Project Acronym:	**BIG**
Project Title:	Big Data Public Private Forum (BIG)
Project Number:	**318062**
Instrument:	**CSA**
Thematic Priority:	**ICT-2011.4.4**

D4.2.2 Final version of IPR, Standardisation recommendations

Work Package:	*WP4 Big Data Public-Private Forum*	
Due Date:	30/04/2014	
Submission Date:	19/11/2014	
Start Date of Project:	01/09/2012	
Duration of Project:	26 Months	
Organisation Responsible of Deliverable:	DFKI	
Version:	1.1	
Status:	final version	
Author name(s):	Tilman Becker	DFKI
Reviewer(s):	Nelia Lasierra Ricard Munné Nuria De-Lama	UIBK ATOS ATOS
Nature:	☒ R – Report ☐ P – Prototype ☐ D – Demonstrator ☐ O - Other	
Dissemination level:	☒ PU - Public ☐ CO - Confidential, only for members of the consortium (including the Commission) ☐ RE - Restricted to a group specified by the consortium (including the Commission Services)	
Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		

Revision history			
Version	**Date**	**Comments**	**Modified by**
0.1	07/10/2013	Final version of first draft	Tilman Becker (DFKI)
0.2	17/04/2014	Updated content with contributions from Ricard Munne and Martin Strohbach	Tilman Becker (DFKI)
0.3	28/05/2014	Update with contributions from plenary meeting	Tilman Becker (DFKI)
0.4	01/06/2014	Finalized for internal review	Tilman Becker (DFKI)
0.5	13/06/2014	Final version addressing internal review	Tilman Becker (DFKI)
0.6	24/06/2014	Formatting issues	Nelia Lasierra (UIBK), Tilman Becker (DFKI)
1.0	29/09/2014	Finalized enhanced version for internal review before re-submission	Tilman Becker (DFKI)
1.1	18/11/2014	Addressed internal review	Tilman Becker (DFKI)

Executive Summary

This deliverable provides an outline of the aspects of IPR and standardisation that apply to Big Data. It identifies the relevant topics in both areas and how they might be addressed. It focuses on important on-going developments that need support and identifies gaps in standards and IPR-related issues that need to be addressed in the near future. These results are summarised in a number of recommendations for IPR issues and for standardisation.

Issues in IPR revolve about data, focussing on ownership and liability, and much less on technology IPR. Although issues in software and hardware IPR are highly relevant, they are often covered by the considerations that apply generally for IT technology.

Standardisation in Big Data applies to hard- and software technology, data interoperability and finally, benchmarking in multiple fields. This report analyses the technical white papers and the use cases from the sector forums to identify key areas in which standardisation issues arise:

- **Data integration:** Data models on all levels of abstraction, from raw data to semantic interpretations must follow standards in data formats as well as terminologies in order to achieve requirements of interchange-ability and transparency. Future data and service marketplaces can only thrive when such requirements are met.
- **Data security and privacy:** In all steps of the Big Data value chain, issues of data security and privacy arise. These must eventually be met by regulatory means, which in turn will rely on the formulation of appropriate standards. Included in this complex are issues of data provenance and trustworthiness.
- **Frameworks:** Some areas of Big Data suffer from a big diversity of architectures, frameworks, tools and, e.g., query languages. Standardisation has proven helpful where de-facto standards such has Hadoop (for architectures) or Linked Data principles (as a framework for publishing and connecting structured data in the web) have developed.
- **Benchmarking:** In multiple areas, the development of standardised benchmarks can have a huge benefit to objectively compare various tools and approaches and support competition. This includes economic impacts such as energy-efficiency of data storage and analytics hardware solutions.

There are a number standardisation bodies and related organisations that have begun to address Big Data by forming working groups, this report includes a comprehensive overview of existing organisation and activities.

This report indicates which areas need further extension and outlines the sources and inputs needed from within the BIG project and beyond.

Table of Contents

Index of Figures

Index of Tables

No table of figures entries found.

1. Introduction

This deliverable addresses the two related aspects of intellectual property rights (IPR) and standards for Big Data. The following second chapter discusses the developing issues where IPR needs to be considered and where regulations will have to be adapted, clarified or newly created.

The third chapter briefly discusses how standardisation is relevant in Big Data and surveys the relevant organisations and standardisations bodies. In many of those, working groups have been formed and have begun the process of identifying relevant topics, roadmaps and potential standardisation activities. The chapter also includes a list of links and contact points for the on-going activities.

The fourth chapter provides general recommendations for supporting activities that are derived from the analysis in the preceding chapters. Some tools and methods that can be used as part of such supporting activities are sketched.

2. IPR

Intellectual Property Rights (IPR) are relevant in Big Data in two major areas: (i) intellectual property related to the IT technology (hardware and software) and the related business processes and (ii) intellectual property related to data.

Big Data technologies are indeed IT technologies and existing IPR approaches for IT technologies thus apply in Big Data. As the following section 2.1 briefly discusses, existing approaches in IT technologies cover also all the important aspects of Big Data and thus we do not foresee the necessity for special efforts approaches.

Beyond the existing challenges around Intellectual Property Rights (IPR) in IT applications, Big Data puts special emphasis on the ownership, protection, security and liability related to the data itself—beyond the procedures and technologies used to acquire, process, curate, analyse and use the data. Section 2.2 presents an analysis of the IPR challenges for data.

2.1. IPR Challenges in Big Data Technology

The goal of the protection of intellectual property is to foster innovation and ensure a proper financial return on the investments involved. Tools that apply in IT technology include Trademarks, Copyright, Patents, and Trade Secrets.

Existing IPR tools and strategies typically cover innovation in technology, products and business processes and thus cover Big Data technology and business processes. Note that there are limits to the patentability of business processes in IT, see (USPTO, 2010). The literature on IPR in Big Data concentrates on aspects of IPR on the data itself, as discussed in the following section of this document. As (Umeh, 2013) puts it: *"who owns the input data companies are using in their analysis, and who owns the output?"* and *"technology is not really that much a differentiator, rather it is the architecture and infrastructure approach that make all the difference."*

In general, we foresee no necessity for special efforts in IPR approaches in Big Data Technologies. The assertion, assignment and enforcement of copyright, design rights, trademarks and patents are applicable for Big Data technologies and face the same challenges as IT technology in general.

2.2. IPR challenges in Data

Since the amount and variability of data is substantially different from other IT technologies, there are a number of areas where IP related to the data (as opposed to the processing technologies) is relevant and different from existing approaches.

These areas include:

- Data ownership
    - Acquisition of data
    - Changing / Curation of data
- Data privacy and protection of data
- Responsibility and liability for data (quality)
    - Disseminating / Selling data

In all areas, we see the (IP) rights to the data themselves as the new challenge, not the Big Data technologies used to acquire, curate, analyse and process the data.

2.2.1 Data Ownership

Data ownership and the rights to use data are covered by copyright and related contracts valid when acquiring data. For Big Data technologies, it is particularly important to understand when and how further processing of big data sets creates new ownership. The collection, curation, combination with other data sets and eventually analysis of data sets derive new rights to the resulting data that must be asserted and enforced. As an example, see the Geospatial Digital Rights Management Reference Model, (GeoDRM) as discussed in (Korn et al, 2007).

Data Ownership is a particular challenge for Big Data and needs support on various levels:

- National vs. European regulations
- Best practice guidelines in Big Data
- Education and support through experts

The BIG project can be a starting point for the development of best practice guidelines that could be seeded from the BIG PPF (Public-Private Forum).

2.2.1.1 Data ownership in the public sector

In the public sector the main issue related to data ownership is that normally the legislation and applicable regulations do not allow the use of data for purposes other than those regulated and for which the data was collected. Thus, the reuse of public sector data in big data processes has to be carefully checked to be sure it matches the uses allowed by the regulations. This issue is strongly related with data protection in the public sector.

Data ownership among public bodies is also an important issue, strongly related to the previous.

In the inverse direction, there is also a lack of legislation for accessing data not generated by Public Sector, so this creates uncertainty.

