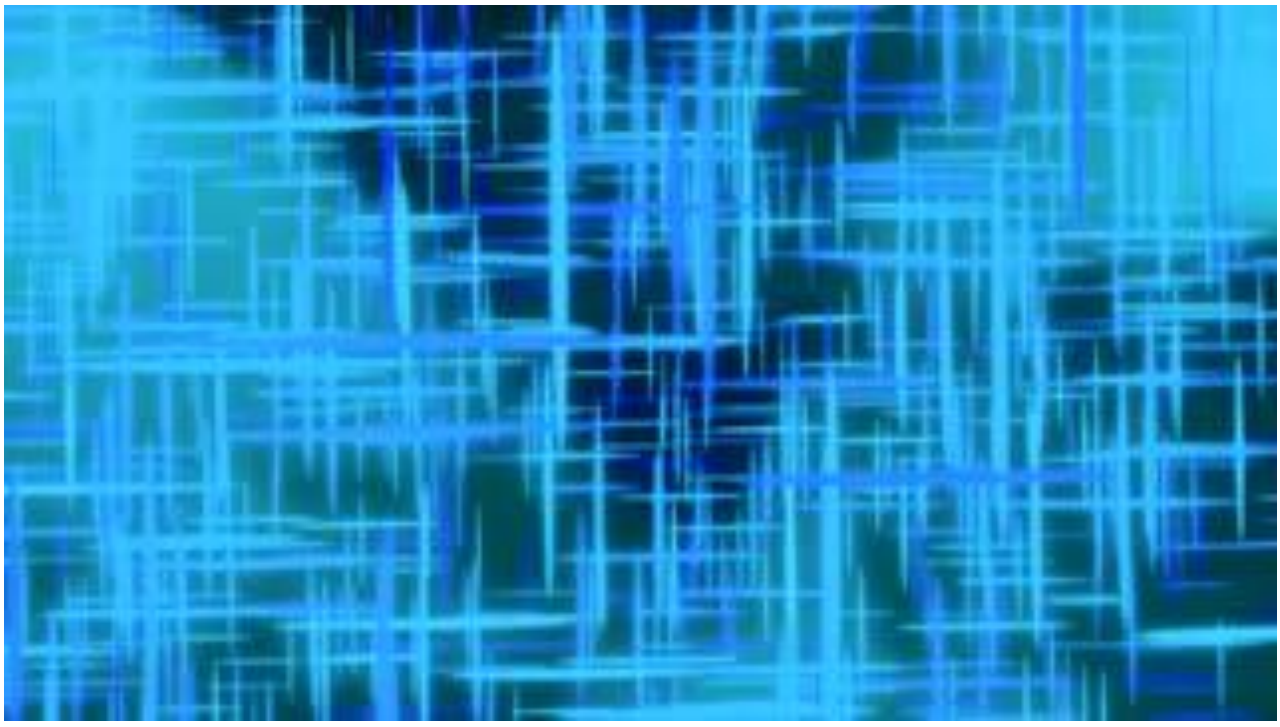
Table of Contents

Big Data Standardization

Ray.Walshe@dcu.ie, Jane.Kernan@dcu.ie



The Need for Standardization

Standards are used by everyone in everyday life all over the world. Standards make it possible to carry out our day to day activities as they impact in communications, technology, media, healthcare, food, transport, construction, furniture and energy.

Some standards have really stood the test of time[1], being around for hundreds if not thousands of years. The railway Standard Gauge is based on the grooves worn into the ground by the wheels of Roman chariots. Centuries later wagon owners found that the ride was more comfortable if the wheels fitted into those grooves and this approach was continued and adopted as the gauge for the first railway carriages and wagons, the track spacing was determined by their wheels. Considerable cost and effort was saved by not having to devise a new 'standard'.

Where can Standards help us:

- Reliability – Adopting standards helps ensure safety, reliability and environmental care. Standardized products and services are perceived as more dependable, raising user confidence, sales and new technology adoption.
- Government policy and legislative support – Standards are used by regulators and legislators for protecting consumer interests, and to support government policies. Standards play a central role in the European Union's policy for a Single Market.
- Interoperability – Standards compliant products and services enable devices to work together.
- Business benefits – standardization provides a solid foundation upon which to develop new technologies and to enhance existing practices.

Specifically, standards:

- Open up market access
- Provide economies of scale

[1] http://www.etsi.org/standards/why-we-need-standards

- Encourage innovation
- Increase awareness of technical developments and initiatives

Consumer choice - standards provide the foundation for a greater variety of new products with new features and options

Without standards we would have:

- Products that might not work, or are dangerous
- Inferior quality products or incompatibility with others
- Customers could be locked in to one supplier
- Manufacturers inventing their own standards for even the simplest problem.

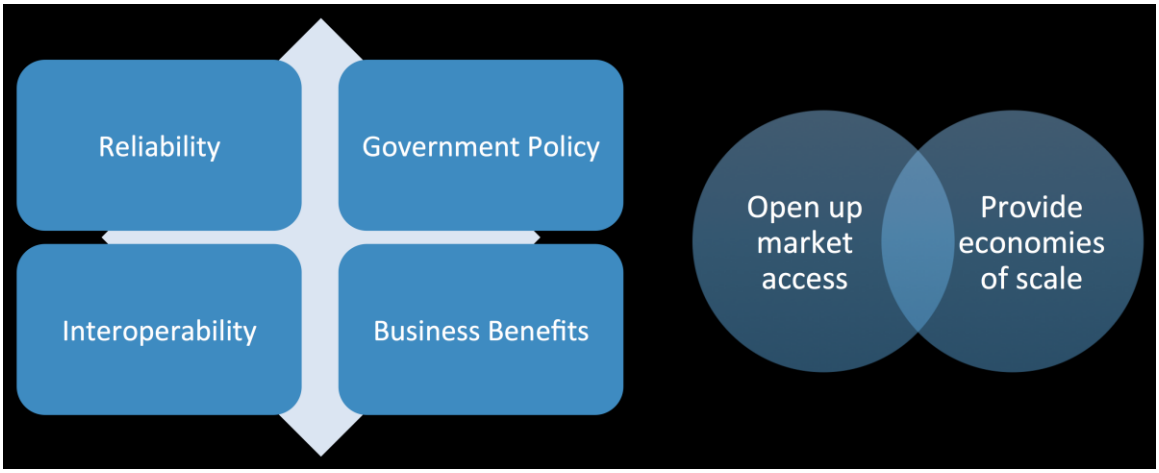


Figure 1 Standards help access new markets

The need for international standardization in the provision of goods and services to consumers should be obvious from the above and is also supported by many factual examples of success based on standards development.

