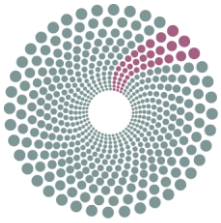

BIG DATA EUROPE

Support Action

Big Data Europe – Empowering Communities with Data Technologies

Project Number: 644564 Start Date of Project: 01/01/2015 Duration: 36 months

Deliverable 2.2: Report on Interest Groups Workshops I

Dissemination Level	Public
Due Date of Deliverable	Month 6, 30/06/2015
Actual Submission Date	Month 7, 28/07/2015
Work Package	WP2, Community Building & Requirements
Task	T2.1
Type	Report
Approval Status	Approved
Version	v0.4, Final
Number of Pages	50
Filename	D2.2_Report-on-Interest-Groups_Workshops_I.pdf

Abstract: This report summarises the organization and derived results from the first three Interest Group workshops organized during the reporting period (Societal Challenges 1 - Health, 3 - Energy, 5 - Climate) and carried out by each group associated with each societal challenges.

Project funded by the European Union's Horizon 2020 Research and Innovation Programme (2014 – 2020)

History

Version	Date	Reason	Revised by
0.0	15.063.2015	Content Provision	Thomas Thurner (SWC)
0.1	20.06.2015	SC1 Report	Simon Scerri (Fraunhofer)
0.2	10.07.2015	SC5 Report	Simon Scerri (Fraunhofer)
0.3	17.07.2015	SC3 Report	Simon Scerri (Fraunhofer)
0.4	24.07.2015	Final Version	Thomas Thurner (SWC)

Author List

Organisation	Name	Contact Information
Fraunhofer	Simon Scerri	simon.scerri@iais.fraunhofer.de
SWC	Thomas Thurner	t.thurner@semantic-web.at
OpenPHACTS	Kiera McNiece	mcneicek@rsc.org
VUA	Victor de Boer	v.de.boer@vu.nl
CRES	Fragiskos Mouzakis	mouzakis@cres.gr
NCSR-D	Mandy Vlachogianni	mandy@ipta.demokritos.gr
NCSR-D	Spyros Andronopoulos	sandron@ipta.demokritos.gr
NCSR-D	Andreas Ikonomopoulos	anikon@ipta.demokritos.gr
NCSR-D	I. Klampanos	iaklampanos@iit.demokritos.gr
NCSR-D	Vangelis Karkaletsis	vangelis@iit.demokritos.gr
NCSR-D	Stasinos Konstantopoulos	konstant@iit.demokritos.gr

Executive Summary

In this deliverable we provide an in-depth report and material associated with the first round of workshops that have taken place in the first 6 months of the BDE project. We provide a report that included information about the participants, the sessions organised, the talks and discussions as well as the gathered results (input for requirement elicitation). In addition, material associated with the workshop, such as the agenda and the original invitation letter, is also included.

Note: In the first Deliverable to be submitted by this work package (D2.1 - Community Building, Coordination and Planning), a section (2.4) describing the characterisation and description of the Transport (Societal Challenge 4) community together with the big data challenges and opportunities was deferred to a later date (due to the subcontracting process for the societal challenge domain representatives). This contribution is now included as an addendum to this deliverable (Appendix A).

Abbreviations and Acronyms

SC	Societal Challenge
EC	European Commission
RE	Requirement Elicitation
RS	Requirement Specification
WP	Work Package

Table of Contents

1. Introduction

This deliverable contains three reports for the first round of BigDataEurope workshops held in the first six months of the project:

1. SC1: Big Data in the H2020 Societal Challenge Health, demographic change and well-being
2. SC3: Big Data in the H2020 Societal Challenge Secure, Clean and Efficient energy
3. SC5: Big Data in the H2020 Societal Challenge Climate action, environment, resource efficiency and raw materials

A summary and a copy of a detailed workshop report (including elicited requirements for the Big Data aggregator platform) is provided for each in the next Section. The report has, or will be circulated to all participants and other identified stakeholders. The communication will take place via multiple channels, including directly by email, project website and the respective W3C interest groups which have been set up.

2. First Round of Societal Workshops

The first three workshops were held in the fifth and sixth month of the project. Invitations were sent to the identified stakeholders, in multiple rounds. The workshops were designed around the blueprint provided in Deliverable 2.1, with minor adjustments to reflect the needs and characteristics of each community. A summary of workshop details, plus the full workshop report, are included below.

2.1 SC1.1 - Health Demographic Change and Wellbeing

The following table includes a summary of the workshop:

Date	21.05.2015
Venue	KoWI, Brussels, Belgium
Invitations Sent	56
Invitations Accepted (Registrants)	38
Attendees (Total)	19
Attendees (Project Consortium & Project Officer)	5
Attendees (Other)	14
Breakout Sessions	3

2.1.1 Agenda

- 10:30 - Welcome and Introductions
- 11:00 – invited speakers introducing BigDataEurope, and highlighting experience of big data exploitation and bottlenecks
    - BigDataEurope Project Introduction, Simon Scerri, Fraunhoffer IAIS (20 mins)
    - Big Data in Drug Discovery - linking data to answer key questions, Bryn Williams-Jones, CEO Open PHACTS Foundation (20 mins)
    - Big Data bottlenecks in Academic Bioscience, Director of Bioinformatics, Williams Harvey Institute QMUL Mike Barnes (20 mins)
    - On the need for intelligent access to big data in life sciences, George Paliouras NCSR Demokritos (15 mins)
- Outline of Breakouts (15 mins) Topics, requirements, and groups
- 12:30 Lunch and networking
- 13:15 Breakouts
    - Working groups to identify key issues and bottlenecks for exploiting big data in this societal challenge
- 15:00 Breakout feedback – 10 mins per group
- 15:30 Q&A, next steps, other meetings and workshops
- 16:00 Close

Workshop summary posted on the BDE website on May 29th 2015, slides are available in the BDE area in slideshare and also linked to this report.

2.1.2 Expectation and Background

The big data challenges in this sector are driven by variety and increasingly volume of data generated, stored, accessed, and analysed in the understanding of biomedical science. In the context of health and wellbeing, the intensive data generation involved in genetic profiling and other technologies used to gather information on health and disease represent significant hurdles for the understanding of disease and health. Indeed the understanding of the biology of the normal situation is mostly lacking, regardless of how this changes in disease, how disease progression or therapeutic intervention can be measured, and how data can be used in new ways to improve health and wellbeing.

The variety of data which is either publicly accessible relating to biomedical science is significant, and represents a significant barrier in the development of understanding of biology and disease. Standardisation of data relating to genetics, genomics, other ‘omic technologies, drugs, drug targets, clinical measurements, diagnostic testing, biomarkers or the development of biomarkers is in many cases lacking. Integration of all of this data into platforms which can be used to explore findings, generate hypotheses or otherwise generate knowledge is complex if even currently possible.

The development of widely applicable interoperable data standards is the key problem which limits the impact of big data approaches in healthcare. The development of interoperable data

standards across the value chain will drive new insights in biomarkers, disease categorisation, and patient segmentation by enabling the integration of diverse and heterogenous data sets. Addressing the fundamental questions in health through big data necessitates the interoperability of diverse and complex data types - which in isolation are arguably not enough to develop new insights into disease.

This workshop was split into two sections. The first involved presentations of project background, followed by illustrative examples of the problems of working with big data in drugs discovery from an industry and then academic perspective. The aim of these presentations was to highlight with real examples some of the key points outlined above. There was also a presentation highlighting experience from another project aiming to deliver the right information to domain experts without deep IT skills.

Within the interactive part of this workshop, participants were guided through selected questions, in order to get input and stable quantitative and qualitative material to feed Requirement Elicitation (RE) and further on to drive the Requirement Specification (RS) activities in work package 2. Attendees were pre-sorted into interest groups to ensure engagement on areas where their experience is relevant. Particular emphasis was placed on the technical/practitioners as in this domain there are probably more resources available than in other societal challenges. Development of inventories of resources and tools *etc* will be key, as well as clear use cases/business questions that can highlight the potential of big data approaches in this societal challenge.

Discussions from the three breakout groups were captured by facilitators on flipcharts which were then used to present the discussion back to the plenary group. Images of the flipcharts were captured, and are summarised in the following sections of this report:

2.1.3 Summary of Breakout Groups

2.1.3.1 SC1.1 - Group 1: Big Data Landscape

1. Background

Group 1 was tasked with discussing the big data landscape in the health sector in Europe as it currently stands: What progress has been made, what other projects exist, what gaps need to be addressed, and what risks and concerns need to be considered. Four people participated in the group discussion: Colm Carroll, Meta Geibel, Kimmo Rossi and___. Group members' knowledge of big data in the health sector was fairly high-level, so the discussion covered mainly broad topics and ideas rather than specific examples of work being done in this sector.

2. General Sentiments in the Health Sector

The group discussed the general attitudes of people involved in the health sector towards big data and its potential. Points raised were:

- There is a sense of great potential
- All parties involved have huge expectations, from patients to industry
- There is a sense of urgency about moving forward
- This sector is about people and their health - it could save lives, not just money or efficiency
- As a result, ethics and emotional aspects are more significant in this sector - people don't like others profiting off their illnesses

- Many different angles need to be considered, including
    - law
    - research
    - societal attitudes
    - industry
- This makes things complicated and confusing, but also exciting
- Privacy and security concerns “hang over everything like a black cloud”

3. The Data Situation

The group then moved on to discuss what is happening with data in the health sector. The general view raised was that the middle of the value chain is being addressed well; there is lots of good analysis of biomedical data being carried out. However the beginning and end of the value chain are lacking.

3.1 Beginning of the Value Chain: Data Collection

The group highlighted this as perhaps the biggest concern in the value chain. They suggested that the solution to better data collection would probably not be technical, but rather a matter of change in attitudes and funding. Points covered in the discussion of data collection were:

- There is a lack of longitudinal data which could help with earlier diagnosis
    - This is difficult to get when project funding usually only lasts a few years at a time
    - However, longitudinal data “exists in a few key cases” where patients have been followed throughout their lives (the group did not name examples)
    - Could this be collected directly from sensors on patients?
- Privacy concerns are the “black cloud” that needs to be addressed
    - Can’t get any data at all from some countries
    - This is partly a political issue, partly a societal one
    - There are some people who “just don’t care” about the privacy of their data
    - People share their own genomes, even in the USA where this could have significant repercussions for their insurance policies
- One group member commented, “We will all voluntarily be wearing sensors soon enough.”

3.2 End of the Value Chain: Learning from Results

The group also pointed out that although good analysis of biomedical data is happening, the results of this are not being fed back into changing the way healthcare works. They highlighted the need to “create a cycle” of results influencing policy and feeding back into better collection and analysis of data. Ideally the sector should be creating new solutions based on big data analysis, and actually changing practices to improve patient outcomes.