2.2.2 Data privacy and security

While data privacy is a huge concern in Big Data, it is not different in principle from other aspects of data privacy and other uses of larger data sets. We thus expect to continuously watch the developments in Europe for data privacy and analyse their impact on Big Data. However, we do not foresee the necessity to contribute to data privacy to address specific Big Data aspects.

A look at “Top 10 Big Data Security & Privacy Challenges” taken from (Cacas, 2013) clarifies this point. The top 10 listed are:

1. Secure computations in distributed programming frameworks
2. Security best practices for non-relational data stores
3. Secure data storage and transactions logs
4. End-point input validation/filtering
5. Real-time security/compliance monitoring
6. Scalable and composable privacy-preserving data mining and analytics
7. Cryptographically-enforced access control and secure communication
8. Granular access control
9. Granular audits

10. Data provenance

All of these Top 10 challenges are derived from general IT challenges. As such, it is imperative to master all of these challenges for the success of Big Data, however, these challenges have a broader impact and will be solved outside of Big Data; with influence and motivation by Big Data at best.

Some of these challenges are discussed in more detail in D2.2.2 on Data Storage. Figure 2-1 shows a finer classification of the 10 challenges.

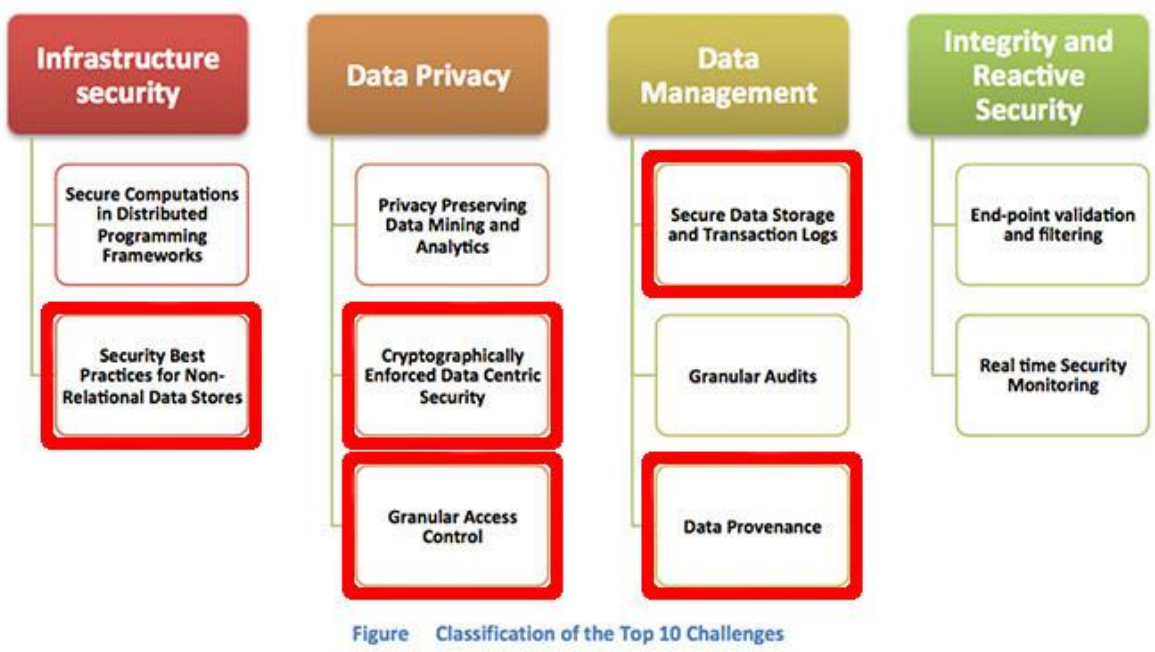

Figure 2-1: CSA Top 10 Security and Privacy Challenges and their relation to Data Storage.

2.2.2.1 Transnational provenance

One notable exception, however, is the growing trend to gather data from international sources and in consequence the need to address data privacy regulations from multiple countries. This ranges from determining the applicable regulations in the first place to the ability to provide varying data privacy guarantees, depending on the nationality of data providers, users, processing, etc.

Large European companies and organisations already face these issues and can contribute through the BIG Public-Private Forum.

2.2.3 Data responsibility and liability

Liability for (negative) consequences of using data derived from Big Data analysis is a big challenge for viability of Big Data in business strategies. As with data ownership, the three main issues are national vs. international laws and regulations, availability of best practice guidelines or even standards and expert services and education. As Big Data encompasses data from many sources (variety), it typically uses data from multiple countries for usage in multiple countries, thus running into the challenge of observing various national and international laws and regulations. Best practices should come from industry bodies and can be seeded from the BIG Public-Private Forum.

2.3. Conclusion

In summary, from a multitude of general IPR issues that are important in any Big Data project, there are three main aspects on intellectual property rights that should be addressed specifically from the point of view of Big Data. They all revolve around the data rather than the processing technologies:

- Data Ownership: raw data and derived data
- Data Privacy: national and international aspects
- Data Responsibility: mapping liability for results from Big Data applications to processing and data sources

Addressing the IPR issues in Big Data is of particular importance for Europe, as a UN study (Falvey et al, 2006) has shown that for most high-income countries, strengthening IPRs raises growth at least partly, due to increased innovation and technology diffusion.

3. Standardisation

Standards play a pivotal role on any market to provide customers with true choice by being able to choose comparable and compatible goods or services from multiple suppliers. In Big Data, this applies to technology and data where technology in turn covers hardware and software. In addition, standards are useful for providing measurements for the results of applying Big Data for ultimate business goals.

This deliverable provides general recommendations regarding standardisation in Big Data and provides an analysis of existing relevant standardisation bodies and their current activities. The analysis is based on the needs and requirements identified in the technical working groups as summarised in sections 3.2 and 3.3. The following sections 3.4 to 3.8 cover various aspects derived from these requirements. The final section 3.9 is a comprehensive collection of on-going activities and an analysis of their relevance. Special attention is given to the relevance of activities to the requirements identified by the technical working groups as well as their relevance to specific sectors or whether they address cross-cutting issues of broad relevance.

As Big Data—besides its revolutionary character in many aspects—is also evolutionary in many other ways, technology development is mainly going ahead without established standards and standardisation is lagging behind. Technologies like Hadoop thus form de-facto standards with limited control. Given the lagging status of standardisation, many organisations have acknowledged the need for standards and formed working groups but most lack critical mass of coverage, partners, and progress. Standardisation activities for special purposes within Big Data or overlapping with Big Data are successfully going on in multiple places, e.g., semantic web standardisation at the W3C. Of the standardisation activities that address Big Data in general, the work at ISO is likely the most successful and should be given special consideration, see the discussion of on-going activities in Section 3.9.

3.1. General Recommendations for Standardisation in Big Data

One aspect of requirements is the perspective and the area of application of standards. We need to view standardisation

- from the perspective of interoperability between Big Data technology providers,
- from the perspective of Big Data users and
- from the perspective of customers of Big Data results.

In line with the context of the work at the ISO/IEC JTC 1 Study Group on Big Data (see section 3.9.1.1 below), we propose the following general recommendations regarding standardisation for Big Data:

1. Use common standards as the basis for an open and successful Big Data market
2. Integrate national efforts on an international (European) level as early as possible
3. Ensure availability of experts for all aspects of Big Data in the standardisation process
4. Provide education and education material to promote developing standards

A global strategy for standards in Big Data will be difficult to follow, as Big Data covers many different technology aspects and is employed in very different market sectors. Nevertheless, ideally, a global strategy should:

1. Follow a modular approach to standards
2. Provide a global view (model) of the relation of these modular standards
3. Ensure maximally formal and certifiable standards

4. Focus on industry standards that address existing practices

In particular a global view, i.e., a formal model of Big Data is a challenge that calls for research to support the Big Data standardisation process. Only with a clear view of the structure of Big Data can individual standards be related to form a coherent big picture.