The GSM™ mobile communication technology and its successors (3G, 4G) which were led by the European Telecommunications Standards Institute (ETSI) is a good example of standardization. GSM was originally envisaged as a telecoms solution for Europe, but the technologies were quickly adopted and have been deployed world-wide. Thanks to standardization, international travellers can communicate and use common services anywhere in the world

Standardization in ICT in the EU

The EU supports an effective and coherent standardization framework, which ensures that standards are developed in a way that supports EU policies and competitiveness in the global market.

Regulations[2] on European standardization set the legal framework in which the different actors in the standardization system can operate. These actors are the European Commission, the European standardization organizations, industry, small and medium-sized industries (SMEs) and societal stakeholders.

The Commission is empowered to identify information and communication technology (ICT) technical specifications[3] to be eligible for referencing in public procurement. Public authorities can therefore make use of the full range of specifications when buying IT hardware, software and services, allowing for greater competition and reducing the risk of lock-in to proprietary systems. The Commission financially supports the work of the three European standardization organizations, ETSI, CEN and CENELEC

ETSI – the European Telecommunications Standards Institute

ETSI, the European Telecommunications Standards Institute, produces globally-applicable standards[4] for Information and Communications Technologies (ICT), including fixed, mobile, radio, converged, broadcast and Internet technologies. These standards enable the technologies on which business and society rely. The ETSI standards for GSM™, DECT™, Smart Cards and electronic signatures have helped to revolutionize modern life all over the world.

ETSI is one of the three European Standards Organizations officially recognized by the European Union, is a not-for-profit organization with more than 800 member organizations worldwide, drawn from 66 countries and five continents. Members include the world's leading companies and innovative R&D organizations.

[2] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012R1025
[3] https://ec.europa.eu/growth/sectors/digital-economy/ict-standardisation/ict-technical-specifications_en
[4] http://www.etsi.org/about/

ETSI are at the forefront of emerging technologies, addressing the technical issues which will drive the economy of the future and improve life for the next generation.

CEN – the European Committee for Standardization [5]

CEN, the European Committee for Standardization, is an association that brings together the National Standardization Bodies of 33 European countries. CEN also one of three European Standardization Organizations (together with CENELEC and ETSI) that have been officially recognized by the European Union and by the European Free Trade Association (EFTA) as being responsible for developing and defining voluntary standards at European level.

CEN provides a platform for the development of European Standards and other technical documents in relation to various kinds of products, materials, services and processes. CEN supports standardization activities in relation to a wide range of fields and sectors including: air and space, chemicals, construction, consumer products, defense and security, energy, the environment, food and feed, health and safety, healthcare, ICT, machinery, materials, pressure equipment, services, smart living, transport and packaging.

CENELEC – the European Committee for Electrotechnical Standardization [6]

CENELEC is the European Committee for Electrotechnical Standardization and is responsible for standardization in the electrotechnical engineering field. CENELEC prepares voluntary standards, which help facilitate trade between countries, create new markets, cut compliance costs and support the development of a Single European Market. CENELEC creates market access at European level but also at international level, adopting international standards wherever possible, through its close collaboration with the International Electrotechnical Commission (IEC)[7], under the Dresden Agreement.

In the global economy, CENELEC fosters innovation and competitiveness, making technology available industry-wide through the production of voluntary standards. CENELEC members, its

[5] https://www.cen.eu/Pages/default.aspx
[6] http://www.cenelec.eu/
[7] https://www.cenelec.eu/aboutcenelec/whoweare/globalpartners/iec.html

experts, the industry federations and consumers, help create European Standards to encourage technological development, to ensure interoperability and to guarantee the safety and health of consumers and provide environmental protection. Designated as a European Standards Organization by the European Commission, CENELEC is a non-profit technical organization set up under Belgian law. It was created in 1973 as a result of the merger of two previous European organizations: CENELCOM and CENEL.

EU-funded research and innovation projects also make their results available to the standardization work of several standards-setting organizations.

The European Multi Stakeholder Platform on ICT Standardization

The European Multi Stakeholder Platform (MSP)[8] on ICT standardization was established in 2011. It advises the Commission on ICT standardization policy implementation issues, including priority-setting in support of legislation and policies, and the identification of specifications developed by global ICT standards development organizations. The Multi Stakeholder Platform addresses:

- potential future ICT standardization needs
- technical specifications for public procurements
- cooperation between ICT standards-setting organizations;
- a multi-annual overview of the needs for preliminary or complementary ICT standardization activities in support of the EU policy activities (the Rolling Plan [9])

The MSP is composed of representatives of national authorities from EU Member States & EFTA countries, by the European and international ICT standardization bodies, and by stakeholder organizations that represent industry, small and medium-sized enterprises and consumers. The MSP meets four times per year and is co-chaired by the European Commission Directorates General Internal Market[10], Industry, Entrepreneurship and SME and CONNECT [11].