3.3 Fragmentation of the Sector

Another major issue with big data in the health sector is fragmentation. The group came back to this point several times, mentioning:

- There are lots of projects in the health sector working on elements of the big data value chain (the group did not name examples)
- In industry, funding is linked to disease areas and projects focus on specific diseases
    - Many separately-funded projects have data sets which could be brought together
    - Many such projects are possessive of proprietary data, and consider it important to their ability to get continued funding
    - How can we incentivise them to share data?
    - One option that has been tried is to set up a separate project to oversee interoperability, but funding for infrastructure-type projects is harder to get than funding for disease-related projects
    - Funding is needed to bridge gaps and bring together data from different projects and sources

4. Potential

The group covered some of the ways in which big data in the health sector has the potential to create positive changes.

4.1 Moving Forward

The group suggested that as top-down big data initiatives have not made much progress, big data in health may be better driven from the bottom up. Comments made were:

- Could we change the ownership of data, give it to patients?
- There will be “huge demand” from the bottom up if people have a share in the value of their data
- Patients want the use and benefits of their data maximised
- Patients should be able to access their own data, decide whom to share it with and for what purposes
    - One group member suggested that “a Scandinavian country” may already be doing this, but could not remember specifics
- Patients get something in return for sharing data - not necessarily money
    - The US website “Patients Like Me” was raised as an example of patients getting together to share information for mutual benefit

4.2 Synergy with Other Sectors

The group raised the idea of overlap with other Societal Challenges, in particular SC6 (Inclusive, innovative and reflective societies). Points raised were:

- Lifestyle is data is very relevant to health - people's shopping habits, exercise habits, living conditions all contribute to health
- There is big potential to linking to datasets from outside the healthcare sector
- Could clearer evidence connecting lifestyle to health change people's behaviour and lead to low-cost, non-pharmaceutical interventions?
- However privacy laws mean the connection between lifestyle and individual health outcomes is lost

5. Risks and Uncertainties

The major risk that the group raised was the unforeseen consequences of releasing personal health data, genome data, etc. They called this the "really scary" part of big data in the health sector.

- Could details from big health data affect employability?
- Combination of genome, phenotype and psychological data could be very powerful, potentially dangerous
- You can cheat a psychological profile, but you can't cheat your genome
- What happens when we can associate a concrete cost with your health risks, or your lifestyle choices?
- This will raise societal questions:
    - Should people pay according to risk, for factors outside their control?
    - Would governments encourage or mandate changes in lifestyle or behaviour?
    - Compare with the car insurance, where discrimination based on sex was made illegal
    - But compare also with tax on cigarettes, alcohol - could there be a fat tax?
- Need to maintain socialised medicine to avoid a situation where your genome affects your health insurance costs

The group also mentioned that a unified healthcare system will have language and communication issues, if there is a single market for healthcare services.

6. Next Steps

Although the group did not suggest many specific examples of projects or groups BDE might approach in this sector, they did suggest some steps forward:

- Carry out landscape mapping of this sector, and what projects are happening
    - Get information from Commission databases, or other public databases?
    - Also look into national-level projects, people in other geographic areas
    - There is no such thing as a "European dataset"; data crosses borders
- Talk to smaller existing projects who are trying to put data together

    - Carry out one-to-one talks? Surveys?
- The physics community has a culture of making data open early
    - This is a good example of a bottom-up change in culture, where people recognised the value to themselves of sharing data
    - How did it come about? Is there anything we can learn from the physics community?
- IBM Watson Health is bringing together lifestyle and health data to help inform medical professionals
    - This has already been rolled out in some hospitals/research facilities - talk to them
- BioASQ is another good example
- Open PHACTS did well by looking at what end users actually want - but who are our "end users" for this project?
    - Healthcare professionals/administrators, or citizens/patients, or someone else?
    - Whose requirements should we be examining, so that we can work on concrete case studies to address their needs?
    - Who is really influencing decisions about policy, and how can we demonstrate the value of sharing data to them?

2.1.3.2 SC1.1 - Group 2: Technology and Data

This section of the report details the breakout session for group 2 (Technology and Data) at the SC1.1 workshop in Brussels. These are based on the transcripts of the notes made on the flipover (photos can be found in this PDF document .

1. General Observations

Group 2 discussed mostly different aspects of data and metadata regarding a potential big data platform. The group consisted of five people (Michael Barnes, Laura Ines Furlong, Christine Chichester, Georgios Paliourias and Jan Kors). All persons were from the biomedical/pharma domain and we discussed mostly this subfield. At least three persons had worked with the Open PHACTS platform and discussed mainly data features and needs that were partially addressed in that project.

2. Veracity/Variance

During the discussion it became clear that variance and veracity of the data are perceived as the most important aspects of the data in this domain. To quote one of the participants: "Link two datasets and you have Big Data". The participants agree that one of the big challenges is to link different datasets into one integrated platform. These datasets will come from different sources:

- Structured Databases
- Text-mining of documents
- Crowd-sourcing / users

Integrating this will pose challenges for identifying the reliability and trustworthiness of the data. A platform would therefore need to have

- ways of validating data and conveying this validation to the user
    - record and display *provenance*
    - record reliability through probabilities (although participants indicate that these are note universal)
    - record whether data is expert-validated or not (user-generated?/crowd sourcing)
    - visualize the provenance/reliability of data and statements
- Ways of curating data, both automatic and semi-automatic

3. Types of Data

We then discussed different types of data that would be linked. For each type of data, we discussed the level of standardization and the existing standards. As a whole, we concluded that there are many standards in this domain and that mapping these standards is one of the major challenges in this field.

- *Data with 'good standardization'*
    - Genome data (Fastq, BAM, "Good" standards)
- *Data with mixed levels standardization*
    - Electronic Health Records (ICD10, CDISC, Snomed: well-defined data standards, also free text with MESH/UMLS tags)
    - Publications (MESH for abstracts, free text)
    - Chemical structures (SD files, SMILES, INCHI, MIABI)
    - Medical imaging (MRI standards, PAX, Raw image data)
    - Pathways (SBML, Biopax)
    - In silico models (SBML, PharmML)
    - Phenotype (Gene ontology, HPO, Snomed)
    - Drug data and pharmacology (ATC)
    - Anatomy [multiscale data] (FMA, UBERON, NCI)
    - Environmental Exposures
- *Data with low levels of standardization*
    - Metabolomics (there are some databases, some identifiers, lot of free text)
    - Assays (BioAssay ontology, although this is a good and singular standard, it is unsupported)

The participants concluded that this list comprises a good overview of the type of data in the biomedical/pharma domain. However, they expressed that it could be very interesting to link this data to all kinds of related datasets, including weather data, traffic and mobility data, social data etc.

4. Important aspects of the data

After this, we discussed aspects of the data that would inform the technologies that are to deal with them. We list them below.

- *Time.* A number of datasets and use cases have temporal dimension. Changes of values in electronic patient records for example change over time. There are no good standards that are used in this domain.
- *Units.*Different units of measure are used. A good standard or mapping between them would help.
- *Multilinguality.* The issue of multilinguality takes two forms:
    - There are a number of multilingual datasets including literature, patents, electronic health records. To have access to these, multilinguality would be needed.
    - Only a small part of the terms in vocabularies (UMLS) have non-english terms. This hinders non-English speakers in their access.
- *Ambiguity / Homonimity:* Mostly in publications, terms are used ambiguously. A system would need to be able to deal with this
- *Data sparseness.* There are often many missing variables when analysing large groups of patients. This is a recurring issue and relates to volume (see next section)

5. The Other V's

We briefly discussed the two remaining "V's of Big Data", velocity and volume.

- *Volume:*
    - Genome files are very large (~500GB for raw or aligned data, ~100MB for a VCF file). Analyses over 100.000 genomes are problematic and require non-relational databases.
    - Raw imaging files are problematic (for example scans)
- *Velocity*
    - Indication of publication speed is that 2 new pubmed articles are published per minute.
    - In a research setting speed is likely not a giant issue, at least compared to other societal domains.
    - In a clinical setting speed could be interesting, especially with new opportunities such as *crowdsensing*

6. Data Value Chain

We finally very briefly discussed what would be expected from a BDE platform based on the data value chain. We here replicate the points brought forward by the participants for each of the times in the data value chain. Because of the short time, this is by no means a complete list.

1. *Generation and Acquisition*
    a. Provide easier access to privacy-protected data (through anonymization services for example)

   b. Access to User generated content or crowdsensing
2. *Analysis and Processing*
   a. Identify outliers
   b. Combine analysis methods, including text mining and sequence analysis
   c. Detect or predict events on both a patient level or at an epidemiological level
3. *Storage and Curation*
   a. Deal with mapping of standards and ontology alignment. Be able to deal with one-to-many mappings
   b. Deal with data provenance and calculate reliability of data
4. *Visualisation and Usage*
   a. Support drug discovery
   b. Do episodic disease prediction (in 24hrs, patient X will have condition Y)
   c. Support verification of information
   d. Support pharmacovigilance
5. *Data-driven Services*
   a. Provide APIs (this was deemed critical to the participants)

2.1.3.3 SC1.1 - Group 3: Legal and Policy Issues

This section of the report details the breakout session for Group 3. This discussion is based on transcriptions of the flipchart notes, photographs of which can be found in the following location.

1. Preamble and Background

Group 3 was tasked to discuss some of the legal and policy issues relating to big data and was made of up the following attendees (Edoardo Camilli, Helena Ursic, Meta Geibel, Mark Goldammer) facilitated by Bryn Williams-Jones. Given the makeup of the group, policy issues were covered in general as there was little practical hands-on experience of the application of big data to healthcare and demographic change practical issues.

In general, there was wide recognition of the issues presented by using patient data whilst maintaining privacy and security, and maintaining anonymisation. This can in part be mitigated by a focus on data which could be safely abstracted from personal data for use in particular business questions, from a central controlling authority – to some extent orthologous to the data management groups in pharma companies.

With respect to policy, a key point to keep in mind is that policy tends to be a retrospective process, and what is most needed are smart ideas that policy makers can highlight and promote. Sharing best practice is fundamental in this domain, and BDE has a chance to tackle some large issues if this route was taken.

Given the size and scale of the issues involved in using big data in the patient setting, care should really be taken on selecting use cases to pilot here, and efforts safely abstracting 'omics data could be used as proof of concept. Success in this domain for BDE should be judged by

the availability of new data, which can be used in smart solutions – new value-added data sets and analytical workflows.