3.2. Standardisation Issues in the Data Value Chain

The Big Data Value Chain as shown in Figure 3-1: The Big Data Value Chain is one of the perspectives that the BIG project takes on Big Data; the second perspective is a selection of five sector forums, collecting a total of 10 industry sectors that are affected by Big Data. Along the Big Data value chain, the BIG project has identified issues that need to be addressed by standardisation efforts as elaborated in the following sections. Upcoming recommendations and projections of future developments in the roadmaps for the sectors and the consolidated cross-sectorial roadmaps (in deliverables D2.4.2 and D2.5, respectively) will allow a weighing of the standardisation issues with respect to their impact in terms of relevant market size and timeliness.

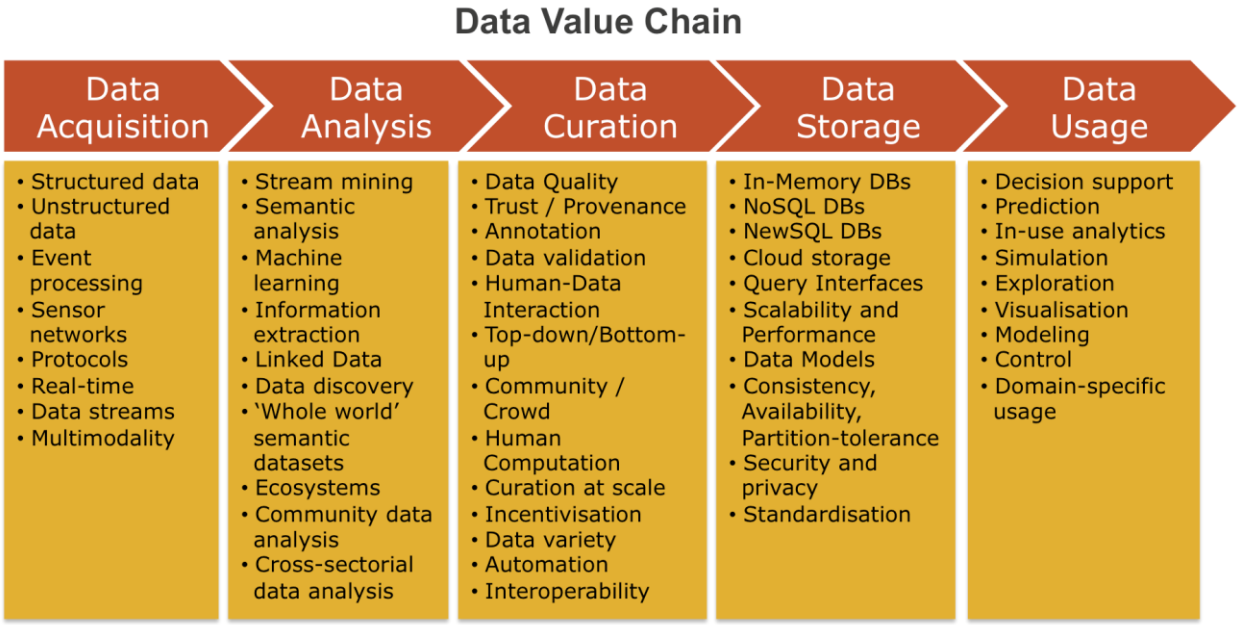

Figure 3-1: The Big Data Value Chain

A critical analysis of the work of the technical working groups of the BIG projects with respect to standardisation issues is summarised in the following sections. Pointers refer to the white papers from the technical working groups, see deliverables D2.2.2.

3.2.1 Data Acquisition

The technical working group on Data Acquisition has analysed a number of industry-leading data acquisition architectures, see the white paper (deliverable D2.2.2) and its analysis of the Big Data architectures of Oracle, Vivisimo, and IBM. There are a large number of frameworks and tools for Data Acquisition, including open source projects, all addressing specific tasks and challenges.

Standardisation could benefit Data Acquisition where it would allow the definition of an encompassing data acquisition framework and the implemented predefined protocols.

3.2.2 Data Analysis

The work of the technical working group on Data Analysis identifies **benchmarking** as an upcoming tool fostering progress in Data Analysis. Benchmarking must have the quality of standards, see e.g., the work by the Linked Data Benchmark Council (LDBC).[1]

The de-facto standard for semantic approaches to Big Data Analysis is **Linked Data** which needs to be developed further to address the important challenges of efficient indexing, entity extraction and classification and the support of search over data found on the web (see section 5.3 of D2.2.2 on Data Analysis).

Semantic approaches to Big Data also address the challenges of multi-lingual data with the prospect of standardised semantic representations and ontologies being language-independent.

3.2.3 Data Curation

One of the key consequences of well-curated data is the often-demanded possibility for data interoperability and the reuse of data. Such demands in eScience and eGovernment are drivers of Data Curation. High quality data-driven models require suitable standards in data formats. Three of the key insights for Data Curation from D2.2.2 that relate to standardisation are:

Data-level trust and permission management mechanisms are fundamental to supporting data management infrastructures for data curation. Provenance management is a key enabler of trust for data curation, providing curators the context to select data that they consider trustworthy and allowing them to capture their data curation decisions. Data curation also depends on mechanisms to assign permissions and digital rights at the data level.

Data and conceptual model standards strongly reduce the data curation effort. A standards-based data representation reduces syntactic and semantic heterogeneity, improving interoperability. Data model and conceptual model standards (e.g. vocabularies and ontologies) are available in different domains. However, their adoption is still growing.

Need for improved theoretical models and methodologies for data curation activities. Theoretical models and methodologies for data curation should concentrate on supporting the transportability of the generated data under different contexts, facilitating the detection of data quality issues and improving the automation of data curation workflows.

Section 7.6 goes into a detailed discussion of standardisation and interoperability, covering general and already existing data model standards such as RDF and the need for standardised terminologies and vocabularies. Specifically, provenance information can be highly relevant in Data Curation to guarantee trustworthiness, see e.g. the W3C PROV standard.

Another noteworthy area in need of interoperability is Natural Language Processing (see the discussion in section 7.8 of D2.2.2 on Data Curation), with the emergence of industrial NLP pipelines such as IBM's Watson and Apache UIMA (Zhu et al., 2014, Götz et al., 2014).

All these challenges fit well into an emerging pattern where standards for privacy and security as well as data models are named as recurring key challenges.

[1] LDBC Homepage, http://www.ldbc.eu/, last visited 13/02/2014

3.2.4 Data Storage

One of the key insights from D2.2.2, the technical white paper on Data Storage is the **lack of standards for NoSQL** data bases, as historically specific technological challenges have been addressed, leading to a wide range of different storage technologies. The large range of choices coupled with the lack of standards for querying the data makes it harder to exchange data stores as it may tie application specific code to a certain storage solution.

The prime need for standardisation arises for NoSQL **query languages**, see the detailed discussion in section 5.1.1 of D2.2.2.

A second challenge identified in D2.2.2 is **privacy and security**. Primarily, this is a technological challenge, calling for required research to better understand how data can be misused, how it needs to be protected and integrated into Big Data storage solutions. However, any solution will need to be supported by standardisation and certification procedures to ensure and guarantee appropriate privacy and security levels.

A similar, third challenge is **open scalability**. Scalability is challenging for graph data models which allow to better capture the semantics and complex relationships with other pieces of information thus improving the overall value that can be generated by analysing the data. Graph databases fulfil such requirements, but are at the same time the least scalable NoSQL databases. Again, any technical solutions will need to be supported by standardisation.

An economic impact of Big Data is the emergence of data platforms such as datamarket.com (Gislason 2013), infochimp.com and Open Data initiatives of the European Union (https://open-data.europa.eu/de/data) and other national portals (e.g. data.gov, data.gov.uk, data.gov.sg, etc.). Also technology vendors are supporting the move towards a data driven economy as can be seen by the positioning of their products and services. The emergence of such data marketplaces will need support from standardised data formats.

Another economic impact of Big Data is that energy consumption becomes an important cost factor with potential negative impact on the environment. According to Koomey (Koomey J.G. 2008) data centres' energy consumption raised from 0.5% of the world's total electricity consumption to 1.3% in 2010. And IDC's 2008 study on the digital universe provided evidence that the costs for power and cooling are raising faster than the costs for new servers (IDC 2008). These increasing costs provide a high incentive to investigate deeper into managing the data and its processing in an energy-efficient way, thus minimizing the impact on the environment. To measure and compare the energy-efficiency of Big Data technology, standards for energy-efficiency such as they are known for consumer products, will be needed.