[8] http://ec.europa.eu/digital-agenda/european-multi-stakeholder-platform-ict-standardisation
[9] https://ec.europa.eu/digital-single-market/en/rolling-plan-ict-standardisation
[10] http://ec.europa.eu/growth/about-us/index_en.htm
[11] http://ec.europa.eu/dgs/connect/

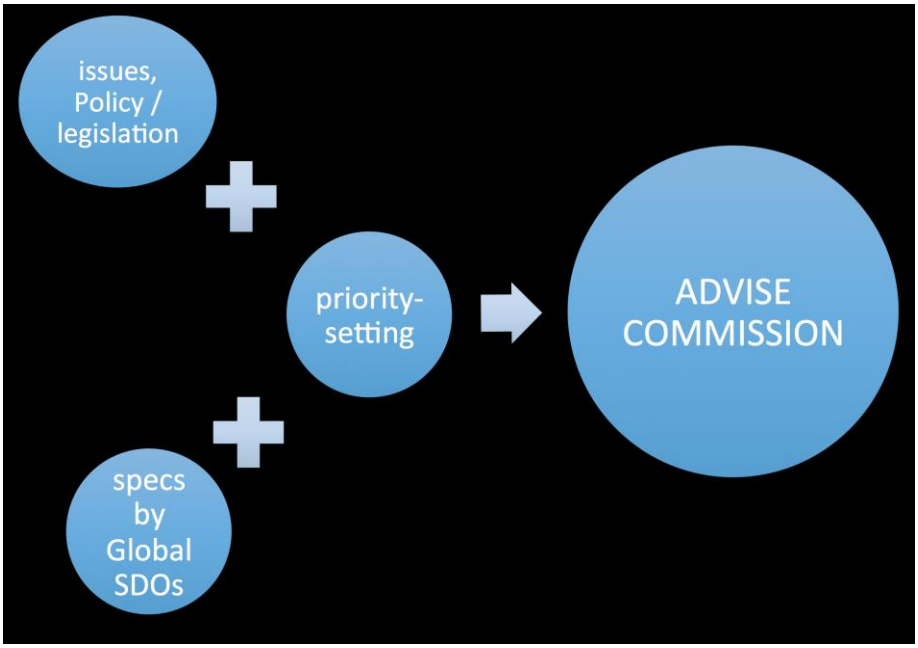


Figure 2 European Commision: Multi Stakeholder Platform

The Platform also advises on the elaboration and implementation of the Rolling Plan on ICT standardization [12].

The Rolling Plan (RP), provides a multi-annual overview of the needs for preliminary or complementary ICT standardization activities to undertake in support of the EU policy activities. It is aimed at the wider ICT community stakeholders and outlines how practically support will be provided. It contains a distinct view of the landscape of standardization activities in a given policy area.

The Rolling Plan

- Puts standardization in the policy context.
- Identifies EU policy priorities where standardization activities.
- Covers ICT infrastructures and ICT standardization horizontals

[12] https://ec.europa.eu/growth/sectors/digital-economy/ict-standardisation_en#rolling_plan_ict_standardisation

The Rolling Plan references legal documents, available standards and technical specifications as well as ongoing activities in ICT standardization. Addenda to the Rolling Plan may be published alongside the Rolling Plan in order to keep current with new developments in the rapidly changing ICT sector.

Mission of the Multi Stakeholders Platform on ICT Standardization [13]

> *Advisory Expert Group on all matters related to European ICT Standardization and its effective implementation: - Advise the Commission on its ICT Standardization workprogramme -Identify potential future ICT Standardization needs - Advise the Commission on possible standardization mandates - Advise the Commission on technical specifications in the field of ICT with regard to its referencing in public procurement and policies - Advise the Commission on cooperation between standards developing organizations*

The 2016 Rolling Plan on ICT Standardization [14] covers all activities that can support standardization and prioritizes actions for ICT adoption and interoperability.

The Plan offers details on the international contexts for each policy:

- Societal Challenges: eHealth, accessibility of ICT products and services, web accessibility, e-Skills and e-Learning, emergency communications and eCall.
- Innovation for the Digital Single Market: e-Procurement, e-Invoicing, card/internet and mobile payments, eXtensible Business Reporting Language (XBRL) and Online Dispute Resolution (ODR).
- Sustainable growth: Smart grids and smart metering, smart cities, ICT environmental impact, European Electronic Toll Service (EETS) and Intelligent Transport System (ITS).
- Key enablers and security: Cloud computing, (Open) Data, e-government, electronic identification and trust services including e-Signatures, Radio Frequency Identification (RFID), Internet of Things (IoT), network and information security (cyber-security) and ePrivacy.

[13] http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=2758
[14] http://ec.europa.eu/DocsRoom/documents/15783/attachments/1/translations

This latest Rolling Plan describes all the standardization activities undertaken by Standard Setting Organizations (SSOs). This ensures an improved coherence between standardization activities in the EU. This is the first time that the European Standardization Organizations and other stakeholders were involved in drafting the RP and this improved process is a stronger guarantee that activities of standardization supporting EU policies in the ICT domain will be aligned.

Big Data Use Cases

In June 2013, the National Institute of Standards and Technology (NIST) Big Data Public Working Group (NBD-PWG) began forming a community of interested parties from all sectors, including industry, academia, and government to develop a consensus on big data definitions, taxonomies, secure reference architectures, security and privacy requirements, and ultimately a standards roadmap. Part of the work carried out by the working group identified Big Data Use Cases in NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements", that would serve as exemplars to help develop a Big Data Reference Architecture (BDRA).

The NBD-PWG defined a use case as "*a typical application stated at a high level for the purposes of extracting requirements or comparing usages across fields"*. They began by collecting use cases from publically available information for various Big Data architecture examples. This process returned 51 use cases across nine broad areas (i.e., application domains). This list was not intended to be exhaustive and other application domains will be considered. Each example of Big Data architecture constituted one use case. The nine application domains were as follows:

- Government Operation
- Commercial
- Defense
- Healthcare and Life Sciences
- Deep Learning and Social Media
- The Ecosystem for Research
- Astronomy and Physics
- Earth, Environmental, and Polar Science

