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

D4.2.1 First Draft of IPR, Standardisation recommendations

Work Package:	*WP4 Big Data Public-Private Forum*	
Due Date:		30/06/2013
Submission Date:		07/10/2013
Start Date of Project:		01/09/2012
Duration of Project:		26 Months
Organisation Responsible of Deliverable:		ATOS
Version:		0.5
Status:		Final version of first draft (a final version of the document is expected in M20)
Author name(s):	Tilman Becker	DFKI
Reviewer(s):	Volker Tresp Nuria de Lama	Siemens 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**	**Modified by**	**Comments**
0.1	10/06/2013	Outline	Tilman Becker (DFKI)
0.2	09/09/2013	Extended Outline and partial content descriptions	Tilman Becker (DFKI)
0.3	30/09/2013	Cleaned out content descriptions	Tilman Becker (DFKI)
0.4	2/10/2013	Finalized content	Tilman Becker (DFKI)
0.5	7/10/2013	Minor modifications by Nuria de Lama (Atos)	Nuria de Lama (Atos)

Executive Summary

This first draft deliverable provides an outline of the aspects of IPR and standardisation that apply to Big Data. It identifies the relevant topics and how they will be addressed. As the main goal of the deliverable is a starting point for recommendations, it will focus on important on-going developments that need support as well as identify gaps in standards or IPR-related issues that need to be addressed in the near future.

This document structures the issues and assesses where existing approaches readily apply to Big Data and where Big Data poses specific challenges that need further support by activities from the community and from the BIG project.

Issues in IPR revolve about data, focussing on ownership and liability, and much less on technology IPR.

Standardisation in Big Data includes hard- and software technology, data interoperability and finally, benchmarking in multiple fields.

As a first draft, it indicates which areas need further extension and outlines the sources and inputs needed from within the BIG project and beyond.

Table of Contents[1]

[1] <In order to work properly with cross-references and caption to figures and tables, please read: http://www.shaunakelly.com/word/numbering/numberingappendixes.html>

1. Introduction

As part of the activities of the BIG project, there will be contributions to standards as well as potential propositions for new ones. This deliverable collates work done in different tasks and WPs of BIG to serve as the outcome point for the general guidelines and recommendations.

This first draft aims at an outline of the final deliverable. It identifies the relevant topics and how they will be addressed. As the main goal of the deliverable is a starting point for recommendations, it will focus on important on-going developments that need support as well as identify gaps in standards or IPR-related issues that need to be addressed in the near future.

Additionally, the overview over IPR and standards-related issues should be utilized in the form of a checklist for Big Data efforts that might be published by the BIG project.

2. IPR

Beyond the existing challenges around Intellectual Property Rights 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.

2.1. IPR Challenges in Big Data Technology

Existing IPR strategies typically cover innovation in technology, products and business processes and easily extend to Big Data. In particular, existing approaches to IT technologies all apply to Big Data. 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 generally suitable for Big Data technologies.

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:

- Acquisition of data
- Changing / Curation of data
- Protection of data
- Disseminating / Selling data
- Liability for data (quality)

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.

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 Public-Private Forum.

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 "*Desperately Seeking Big Data Standards, January 1, 2013, By Max Cacas"*

(http://www.afcea.org/content/?q=node/10487) 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 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.

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/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

The final version of this deliverable will concentrate on exploring these three areas in more detail and propose mechanisms to address these issues (e.g., forming interest groups to generate best practice guidelines, proposing research topics) in a small roadmap. It will lay out the role that these issue play in the BIG Public-Private Forum.

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 first draft identifies areas in Big Data where standards are required and provides a first analysis of existing relevant standardisation bodies and their current activities.

The final version of this deliverable will provide more comprehensive standardisation recommendations by identifying all areas of Big Data that require new or extended standards and by drafting a standardisation roadmap and identifying relevant players in Europe, perhaps including the BIG Public-Private Forum.

3.1. Requirements for Standardisation in Big Data

The final version of this deliverable will set out the requirements for standards 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.

For standards in technology, input from BIG's technology working groups (TWG) will provide the basis for a roadmap in standardisation for Big Data.

For standards in data and benchmarks, input from BIG's sector forums (SF) will provide the basis for a roadmap in standardisation for Big Data.