In summary, the three prime challenges for Data Storage that involve standardisation are:

- NoSQL query languages
- Security and privacy
- Common data models

3.2.5 Data Usage

From the key insights summarised in the white paper on Data Usage (deliverable D2.2.2), one relates specifically to standardisation: Smart Data and Service Integration. The corresponding use cases demonstrate the depth of the corresponding challenges:

To enable the application of smart services to deal with the Big Data Usage problems, there are technical and organisational matters. Data protection and privacy issues, regulatory issues and new legal challenges, e.g., wrt. ownership issues for derived data, must all be addressed.
On a technical level, there are multiple dimensions along which the interaction of services must be enabled: on a hardware level from individual machines to facilities to networks; on a

conceptual level from intelligent devices to intelligent systems and decisions; on an infrastructure level from IaaS to PaaS and SaaS to new services for Big Data Usage and even Business Processes and Knowledge as a Service.

A particularly challenging use case is the data integration needed for Industry 4.0 (industrial internet), where the digital integration of the complete manufacturing chain needs to apply standards from the Internet of Things as well as the Internet of Services.

3.2.6 Criteria for Critical Issues in Standardisation

The following aspects should be taken into consideration when weighing the importance of the issues listed in the previous sections:

- Identified as blocker in technology
- Identified as blocker in (multiple) industrial sectors
- No or immature standardisation efforts so far
- European dimension
- Absence of standards causes high costs
- Absence of standards prevents common solutions/products/markets
- Standardisation involves large number of stakeholders

In addition, the estimates for timelines and market size from the upcoming roadmaps in deliverables D2.4.2 and D2.5 must be taken into consideration when weighing standardisation issues.

3.3. Summary of Standardisation Requirements

Some of the challenges in standardisation are recurring in all steps of Big Data value chain (see Figure 3-1) and thus represent the core of requirements for standardisation in Big Data and are depicted in Figure 3-2:

- **Data integration:** Data models on all levels of abstraction, from raw data to semantic interpretations must follow standards in data formats as well as terminologies in order to achieve requirements of interchange-ability and transparency. Future data and service marketplaces can only thrive when such requirements are met.

  Data integration and standards for data modelling are a cross-cutting issue that arises in **all steps of the Big Data value chain** and in use cases **in all sector forums**.

- **Data security and privacy:** In all steps of the Big Data value chain, issues of data security and privacy arise. These must eventually be met by regulatory means, which in turn will rely on the formulation of appropriate standards. Included in this complex are issues of data provenance and trustworthiness.

  Data security and privacy are cross-cutting issues that arise in **all steps of the Big Data value chain** and in use cases **in all sector forums.**

- **Frameworks:** Some areas of Big Data suffer from a big diversity of architectures, frameworks, tools and, e.g., query languages. Standardisation has proven helpful where de-facto standards such has Hadoop (for architectures) or Linked Data principles (as a framework for publishing and connecting structured data in the web) have developed.

Specific steps in the Big Data value chain include **Data Acquisition, Data Curation, Data Usage,** and **Data Storage.**

- **Benchmarking:** In multiple areas, the development of standardised benchmarks can have a huge benefit to objectively compare various tools and approaches and support competition. This includes economic impacts such as energy-efficiency of data storage and analytics hardware solutions.

  This requirement is particularly relevant for **Data Analysis** but applies in all steps.

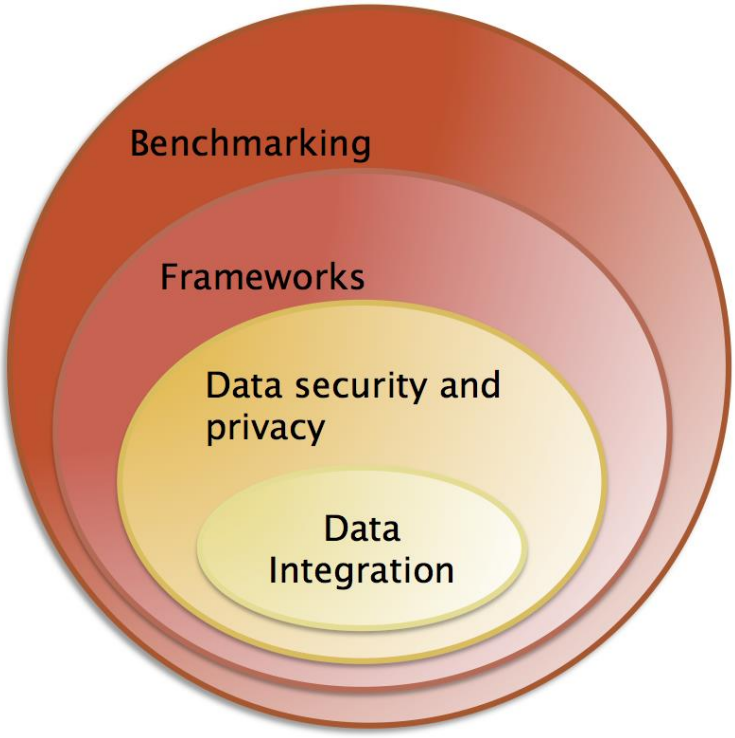

Figure 3-2: Key requirements for Standardisation in Big Data

Besides these four core requirements, there are a number of technology- and sector specific challenges related to standardisation as identified in the preceeding sections.

3.4. Existing Activities for Big Data Standards in Technology

This section addresses issues in technology standardisation that are derived from the results of the technology working groups as described in the previous two sections. How technology relates to the various economic sectors is elaborated in sections 3.5 and 3.9.

Technology for Big Data covers hardware as well as software. The most prevalent hardware aspects are storage technology, processing hardware and networking hardware. Software is understood to include data formats that are particularly relevant for interoperability. In addition, a following section (3.6) will address benchmarking standards.

3.4.1 Standards in Hardware Technology

Big Data database software, file systems and algorithms are typically taunted as “running on commodity hardware” to set Big Data apart from legacy applications. For a survey of storage technology and technology stacks, see BIG Deliverable D2.2.2, in particular the parts on Data Storage and Data Usage.

A notable development in hardware standardisation is driven by the “Open Compute Project” (Frankovsky, 2014) that was initiated by Facebook in 2011 and aims at providing open hardware specifications for servers, storage systems, networks, data centre design, etc.

3.4.2 Standards in Software Technology

Most technology standards in Big Data processing software are *de facto* standards that are not prescribed (but at best described after the fact) by a standards organisation. The most prominent example is, of course, Hadoop and Map/reduce.

As far as Big Data relies on Open Source software, existing mechanisms for commercial success are proven and readily apply to Big Data in the same manner as for other software technologies.

Technology areas that are difficult to capture for standardisation are:

- In-memory implementation of Big Data databases and software
- Complex event processing for real-time Big Data applications
- Vendor platforms that promise efficiency through coherent and integrated approaches but might stifle competition and interoperability

3.4.2.1 Standards for Query Languages

As formulated in section 3.2.4, a key requirement for Data Storage is the development of standards for query languages beyond SQL, the so-called NoSQL (“Not only SQL”) query languages. There are two successful developments in this area with wide acceptance:

SPARQL[1], “*(a recursive acronym for: SPARQL Protocol and RDF Query Language) is an RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format. It was made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is recognized as one of the key technologies of the semantic web. On 15 January 2008, SPARQL 1.0 became an official W3C Recommendation, and SPARQL 1.1 in March, 2013.”*

At XLDB, a new query language, ArrayQL[2], is being developed: *“At XLDB 2012 we announced that two major databases that support arrays as first-class objects (MonetDB SciQL and SciDB) have formed a working group in conjunction with XLDB. This working group is proposing a common syntax (provisionally named “ArrayQL”) for manipulating arrays, including array creation and query. In addition, we are working with a third major array-supporting implementation (Rasdaman) to propose an algebra for common array operations. A common algebra and syntax are expected to benefit array users in the same way that the relational algebra and SQL benefit relational users.”*

However, for other applications, standardisation is still missing, e.g., there is a lack of standardized query languages of NoSQL stores.