- Energy

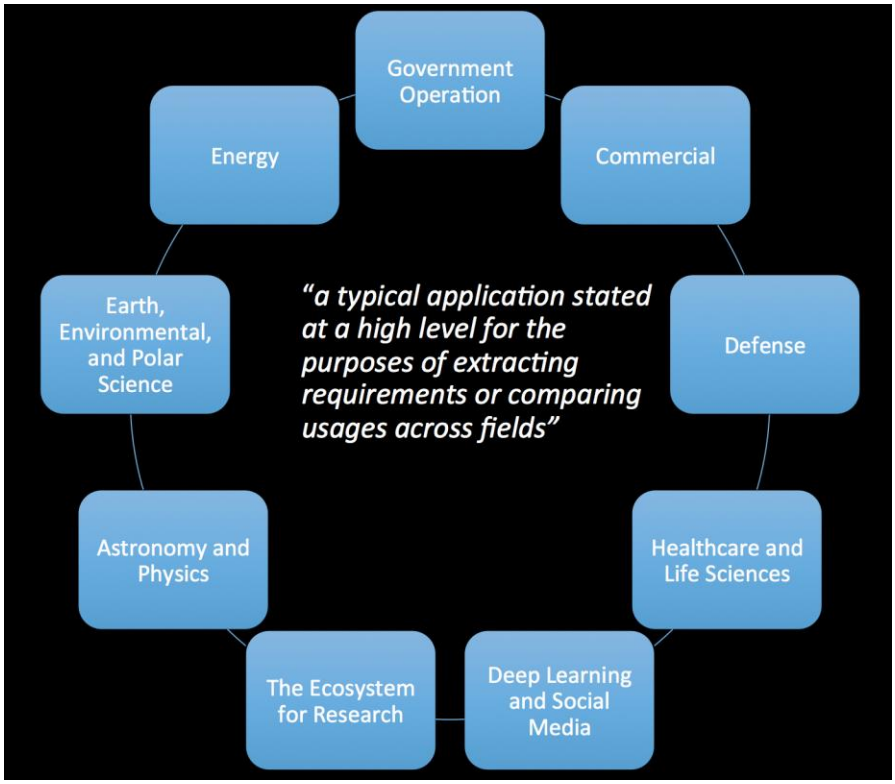


Figure 3 Big Data Use Cases

Use Case Summaries

The initial focus of the NBD-PWG Use Case and Requirements Subgroup was to form a community of interest from industry, academia, and government, with the goal of developing a consensus list of big data requirements across all stakeholders. This included gathering and understanding various use cases from diversified application domains. Tasks assigned to the subgroup include the following:

- Gather input from all stakeholders regarding big data requirements. A goal that turned into gathering use cases
- Analyze/prioritize a list of challenging general requirements derived from use cases that may delay or prevent adoption of big data deployment.
- Develop a comprehensive list of big data requirements.

The report was produced by an open collaborative process involving weekly telephone conversations and information exchange using the NIST document system. The 51 use cases came from participants in the calls (subgroup members), and from others informed of the opportunity to contribute. The use cases are organized into the nine broad sectors/areas (application domains) listed below with the number of use cases in parentheses and sample examples:

- Government Operation (4): National Archives and Records Administration, Census Bureau
- Commercial (8): Finance in Cloud, Cloud Backup, Mendeley (Citations), Netflix, Web Search, Digital Materials, Cargo Shipping (as in UPS)
- Defense (3): Sensors, Image Surveillance, Situation Assessment
- Healthcare and Life Sciences (10): Medical Records, Graph and Probabilistic Analysis, Pathology, Bioimaging, Genomics, Epidemiology, People Activity Models, Biodiversity
- Deep Learning and Social Media (6) Self-driving cars, Geolocate Images, Twitter, Crowd Sourcing, Network Science, NIST Benchmark Datasets
- Ecosystem for Research (4): Metadata, Collaboration, Language Translation, Light Source Experiments
- Astronomy and Physics (5): Sky Surveys (and comparisons to simulation), LHC at CERN, Belle Accelerator II in Japan
- Earth, Environmental, and Polar Science (10): Radar Scattering in Atmosphere, Earthquake, Ocean, Earth Observation, Ice Sheet Radar Scattering, Earth Radar Mapping, Climate Simulation Datasets, Atmospheric Turbulence Identification, Subsurface Biogeochemistry (microbes to watersheds), AmeriFlux and FLUXNET Gas

Evolution of Big Data Standards

Achieving the big data goals set out by business and consumers will require the interworking of multiple systems and technologies, legacy and new. Technology integration calls for standards to facilitate interoperability among the components of the big data value chain[15]. For instance, UIMA, OWL, PMML, RIF and XBRL are key software standards that support the interoperability of data analytics with a model for unstructured information, ontologies for information models, predictive models, business rules and a format for financial reporting. The standards community has launched several initiatives and working groups on big data. In 2012, the Cloud Security Alliance established a big data working group with the aim of identifying scalable techniques for data-centric security and privacy problems. The group's investigation is expected to clarify best practices for security and privacy in big data, and also to guide industry and government in the adoption of those best practices. The U.S. National Institute of Standards and Technology (NIST) kicked-off its big data activities with a workshop in June 2012 and a year later launched a public working group. The NIST[16] working group intends to support, secure and effective adoption of big data by developing consensus on definitions, taxonomies, secure reference architectures and a technology roadmap for big data analytic techniques and technology infrastructures.

NIST Big Data Public Working Group

NIST developed a Big Data Interoperability Framework[17] which consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

---

[15] http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000220001PDFE.pdf
[16] https://www.nist.gov/el/cyber-physical-systems/big-data-pwg
[17] Big Data Definitions: http://dx.doi.org/10.6028/NIST.SP.1500-1
Big Data Taxonomies: http://dx.doi.org/10.6028/NIST.SP.1500-2
Big Data Use Cases and Requirements: http://dx.doi.org/10.6028/NIST.SP.1500-3
Big Data Security and Privacy: http://dx.doi.org/10.6028/NIST.SP.1500-4
Big Data Architecture White Paper Survey: http://dx.doi.org/10.6028/NIST.SP.1500-5
Big Data Reference Architecture: http://dx.doi.org/10.6028/NIST.SP.1500-6
Big Data Standards Roadmap: http://dx.doi.org/10.6028/NIST.SP.1500-7

Volume 1, Definitions

The Definitions volume addresses fundamental concepts needed to understand the new paradigm for data applications, collectively known as Big Data, and the analytic processes collectively known as data science. Big Data has had many definitions and occurs when the scale of the data leads to the need for a cluster of computing and storage resources to provide cost-effective data management. Data science combines various technologies, techniques, and theories from various fields, mostly related to computer science and statistics, to obtain actionable knowledge from data.

