÐA Systematic Review. PLoS ONE 11(10): e0163477. doi:10.1371/journal.pone.0163477

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. http://doi.org/10.1002/asi.23730

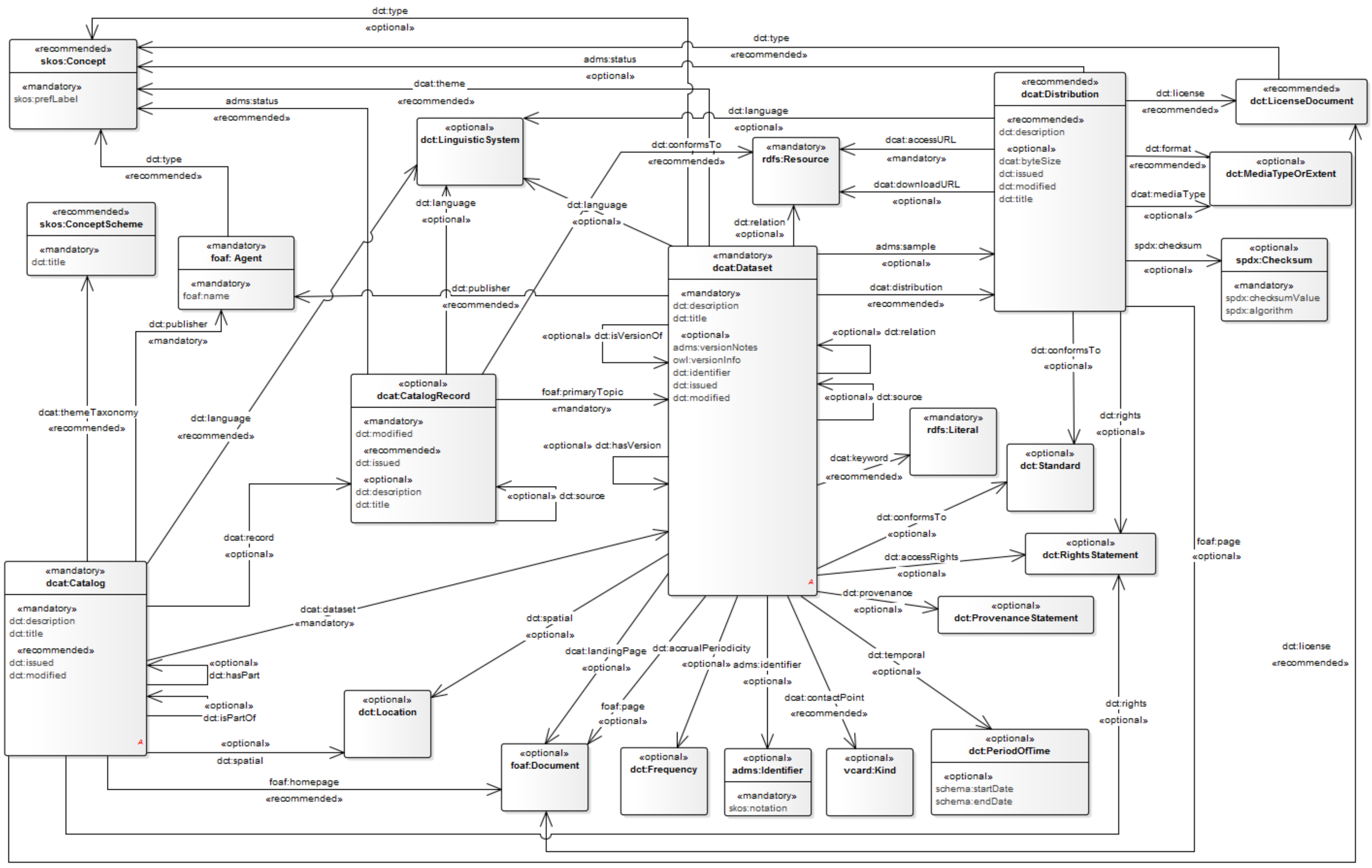
Zafar, F., Khan, A., Suhail, S., Ahmed, I., Hameed, K., Khan, H. M., … Anjum, A. (2017). Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes. *Journal of Network and Computer Applications, 94*, 50–68. http://doi.org/http://dx.doi.org/10.1016/j.jnca.2017.06.003

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality Assessment for Linked Data: A survey. *Semantic Web Journal*, *7*(1), 63–93. Retrieved from http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data

Zibin Zheng, Shaoan Xie, Hongning Dai, Xiangping Chen, and Huaimin Wang (2017). An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends. *2017 IEEE 6th International Congress on Big Data, DOI 10.1109/BigDataCongress.2017.85*

Zhu, T., Li, G., Zhou, W., & Yu, P. S. (2017). Differentially Private Data Publishing and Analysis: a Survey. *IEEE Transactions on Knowledge and Data Engineering*, *14*(8), 1–1. http://doi.org/10.1109/TKDE.2017.2697856

ANNEX 1: DCAT APPLICATION PROFILE UML CLASS DIAGRAM [15]



[15] DCAT Application Profile for data portals in Europe, Version 1.1, https://joinup.ec.europa.eu/system/files/project/dcat-ap_version_1.1_0.pdf