Beyond the special importance of NoSQL query languages for Data Storage, they are a cross-cutting issue with relevance for all technology field of the data value chain and applications in all economic sectors.

[1] http://en.wikipedia.org/wiki/Sparql
[2] http://www.xldb.org/arrayql

3.4.3 Testing and Integration Frameworks

With reference to Figure 3-2, Data Integration is covered by standards for data (following section) and query languages (previous section). The next layer, Data security and privacy is discussed in the previous chapter. The third layer, then, are standards for frameworks. On the technology level, frameworks are a cross-cutting issue, relevant for all steps in the data value chain.

There are a number of activities aimed at developing testing and integration frameworks for specific software ecosystems.

Apache **Bigtop**[1] is a *"project for the development of packaging and tests of the Apache Hadoop ecosystem.*

The primary goal of Bigtop is to build a community around the packaging and interoperability testing of Hadoop-related projects. This includes testing at various levels (packaging, platform, runtime, upgrade, etc…) developed by a community with a focus on the system as a whole, rather than individual projects."

BDAS, the Berkeley Data Analytics Stack[2] is "*an open source software stack that integrates software components being built by the AMPLab to make sense of Big Data."*

Blueprints[3] is *"a collection of interfaces, implementations, […] and test suites for the property graph data model. Blueprints is analogous to the JDBC, but for graph databases. As such, it provides a common set of interfaces to allow developers to plug-and-play their graph database backend. Moreover, software written atop Blueprints works over all Blueprints-enabled graph databases. Within the TinkerPop software stack, Blueprints serves as the foundational technology for:*

- *Pipes: A lazy, data flow framework*
- *Gremlin: A graph traversal language*
- *Frames: An object-to-graph mapper*
- *Furnace: A graph algorithms package*
- *Rexster: A graph server"*

3.4.4 Big Data Standards for Data

Attempting to standardize data formats for Big Data is, in general, a hopeless enterprise as "variety" as one of the Vs is a hallmark of Big Data. Nevertheless, data exchange and data interoperability is one of the core challenges for Big Data and is addressed in multiple ways. These range from domain (industry sector) specific solutions, like domain ontologies to general concepts such as Linked Open Data.

Standardisation would be needed for (i) data integration, but also for (ii) data security and privacy, (iii) frameworks and tools, and (iv) benchmarks for comparison of tools and solutions. Currently, there is scarcely any standardisation directly related to Big Data or specifically Big Data Usage. As far as data integration is concerned, a representation framework for the semantics of data exists: RDF (Resource Description Framework) and Linked Data. However, the actual instances, i.e., ontologies and terminologies are not standardised. A very big data base would be needed as a comprehensive, normative standard. Such de-facto standards exist only in a few sectors, e.g., in the Health and Life-Sciences sectors. The Google Knowledge

[1] http://bigtop.apache.org
[2] https://amplab.cs.berkeley.edu/software
[3] https://github.com/tinkerpop/blueprints/wiki

Vault (Graph) is such a very big collection of some 1.6 billion general facts that would have the potential as a standard, however, it is not freely available.

Data exchange is a very active area in standardisation efforts, see as an example NIEM (US government) or CMIS (Content Management Interoperability Services, OASIS) as described in section “Standards organisations and subgroups” below.

Standards for data touch upon all steps of the data value chain, with direct importance for Data Acquisition and Data Quality. With respect to the industrial sectors, they are of immediate relevance for Data Management Engineering and Analytics (see Figure 3-3) and as such represent another cross-cutting issue, relevant for all sectors.

3.5. Sector specific views on standards

The most prominent standards-related requirement across all sectors in BIG is the common sentiment in the public sector that there is a general lack of business standards for open data and the reuse of data for big data purposes, resulting in a lack of interoperability. Standards for data formats might evolve through the increase of available open data.

The same sentiment is generally shared in other sectors, however, some sectors have a history of developing sector-wide standards in pre-competitive collaborations, e.g., the manufacturing sector. Thus, some sectors will be lagging behind in the development and adoption of standards unless support actions are undertaken.

The work on roadmapping in BIG, as described in Deliverables D2.4.2 and consolidated in D2.5 has collected the key requirements from each of 10 economic sectors (Health, Public, Finance, Insurance, Telecom, Media & Entertainment, Energy, Transport, Manufacturing, and Retail). They are summarised in five groups of technology requirements which are: Data Management Engineering, Data Quality, Data Security and User Experience, and Deep Data Analytics.

Figure 3-3 shows the distribution of technology requirements across these sectors and is relevant as the basis for relating standardisation activities to sectors. This linking is made explicit for each applicable on-going standardisation activity in Section 3.9.

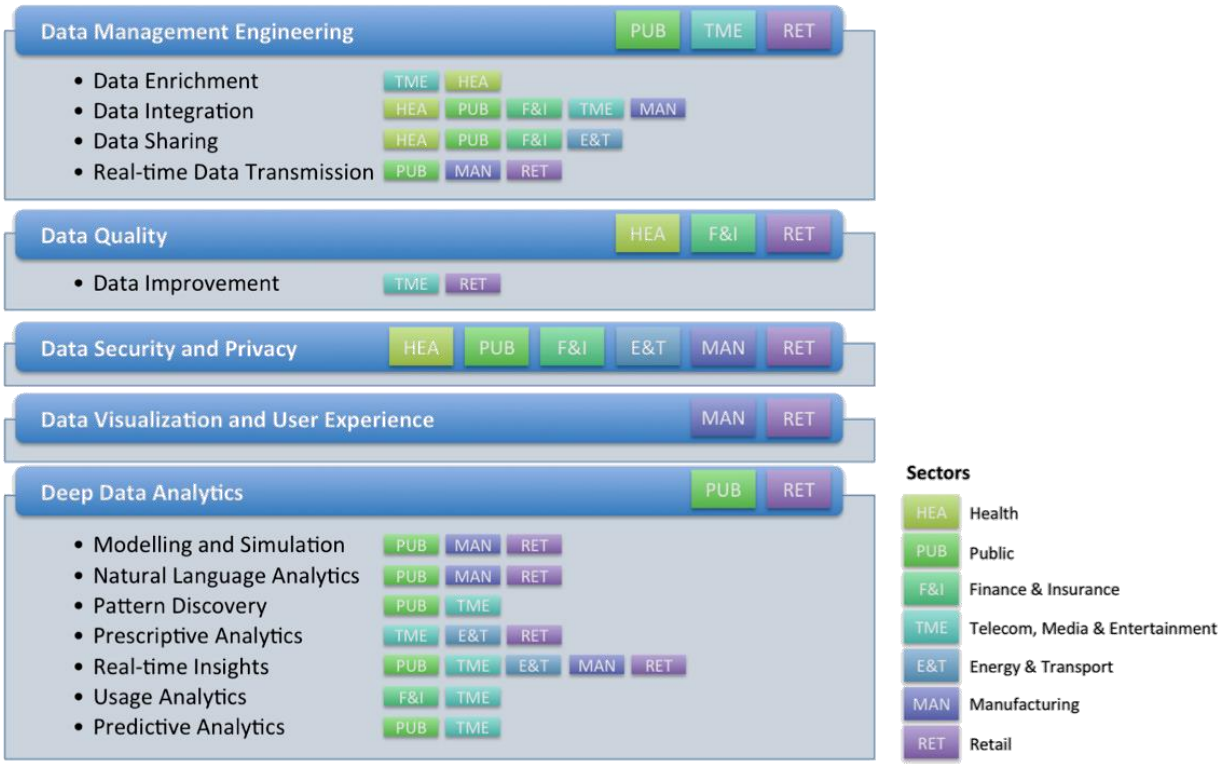

Figure 3-3: Distribution of requirements across sectors as taken from D2.5.