Volume 2, Taxonomies

Taxonomies were prepared by the NIST Big Data Public Working Group (NBD-PWG) Definitions and Taxonomy Subgroup to facilitate communication and improve understanding across Big Data stakeholders by describing the functional components of the NIST Big Data Reference Architecture (NBDRA). The top-level roles of the taxonomy are System Orchestrator, Data Provider, Big Data Application Provider, Big Data Framework Provider, Data Consumer, Security and Privacy, and Management. The actors and activities for each of the top level roles are outlined as well. The NBDRA taxonomy aims to describe new issues in Big Data systems but is not an exhaustive list. In some cases, exploration of new Big Data topics includes current practices and technologies to provide needed context.

Volume 3, Use Cases and General Requirements

Use Cases and General Requirements document was prepared by the NIST Big Data Public Working Group (NBD-PWG) Use Cases and Requirements Subgroup to gather use cases and extract requirements. The Subgroup developed a use case template with 26 fields that were completed by 51 users in the following broad areas:

- Government Operations (4)
- Commercial (8)
- Defense (3)

- Healthcare and Life Sciences (10)
- Deep Learning and Social Media (6)
- The Ecosystem for Research (4)
- Astronomy and Physics (5)
- Earth, Environmental and Polar Science (10)
- Energy (1)

The use cases are, of course, only representative, and do not represent the entire spectrum of Big Data usage. All of the use cases were openly submitted and no significant editing was performed. While there are differences in scope and interpretation, the benefits of free and open submission outweighed those of greater uniformity.

Volume 4, Security and Privacy

Security and Privacy document was prepared by the NIST Big Data Public Working Group (NBD-PWG) Security and Privacy Subgroup to identify security and privacy issues that are specific to Big Data. Big Data application domains include healthcare, drug discovery, insurance, finance, retail and many others from both the private and public sectors. Among the scenarios within these application domains are health exchanges, clinical trials, mergers and acquisitions, device telemetry, targeted marketing, and international anti-piracy. Security technology domains include identity, authorization, audit, network and device security, and federation across trust boundaries.

Volume 5, Architectures White Paper Survey

Architectures White Paper Survey, was prepared by the NIST Big Data Public Working Group (NBD-PWG) Reference Architecture Subgroup to facilitate understanding of the operational intricacies in Big Data and to serve as a tool for developing system-specific architectures using a common reference framework. The Subgroup surveyed published Big Data platforms by leading companies or individuals supporting the Big Data framework and analyzed the material. This effort revealed a remarkable consistency of Big Data architecture. The most common themes occurring across the architectures surveyed are outlined below.

- Big Data Management

    - Structured, semi-structured, and unstructured data
    - Velocity, variety, volume, and variability
    - SQL and NoSQL
    - Distributed file system
- Big Data Analytics
    - Descriptive, predictive, and spatial
    - Real-time
    - Interactive
    - Batch analytics
    - Reporting
    - Dashboard
- Big Data Infrastructure
    - In-memory data grids
    - Operational database
    - Analytic database
    - Relational database
    - Flat files
    - Content management system
    - Horizontal scalable architecture

Volume 6, Reference Architecture

The NIST Big Data Public Working group (NBD-PWG) Reference Architecture Subgroup prepared this NIST Big Data Interoperability Framework: Reference Architecture, to provide a vendor-neutral, technology- and infrastructure-agnostic conceptual model and examine related

issues. The conceptual model, referred to as the NIST Big Data Reference Architecture (NBDRA), was crafted by examining publicly available Big Data architectures representing various approaches and products. Inputs from the other NBD-PWG subgroups were also incorporated into the creation of the NBDRA. It is applicable to a variety of business environments, including tightly integrated enterprise systems, as well as loosely coupled vertical industries that rely on cooperation among independent stakeholders. The NBDRA captures the two known Big Data economic value chains: information, where value is created by data collection, integration, analysis, and applying the results to data-driven services, and the information technology (IT), where value is created by providing networking, infrastructure, platforms, and tools in support of vertical data-based applications.

Volume 7, Standards Roadmap

Standards Roadmap, summarizes the deliverables of the other NBD-PWG subgroups (presented in detail in the other volumes of this series) and presents the work of the NBD-PWG Technology Roadmap Subgroup. In the first phase of development, the NBD-PWG Technology Roadmap Subgroup investigated existing standards that relate to Big Data and recognized general categories of gaps in those standards.

ISO/IEC JTC1's data management and interchange standards committee (SC32)

ISO/IEC JTC1's data management and interchange standards committee (SC32) has a study on next generation analytics and big data. 18 The W3C has created several community groups on different aspects of big data.

At the June, 2012 SC32 Plenary in Berlin, the SC32 Chair, Jim Melton, appointed an Ad Hoc committee from all four SC32 working groups: WG1 E-business, WG2 Metadata, WG3 Database Languages and WG4 Multimedia.

Since the original request from JTC1 referenced a report by the US industry analyst Gartner Group, it is useful to review what Gartner has to say. Both "Next Generation Analytics" and "Big Data" are identified as Strategic Technologies for 2012:

[18] http://www.jtc1sc32.org/doc/N2351-2400/32N2388b-report_SG_big_data_analytics.pdf

- Next-Generation Analytics.
    - Analytics is growing along three key dimensions:
        - From traditional offline analytics to in-line embedded analytics. This has been the focus for many efforts in the past and will continue to be an important focus for analytics.
        - From analyzing historical data to explain what happened to analyzing historical and real-time data from multiple systems to simulate and predict the future.
        - Over the next three years, analytics will mature along a third dimension, from structured and simple data analyzed by individuals to analysis of complex information of many types (text, video, etc…) from many systems supporting a collaborative decision process that brings multiple people together to analyze, brainstorm and make decisions.

    Analytics is also beginning to shift to the cloud and exploit cloud resources for high performance and grid computing.

    In 2011 and 2012, analytics will increasingly focus on decisions and collaboration. The new step is to provide simulation, prediction, optimization and other analytics, not simply information, to empower even more decision flexibility at the time and place of every business process action.