2. Observations

- Abstracting from patient data
    - Can there be different levels of safety/security
    - Is age/sex the only thing we're left with?
- Companies are starting to work directly with patient organisations to circumvent these concerns
- Many patients will give data for free
    - How can you maintain anonymity
    - Do patients really understand the full reach/implications of open/free data?
- Is this more about signal detection if mining social media (as opposed to hard data)
- Health data
    - Personal/sensitive
        - Safety
        - Robust vs hack
- Impact of public health on privacy expectation
    - Eg transmissible disease
- Concerns over monetisation by pharma/big corporate
    - Reputational damage
    - Open data can't be closed again
- Aggregation effects, benign data in combination could become dangerous
- Bringing the debate to a higher level as healthcare is very individual focused
    - The vaccines argument – negligible personal benefit vs societal protection
- Policy influence is really about smart ideas
    - Come with a solution
    - Promote best practices
    - Implement best practice
    - Evaluation should be bottom-up
- Standards need to be interoperable too
    - Sharing experience with policy implementation
- How does retail/food sector deal with personal/private data?
    - Is there a different expectation of privacy in healthcare?
- Veracity
    - And data quality is vital

    - What is the expectation of precision?
        - Probably varies by use case
- Is expectation of privacy different for the facebook generation?
    - The google flu trends example
    - Is it sustainable?
    - Will policy bring protection?
- Useful tools to give signals with existing technologies
    - Policy can encourage but not create
- Success factors
    - New data
    - Smart solutions
    - New assets/value propositions
    - Entrepreneurship

2.1.4 Appendices

2.1.4.1 Workshop Invitation and Background Requirements

Stakeholder engagement will be focussed around the yearly workshops organised (at least one) per each of the seven SC communities. As per the DoW in the first round (Year 1), the workshops' main objective will be to elicit requirements for the big data integrator platform. In the second and third year, the focus will be on reviewing the architecture for prototype implementation, and platform evaluation and showcasing, respectively. To facilitate the organisation of the resulting (at least) 21 workshops and ensure a consistency in results across all SCs, we have established a workshop blueprint, described below.

a. Short paragraph explaining BDE in general and the impact to the specific domain.
b. Short paragraph why this workshop was initiated, and what the expected outcome should be
c. invitation letter text (see appendix 2.1.4.2) which emphasizes that the need in this first instance is to focus on the workshop on May 21st, 2015
d. http://w3.org/community/bde-health/ group now setup to serve highlight visibility
e. See the guidelines to a SC Workshop in Deliverable D2.1
f. See the list of stakeholders and EC names in 2.1.4.4

Along with the invitation letter, we need to ensure that the focus is on invitee's who are technically dealing with big data and informatics systems in biomedical data already. Given the public resources already available in this space, practical implementations and a focus on real use cases is needed. Invited speakers active in the domain will give an introduction and outline of their activities, and will be briefed to expand on use cases which will be of relevance to BDE. Special effort will be made to ensure that the informatics perspective highlights domain difficulties which may be unfamiliar to the BDE technical resources.

2.1.4.2 Text of Invitation Letter

Dear Sir/Madam,

*We would like to invite you to a workshop on **Big Data in the H2020 Societal Challenge Health, demographic change and well-being** in **Brussels** on the **21st May 2015**. As a recognised stakeholder in the **Health** sector, you will have the opportunity to influence the design, and ultimately benefit from, the Big Data platform that the BigDataEurope project will deliver. This platform aims to facilitate Big Data usage in real world examples, and will consist of an architecture, components, guidelines and best practices to make the best of Big Data in this case in the setting of health.*

This workshop will introduce the background, cover the main challenges, and seek real examples of the potential, challenges and complexities of using Big Data in the health domain. The results will be used to design and realise the required ICT infrastructure and support the use and deployment of the platform – maximising the opportunities of the latest European RTD developments, including multilingual data harvesting, data analytics, and data visualisation.

***Event Registration**: More information about workshop registration, agenda, venue, etc. can be found or will be provided here: http://www.big-data-europe.eu/event/sc1-brussels-2015/*

*In addition, as a representative of one of our target communities, we would like to engage with you in the long term, in order to **identify your big data technology needs, challenges and requirements**. We therefore invite you to give due consideration to our participation offer by considering one or more of the following different levels of stakeholder engagement:*

- *Subscribe to the BigDataEurope Newsletter;*
- *Join your respective W3C Community Group;*
- *Participate in the planned key Stakeholder Workshops, including the one announced above;*
- *Participate in one of our BigData Pilots (subscribe for more information);*
- *Follow us on the Social Media (Twitter, LinkedIn,SlideShare).*

*This invitation is **extended to others in your network** with an interest in the challenges of Big Data in your sector. Where possible we would be grateful if you could focus on informatics professionals who have current experience of working with Big Data on a daily basis. When forwarding the invitation, please include bryn@openphactsfoundation.org in the carbon copy field.*

We look forward to your participation!

Kind regards,

Bryn Williams-Jones, CEO The Open PHACTS Foundation
Societal Challenge Representative on behalf of the BigDataEurope Consortium
bryn@openphactsfoundation.org
www.big-data-europe.eu

2.1.4.3 Advertised Workshop Description and Agenda

The big data challenges in this sector are driven by variety and increasingly volume of data generated, stored, accessed, and analysed in the understanding of biomedical science. In the context of health and well being, the intensive data generation involved in genetic profiling and

other technologies used to gather information on health and disease represent significant hurdles for the understanding of disease and health. Indeed the understanding of the biology of the normal situation is mostly lacking, regardless of how this changes in disease, how disease progression or therapeutic intervention can be measured, and how data can be used in new ways to improve health and well being.

The variety of data which is either publicly accessible relating to biomedical science is significant, and represents a significant barrier in the development of understanding of biology and disease. Standardisation of data relating to genetics, genomics, other 'omic technologies, drugs, drug targets, clinical measurements, diagnostic testing, biomarkers or the development of biomarkers is in many cases lacking. Integration of all of this data into platforms which can be used to explore findings, generate hypotheses or otherwise generate knowledge is complex if even currently possible.

The development of widely applicable interoperable data standards is the key problem which limits the impact of big data approaches in healthcare. The development of interoperable data standards across the value chain will drive new insights in biomarkers, disease categorisation, and patient segmentation by enabling the integration of diverse and heterogenous data sets. Addressing the fundamental questions in health through big data necessitates the interoperability of diverse and complex data types - which in isolation are arguably not enough to develop new insights into disease.

- **Welcome & Introduction**, 30 mins
    - Tour de Table
        - Name and affiliation
        - Role in organisation
        - Connection to big data & data management
        - Expectations for the workshop
        - Expectations for the BigDataEurope project
- **Introductory Talks**, 1 hour
- Setting the scene with a background to the BigDataEurope project, the approach taken, and some explanation of the strategic fit of the external speakers
        - BigDataEurope - overview from public launch (20 mins)
        - What is big data in drug discovery? (20 mins invited speaker) – an example project which combines many of the challenges seen in big data in a health setting
        - Which public big data sources are available, or coming soon? (20 mins invited speaker) an introducing to life science computing infrastructure, and what data interoperability really means in this space
- **Outline** of afternoon session - interactive breakouts (15 mins)
    - Introduce the requirements process with some example questions, and ensure the rational for people grouping is explained
    - split attendees into groups and introduce the topics they'll be working on
- **Lunch,** 45 mins
- **Interactive Sessions,** 2 hour

Using the requirements elicitation process outlined in Deliverable 2.1, (Stories, Persona's, Data, technology) Within the workshops, participants will be guided through selected questions, to get input and stable quantitative and qualitative material to feed Requirement Elicitation (RE) and further on to drive the Requirement Specification (RS). Attendees will be pre-sorted into the right groups to ensure engagement on areas where their ep. Particular emphasis will be placed on the technical/practitioners as in this domain there are probably more resources available in this SC. Development of inventories of resources and tools *etc* will be key, as well as clear use cases/business questions that can highlight the potential of big data approaches in this societal challenge. Timings indicated are for guidance only and participants will be guided by a facilitator who will ensure that the right topics are covered. The aim is to capture all perspectives, there are no right and wrong answers, and long in depth debates on technical solutions/problems should be avoided.

- Group 1 -
    - [30'] **Data-centric initiatives in the SC –** identify other project and resources that BDE should be aware of, engaged and/or collaborating with.
    - [30'] **Stories and persona –**describing current status, and the different types of people involved
    - [30'] **Big Data use-cases in the SC –** which big data would be needed, does it exists, what questions could be answered
- Group 2
    - [30'] **Technologies and tools used and envisaged –** gathering information on the currently available/used technologies, feedback on approaches that have been tried but might not work, are there known gaps or pitfalls?
    - [30'] **Data requirements -** accessibility, availability, licensing and consent, geographic restrictions?
    - [30'] **Technology requirements – f**unctional and non-functional requirements of the platform
- Group 3
    - [30'] **Industrial session/ EU policy requirements -**
    - [30'] **Legal issues around (big) data, Governance, Data portability –** with emphasis on any already know restrictions on data access, availability, distribution *etc* with particular types of data
    - [30'] **Other requirements –** areas of this SC where big data solutions are needed but have not been covered in the discussion today
- **Summary, outreach & farewell**, 1 hour
    - feedback per group (10 mins each)
    - Q&A session – leveling session to ensure other points have not been missed for each group
    - Give participants clear picture of the workshop's outcomes
        - set scene for follow ups through interviews, and highlight plan for workshop in years 2 and 3. Early testers/adopters particularly welcome and should be captured for follow-up

2.1.4.4 Attendees

The following table lists registered attendees for the workshop:

Last Name	First Name	Institution/Company	Country
Barnes	Michael	Queen Mary University of London	UK
Camilli	Edoardo	Hozint	Belgium
Carroll	Colm	Innovative Medicines Initiative	EU/Brussels
Chichester	Christine	Swiss Institute of Bioinformatics	Switzerland
Chuma	Andreattah		
CUPERS	Philippe	EC	EU/Brussels
de Boer	Victor	VU Amsterdam	Netherlands
Durst	Ludovica	Lynkeus	Italy
Estival	Stephane	CVSE.be	Belgium
Furlong	Laura Ines	IMI Barcelona	Spain
Geibel	Meta	EC	EU/Brussels
Goldammer	Mark	EC	EU/Brussels
Jenko	Sasa	EC	EU/Brussels
Kerstiens	Barbara	EC	EU/Brussels
Kloots	Rob	TrustingTheCloud	Netherlands
Kors	Jan	Erasmus Med. Centre	Netherlands
Larrabide	Ibai	Zabala Consulting	Belgium
McNeice	Kiera	Open PHACTS Foundation	UK
Morley-Fletcher	Edwin	Lynkeus	Italy
Paliouras	Georgios	NCSR Demokritos	Greece
Palmirani	Monica	U Bologna	Italy
Peetso	Terje	EC	EU/Brussels
Ranjan	Ravi	GlobalVein	
Renmans	Bram	Tenforce	Belgium
Roesems	Gisele	EC	EU/Brussels
Rossi	Kimmo	EC	EU/Brussels
Scerri	Simon	IAI U Bonn	Germany
Schee genannt Halfmann	Sebastian	U Maastricht	Netherlands
Sellez	Adrien	Birmingham University	UK
Spek	Wouter	TIB Development	Netherlands
Ursic	Helena	EC	EU/Brussels
van Bochove	Kees	The Hyve	Netherlands
Williams-Jones	*Bryn*	Open PHACTS Foundation	UK

This table details the breakout groups for those who attended the working session, names in italics facilitated the discussion:

Group 1	Participants	Group 2	Participants	Group 3	Participants
Big Data Landscape	*McNeice, Kiera*	**Technical/ Data**	*de Boer, Victor*	**Policy**	*Williams-Jones, Bryn*
	Carroll, Colm		Barnes, Michael		Camilli, Edoardo
	Spek, Wouter		Furlong, Laura Ines		Ursic, Helena
	Rossi, Kimmo		Paliouras, Georgios		Geibel, Meta
	Renmans, Bram		Chichester, Christine		Goldammer, Michael
			Kors, Jan		

2.1.4.5 Follow-Up Message

Follow-up message sent to stakeholders highlighting comms channels, survey and workshop summary:

Dear Friends,

*We would like to thank you for your participation in our workshop on **Big Data in the H2020 Societal Challenge Health, demographic change and well-being** held on the **21st May 2015**.*

We are working on the full summary report of the workshop and will be in touch with a link to the report soon. For now we'd like to draw your attention to other ways you can engage with, and help the BigDataEurope project.