3.6. Big Data Benchmarks

Another aspect for standardisation is the evaluation of performance aspects of Big Data, ranging from intrinsic benchmarks (e.g., processing speed or storage capacity as covered under hardware technology standards) to extrinsic benchmarks. An example of the latter is the *Complete Social Media Measurement Standards* as agreed by a large number of industry groups and customers ("The Conclave"), see the description below.

Another approach are combined benchmarks, such as Intel's HiBench suite of 10 benchmarks in four categories (Intel, 2013).

3.7. Laws and Regulations

Laws and regulations can be seen as a type of standard, however they are beyond the coverage of the BIG project. It is however conceivable that BIG's Public-Private Forum will eventually address legislative aspects of Big Data technology.

Note that in some sectors regulations have a huge impact on the uptake and success of Big Data, see, e.g., the situation in the Health sector as described in D2.3.2 and the upcoming D2.4.2..

3.8. A European perspective

New standards for Big Data should be international. While this is uncontroversial, the question must be explore what the specific European interests are that need to be pushed on the world-wide level. These can be economic sectors of particular relevance to Europe and those are pointed out in section 3.9 where applicable. A more direct influence is available in the public sector and through European policies and regulations applying to Open Data.

3.8.1 Open data on a European level

Repackaging and standardizing of Open Data could conceivably be supported specifically at the European level as a by-product of initiatives for collecting open data sets. The consolidated roadmap of BIG (Palmetshofer et al., 2014) lists Europe to be leading in Open Data by 2019 as a medium priority goal.

To embrace a clear open data strategy at European level, the following aspects must be addressed:

- **Accessibility**: Common list of datasets at European level, common and standardised access (formats, APIs, storage services).

- **Quality**: Data quality must be transparent, i.e. marked-up in a standardised and guaranteed manner

- **Comprehensive**: Sufficient data must be available across countries (avoid checker-board coverage).

- **Timeliness**: Data must be current and update cycles must be transparent and assurances for future updates must be available.

- **Transparent market**: Common licensing models for the use of open data sets.

3.9. Standards organisations and relevant subgroups

The BIG project through its technical working groups, its sector forums and eventually through its Public-Private Forum shall network with the relevant subgroups in (multiple) standardisation organisations that contribute—at least partially—to Big Data aspects. Thus, we have collated a comprehensive list of relevant organisations, their relevant subgroups and contact points.

As pointed out in the introduction to this chapter, Big Data is evolutionary in many ways and technology development is mainly going ahead without established standards and standardisation is lagging behind. Given the lagging status of standardisation, many organisations have acknowledged the need for standards and formed working groups but most lack critical mass of coverage, partners, and progress as pointed out in the following sections.

Some standardisation activities for special purposes within Big Data or overlapping with Big Data are successfully going on in multiple places, e.g., semantic web standardisation at the W3C. Of the standardisation activities that address Big Data in general, the work at ISO is likely the most successful and should be given special consideration, see the discussion of on-going activities in Section 3.9.

3.9.1 ISO

ISO, the "International Organisation for Standarization"[1] is possibly the biggest standardisation organisation. Many national standards are related to ISO standards in some way.

3.9.1.1 ISO/IEC JTC 1 Study Group on Big Data

Although there are some 19,000 standards administered at ISO, none is designed specifically for Big Data.

However, in collaboration with the NIST Big Data Working Group (see section 3.9.4.1 below), ISO has established the "ISO/IEC JTC 1 Study Group on Big Data (BD-SG)" in November 2013. Given the potential impact of standards published by ISO, and the decent level of participation by stakeholders (including industry), this activity is the most promising of those discussed here and most likely to produce relevant output with high impact.

As this work is now carried out under the ISO umbrella, NIST (a U.S. government organisation) has a strong influence that should be counter-balanced by EU participation.

Given the initial broadness of its topics, it cuts across all of the sectors considered in the BIG project. Future specialisation might change this, but currently **all sectors** need to pay attention to and should participate in these efforts.

ISO itself states:

"JTC 1 is recognizing the Big Data:

- *Has been identified by SWG Planning as an important future area for JTC 1 focus,*
- *Is a topic of consideration within SC 32 as reported to the Plenary, and*
- *Continues to be of interest to other JTC 1 Subcommittees including SC 27, SC 34 and SC 38*

Therefore, JTC 1 establishes a Study Group on Big Data for consideration of Big Data activities across all of JTC 1 with the following terms of reference:

[1] http://www.iso.org

1. *Survey the existing ICT landscape for key technologies and relevant standards /models/studies /use cases and scenarios for Big Data from JTC 1, ISO, IEC and other standards setting organizations,*
2. *Identify key terms and definitions commonly used in the area of Big Data,*
3. *Assess the current status of Big Data standardization market requirements, identify standards gaps, and propose standardization priorities to serve as a basis for future JTC 1 work, and*
4. *Provide a report with recommendations and other potential deliverables to the 2014 JTC 1 Plenary.*

Membership in the SG on Big Data is open to:

1. *JTC 1 National Bodies, JTC 1 Liaisons and approved JTC 1 PAS Submitters;*
2. *JTC 1 /SCs, JTC 1/WGs, relevant ISO and IEC TCs;*
3. *Members of ISO and IEC central offices; and*
4. *Invited standards setting organizations that are engaged in Big Data standardization as approved by the SG on Big Data.*

Given the potential of these efforts, the BIG project has participated in the "2nd Big Data Interoperability Framework Workshop: Building Robust Big Data Ecosystem" of the ISO/IEC JTC 1 Study Group on Big Data on May 13 – 16, 2014 at the University of Amsterdam, Amsterdam, Netherlands through a presentation titled "*Towards a Big Data Roadmap for Europe*", coordinated and given by Martin Strohbach of AGT. An extended version of this presentation, authored by several BIG partners, has been submitted and is under review by ACM.

3.9.1.2 General ISO Standards

Many ISO standards are relevant for aspects of IT operations, in general and in detail, and quite a number of ISO standards are relevant for business operations and processes.

As an example for a technical standard, consider *"ISO/IEC 17826:2012 Information technology – Cloud Data Management Interface (CDMI)",* an open standard specifically for data storage as a service as part of cloud computing. It was developed jointly with the Storage Networking Industry Association (SNIA) by the Joint Technical Committee 1 (JTC 1) of ISO and the International Electrotechnical Commission (IEC).

As an example for a business process-oriented standard, the ISO standard *"ISO/IEC 27001:2005 – Information technology – Security techniques – Information security management systems – Requirements"* is concerned with IT security. Like a number of ISO standards it is meant to be used with an *accreditation.* This includes checking of adherence to the standard through a checklist by an approved, independent third party.

Such standards have **no specific relation to Big Data**, we foresee no necessity for direct participation by the BIG project or its partners.

3.9.2 IEEE

IEEE and its IEEE Standards Association are concerned with—from our perspective on Big Data—low level technical standards.

Eg., there is a standards working group in the IEEE called "IEEE P2302 – Standard for Intercloud Interoperability and Federation (SIIF)" with some technical relevance for Big Data.

Also loosely related to Big Data, the IEEE Standards Association (IEEE-SA) has formed two Working Groups (WGs) around IEEE P2301 and IEEE P2302, see (IEEE, 2011). IEEE P2301

provides profiles of existing and in-progress cloud computing standards in critical areas such as application, portability, management, and interoperability interfaces, as well as file formats and operation conventions. IEEE P2302 defines essential topology, protocols, functionality, and governance required for reliable cloud-to-cloud interoperability and federation.

An overview of IEEE activities related to cloud computing can be found at (IEEE, 2013) and shows the low-level technical nature of IEEE standards.

We are not aware of any IEEE standardisation efforts aimed specifically at Big Data. IEEE does organise a yearly conference on big data. The 2014 International Congress on Big Data takes place on June 27 – July 2, 2014 in Anchorage, Alaska, USA[1].

Given the background of IEEE, all activities are on the **technology** level, with a cross-cutting nature and without a specific interest to selected sectors.

As these activities center around **cloud computing**, they have higher relevance in sectors with smaller players (SMEs) or companies with diverse infrastructure, and thus are of higher relevance for the **Manufacturing, Retail, Media & Entertainment, Transportation, and Health** sectors.