- Big Data. The size, complexity of formats and speed of delivery exceeds the capabilities of traditional data management technologies; it requires the use of new or exotic technologies simply to manage the volume alone. Many new technologies are emerging, with the potential to be disruptive (e.g., in-memory DBMS). Analytics has become a major driving application for data warehousing, with the use of MapReduce outside and inside the DBMS, and the use of self-service data marts. One major implication of big data is that in the future users will not be able to put all useful information into a single data warehouse. Logical data warehouses bringing together information from multiple sources as needed will replace the single data warehouse model.

ISO Big Data Standards Work

ISO IEC JTC 1 WG9 Big Data Working Group

Standard ecosystems are required to perform analytics processing regardless of the dataset's needs in relation to the Vs (volume, velocity, variety, etc.) characteristics, underlying computing platforms, and how Big Data analytics tools and techniques are deployed. Unified data platform architecture will support Big Data strategy across information management, analysis, and search technology.

A standard ecosystem provides vendor, technology, and infrastructure agnostic platforms that will enable data scientists and researchers to share and reuse interoperable analytics tools and techniques. WG 9 works with academic, industry, government, and various other stakeholders to understand the needs and foster such a standard Big Data ecosystem.

WG 9 has a 3-pronged technical approach to achieve this standard ecosystem:

a. Identify standard Big Data Reference Architecture (RA): this approach has already been captured in ISO/IEC 20547 to identify overall RA components and their interface descriptions.
b. Identify standard Big Data Reference Architecture Interfaces: this would be a new project to investigate how data flows between RA components and define standard interfaces for such interactions. The goal is to use these validated standard interface to build Big Data applications.
c. Identify standard Big Data Management Tools: this would be another new project to investigate how collection of analytics tools and computing resources can be efficiently and effectively managed to enable standard Big Data enterprise computing. The goal is to provide system management tools to manage, monitor, and fine tune Big Data applications

WG 9 produced ISO/IEC 20546 (IS) Big Data Overview and Vocabulary committee draft (CD) in March 2016 with balloting results from 9 countries approved as presented, 5 countries approved with comments, 2 countries disapproved with comments, and 15 countries abstention. WG 9 spent two teleconferences (Aug. 15 and Aug. 30 to review, discuss, and resolute all comments and generated the DOC and revised text for further contribution. We are hoping to have good contributions to enhance the quality of this standard and generate CD-2 at the 6th WG 9 November – December 2016 meeting.

WG 9 produced ISO/IEC 20547-2 Big Data Use Cases and Derived Requirements PDTR (51 use cases, 300+ pages) in July 2016 with a two-month balloting period. All comments are expected to be reviewed, discussed, and resolved at the 6th WG 9 November – December 2016 meeting.

WG 9 is currently actively recruiting Big Data experts and promoting JTC 1 Big Data standardization development by hosting workshops ahead of the WG 9 standards meetings.

For the 4th WG 9 meeting (7 March 2016, Ireland) WG 9 hosted a full day program with 16 speakers, one panel discussion, with over 50 participants. For the 5th WG 9 meeting (11 July 2016, China) a half-day program (see ANNIX-B) with 8 speakers and over 80 participants were attended. Through the outreach effort, and in addition to recruiting more Big Data experts, it hopes to explore new opportunities and expand the Big Data standard foundation technologies such as Big Data Reference Architectures Standard Interface and Big Data Reference Architecture Standard Management.

Title	Lead Editor	Co-Editors
ISO/IEC TR 20547-1, Information technology – Big Data Reference Architecture -- Part 1: Framework and Application Process	David BOYD (US)	Suwook HA (KR), Ray WALSHE (IE)
ISO/IEC TR 20547-2, Information technology – Big Data Reference Architecture - - Part 2: Use Cases and Derived Requirements	Ray WALSHE (IE)	Suwook HA (KR)
ISO/IEC 20547-3, Information technology -- Big Data Reference Architecture -- Part3: Reference Architecture	Ray WALSHE (IE)	David BOYD (US), Liang Guang (CN), Toshihiro Suzuki (JP)
ISO/IEC TR 20547-5, Information technology – Big Data Reference Architecture -- Part 5: Standards Roadmap	David BOYD (US)	Toshihiro SUZUKI (JP), Abdellatif Benjelloun TOUIMI (UK)

Trends and Future Directions of Big Data Standards[19]

Public Sector Information, Open Data and Big Data

With the continuously growing amount of data (often referred to under the notion Big Data) and the increasing amount of Open Data, interoperability ever more becomes a key issue for leveraging the value of this data. Standardization at different levels (such as metadata schemata, data representation formats and licensing conditions of Open Data) is essential to enable broad data integration, data exchange and interoperability with the overall goal to foster innovation on the basis of data. This refers to all types of (multilingual) data, including both structured and unstructured data, as well as data from different domains as diverse as geospatial data, statistical data, weather data, Public Sector Information (PSI) and research data (see also the Rolling Plan contribution on 'e-Infrastructures for Data and Computing-Intensive Science'), to name just a few.

European Standardization ongoing activities

Overall, the application of standard and shared formats and protocols for gathering and processing data from different sources in a coherent and interoperable manner across sectors and vertical markets should be encouraged, for example in R&D&I projects and in the EU Open Data Portal and the Pan-European Open Data Portal. Studies conducted on behalf of the European Commission show that businesses and citizens were facing difficulties in finding and re-using public sector information. In its communication on Open Data of December 12th 2011, the European Commission states that the availability of the information in a machine-readable format as well as a thin layer of commonly agreed metadata could facilitate data cross-reference and interoperability and therefore considerably enhance its value for reuse. A common standard for the referencing of Open Data in the European Open Data portals would be useful. The candidate for a common standard in this area is the Data Catalog Vocabulary (DCAT) in collaboration with FIWARE (Future Internet Middleware Platform) open stack-based specification and open standards APIs. The DCAT Application Profile has been developed as a common project from the ISA (Interoperability Solutions for European Public Administrations) Programme, the Publications Office (PO) and Directorate General for Communications Networks, Content and Technology-European Commission (DG CONNECT) to describe public sector data catalogues and datasets and