- *To keep in touch with the latest BigDataEurope project developments, please visit our website and subscribe to our Newsletter;*
- *the slides presented at the workshop will be available on our SlideShare*
- *We have established W3C Community Groups for each Horizon 2020 societal challenge, please join us there*
- *Participate in the other planned key Stakeholder Workshops,*
- *Participate in one of our BigData Pilots - please contact us directly for details*
- *Follow us on the Social Media (Twitter, LinkedIn,SlideShare)*

If you have 12 minutes to spare, we would love to have your input to our big data survey and help set the requirements for our big data aggregator platform. Alternatively if you have a little more time to spare, please me directly for a more in depth discussion.

Thanks for your participation!

Kind regards,

Bryn Williams-Jones, CEO The Open PHACTS Foundation
Societal Challenge Representative on behalf of the BigDataEurope Consortium
bryn@openphactsfoundation.org

www.big-data-europe.eu

2.1.4.6 Pictures

Pictures from the workshop are available in the BDE Flickr stream, this link is specific for those captured at this workshop.

2.1.4.7 Presentations

Slidesshare links to the presentations used in the workshop are as follows:

- BigDataEurope Project Introduction, Simon Scerri, Fraunhoffer IAIS(20 mins)
- Big Data in Drug Discovery - linking data to answer key questions, Bryn Williams-Jones, CEO Open PHACTS Foundation (20 mins)
- Big Data bottlenecks in Academic Bioscience, Director of Bioinformatics, Williams Harvey Institute QMUL Mike Barnes (20 mins)
- On the need for intelligent access to big data in life sciences, George Paliouras NCSR Demokritos (15 mins)

2.1.4.8 Other Workshop Opportunities in this Societal Challenge

More granular requirements are needed to establish BDE data integrator in this domain, the Open Bridges meeting in November 2015 in Hinxton, UK brings together 2-300 domain experts who will discuss data interoperability and infrastructure for life sciences. Opportunities for BDE should be explored. Particular emphasis is expected on name space and ontologies to facilitate data interoperability from multiple domains.

2.2 SC3.1 - Secure, Clean and Efficient Energy

The following table includes a summary of the workshop:

Date	16.06.2015
Venue	NRW Representation, Brussels, Belgium
Invitations Sent	110
Invitations Accepted (Registrants)	~25
Attendees (Total)	25
Attendees (Project Consortium & Project Officer)	7
Attendees (Other)	18
Breakout Sessions	2

2.2.1 Agenda

10:00 – 10:10 Welcome & Introduction
- Introducing attendees & workshop goals (F.Mouzakis)

10:10 – 10:40 Introductory Talks
- Project Scope and Community opportunities (S. Auer Project coordinator)
- Data management and Big Data in Energy domain (F. Mouzakis)

10:40 – 11:00 Energy Industry
- Data: The Gate to a Smart Energy system – Views of the Electricity Industry (Mr. Hans ten Berge EurElectric Secretery General)

Coffee break

11:15 – 11:35 BigDataEurope Technology background
- Tools, methods and examples (Dr. A. Ikonomopoulos NCSR Dimokritos)

11:35 – 12:15 Resource Forecasting
- BigData applications for RES siting and forecasting (Mr. Martin Qvist VESTAS S.A.)
- Resource forecasting, data management challenges and BigData opportunities (Prof. G.Kallos UoA Forecasting Unit)

12:15 – 12:30 System monitoring
- Data management for a RES developer – use case (Mr. A.Papoutsakis TERNA S.A.)

12:30 – 12:45 Smart grids
- Smart grid community, data management challenges and BigData opportunities (Dr. S. Tselepis SmartGrid.eu Platform)

Lunch break

13:10 – 14:45 Break-out sessions
- BigData platform Requirement elicitation
- Interviews and Community building

14:45 – 15:00 Summary, Outreach & Farewell

15:00 End

2.2.2 Workshop Scope and Structure

The aim of the workshop, the first of a scheduled series on the domain, is the identification of current and future challenges for data management and analysis in the energy domain; challenges to be tackled with the evolving Big Data technology. In the workshop real examples of the potential, challenges and complexities of using big data in the energy domain will be discussed. The workshop addresses a wide audience comprising data users from a variety of fields in the energy domain.

The outcome of the workshop will support the design and realisation of the necessary ICT infrastructure on which the deployment and use of the BigDataEurope platform will be based. The platform targets the facilitation of big data usage in real world examples and will consist of the architecture, components, guidelines and best practices.

BigDataEurope platform will offer to the interested participating third parties the opportunities of the latest European RTD developments, including real time streaming, multilingual data harvesting, data analytics and data visualisation.

The workshop was divided in three parts, described in the following paragraphs.

- *Part I:* General introduction to the BDE background, objectives and targets, as well as an overview of the tools and technologies envisaged within the project. An extended summary of the BDE technology overview is provided in the next section of this report (section 3). The workshop participants were further on introduced to three sets of

questions selected to map the user requirements and contribute to the Requirement Elicitation (RE) and the Requirement Specification (RS) activities in WP2.

- *Part II:* Keynote presentations were given by invited speakers in selected data management related topics.

Topic A: Electricity Industry

The views of the Electricity Industry on Data value was presented by Mr Hans ten Berge (Secretary General of EurElectric).

Topic B: Resource Forecasting

The data management challenges in energy resource forecasting were presented by Mr Martin Qvist (Head of Super-Computing and BigData applications of VESTAS, the leading company in wind energy sector) and Prof. George Kallos (Forecasting Unit of UoA)

Topic C: System monitoring

The data management challenges in system monitoring and a candidate use case on the topic were presented by Mr. A.Papoutsakis of TERNA S.A., a RES developer and operator.

Topic D: Smart grids

The data challenges in Smart Grids field were presented by Dr. S. Tselepis (SmartGrid.eu platform) and Mr Tierry Pollet (Landis+Gyr).

- *Part III:* The third section constituted the interactive part of the workshop during which the participants were split into two breakout sessions according to two types of stakeholder groups to serve the user requirements elicitation process.
    1. In the first group the candidate use cases were discussed namely the asset monitoring (CRES, TERNA and AlSol) and resource forecasting (TERNA and UoA).
    2. In the second group the technology aspect were discussed (NSCRD, Franhaufer, VESTAS).

2.2.3 Domain Topic Reviews

2.2.3.1 Topic A: Electricity Industry

Mr Hans ten Berge (Secretary General of EurElectric) presented the views of the Electricity Industry related to Data Value. EurElectric is the sector association which represents the common interests of the electricity industry at pan-European level, with full members from 32 European countries. EurElectric (www.eurelectric.org) is invited as one of the major stakeholders for BigDataEurope. Its work (through the maintenance of numerous working groups) is focused towards the main objectives that are: delivering carbon-neutral electricity in Europe by 2050, ensuring a cost efficient and reliable supply through an integrated market as well as developing energy efficiency and the electrification of the demand-side for climate change mitigation.

The key data related aspects of the view of the electricity industry towards a smart energy system were presented, including data handling, role of actors, data hub models, data protection and privacy, data transparency, relation with TSOs on data sharing and the role of ICT players and Telcos. The principal role of the individual consumer was emphasized.

An analysis of the data flows for the business of the electricity suppliers and the electricity distribution players (market facilitators and system operators) as well as the data collection and data process from the DSO point were presented. Of primary importance was the presentation of three different cases of data provision hubs, namely the DSO market facilitator, the Third party facilitator and the Data Access Point Manager. Information about the joint platform (EurElectric, ENSTO-e, EDSO, CEDEC, GEODE, DG Energy etc) on data exchange for electricity system purposes was also delivered.

In the concluding messages the following were pointed out in relation to BigData applications:

- Digitalization and exploitation of consumption data is the key area for new tailor-made services and products. Customer related data considered as 'personal' are covered by current EU policies.
- Smart metering data is the cornerstone of DSO optimization, operation and planning of distribution networks. There is a variety of data hubs, formats and market models among Member States.
- The importance of the cooperation of DSOs and Telcos for data acquisition and system control is increasing as we head to smart grids and new business opportunities.

In Appendix 2.2.5.1the link for the presentation is given for full and accurate reference.

2.2.3.2 Topic B1: Asset Siting and Resource Forecasting

The views of the leading manufacturer in Wind Energy on BigData applications were presented by Mr. Martin Qvist (Head of Super-Computing and BigData applications of VESTAS, the leading company in wind energy sector) from VESTAS Wind Systems A/S.

Wind Industry meets the challenge of providing highly competitive renewable energy in a highly stochastic resource field. Its success was supported by data management in system design and production, resource assessment, system optimisation as well as asset and market management.

The data management challenges, for a leading manufacturer providing ~6GW globally installed capacity per year (2014 data), in the fields of project development and operation support were presented.

The asset siting is supported by a global climate library (data size order of 10^{14}) along with geospatial (terrain, land use etc) and electrical grid related data bases. The climate library extends for more than 15 years with a time and spatial discretisation of 1hour and 3km, respectively. The applications based on these BigData support, except from the manufacturer, the wind power plant developers, the operators and decision makers in energy policy.

The VESTAS resource forecasting (i.e. extremely localised wind forecasting) system was presented, including descriptions of the data analytics and output volume, computational effort and data layer interactions.

Finally the challenges of BigData technology and energy related applications were presented from a leading manufacturer point of view:

- BigData technology momentum effects on application pyramid
- Visualization tools and interactivity
- Deep integration

In Appendix 2.2.5.1the link for the presentation is given for full and accurate reference.

2.2.3.3 Topic B2: Resource Forecasting

A thorough review of resource forecasting was presented by Prof. George Kallos from Atmospheric Modelling and Weather Forecasting Group of UoA. The presentation covers the data management challenges in the service provision and in the research field.

The Numerical Weather Prediction (NWP) current practices were presented in detail for wind, solar and tidal/wave energy resources. Details were given related to time and special model discretisation and resource requirements. The example of the MARINA PLATFORM (an FP7 project) for the mapping of wind, wave, tidal and currents resources for NE Atlantic and Mediterranean was discussed in detail.

In the sequel, an overview for the resource assessment and forecasting was presented, describing the analytics requirements of LAM (Limited Area Modelling) either by Computational Fluid Dynamics (CFD) or Mesoscale Meteorological (MM) modelling and commenting on data volume, computational effort and output requirements.

The current practice challenges of NWP in RES and energy management operations related to with LAM, RCM & GCM (regional/global models) and the future perspectives were identified. In the latter direction the integrated model RAMS/ICLAMS (a high resolution Numerical Weather Prediction system).

