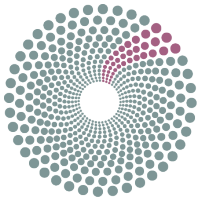
WORKSHOP REPORT



BIG DATA EUROPE

First Community Workshop for Societal Challenge 1 - Health, Demographic Change and Wellbeing

Location - Kowi Brussels, 21st May 2015

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, William 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 May 29th 2015, slides are available in the BDE area in slideshare and also linked to this report.

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 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.

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:

Summary of Breakout Group - 1

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. 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?

Summary of Breakout Group - 2

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 flipchart.

General observations
Group 2 discussed mostly different aspects of data and metadata regarding a potential big data platform. 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.

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

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.

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:
    a. There are a number of multilingual datasets including literature, patents, electronic health records. To have access to these, multilinguality would be needed.
    b. 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)

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*

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)

Summary of Breakout Group - 3

Group 3 – Report of the Discussion Regarding Big Data 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.

Preamble and Background

Group 3 was tasked to discuss some of the legal and policy issues relating to big data. 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.

- 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?
- Veractity
    - 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

Appendices

Background Material and Preparation:

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. http://w3.org/community/bde-health/ group now setup to serve highlight visibility

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.

Draft Workshop Agenda

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 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
- what 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 hours

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 area 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

Links to other Material

Photos from the workshop (on BDE Flickr site)
Material for agenda, slides etc (available in BDE Slideshare and from BDE website)

Other Workshop Opportunities in this SC

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.

Report by Bryn Williams-Jones (OPF), Victor de Boer (VU), Kiera McNeice (OPF)