[19]: http://ec.europa.eu/DocsRoom/documents/15783/attachments/1/translations/en/renditions/pdf

to promote the specification to be used by data portals across Europe. By agreeing on a common application profile and promoting this to the Member States (MSs), the interoperability amongst data catalogues and the exchange of data between MSs will be substantially improved. The DCAT-AP is the specification that will be used by the Pan-European Open Data Portal, which is part of the Connecting Europe Facility infrastructure. FIWARE CKAN is an open source solution for the publication, management and consumption of Open Data. FIWARE NGSI is an API that provides a lightweight and simple means to gather, publish, query and subscribe to context information. [20]

The mapping of existing relevant standards for a number of big data areas would be beneficial. Moreover, it might be useful to identify European clusters of industries that are sufficiently homogeneous in their activities to develop data standards. Especially in the context of Open Data, the subjects of data provenance and licensing (for example the potential of machine-readable licenses) need to be addressed, as encouraged by the revised Public Sector Information (PSI) Directive This directive (2013/37/EU) encourages the use of standard licenses which must be available in digital format and be processed electronically (Article 8(2)). Furthermore, the Directive encourages the use of open licences available online, which should eventually become common practice across the EU (Recital 26). In addition, to help Member States in the transposition of the revised provisions, the Commission adopted guidelines that, amongst others, recommend the usage of such standard open licenses for the re-use of PSI.

There are however many organizations who are actively pushing the Big Data Standardization issue forward by supporting and developing Big Data projects and programmes across multiple stakeholder groups. Exemplar projects and programs are listed below.

Exemplar Big Data projects and programmes

SHARE-PSI 2.0, PROJECT FUNDED BY DG CONNECT AND LED BY GEIE ERCIM (EUROPEAN HOST OF W3C)
Re-use of public sector information and harmonisation of the implementation of the new PSI Directive (Directive 2013/37/EU) across Europe
EU COMMISSION
Smart Open Data project of DG ENV for contributing to standards developments
G8 OPEN DATA CHARTER

[20] http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/final_version_study_psi.docx
http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive_proposal/2012/open_data.pdf
http://ec.europa.eu/digital-agenda/overview-2003-psi-directive
http://ec.europa.eu/information_society/policy/psi/rules/eu/index_en.htm
http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/final_version_study_psi.docx for an overview and
http://ec.europa.eu/ information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/en.pdf

Exemplar Big Data projects and programmes

In 2013, the EU endorsed the G8 Open Data Charter and, with other G8 members, committed to implementing a number of Open Data activities in the G8 members' Collective Action Plan (publication of core and high quality datasets held at EU level, publication of data on the EU Open Data Portal and the sharing of experiences of Open Data work)
FUTURE INTERNET PUBLIC PRIVATE PARTNERSHIP PROGRAMME
Specifications developed under the Future Internet Public Private Partnership Programme (FP7): FIWARE NGSI is an API for context information management that provides a lightweight and simple means to gather, publish, query and subscribe to context information. FIWARE NGSI can be used for re-al-time Open Data management. FIWARE CKAN: Open Data publication Generic Enabler. FIWARE CKAN is an open source solution for the publication, management and consumption of Open Data, usually, but not only, through static datasets. FIWARE CKAN allows to catalogue, upload and manage open datasets and data sources, while it supports searching, browsing, visualizing or accessing Open Data 22 see http://www.europeandataportal.eu/en/content/edp-and-fiware-launch-new-partnership
ISA AND ISA SQUARE PROGRAMME OF THE EUROPEAN COMMISSION
The DCAT application profile (DCAT-AP) has been defined. DCAT-AP is a specification based on DCAT (a RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web) to enable the interoperability between data portals, for example to allow for meta-searches in the Pan-European Open Data Portal that harvests data from national Open Data portals. https://joinup.ec.europa.eu/ asset/dcat_application_profile/ asset_release/dcat-application-profile-data-portals-europe-draft-1 Under the framework of the Connecting Europe Facility Programme tools for the interoperability of metadata and data at national and EU level will be developed.
ITU-T
Recommendation Y.3600 provides requirements, capabilities and use cases of cloud computing based big data as well as its system context. Cloud computing based big data provides the capabilities to collect, store, analyze, visualize and manage varieties of large volume datasets, which cannot be rapidly transferred and analysed using traditional technologies. http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9853 The ITU workshop on "Big Data" (June 2014) discussed standards needs for big data in the telecommunications sector and adopted an outcome document. http://itu.int/en/ITU-T/Workshops-and-Seminars/bigdata SG13 is developing a definition for Big Data and most importantly a roadmap for big data standardization in ITU-T, including standardization landscape, identification/ prioritization of technical areas and possible standardization activities.
W3C
The project Multilingual Web-LT funded by the CSA grant LT-WEB, standardization work coordinated and managed by W3C Working Group "Multilingual Web-LT addressed standardization and promotion of best practices in language processing, exchange and interoperability of multilingual data, and on multilingual Web content management and was funded by the CSA grant LT-WEB. This group is part of the Internationalization (I18N) Activity of W3C with the main task to implement an Internationalisation Tag Set (ITS) that provides a standardized set of metadata for web content and "deep web" content that facilitates its interaction with multilingual technologies and translation/localization processes, ensuring smooth automated multilingual processing of web content. Version 2.0 of ITS has on 29 October 2013 been published as a W3C Recommendation. In the multilingual open data track of the Multilingual Web initiative, which is driven by the World Wide Web Consortium (W3C), there is an ongoing discussion about the standardization of multilingual URIs and localisation of URIs. Moreover, a W3C community group on "Best Practices for Multilingual Linked Open Data" has been created, where this topic is also discussed. http://www.multilingualweb.eu, http://www.w3.org/International/multilingualweb/lt/
OASIS
The project addresses the querying and sharing of data across disparate applications and multiple stakeholders for re-use in the enterprise, Cloud, and mobile devices. Specification development in the OASIS OData TC builds on the core OData Protocol V4 released in 2014 and addresses additional requirements identified as extensions in four directional white papers: data aggregation, temporal data, JSON documents, and XML documents as streams. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=odata ODF is an open, standardized format for reports, office documents and free-form information, fully integrated with other XML systems, and increasingly used as a standard format for publicly-released government information. Link: https://www.oasis-open.org/ committees/ office https://www.oasis-open.org/committees/odata OASIS XML Localisation Interchange File Format (XLIFF): https://www.oasis-open.org/committees/xliff
W3C
DCAT vocabulary (done in the Linked Government Data W3C Working Group) http://www.w3.org/TR/vocab-dcat/
ISO/IEC JTC1 WG 9 – Big Data.
This working group was formed at the November 2014 JTC1 Plenary. They have begun working on requirements, use cases, vocabulary and a reference architecture for Big Data