In conclusion the tendencies in NWP were summarized:

- GCMs will run at grid resolutions of the order of ~5km for the entire planet for both weather and climate predictions
- LAMs will run at grid resolution of the order of ~100 m and time steps of a fraction of a second and forecasting horizon of 5-10 days
- Both modelling categories will become of the concept of “Earth simulators” where atmospheric (physical & chemical), ocean and land processes will be directly coupled
- All modelling categories require new methodologies for data handling. Ensemble modelling requires at least an order more of data handling
- A large portion of such applications is directly and indirectly related to energy production and management
- Data visualisation is emerging as an important requirement

In Appendix 2.2.5.1the link for the presentation is given for full and accurate reference.

2.2.3.4 Topic C: System (Asset) Monitoring

The data management requirements for system (asset) monitoring were presented by Mr. A.Papoutsakis of TERNA S.A., a RES developer and operator activated in several countries with a variety of RES power production assets.

The data acquisition network was described, comprising resource measurements, power plant SCADA data as well as specific condition monitoring system data.

The asset performance evaluation procedures (the main objective of the data management and analysis from the operator’s point of view) were briefly presented.

The concluding comments were:

- Data management requirements are exploding as asset fleet is expanding; except from coping with the new data streams there is also the need for revisiting old data sets for model optimisation and research
- New technologies in condition monitoring field and performance evaluation will demand higher streaming and storage capacity as well as computational efforts
- Forecasting developments (for market support) are expected to increase requirements for analytics and their execution time
- Information is gathered by a variety of sources. Although the data variety itself is limited, there is variety in data storage and transfer.

Supportive information regarding condition monitoring data streaming is found in presentation [8] (see 2.2.5.1).

2.2.3.5 Topic D: Smart Grids

The community, the metering data management challenges and Big Data application opportunities in Smart Grids field were presented by Dr. S. Tselepis (SmartGrid.eu & European Technology platforms).

The European Technology Platform on SmartGrids was presented, focusing on its mission, projects, members and stakeholders. In the sequel, the fields where collected data would need processing and analysis, offering decision support towards the Smartening of Electricity Grids were discussed.

The Smart metering data utilisation was presented in detail. The analysis regarded Grid Internal Data which deal with the observability of the electricity network and Grid External Data which deal with prosumers issues.

The concluding remarks, focusing on metering opportunities where BigData technology may apply:

- Large data mining processes considering operational and planning applications
- Data protection tools (access, authentication and encryption)
- Distributed online analytical stream processing system with spatial and temporal dimensions
- Development of specialised analytics (consumption behaviour, metwork modelling etc)
- Standardization of data models
- New IT solutions to process large data streams (in cooperation with the bank industry)
- Data publishing systems
- Data storing systems (E.g. web dashboards for managing data, etc.).
- Best practices and recommendations for data privacy and data use by the different stakeholders of the electric system.

In the sequel Mr Tierry Pollet (Landis+Gyr, ETP) presented the results of a survey studying the impact of specific technology drivers (such as power type, technology advances like BigData, grid type etc) on the Utility of the future (see ref [10] in 2.2.5.A).

2.2.3.6 Round Table Discussion

During the presentations the following were commented:

- The importance of citizen involvement (smart metering)
- The absence of DSO/TSO and major ICT providers from BigDataEurope consortium (only in stakeholder level)
- BigDataEurope platform specifications (will data and analytics be open, data processing is local or distributed)
- Cyber security aspects in relation to distributed analytics
- Importance of consumer options; consumer need for new services from the Utilities and this drives the developments
- Consumer options effects on market regulation
- Variety of business models in EU countries
- DSO role changes (market dynamics, consumer choices)
- Business opportunities in utility new services
- EC (DG Energy) views on Market Design and Grid operation in relation to data

2.2.4 Summary of Breakout Groups

2.2.4.1 SC3.1 - Group 1: Report of the Discussion Regarding Use cases

1. Preamble

Participants:

- F. Mouzakis (CRES)
- S. Tselepis (CRES)
    - Papoutsakis (TERNA)
- T. Kyritsis (ALSOL)
- G. Kallos (UoA)

The basic task of Group 1 was to elaborate on the current status of data management in the fields of system monitoring and resource forecasting within SC3 domain and the prospects for developing use cases for BigDatEurope platform.

2. General Considerations

The group discussed the general attitudes of people involved in the energy sector towards big data usage and potential. The key points that were raised and discussed were the following:

- In the industrial sector the data are generated internally and as such they are proprietary
- Large companies develop in-house BigData applications or the rely on available commercial tools provided by the major ICT companies
- The majority of the energy industry related companies they do not exploit the full value of their data, as they do not invest in BigData solutions
- The convergence of Information Technology (IT) with Operation Technology (OT) is of primary importance; this is a field for BigData applications

- Available standards in data exchange in energy domain (i.e. IEC 61400-25).

3. Candidate Use-case

The group considered two options for BigDataEurope platform:

System monitoring

The scope is the development of a platform capable to provide a complete asset fleet operational and condition monitoring featuring:

- On line analysis
- Merging of conventional SCADA data with advanced (research) condition monitoring systems
- Optimisation capabilities
- Visualisation capabilities

Data description can be provided by CRES, TERNA and ALSOL.

Resource forecasting

The scope is the development of a platform capable to provide the data management of localised (point) weather prediction in country level featuring:

- Conversion of resource to energy forecasts
- Inclusion of operator on line weather measurements and power production data
- Periodical optimisation of models based on stored simulation data

The simulation data could be provided by UoA.

2.2.4.2 SC3.1 - Group 2: Report of the Discussion regarding Technologies and Data

1. Preamble

Participants:

- M. Qvist (VESTAS)
- S. Auer (Fraunhofer)
- V. Karkaletsis (NCSRD)
- N. Ikonomopoulos (NCSRD)

The basic task of Group 2 was to discuss functional and non-functional requirements for the BDE platform. Group members' knowledge and experience of Big Data and data management in the energy sector was such so that the discussion covered the following topics.

2. Data Acquisition

During the data acquisition phase, energy domain experts identified data heterogeneity as something that BDE could help with. Experts typically work with streams of data originating at sensors located on wind turbines and other devices. Other data of interest include more traditional, yet still streaming, multimedia data, such as video. These data are analysed both on the fly as well as in-situ, i.e. after having been stored in data centres. For device-originated data, standards of interest include the common information model (CIM), which "provides a common definition of management information for systems, networks, applications and

services, and allows for vendor extensions"[1], as well as the IDIS data object model for data captured by smart meters[2]. Data transmission from and between energy meters is typically handled via the DLMS/COSEM (Design Language Message Specification/Companion Specification for Energy Metering)[3] or via the Distributed Network Protocol DNP3 SCADA[4]. For accessing data on a large scale some federated querying and aggregation solution would be required. This solution also needs to be able to convert the incoming streams into the desirable format, by making use of existing, standard mappings.

3. Analysis and Processing

Energy experts indicated that they typically make use of in-house analysis tools, with R being the de facto standard. Other, commercial, software packages are also in use. Regarding processing and analysis, it often needs to take as soon as the data arrives, in a streaming fashion, as delays may incur costs, for instance when such analysis is used for the purpose of maintaining remote devices.

4. Storage and Curation

Regarding storage and curation, a number of items were raised, such as the need for mapping between standards used. It may be the case that these transformations take place on the fly in order to support streaming analysis, before results and byproducts are optionally transferred onto disk for longer-term storage. For the reasons above energy professionals make use of Apache Hive[5], which allows them to query very large datasets using an SQL-like interface, to query data expressed primarily in the Optimised Raw Columnar (ORC) data format[6]. Columnar data formats such as ORC are useful in large datasets and fit the streaming nature of relevant energy data.

BDE will need to cover the needs above by making use of the HortonWorks solution, which encapsulates technologies overlapping these required by the energy community, such as Hive+ORC. It will further need to provide relevant data-transformations in order to support the chosen pilot use-cases.

2.2.5 Appendices

2.2.5.1 Slides & Presentations

1. BigDataEurope Project Introduction (BDE Coordinator)
2. BDE Energy Societal Challenge (CRES)
3. Data the Gate to a Smart Energy System (EurElectric)
4. BDE technology Overview (NCSR Demokritos)
5. Big Data Applications For Siting And Forecasting (VESTAS)

[1] http://www.dmtf.org/standards/cim
[2] http://idis-association.com/about.html
[3] http://www.dlms.com/information/whatisdlmscosem/index.html
[4] http://www.dnp.org/
[5] https://hive.apache.org
[6] https://cwiki.apache.org/confluence/display/Hive/LanguageManual+ORC

6. Data Management In Resource Forecasting (University of Athens - Forecasting Unit)
7. Data Management in Wind Energy (TERNA)
8. System_Monitoring_Case (CRES)
9. SmartGrid data management challenges (CRES, ETP Smart Grids)
10. ETP_Smart_Grids_Utility_Survey (ETP Smart Grids)

2.2.5.2 Pictures

Link to the Photos slide set on the public BDE Flickr account.

2.2.5.3 Workshop Invitation

Dear Sir/Madam,

*We would like to invite you to a workshop on **Big Data in the H2020 Societal Challenge Secure, Clean and Efficient Energy,** in **Brussels** on the **16th of June 2015**. As a recognized stakeholder in the **Energy** sector, you will have the opportunity to influence the design, and ultimately benefit from, the Big Data platform that the BigDataEurope project will deliver. This platform aims to facilitate Big Data usage in real world examples, and will consist of an architecture, components, guidelines and best practices to make the best of Big Data in this case in the setting of Energy.*

The workshop will address topics including monitoring and control of complex electromechanical systems, transmission systems, smart grids, forecasting and policy making.

This workshop will introduce the background, cover the main challenges, and seek real examples of the potential, challenges and complexities of using Big Data in the Energy domain. The results will be used to design and realise the required ICT infrastructure and support the use and deployment of the platform – maximising the opportunities of the latest European RTD developments, including multilingual data harvesting, data analytics, and data visualisation.

***Event Registration**: More information about workshop registration, agenda, venue, etc. can be found or will be provided here: http://www.big-data-europe.eu/event/sc3-brussels-2015/*

*In addition, as a representative of one of our target communities, we would like to engage with you in the long term, in order to **identify your big data technology needs, challenges and requirements**. We therefore invite you to give due consideration to our participation offer by considering one or more of the following different levels of stakeholder engagement:*

- *Subscribe to the BigDataEurope Newsletter;*
- *Join your respective W3C Community Group;*
- *Participate in the planned key Stakeholder Workshops, including the one announced above;*
- *Participate in one of our BigData Pilots (subscribe for more information);*
- *Follow us on the Social Media (Twitter, LinkedIn,SlideShare).*

*This invitation is **extended to others in your network** with an interest in the challenges of Big Data and data management in your sector. Where possible we would be grateful if you could focus on informatics professionals who have current experience of working with data management. When forwarding the invitation, please include mouzakis@cres.gr in the carbon copy field.*

We look forward to your participation!