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. Big Data Standards in Technology

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 which are particularly relevant for interoperability. In addition, a final section will address benchmarking standards.

3.2.1 Standards in Hardware Technology

This section will concentrate on storage technology and processing hardware. Storage technology expertise is concentrated in BIG's technical working group on "Data Storage". Chapter 4 of Deliverable D2.2.1 provides a good introduction to the field. The technical working group will extend this section for the final version of this deliverable. Processing hardware expertise is available across multiple of BIG's technical working groups who will collaborate to extend this section for the final version of this deliverable.

3.2.2 Standards in Software Technology

Most technology standards in Big Data processing software are *de facto* standards that are not prescribed (but at best *de*scribed 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

The final version of this deliverable must identify for each of these technology fields, how standardisation can be achieved and applied successfully.

3.3. 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.

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.

3.4. 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

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

3.5. 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 D2.3.1 and D2.4.1.

The sector forums of BIG might extend this section in the final deliverable with specific sector requirements on legislation and regulations.

3.6. 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 an initial list of relevant organisations, their relevant subgroups and contact points.

The final version of the deliverable will extend this list aiming at comprehensive coverage.

3.6.1 ISO

ISO, the "International Organisation for Standarization" (www.iso.org) is possibly the biggest standardisation organisation. Many national standards are related to ISO standards in some way.

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

3.6.2 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. 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.

(see http://standards.ieee.org/news/2011/cloud.html)

An overview of IEEE activities related to cloud computing can be found at http://standards.ieee.org/develop/msp/cloudcomputing.pdf and shows the low-level technical nature of IEEE standards.

We are not aware of any IEEE standardisation efforts aimed specifically at Big Data.

3.6.3 W3C

W3C has formed a *"Big Data Community Group"* in April 2012. From its charter at http://www.w3.org/community/bigdata :

"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*" (see http://www.w3.org/Submission/cedda1/) 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 behavioral 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 behaviors."

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.

3.6.4 Oasis

OASIS (Organisation for the Advancement of Structured Information Standards, see https://www.oasis-open.org/org) "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 (see https://www.oasis-open.org/committees/tc_cat.php?cat=bigdata):

- 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".

See https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=cmis

3.6.5 Cloud security alliance

The Cloud Security Alliance (CSA, see https://cloudsecurityalliance.org) "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 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.

3.6.6 Cloud Standards Customer Council (CSCC)

The Cloud Standards Customer Council (CSCC, see http://www.cloud-council.org/) "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”.

3.6.7 OMG

OMG, the Object Management Group (see http://www.omg.org/) “is an international, open membership, not-for-profit computer industry standards consortium.”

Its EDM Council (see http://edmcouncil.org/financialbusiness) 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).

3.6.8 NIEM

In the US, NIEM—the National Information Exchange Model—is a community-driven, government-wide, standards-based approach to exchanging information. See: https://www.niem.gov.

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

3.6.9 Contacts at Standardisation Organisations

W3C Big Data Community Group

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

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. Further Work / Next Steps

Throughout this first draft document, necessary extension and next steps have been noted. The goal of the final version is a global view of Big Data from which all areas of Big Data can be identified where IPR and standardisation are crucial. We aim at outlining and structuring the requirements for standardisation and providing an overview over existing standardisation efforts, resulting in a Big Data standardisation roadmap.

4.1. IPR and Standards Checklist for Big Data Projects

The final version of this document proposes to provide an outline for a checklist for Big Data projects that covers the aspects of IPR and standardisation that are collected in this deliverable. Ideally, this can be created in collaboration with an external partner or partners in BIG's Public-Private Forum.

Such a checklist could eventually be further developed into best practice guides for Big Data.

5. Summary

This first draft deliverable provides an outline of the aspects of IPR and standardisation that apply to Big Data. It structures the issues and assesses where existing approaches readily apply to Big Data and where Big Data poses specific challenges that need further support by activities from the community and from the BIG project.

Issues in IPR revolve about data, focussing on ownership and liability, and much less on technology IPR.

Standardization in Big Data includes hard- and software technology, data interoperability and finally, benchmarking in multiple fields.

As a first draft, it indicates which areas need further extension and outlines the sources and inputs needed from within the BIG project and beyond.