Table 1 Ongoing Standards Development Initiatives Ongoing standards development

Stakeholder feedback

Existing standards should be checked for account to the protection of individuals with regards to the processing of personal data and the free movement of such data in the light of Data Protection principles. Identification and where needed development of specific Privacy by Design standards should be done. Since early 2014, French companies and public entities have been working in the context of the French Association for standardization (AFNOR) on a white paper on expectations regarding standards for Big Data. The white paper is publicly available [21]. Several priorities have been identified:

- Data access including open data and governance of data within companies (Enhanced exploitation, data quality, security): mix the requirements of Big Data into the existing management standards. The development of a standard regarding data management could be considered.
- Data transformation where three elements are identified Processes and methods of reversibility in pseudonymisation algorithms, evaluation of system performance (ex: Hadoop), NoSQL query language, or visualization and manipulation process of Big Data results ;
- Adapt infrastructures to Big Data, like cloud computing for storage and massively parallel architectures.
- Data quality and data identification
    - criteria and methods to characterize sources and information, in terms of perceived quality and trust in a specific context ;
    - indexing of unstructured data coming from social networks and data associated with mobility and sensors ;
- Big data use cases: the last step covers the need of normalization to develop big data uses. Highly visible issues of end users that were presented in the introduction should be addressed: technical interoperability, SLAs, traceability of treatment, data erasure, regulatory compliance, data representation, APIs, etc

[21] http://www.afnor.org/liste-des-actualites/actualites/2015/juin-2015/big-dataimpact-et-attentes-pour-la-normalisation-decouvrez-le-livre-blanc-afnor

Proposed new standardization actions

ACTION 1: invitation to the CEN to support and assist DCAT-AP standardization process. DCAT-AP is based on the Data Catalogue vocabulary (DCAT). It contains the specifications for metadata records to meet the specific application needs of data portals in Europe while providing semantic interoperability with other applications on the basis of reuse of established controlled vocabularies (e.g. EuroVoc23) and mappings to existing metadata vocabularies (e.g. SDMX, INSPIRE metadata, Dublin Core, etc.). DCAT-AP has been developed by a multi-sectorial expert group. Experts from international standardization organizations as well as open data portal owners participated in the group to ensure the interoperability of the resulting specification and to assist in its standardization process.

ACTION 2: promote standardization in/via the Open Data infrastructure, especially the Pan-European Open Data Portal deployed in the period 2015-2020 as one of the Digital Service Infrastructures under the Connecting Europe Facility programme,

ACTION 3: support of standardization activities at different levels: H2020 R&D&I activities ; support internationalisation of standardization, in particular for the DCAT-AP specifications developed under the ISA programme, as well for specifications developed under Future Internet Public Private Partnership, such as FIWARE NGSI and FIWARE CKAN

ACTION 4: involvement of stakeholders in a dialogue about standards for Open Data and Big Data.

ACTION 5: For standardizing the DCAT - Application Profile CEN should coordinate with the relevant W3C Groups to avoid making incompatible changes as well as on the conditions for availability of the standard(s)

Summary

This chapter has outlined the case for Standardization in Big Data and described some of the activities ongoing in the Big Data standards ecosystems. Numerous projects completed and underway Nationally, within Europe and Global Initiatives have been mentioned. Sample Big Data use case scenarios are listed and some of the initiatives in the evolution of Big Data standards are described. Finally

Ray Walshe – Brief Biography

Ray Walshe is the NSAI Head of Delegation Big Data on ISO JTC1 WG9. He is a Science Foundation Ireland Funded Investigator for the largest Data Analytics research centre in Europe (Insight National Centre for Data Analytics) where he runs Big Data projects in Personal Sensing and Media Analytics Groups. Ray is Co-Director of the STRICT Network (STandards Research in

ICT) and leading Multiple European H2020 Initiatives in ICT Standards. Currently serving as the ISO Standards Lead Editor on the ISO/IEC JTC1/WG 9 20547 Big Data Reference Architecture and involved in multiple other standards SDOs including IEEE, NIST, ETSI and CENELEC. Ray has over 30 years' experience in electronics, software and telecommunications industries having worked for Electric Ireland, Ericsson, Siemens, Siemens Nixdorf and Software & Systems Engineering Ltd before joining Dublin City University. He works with European Commission in the area of entrepreneurship, representing Ireland on the Startup Europe University Network and coordinates StartUp Europe Week Dublin, part of European Commissions Startup Europe Week 2016 (Largest Entrepreneurship event in Europe). Ray also serves as Chief Architect ReskiTV, Chief Architect Performance Tracking Solutions, Chair of Graduate Diploma Information Technology and Director of CloudCORE Research Institute in Cloud Computing

Jane Kernan – Brief Biography

Ms Jane Kernan (female): Ms Kernan has over 20 years' experience as Lecturer in the School of Computing in Dublin City University, Ireland, She is coordinator for 1st and 2nd Years Programme in Computing, former Chair of MSc. in IT and Chair of the Graduate Diploma in IT. Jane delivers graduate and postgraduate course in Business Database Management and Business Applications and has extensive experience supervising student projects. Jane conducts research in

the CloudCORE Research Centre and the European Industry University Research Association (EIURA). She organised the EIURA Cloud Forum, the 23rd Irish Conference on Artificial Intelligence & Cognitive Science and has been involved most recently with the European Commission's Start-up Europe initiative including Start-up Europe Week, Start-up Europe University Network and the SEC2U initiative.