Kind regards,

Fragiskos Mouzakis
Head Wind Energy Department
Center for Renewable Energy Sources and Saving
Societal Challenge Representative on behalf of the BigDataEurope Consortium
mouzakis@cres.gr
www.big-data-europe.eu

2.2.5.4 Attendees

No	Name		Organisation
1	Andronopoulos	Spyros	NCSR Dimokritos
2	Auer	Soren	Franhaufer
3	Bautista	Antonio	Cleanergetic
4	ten Berge	Hans	EURELECTRIC
5	Cofino	Antonio S.	Uunican
6	Costa-Soria	Cristobal	ITI
7	Denos	Remy	EC
8	Folkmanis	Janis	EC
9	Getsiou	Maria	EC
10	Guilloud	Geraud	NCP Wallonie
11	Ikonomopoulos	Andreas	NCSR Dimokritos
12	Kallos	George	UoA
13	Karkaletsis	Vagelis	NCSR Dimokritos
14	Kolman	David	Helmholtz
15	Kyritsis	Athanasios	Altsol
16	Larrabide	Ibai	Zabala
17	Mouzakis	Fragiskos	CRES
18	Pollet	Thierry	ETP Smart Grids
19	Post	John	TKI SWITCH2SmartGrids
20	Qvist	Martin	VESTAS

21	Rossi	Kimmo	EC
22	Schulte-Derne	Sebastian	EWE
23	Tselepis	Stathis	CRES
24	Van Hove	Patrick	EC
25	Vlachogianni	Mandy	NCSR Demokritos

2.3 SC5.1 - Climate Action, Environment, Resource Efficiency and Raw Materials

The following table includes a summary of the workshop:

Date	15.05.2015
Venue	NRW Representation, Brussels, Belgium
Invitations Sent	184
Invitations Accepted (Registrants)	6
Attendees (Total)	21
Attendees (Project Consortium & Project Officer)	6
Attendees (Other)	15
Breakout Sessions	3

2.3.1 Agenda

- 12:00 - 13:30 Welcome
  - Round Table Introduction (M. Vlachogianni)
  - Introductory Talks
    - Big Data Europe (S. Auer)
    - Big Data Europe Technology Overview (A. Ikonomopoulos)
    - Breakout Group Sessions (S. Andronopoulos)
- *13:30-14:00 Lunch Break*
- 14:00-15:15 Keynote Presentations
  - *Big Data in ECMWF* - Ioannis MALLAS (ECMWF)
  - *Climate Analysis and Big data* - Rasmus E. BENESTAD (Norwegian Meteorological Institute)
  - *End-user gateway for climate services and data initiatives* - Antonio S. COFIÑO (Group of Meteorology and Computing. University of Cantabria)
- *15:15-15:30 Coffee Break*
- 15:30-16:20 Keynote Presentations
  - *Big, Linked and Open Earth Observation Data: the projects TELEIOS and LEO* - Manolis KOUBARAKIS (Department of Informatics and Telecommunications, University of Athens)

- *Improving climate research societal benefit: JPI Climate strategy on Open Knowledge* - Alexis SANCHO REINOSO (Centre for Global Change and Sustainability, BOKU University)
- 16:20-18:00 Breakout Sessions
  - Data-centric initiatives in Climate
  - Technologies and data
  - EU Policy Requirements
- 18:00-18:30 Summary, Outreach & Farewell

2.3.2 Expectation and Background

Climate research is heavily based upon computer models that simulate the earth's climate for time periods spanning several decades. These three-dimensional global models discretize the entire earth surface and atmosphere to a resolution that has recently gone down to a few tenths of kilometres resulting in billions of grid cells. In this framework, millions of weather observations (including Earth Observation data) are collected and assimilated on a daily basis, while past observational data are re-analysed and climate simulations are performed producing massive amounts of data of the order of terabytes per day. Repeated climate simulations are carried out considering different scenarios of world-wide emissions of anthropogenic and natural pollutants to study their effects in the climate and the climate change societal impacts.

As the atmosphere and its interactions with the land and surface constitute a complex dynamical system, a lot remains to be understood about the earth's physical processes. Given the abundance of climate data from model simulations, Earth orbiting satellites and in situ observations, scientific efforts are heavily placed to narrow the knowledge gaps by directly handling these large data sets. Currently, the progress in climate science induced by Big Data is slow although this field has become one of the most data rich domains in terms of volume, velocity and variety.

Therefore management and manipulation of climate models simulations' results is a Big Data challenge for the organizations engaged in climate research and services and involves techniques and tools for storage, analysis and visualisation in order to extract useful conclusions. It also requires techniques and tools for combination of climate models results with data from other areas, e.g., agricultural production, population distribution, economic activities, etc. Big Data management and analytics of global climate models' results can be used to address real world impacts of climate change.

The aim of the first BigDataEurope SC5 workshop was the identification of current and future challenges for Big Data and data management in the Climate and integration of Earth Observation data domain. The Big Data of SC5 focus mainly on real-time monitoring, stream processing and data analytics. In the workshop, real examples of the challenges and complexities of using Big Data in the domain were presented and discussed.

The workshop focused on the elicitation of the user requirements to support the design and realization of the necessary ICT infrastructure on which the deployment and use of the BigDataEurope platform (aggregator) could be based. The platform targets the facilitation of Big Data usage in real world examples. The BigDataEurope platform will offer the opportunities of the latest RTD developments to the interested participating third parties, including real time streaming, multilingual data harvesting, data analytics and data visualisation.

The workshop was structured with three sections. The first involved a general introduction to the BDE background, objectives and targets, as well as an overview of the tools and technologies envisaged within the project. An extended summary of the BDE technology

overview is provided in the next section of this report (Section 3). The workshop participants were further on introduced to three sets of questions selected to map the user requirements and contribute to the Requirement Elicitation (RE) and the Requirement Specification (RS) activities in work package 2.

In the second workshop section, five keynote presentations were given by invited speakers, illustrating Big Data and data management activities and experiences from the data services, academia and research sectors:

- The first presentation (by I. Mallas) introduced the audience to the current Big Data activities at the European Medium Weather Forecast (ECMWF) involving data acquisition, processing and provision to a great number of weather and climate services worldwide.
- The second presentation (by R. Benestad) discussed the climate and weather analysis carried out currently at the Norwegian Meteorological Institute and presented an in-house built open source Big Data tool for data analysis, statistical downscaling and visualization.
- The third presentation (by A. S. Coffino) provided information on the End User Gateway for climate services from the Meteorology Group of the University of Cantabria and other data Initiatives (e.g. ECOM, COST “VALUE”, WCRP-CORDEX) operating at European and World-wide level.
- The fourth presentation (by M. Koubarakis) presented highlights of the two EC projects TELEIOS and LEO for managing big, linked and open earth observation data, coordinated by the University of Athens.
- The fifth presentation (by A.S. Reinoso, BOKU University, Vienna), discussed the Joint Initiative Programme (JPI) of Climate Strategy on open knowledge and data transparency for improving the climate research societal benefit.

The third section constituted the interactive part of the workshop during which the participants were split into different breakout sessions according to three types of stakeholder groups to serve the user requirements elicitation process. The results of the discussions are summarized in the following sections of this report (Sections 4, 5 and 6).

2.3.3 Summary of Breakout Groups

2.3.3.1 SC5.1 - Group 1: Data Centric Initiatives in Climate

1. Preamble

The basic task of Group 1 was to elaborate on the current state of the art of Big Data in SC5, discussing current and past projects, the gaps that need to be addressed and use cases.

Seven people participated in the group discussion: I. Mallas (ECMWF), A. Cofino (Univ. Cant), R. Benestad (MetNO), G. Kallos (Univ. Athens), M. Vlachogianni (NCSRD), A. Ikonomopoulos (NCSRD) and F. Mouzakis (CRES). Group members’ knowledge and experience of Big Data and data management in the climate sector was fairly high-level, so the discussion covered mainly broad topics and ideas rather than specific examples of work being done in this sector.

The group discussed the general attitudes of people involved in the climate sector towards big data usage and potential. The points raised were the following:

- Access of users to Big Data is an issue of concern, in terms of their ability to access in a form that they can integrate in their processes and ability to store and manipulate.
- There is a need to develop a common language based on what the users need.
- The users must have the ability to build their own case using predefined questions.
- The big data users can be categorized into different levels depending on their use cases.
- Building a small inventory of questions of what people need can help the user to access the data needed.
- There is a general sense of urgency about moving forward.
- From a knowledge worker's view what should be realized is the facilitation of the data collection and integrated management of large (Tb to Pb) datasets from diverse knowledge domains for interdisciplinary research, thus allowing researchers to concentrate on knowledge rather than on data management.
- Various tools and management practices are common:
    - GIS data management and analysis tools, statistical tools
    - Custom-developed data processing tools and procedures developed in R, Python, Matlab
    - General purpose libraries and suites, e.g., NetCDF Tools, NCO, CDO
- Tasks involved in handling large datasets from users were stated to be:
    - Data collection from more than 30 global and regional models and observational data
    - Data processing for impact modelling analysis
    - Remote sensing research, agro-environmental modelling,
    - GIS analysis
- What is lacking are reliable methods and (semi) automatic procedures to discover and integrate heterogeneous data from different domains and also, as a basis for that, the integration of currently scattered semantics / ontologies from diverse domains with sufficient detail to be able to describe sources on the data level.
- There is a great benefit for improving and incorporating the Big Data approach as it is a technology challenge awareness and assessment.
- The Big data management solution can be used by:
    - Internal researchers and external project partners and collaborators
    - Researchers and modellers in data-intensive research areas (e.g. life-sciences)
    - The climate change impact assessment community.

  Such a solution would dynamically collect, re-format and pre-process diverse datasets at a large scale in order to prepare a single integrated dataset appropriately formatted for custom-made analysis tools. If such pre-processing can be reliably and efficiently replicated, this alleviates the need to persist large-scale derivative datasets for the purpose of reproducing experiments.
- In some organizations, there is no procedure or best practice for administrating Data Management procedures.

2. The Data Situation

The group then discussed the data situation.

- The kinds of data of the users mentioned were the following:
  - Climate Data. i.e. CMIP5 /CMIP6, CORDEX, SPECS and many others from different providers.
  - Agro-environmental data, remote sensing data, model input/output
  - Observational data
- Data for users from partners, external service providers, open data.
- The data format is important. The tendency is towards GRIBB. NetCDF is preferable for analysis. Tables of parameters, if local (e.g. from integrated models), constitute a problem. The World Meteorological Organization (WMO) is responsible for such guidelines.
- Popular formats are: Structured e.g. Multidimensional arrays. Netcdf3, Netcdf4, HDF5, GRIB1, GRIB2, and ASCII formats, GIS-databases, shape, JSON. Unstructured like documents.
- Data encoding must draw special attention.
- The format conversion is a problem to the users and not as such the format itself. It was mentioned that ECMWF is working on encoding, as it has been requested by users, on Buffer data.
- Regarding the metadata, the data exchange between different geographical regions and communities must be standardized.
- Currently there are metadata and multilingual problems. In this context, libraries or services should facilitate data access.
- Pre-knowledge must be available for metadata. Routines must be built-up to test if the data makes sense. Visualization tools need metadata not only at the final stage of processing to the users.
- The tools used for presenting the output were mentioned to be:
  - client tools like R, Python or Matlab
  - maps, graphs, integrated viewers and dashboards

The group then discussed possible pilot use cases as a test bed for the BDE platform. A typical example for a climate researcher was proposed for the geographic area of the Iberia peninsula. Large datasets from modelling experiments like CORDEX, CIMP5 are available to set up a use case touching upon issues like infographics (analysis and visualization), metadata and formats.