3.9.3 W3C

W3C has formed a *"Big Data Community Group"* in April 2012. From its charter[2]:

"This group will explore emerging BIG DATA pipelines and discuss the potential for developing standard architectures, Application Programming Interfaces (APIs), and languages that will improve interoperability, enable security, and lower the overall cost of BIG DATA solutions.

The BIG DATA community group will also develop tools and methods that will enable: a) trust in BIG DATA solutions; b) standard techniques for operating on BIG DATA, and c) increased education and awareness of accuracy and uncertainties associated with applying emerging techniques to BIG DATA."

The community group appears to suffer from **low participation**, there are no efforts at identifying relevant topics for standardisation yet.

Related to Big Data, a W3C standard on "*Customer Experience Digital Data Acquisition*" (Srikanth et al., 2012) describes a rather technical "method for surfacing Customer Experience Digital Data on a Web/Digital resource as a set of JavaScript Objects, and also specifies the parameters for communicating this data to digital analytic and reporting servers". From its introduction: "Collection and analysis of visitor behavioural and demographic data has become an integral part of web design and website success. This data is central to site performance analysis, dynamically tailoring site content to visitor activity and interest and retargeting visitors based on behaviours."

Also related to Big Data, the W3C has standardized many aspects of **semantic technology**, including RDF, RDFS, SPARQL, OWL and Linked Data. These technologies will become highly relevant for Big **Data analysis** in the future.

In summary, cooperation with W3C is relevant mainly for its engagement in semantic technologies which are relevant to all sectors, and of particular importance for all sectors that have been identified as relying on semantic technologies, including **Health, Public, Telecom, Media & Entertainment, and Manufacturing.**

[1] http://www.ieeebigdata.org/2014/

[2] http://www.w3.org/community/bigdata

3.9.4 NIST

3.9.4.1 NIST Big Data Public Working Group

NIST, the US "National Institute of Standards and Technology" has established a "NIST Big Data Public Working Group" (NBD-PWG).

From their charter: *"NIST is leading the development of a **Big Data Technology Roadmap**. This roadmap will define and prioritize requirements for interoperability, portability, reusability, and extensibility for big data usage, analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data. To help develop the ideas in the **Big Data Technology Roadmap**, NIST is creating the Public Working Group for Big Data."*

The working group has established subgroups on definitions and taxonomies, requirements, security and privacy, reference architecture, and technology roadmap. Chairs and co-chairs of the subgroups come from a variety of academic and industrial organisations. Participation in the WG and Subgroups are open to all interested parties. There are no membership fees. WG results will be available to all stakeholders in the commercial, academic, and government sectors.

3.9.4.2 NIST Data Science Program

NIST has also established a cross-cutting "Data Science Program" which it describes as follows: "*The NIST Information Technology Laboratory (ITL) has over 25 years of experience measuring and evaluating technology that is used to process, analyse, and derive knowledge from various structured and unstructured data, including text, audio/voice, imagery, video, and multimedia. ITL and the Information Access Division (IAD) within ITL have extended our current efforts by forming a new, cross-cutting data science program, focused on driving advancements in big data analytics. This focus, alongside our collaboration with industry, academia, and government, will accelerate development of these technologies as well as enable faster transition into government applications and the commercial market place.*

As part of this program, ITL and IAD will create and manage a big data analytics evaluation series and develop a model of data science through which we can measure various attributes, e.g., component classes, component-level performance, end-to-end performance, flow analyses, and propagation of uncertainty. This approach will define metrology supporting objective comparison of approaches, identify key areas for improvement in big data analytics, and focus research community critical mass on the hard problems.

[…]

Current Data Science measurement and evaluation activities:

Modeling Data Science: The complexities inherent to big data analytics require new methods and models to inform novel evaluation paradigms and measurement methods. NIST is collaborating with data science researchers and practitioners to develop models that will support the metrology necessary to drive data science forward.

Data Analytics Evaluation Series: In 2013, the NIST Information Technology Laboratory and Information Access Division began investigation into requirements for a new big data analytics evaluation series.

Data Science Symposium Series: The inaugural NIST Data Science Symposium took place on March 4 2014."

In summary, there are a number of active communities organised by NIST that are worthwhile observing. Being a U.S. national organisation, it is of lesser relevance to EU stakeholders.

However, many of the activities are influencing, of have become officially a part of ISO activities as outlined in section 3.9.1, putting them on the international stage. EU stakeholders should attempt to work with NIST through **ISO** as the appropriate platform.

3.9.5 Oasis

OASIS (Organisation for the Advancement of Structured Information Standards[1], *"is a non-profit consortium that drives the development, convergence and adoption of open standards for the global information society. OASIS members broadly represent the marketplace of public and private sector technology leaders, users and influencers. The consortium has more than 5,000 participants representing over 600 organisations and individual members in more than 65 countries."*

OASIS themselves claim the following technical committees as directly relevant to Big Data[2]:

- OASIS Advanced Message Queuing Protocol (AMQP) TC. Defining a ubiquitous, secure, reliable and open internet protocol for handling business messaging.
- OASIS Message Queuing Telemetry Transport (MQTT) TC. Providing a lightweight publish/subscribe reliable messaging transport protocol suitable for communication in M2M/IoT contexts where a small code footprint is required and/or network bandwidth is at a premium.
- OASIS XML Interchange Language (XMILE) for System Dynamics TC. Defining an open XML protocol for sharing interoperable system dynamics models and simulations.

In addition, there is a longer list of committees related to cloud computing that also have relevance to Big Data. These include

- OASIS Identity in the Cloud TC (ID-Cloud)
- Open Data Protocol (Odata)
- Topology and Orchestration Specification for Cloud Applications (TOSCA)
- Advanced Message Queuing Protocol (AMQP)
- Cloud Application Management for Platforms (CAMP)
- Cloud Authorisation (CloudAuthZ)

In the area of data exchange and interoperability, OASIS has standardized CMIS, the "Content Management Interoperability Services"[3].

In summary, Oasis activities bear only **indirect relations to Big Data** and are not further analysed here.

3.9.6 Cloud security alliance

The Cloud Security Alliance (CSA[4]) *"is a not-for-profit organisation with a mission to promote the use of best practices for providing security assurance within Cloud Computing"*.

It has established a Big Data Working Group (BDWG). Its charter formulates *"The Big Data Working Group (BDWG) will be identifying scalable techniques for data-centric security and*

[1] https://www.oasis-open.org/org
[2] https://www.oasis-open.org/committees/tc_cat.php?cat=bigdata
[3] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=cmis

[4] https://cloudsecurityalliance.org

privacy problems. BDWG's investigation is expected to lead to crystallisation of best practices for security and privacy in big data…"

Thus, the working group will address standardisation only through best practice documentation. Thus, CSA activities are **not in the focus of Big Data** standardisation issues.

3.9.7 Cloud Standards Customer Council (CSCC)

The Cloud Standards Customer Council (CSCC[1]) *"is an end user advocacy group dedicated to accelerating cloud's successful adoption, and drilling down into the standards, security and interoperability issues surrounding the transition to the cloud.*

Cloud Standards Customer Council founding enterprise members include IBM, Kaavo, Rackspace, Software AG. The world's leading organisations including Lockheed Martin, Citigroup, State Street and North Carolina State University have already joined the Council."

CSCC has been formed by OMG (see next section) and it has established a "Big Data in the Cloud Working Group".

CSCC has a strong U.S. focus and concentrates on cloud computing and thus its activities are **not in the focus of Big Data** standardisation issues.

3.9.8 OMG

OMG, the Object Management Group[2] *"is an international, open membership, not-for-profit computer industry standards consortium."*

Its EDM Council[3] is the author and steward of the Financial Industry Business Ontology (FIBO).

FIBO is being released as a series of standards under the technical governance of the Object Management Group (OMG).

Given its focus, FIBO will be of special interest to EU stakeholders from the **Finance and Insurance** sectors.

3.9.9 NIEM

In the US, NIEM[4]—the National Information Exchange Model—is a community-driven, government-wide, standards-based approach to exchanging information.

As an example for the growing adoption of NIEM in U.S. government organisations, see http://www.fiercegovernmentit.com/story/dod-adopts-niem-will-no-longer-support-ucore-development/2012-10-27.

Comparable to NIST, U.S. government activities have **no direct impact on EU** stakeholders. However, they might well influence international standardisation activities, albeit much less than NIST. Also, similar activities might be beneficial for the EU arena, with lessons to be learned from NIEM's efforts.

[1] http://www.cloud-council.org/
[2] http://www.omg.org/
[3] http://edmcouncil.org/financialbusiness
[4] https://www.niem.gov

3.9.10 Contacts at Standardisation Organisations

This section summarizes the list of contacts at standardisation organisations.

ISO/IEC JTC 1 Study Group on Big Data

http://jtc1bigdatasg.nist.gov/home.php

W3C Big Data Community Group

http://www.w3.org/community/bigdata/

NIST Big Data Public Working Group

http://bigdatawg.nist.gov/home.php

Cloud Security Alliance – Big Data Working Group Leadership

https://cloudsecurityalliance.org/research/big-data/

Chair: Sreeranga Rajan, Fujitsu

Co-Chairs: Neel Sundaresan, eBay and Wilco van Ginkel, Verizon

Cloud Standards Customer Council – Big Data in the Cloud Working Group

http://www.cloud-council.org/workinggroups.htm

or contact becky@omg.org.

4. Recommendations

This deliverable has identified a discrepancy between Big Data as a rapidly developing technology and a slow process of addressing IPR and standardisation issues. IPR issues that are specific to Big Data relate mainly to aspects of data and less so to software and hardware.

Over the last year, a number of standardisation organisations have begun to address Big Data through working groups that aim at investigation the relevant issues and eventually begin work at new standards, e.g., at NIST and ISO/IEC. These activities need support and coordination.

4.1. Clarification of data IPR

With respect to IPR issues in Big Data, the situation as described above leads to the following recommendations:

- Clarification of data ownership **rights** when data is obtained from a data market or other source and when data is **generated** through algorithms and/or combination with other data
- Clarification of **liabilities** when data is used in further processing
- Clarification of possibilities and potential requirements for meta data, describing data **provenance.**
- Clarification of **international** aspects of data **privacy.**

Supporting activities can include legal considerations, specific sector requirements or use case scenarios and will require a co-operation on the international level, including the participation of national stakeholders. These issues can have a place in the Big Data PPF.

4.2. Support for beginning standardisation activities

As described above, a number of standardisation organisations have formed Big Data working groups that aim at an understanding of the current situation, the requirements, roadmaps and eventually aim at identifying areas in needs of standardisation and the developments of such standards. Looking at the variety of organisations involved, this leads to the following recommendations:

- Support **co-operation** between standardisation activities. A prime example is the close cooperation between NIST and ISO/IEC which has led the BIG project to participate in these activities. With NIST being a U.S. government body, European input and influence is especially important to balance the international influence and character of this ISO activity.
- Support **sector specific** activities, in particular with respect to considerations from the European perspective.
- Support **open source developers** to participate in standardisation. As there are many crucial software packages for Big Data (e.g., Hadoop) being developed in an open source model, the developers and organisations (e.g., Apache) will need support to participate in standardisation groups at the same level of activity as industry participants get support from their corporations.
- Support standardisation activities in publicly supported projects by clarifying and making the **eligibility of costs** for such activities explicit.
- Support standardisation activities for **data integration, data security & privacy, frameworks and benchmarks** as outlined in section 3.3**.** These are the crucial topics in

Big Data whose standardisation issues are not well understood and will vary across sectors.

- Support standardisation activities that address **Open Data** activities. In particular, the public sector names the lack of (business) standards for the use of Open Data as a specific blocker.
- Support standardisation activities in all areas that are relevant for **data markets.** This extends into the realm of rules and regulation for new concepts such as data marketplaces.

4.3. Enabling tools

In order to support new activities in the field of Big Data, suitable tools should be developed that help a new player—be it a software developer, a company using Big Data in their business processes, a service provider or a regulation body—to understand the issues surrounding IPR and standardisation aspects of Big Data.

Such tools will differ for the various target groups and the different sectors; they can include tools such as:

- Patent research giving an overview of existing patents and patentable ideas
- Checklists presenting applicable regulations and standards for relevant use case scenarios. Such a checklist could eventually be further developed into best practice guides for Big Data.
- A standardisation roadmap, as soon as the newly formed Big Data working groups begin to identify concrete targets for standardisation activities.

5. Abbreviations and acronyms

IPR	Intellectual property rights
IP	Intellectual property
IT	Information technology
PPP	Public-private partnership
TWG	Technical working groups
SF	Sector forum
SPARQL	SPARQL Protocol and RDF Query Language
RDF	Resource description framework
SQL	Structured query language
NoSQL	Not only SQL
IaaS	Infrastructure as a service
PaaS	Platform as a service
SaaS	Software as a service
IoT	Internet of Things
IoS	Internet of Services
BDAS	Berkeley data analytics stack
ISO	International organisation for standardisation
IEC	International Electrotechnical Commission
SC	Sub-committee
JTC	Joint technical committee
IEEE	Institute of electrical and electronics engineers
W3C	World Wide Web Consortium
NIST	National institute of standards and technology (USA)
OASIS	Organisation for the advancement of structured information standards
XML	eXtensible markup language
CSA	Cloud security alliance
CSCC	Cloud standards customer council
OMG	Object management group
NIEM	National Information exchange model (USA)

6. References

Cacas, Max esperately Seeking Big Data Standards, January 2013.

Available at: http://www.afcea.org/content/?q=node/10487

Falvey, Rod; Foster, Neil; Memedovic, Olga, The Role of Intellectual Property Rights in Technology Transfer and Economic Growth: Theory and Evidence, UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION, Vienna, 2006.

Frankovsky, Frank, The Open Compute Project: 2014 and Beyond, Keynote, OPC Summit V, San Jose, 2014.

Gislason, Hjalmer, interview by John Dominque. BIG Interview (3 May 2013).

Götz, T, Kottmann, J, Sandstone, SA, Lang, A; Quo Vadis UIMA?, COLING 2014, Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT, pages 77–82, Dublin, Ireland, August 23rd 2014.

IEEE, IEEE LAUNCHES PIONEERING CLOUD COMPUTING INITIATIVE, press release, 2011.

Available at: http://standards.ieee.org/news/2011/cloud.html

IEEE, IEEE Standards Activities in Cloud Computing, 2013.

Available at: http://standards.ieee.org/develop/msp/cloudcomputing.pdf

Intel IT Center, Planning Guide, Getting Started with Big Data, February 2013.

Available at: http://www.intel.de/content/dam/www/public/us/en/documents/guides/getting-started-with-hadoop-planning-guide.pdf

Korn, Naomi;, Oppenheim, Charles; Duncan, Charles, IPR and Licensing issues in Derived Data, JISC IPR Consultancy scoping report, May 2007.

Palmetshofer, Walter et al.; BIG Deliverable D2.5, Cross-sectorial roadmap consolidation, BIG project, 2014.

SPARQL, http://en.wikipedia.org/wiki/Sparql, last retrieved on May 28, 2014.

Srikanth, Viswanath; Niemann, Michael; White, Hutch; Towb, Eliot; Customer Experience Digital Data Acquisition 1.0, W3C Member Submission (IBM), September 2012.

Available at: http://www.w3.org/Submission/cedda1/

Umeh, Jude, Big Data, Privacy and Intellectual Property, ITNow magazine, September 2013.

USPTO, Interim Guidance for Determining Subject Matter Eligibility for Process Claims in View of Bikski v. Kappas, July 2010.

Available at: http://www.uspto.gov/patents/law/exam/bilski_guidance_27jul2010.pdf

Wei-Dong (Jackie) Zhu, Bob Foyle, Daniel Gagné, Vijay Gupta, Josemina Magdalen, Amarjeet S Mundi, Tetsuya Nasukawa, Mark Paulis, Jane Singer, Martin Triska; IBM Watson Content Analytics: Discovering Actionable Insight from Your Content, IBM Redbooks, July 2014