Finally, there was the comment that the Big Data efforts can primarily provide the means to support researchers in data-intensive science fields, in shifting the majority of their work from the data and information level towards the level of domain knowledge.

2.3.3.2 SC5.1 - Group 2: Technologies and Data

1. Preamble

The basic task of Group 2 was to discuss functional and non-functional requirements for the BDE platform. Group members' knowledge and experience of Big Data and data management

in the climate sector was such so that the discussion covered the following: Data Acquisition, Analysis and Processing, Storage and Curation (see sections below).

2. Participants

1. M. Koubarakis (Univ. Athens)
2. L. Blomme (EVERIS)
3. D. Quintart (EC)
4. I. Larrabide (ZABALA INNOVATION CONS.)
5. C. Costa-Soria (ITI INSTITUTO TECHNOLOGIC INFORMATION)
6. S. Auer (Fraunhofer)
7. V. Karkaletsis (NCSRD)

3. Data Acquisition

During the data acquisition phase, climate domain experts identified data heterogeneity as something that BDE could address. Experts typically work with data maintained and provided in remote repositories and in different formats. The experts then are typically required to manually discover and download datasets of interest, before they process them to match their procedures, locally. Federated querying tools, such as SemaGrow[7], as well as their integration with tools commonly used for pre- and post-processing (e.g. Hadoop[8] MapReduce) could aid experts in this regard.

4. Analysis and Processing

Climate experts indicated that they typically make use of in-house analysis tools in order to carry out their work. As a follow-up activity to that of data acquisition, this requires that data are shipped over to a local computing facility before analysis can take place. A question that was raised was whether it would be possible for analysis to take place next to the data, and therefore either avoid shipping data locally, or the expert having to cater for differences in types of data and remote or local infrastructures. Whereas BDE is not suited to adapting analysis software to domain-specific requirements, it could potentially provide for more streamlined data analysis procedures by including and, wherever possible, integrating analysis software in the tools aggregator. Solutions such as Hadoop are used in order to process massive amounts of data in-situ. BDE can therefore investigate whether and which analysis tools can be integrated with a Hadoop ecosystem – integrating components such as DFS, MapReduce, HBASE[9], Parquet[10], etc – and therefore be made available within the BDE aggregator.

5. Storage and Curation

Regarding storage and curation, a number of items were raised, such as the frequent need for standards and ontology alignment. Already from the data-acquisition phase of the discussion it was made clear that alignment of data takes places manually, after data from different sources have been transferred locally. Coupled with a variety of analysis tools, experts end up having to cope with heterogeneous datasets which cannot be readily processed and analysed further or compared, etc. BDE is therefore called to investigate ways to bridge such differences via storage and curation solutions, in order to facilitate further processing, analysis and

[7] http://semagrow.eu
[8] http://hadoop.apache.org
[9] http://hbase.apache.org
[10] https://parquet.apache.org

archiving more effectively. Semantic maps, as these will result from further discussions within BDE, in conjunction with the use of distributed querying and annotation tools is a field BDE should investigate.

Further to the semantic alignment of datasets, the climate community also discussed the issues of data versioning and data quality. Due to the nature of climate data, being streamed time-series, as well as to the diverse processing and variety of data products, data versioning was brought forward as something desirable by the community. Identifying and dealing with incomplete and temporarily erroneous data sets was also discussed. These seemingly unrelated issues are important to big data scientific processing as well as to other fields and can be addressed by a combination of technologies and approaches, such as the use of persistent identifiers, semantic annotation and metadata as well as by the general field of data provenance. As these are still large and research-active fields and the solutions that exist are typically domain-specific, in BDE we propose to address versioning and data annotation incrementally, depending on priority within as well as the overlap across BDE communities.

2.3.3.3 SC5.1 - Group 3: Big Data Legal and Policy issues

1. Preamble

The basic task of Group 3 was to discuss legal and policy issues related to Big Data. The discussion was oriented around the five questions addressed below (3-8).

2. Participants

1. Alexis-Sancho Reinoso (JPI Climate)
2. Kimmo Rossi (EC)
3. Géraud Guilloud (NCP-Wallonie)
4. Spyros Andronopoulos (NCSRD)

3. What is your domain strategic vision regarding data management?

In H2020 an Open Research Data Pilot is applied as an option to all EU-funded projects (European Commission, Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, v. 1.0, 11 December 2013, pp. 14). In this respect the EU-funded projects are asked to apply an open-access policy to research data (data produced or collected by the project). This will be optional, so that projects may opt out (totally or partially). Participating projects are asked to develop a Data Management Plan, which should specify what data will be open (European Commission, Guidelines on Data Management in Horizon 2020, Version 16 December 2013). More specifically the DMP should describe: (a) what types of data are produced or collected, (b) what standards are used, (c) whether and how it will be exploited or made accessible for verification and re-use, and (d) how it will be curated and preserved. The data access policy will depend on the owners. A template of a DMP is described in the document A.CARTIER, M.MOYSAN, N.REYMONET, Réaliser un plan de gestion de données : guide de rédaction (V1, 09/01/2015).

The value of data was pointed out as a factor that affects the data management plan. Usually “academic” and “industrial” stakeholders follow different approaches. In general academics favour the open access policies whereas industrial people are in favour of restrictions in data availability in order to protect their commercial value. However this is not always true, as in many cases academic people impose restrictions to access to data, in order to gain some

advantage due to earlier publications and competitiveness for funding (a usual data management strategy is the “embargo” on data for a short period of time before they become freely available).

Satellite images have been mentioned as an example of data with great commercial potential. The challenge would be to combine an open-data policy with commercial benefit. This could be seen in analogy to open source software that has given rise to an entire economy.

4. Which political framework or domain specific strategic factors can influence your decisions on your data management strategy/systems in use?

- The EC Digital Single Market
- The existing legal obstacles, IPR, copyright legislation
- The incentives to promote open data policies
- The capacity / potential given by legislation to profit

5. Can data be private and secure yet useable for research?

- Yes, it is possible but difficult. Issues about anonymity must be considered.

6. Which legal concerns could be encountered in the usage of (Big) Data management solutions?

- There are big issues with the copyright and the IPR. Legislation is currently in preparation. It was mentioned, as an example, that use of a model in cloud computing may be violating IPR.

7. What needs to be done to influence policy requirements?

Lobbying, public consultation – in connection to Digital Single Market.

2.3.4 Appendices

2.3.4.1 Invitation Letter

Dear Sir/Madam,

In the framework of the BigDataEurope H2020-ICT-2014 CSA, we would like to invite you to a Workshop on ***Big Data and Data Management in the H2020 Societal Challenge (SC5) Climate action, environment, resource efficiency and raw materials,*** *in* ***Brussels,*** *on* ***June 15, 2015****. As a recognised stakeholder in the sector, you will have the opportunity to influence the design, and ultimately benefit from, the Big Data platform that the BigDataEurope project will deliver.*

The aim of this workshop is the identification of current and future challenges for Big data and data management in the Climate and integration of Earth Observation data domain. In the workshop, real examples of the potential, challenges and complexities of using big data in this domain will be discussed.

The workshop output will support the design and realization of the required ICT infrastructure on which the deployment and use of the BigDataEurope platform will be based. The platform

targets the facilitation of Big Data usage and Data Management in real world examples and it will consist of an architecture, components, guidelines and best practices. BigDataEurope platform will offer to the interested participating third parties the opportunities of the latest European RTD developments, including real time streaming, multilingual data harvesting, data analytics and data visualisation.

***Event Registration**: More information about the Workshop registration, agenda, venue, etc. can be found on: http://www.big-data-europe.eu/event/sc5-brussels-2015/*

*In addition, as a representative of one of our target communities, we would like to engage with you in the long term, in order to **identify your big data technology needs, challenges and requirements**. We therefore invite you to give due consideration to our participation offer by considering one or more of the following different levels of stakeholder engagement:*

- *Subscribe to the BigDataEurope Newsletter;*
- *Join your respective W3C Community Group;*
- *Participate in the planned key Stakeholder Workshops, including the one announced above*
- *Participate in one of our BigData Pilots (please contact mandy@ipta.demokritos.gr for more information)*
- *Follow us on the Social Media (Twitter, LinkedIn, SlideShare).*

*This invitation is **extended to others in your network** with an interest in the challenges of Big Data and data management in your sector. Where possible, we would be grateful if you could focus on informatics professionals, who have current experience on Big Data and Data Management on a daily basis. When forwarding the invitation, please include mandy@ipta.demokritos.gr in the carbon copy field.*

We look forward to your participation!

Kind regards,

Diamando Vlachogiannis,
Senior Researcher, Environmental Research Laboratory, NCSR "Demokritos"
Societal Challenge 5 Representative on behalf of the BigDataEurope Consortium
mandy@ipta.demokritos.gr
www.big-data-europe.eu

2.3.4.2 Advertised Workshop Description and Agenda

The aim of this workshop, the first of a scheduled series in the BigDataEurope project, is the identification of current and future challenges for Big Data and data management in the Climate, integration of Earth Observation data and Smart Cities and Sustainability domain. In the workshop, real examples of the potential, challenges and complexities of using big data in this domain will be discussed.

The workshops output will support the design and realization of the necessary ICT infrastructure on which the deployment and use of the BigDataEurope platform will be based. The platform targets the facilitation of big data usage in real world examples and will consist of the architecture, components, guidelines and best practices.

BigDataEurope platform will offer to the interested participating third parties the opportunities of the latest European RTD developments, including real time streaming, multilingual data harvesting, data analytics and data visualisation.

12:00 -13:30 Welcome and Introduction

- Round Table Introduction
    - Name & Affiliation
    - Role in organisation
    - Connection to big data & data management
    - Expectations for the workshop (what to take home)
- Introductory Talks
    - Big Data Europe
    - Big Data technology for Climate applications
    - Big Data & Data Management in Climate domain

13:30-14:00 Lunch Break

14:00-17:30 Interactive Sessions

- 14:00-15:30 Session 1: Data-centric initiatives in Climate
    - Identification of projects and initiatives addressing the issue of Big Data in relation to Climate Research & Applications
    - Big Data use-cases in Climate, Pilots: discussion on potential use cases and selection of pilot cases for BDE

15:30-15:45 Coffee Break

- 15:45-16:45 Session 2: Technologies and data
- 16:45-17:30 Session 3: EU Policy Requirements
    - Legal issues around (Big) data, Governance, Data Portability
    - Other Requirements

17:30-18:00 Summary, Outreach & Farewell

- Summary, Q&A session
- Closing note, outreach plans

2.3.4.3 Attendees

PARTICIPANT LIST OF THE 1ST BDE SC5 WORKSHOP BRUSSELS, JUNE 15, 2015			
Last Name	**First Name**	**Organisation**	**Country**
Andronopoulos	Spyros	NCSR DEMOKRITOS	Greece
Auer	Sören	Fraunhofer	Germany
Benestad	Rasmus	NORWEGIAN METEOROLOGICAL INSTITUTE	Norway
Blomme	Lode	EVERIS	
Cofino	Antonio S.	UNIV. OF CANTABRIA	Spain

Costa-Soria	Cristobal	ITI INSTITUTO TECHNOLOGIC INFORMATIO	Spain
Guilloud	Geraud	NCP - WALLONIE	Belgium
Ikonomopoulos	Andreas	NCSR DEMOKRITOS	Greece
Kallos	George	UNIV. OF ATHENS	Greece
Karkaletsis	Vangelis	NCSR DEMOKRITOS	Greece
Kolman	David	HELMHOLTZ ASSOCIATION	Germany
Koubarakis	Manolis	UNIV. OF ATHENS	Greece
Larrabide	Ibai	ZABALA INNOVATION CONSULTING	Spain
Mallas	Ioannis	ECMWF	UK
Mouzakis	Fragiskos	CRES	Greece
Papoutsakis	Antonis	TERNA-ENERGY	Greece
Petrowski	Andrea	EUROPEAN COMMISSION DG RTD	EC
Quintart	Daniel	EUREPEAN COMMISSION DG. COPERNICUS	EC
Reinoso	Alexis Sancho	BOKU UNIVERSITY	Austria
Rossi	Kimmo	EUROPEAN COMMISSION CNECT.G3	EC
Vlachogianni	Mandy	NCSR DEMOKRITOS	Greece

2.3.4.4 Follow-Up Message

Follow-up message sent to stakeholders highlighting comms channels, survey and workshop summary:

Dear Friends,

We would like to thank you for your participation in our workshop on Big Data in the H2020 Societal Challenge Climate action, environment, resource efficiency and raw materials calls held on ***June 15, 2015****.*

We are working on the full summary report of the workshop and will be in touch with a link to the report soon. For now we'd like to draw your attention to other ways you can engage with, and help the BigDataEurope project.

- *To keep in touch with the latest BigDataEurope project developments, please visit our website and subscribe to our Newsletter;*
- *the slides presented at the workshop will be available on our SlideShare*
- *We have established W3C Community Groups for each Horizon 2020 societal challenge, please join us there*
- *Participate in the other planned key Stakeholder Workshops,*
- *Participate in one of our BigData Pilots - please contact us directly for details*

- *Follow us on the Social Media (Twitter, LinkedIn, SlideShare)*

If you have 12 minutes to spare, we would love to have your input to our big data survey and help set the requirements for our big data aggregator platform. Alternatively if you have a little more time to spare, please me directly for a more in depth discussion.

Thanks for your participation!

Kind regards,

Diamando Vlachogiannis,
Societal Challenge 5 Representative on behalf of the BigDataEurope Consortium
mandy@ipta.demokritos.gr
www.big-data-europe.eu

2.3.4.5 Pictures

Photos from the workshop (BDE Flickr site)

2.3.4.6 Slides & Presentations

Material for slides (Slideshare)

2.3.4.7 Group Questions

Questions per group based on the WP2 elicitation spreadsheet:

Group 1 Data centric initiatives in Climate

1. Which data management tools and practices are currently common in your domain/your organisation? Are they future proof?
2. What tasks using large (internal and/or external) datasets would you like to perform? What are the problems in getting, processing, analysing and storing the data?
3. What use cases would be of interest to you? Can we setup 2 pilot cases?
4. What are the most important gaps?

   What could be accomplished if those gaps were filled?

Group 2 Technology and Data

1. What are the requirements for the big data platform?

   Both functional and non-functional
2. What kind of data do you need to process?
3. What tools and approaches have been applied?

. Tools/systems/databases/resources -gaps and pitfalls

   Which of them are future proof?
4. What restrictions and limitations might impede access and processing of the data?

Practical or technical restrictions

Group 3 EU Policy Requirements (Legal issues around (Big) data, Governance, Data Portability, other requirements)

- What is your domain strategic vision regarding data management?
- Which political framework or domain specific strategic factors can influence your decisions on your data management strategy/systems in use?
- Can data be private and secure yet useable for research?
- Which legal concerns could be encountered in the usage of (Big) Data management solutions?
- What needs to be done to influence policy requirements?

  Identify key roadblocks, and how to highlight their impact

3. Summary

The workshop reports provided in this deliverable cover the BDE WP2 activities in the first six months of the project. Additional workshops (4) will be carried out between month 7 and month 11 and will be similarly covered in the second deliverable in this series.

4. Appendix

Note: See note in introduction about inclusion of Appendix A in this deliverable.

4.1 D2.1-SC4: Smart, Green and Integrated Transport - Data Community

Originally planned to be contained in D2.1 - Community Building, Coordination and Planning (Section 2.4), a description of the characteristics of the Transport (Societal Challenge 4) community together with the big data challenges and opportunities is instead being provided in this appendix (as outlined in D2.1).

4.1.1 General Description of the Sector

The transport and mobility community is unique: it is entirely about moving things. As with many other sectors, it is seeing rapid growth in the application of information and communication technologies (ICT), and the exploitation of mobile Internet in particular. Still, much of the transport system is managed in the same way since many years, and much

infrastructure has a lifetime of decades or more. In some respects change is slow, whereas the new technologies can hugely accelerate the rate of change.

Two worlds share the same transport systems and infrastructure: the mobility of people and that of goods. Data are essential for running transport systems and services, and information is needed by users for planning and executing trips, both static and dynamic, and both before and during the trip.

Traffic information is now well established in terms of data collection, processing and delivery as a user service. However, it is not always accurate or complete in coverage. Traditionally these data are captured by infrastructure (e.g. road loops, radar, cameras), but of course only where this infrastructure is installed. Nowadays the quality and coverage of road traffic data are highest for vehicle-sourced data, also called “floating” or “probe” car/vehicle data. These are collected from an in-vehicle device via a mobile data connection. The device can be built-in or else a portable (e.g. personal navigation device) or mobile (smartphone) unit. The Waze service collects traffic data from members’ vehicles, together with user-sourced data such as comments explaining delays, and even allowing communication between users.

For traffic data, the volumes needed for a reasonable information service are relatively small. Public transport real-time information services need even fewer data, and these are generally provided by the service operators, who generally track all their vehicles. However, two new trends are likely to disrupt this sector, and will drastically increase the amount of data in the ecosystem:

- The “Internet of Moving Things”, where large numbers of sensors may be deployed in both fixed and mobile “things”, that would communicate their data to information and service centres; the data will go well beyond what is collected today (e.g. location, speed, heading) and could include air quality data, vehicle sensor data (rain, darkness, lighting) and user-sourced data (potholes, incident/accident reports);
- The traveller as data platform, via the widespread use of always-connected smart devices (smartphones but also wearables). Here it is possible to collect new kinds of data, for all kinds of journeys and transport means, including car but also public transport and alternative modes (2-wheelers, bikes, walking). Potential new data include trip destination and route (waypoints). Already new sources of data are being analysed for their information about mobility, e.g. bank card transactions that indicate the location and movements of users.

For goods transport a great volume of data may become available once packets and consignments carry RFID tags or other sensors, to add to vehicle-sourced data. When shared with the road network management, a better optimised routing can be achieved through cooperative vehicle-infrastructure systems.

Lastly, a new concept taking off is “mobility as a service” (MaaS). In this paradigm a service provider offers a personalised package of mobility (e.g. vehicle, information, payment…) comprising a mix of transport means adapted to the particular needs of a customer at a time and place, and paid for by subscription for the package as a whole. The MaaS provider takes care of distributing the revenue to individual service operators.

4.1.2 Sectoral Structure of the Community

The key players in the domain can be governmental and non-governmental organisations, an indicative list of which, is the following:

The ITS transport community consists mainly of the following sectors:

Public Authorities: including ministries of transport or ICT, road operators and agencies, city administrations and traffic managers, municipal public transport operators etc.; are involved in ITS because ITS can help increase road safety and capacity at lower cost than building new infrastructure or conventional measures.

Mobile Network operators: provide connectivity for vehicles and travellers, and offer mobility-related services to their customers (e.g. fleet management, M2M for vehicle and unit tracking, mobile ticketing and payment…).

Vehicle manufacturers: car connectivity is a desirable feature, giving access to mobile internet and enabling vehicle telematics services (e.g. emergency call, traffic information and guidance, security services, pay-by-use insurance…). New applications of Cooperative ITS (C-ITS) allow vehicle manufacturers to enhance comfort and safety (e.g. by in-vehicle warning of roadworks, queuing traffic, fog, accident, emergency vehicle, red light violation…), and to enhance services for their customers.

Traffic and transport industry: supply products for managing traffic, parking, transport systems and demand to public authorities, transport operators and commercial transport providers.

Suppliers: This sector includes automotive suppliers as well as others that supply products that may be used for transport and mobility purposes, such as ICT systems and devices. By ITS we understand the application of information and communication technologies for transport, so in the car or commercial vehicle this means safety systems using sensors, radar or communication; information systems for warnings, advice or guidance; systems for automated driving; systems for payment or access control etc.

Service providers: work with vehicle manufacturers, transport operators or public authorities to provide mobility services. Examples are providers of digital maps, traffic and travel information, telematics applications, payment and ticketing, parking space finding & booking, vehicle breakdown assistance, tourism information etc.

Users: user associations (e.g. automobile clubs, transport operators' associations, associations for road safety, cyclists, motorcyclists, vulnerable road users, public transport users etc.) that may lobby for fair treatment for their members, promote their cause in legislative assemblies and advocate data protection and privacy issues for end users.

Research: there is a large community of research and innovation bodies, including university ITS departments or institutes; independent industrial research centres; and research departments of ITS-related industry. They often focus on developing and proving future technology, and enhancement of existing ITS services and solutions. Increasingly research is carried out on social or business-related issues associated with ITS deployment.

4.1.3 Size of the Community

The size of the ITS community is difficult to assess as ITS is not really a free-standing domain. Certainly in Europe there are many tens of thousands of people directly and indirectly buying, supplying, operating and using ITS products and services. And ultimately, virtually every person is an ITS user, since traffic and travel information, in-vehicle safety & infotainment systems and personal navigation devices (not to mention smart mobility apps for smartphones) can all be considered ITS.

4.1.4 Formal Networks

- ERTICO – ITS Europe (multi-sector partnership for ITS deployment)
- Network of ITS Nationals (coordinated by ERTICO)
- Individual national ITS associations
- ERTRAC (research)
- CLEPA (vehicle suppliers)
- POLIS (ITS cities)
- ACEA (automotive vehicle manufacturers)
- EUCAR (automotive research, under ACEA)
- FIA (world federation of motoring associations)
- IRU (world union of national freight and commercial vehicle associations)
- UITP (world union for public transport)
- EARPA (European association for automotive research)
- CEDR (European road directors)
- ITF (Europen ministers of transport)
- Car-to-Car Communication association (vehicle manufacturers and suppliers)
- GSMA (GSM operators' association)
- TISA (traveller information services association)
- TM2.0 (platform for cooperative traffic management)
- ADASIS (platform for advanced driver assistance map data standards)

4.1.5 Informal or Upcoming Networks

- Tri-lateral working group USDoT – Japan - EU
- iMobility Forum
- C-ITS Platform