Big Data Europe: Societal Challenge 1

Health, Demographic Change and Wellbeing

3rd Workshop Report

Objectives

The third workshop for SC1 (Health, Demographic Change and Wellbeing) was also the final workshop in the Big Data Europe project. As such our aims for this workshop were twofold.

Firstly, as the final workshop for SC1, our aim was to showcase the final version of our SC1 pilot and its applications to early bioscience research data. Beyond this, we aimed to engage with ongoing and future projects to discuss the future of big data in health. Health is a complex domain with a large number of academic, private and global entities operating in this space; there are many challenges still to be addressed in this domain, and we aimed to host a lively discussion about where big data in health may lead us in the years to come, and what roadblocks it might run into.

Secondly, as the final workshop for the BDE project as a whole, our aim was to showcase the BDI itself, as the primary technical result (support) of this 3-year Coordination and Support Action. We invited a core technical project member to present a live demonstration of the functionality of the BDI, and to engage attendees in discussion about how they might deploy it for their own needs.

A summary of the day's events was posted to the Big Data Europe blog.

Participants

The workshop was attended by 16 participants from a variety of backgrounds.

	Name	**Surname**	**Organisation**
1	Kiera	McNeice	BDE (Open PHACTS Foundation)
2	Simon	Scerri	BDE (Fraunhofer IAIS)
3	Jonathan	Langens	BDE (Tenforce)
4	Michaela	Black	Speaker (University of Ulster)
5	Supriyo	Chatterjea	Speaker (Philips Research Europe)
6	Guillermo	Palma	Speaker (L3S Research Center)
7	Aldo	Camargo	Technopark Peru
8	Vasily	Epishkin	Permanent Mission of the Russian Federation to NATO
9	Ilias	Iakovidis	EC/DG CONNECT

10	Violeta	Isabel Perez Nueno	EC/DG CONNECT
11	Hans-Joerg	Lutzeyer	EC/DG RTD.F3
12	Jana	Makedonska	EC/DG RTD
13	Cédric	Peeters	Vrije Universiteit Brussel
14	Saila	Rinne	EC/DG CONNECT
15	Paola	Saura	Zabala Innovation Consulting
16	Gregor	Schaffrath	EC

Report

11:00-11:10 Welcome and Coffee

Kiera McNeice, Open PHACTS Foundation

11:10 – 11:25 Introduction: Big Data Europe

Simon Scerri (Slides)

The workshop began with an overview of the Big Data Europe project. Simon Scerri presented an overview of the big data landscape when the project started, and the needs of different stakeholders with regards to volume, velocity, and variety of data, as well as infrastructures and data value chain requirements. Several common requirements were identified in the health domain, which implied that a generic data solution might be valuable in this domain.

Simon then presented the major results of the BDE project, from both a community and infrastructure perspective. He gave an overview of the Big Data Integrator and its seven implemented instances across the seven societal challenges, as well as a view to what could be done with the BDI in follow-up projects and other initiatives – for example in standardisation efforts within BDVA.

11:25 – 12:00 Live Demo: The Big Data Integrator

Jonathan Langens, Tenforce

A key aim of this workshop was to present a working live demo of the Big Data Integrator to our attendees, to demonstrate how BDE has worked to lower barriers to entry for people interested in working with big data. Jonathan first explained how the various components of the BDI are designed to help users build, setup, deploy, and monitor big data pipelines. He then gave a live demonstration of how to build a big data pipeline using the Stack Builder, which allows users to drag and drop components into a stack to build a Docker Compose file; the Workflow Builder, which allows users to create steps and conditions to determine which Docker Compose files are initiated in which order; and the Swarm UI, which facilitates monitoring of the pipeline, including options such as live monitoring with a Kibana dashboard.

12:00 – 12:15 The SC1 Pilot: Open PHACTS

Kiera McNeice, Open PHACTS Foundation (Slides)

The final version of the SC1 Pilot was presented by Kiera McNeice, who gave an overview of the background of the Open PHACTS Discovery Platform and its usefulness for early stage drug discovery. By semantically linking data across multiple open pharmacological databases, focussing on questions asked by real researchers, the platform can help significantly reduce the time and cost of researchers' queries.

As the SC1 pilot, the Open PHACTS Discovery platform has been successfully re-built for BDE, using all open components and allowing users to install the entire platform on a local machine. This makes the Open PHACTS functionality even more accessible to researchers from academia, SMEs and industry, as well as allowing for integration with wider platforms using BDE architecture, meaning the platform has increased flexibility, scalability and extensibility.

Attendees were particularly interested in whether the platform had been extended or connected to other data across different Societal Challenges. Although this was not possible to achieve during the lifetime of the project, it could well be a possibility in the future. We briefly discussed the effort required to create the semantic mappings within Open PHACTS, particularly the difficulty of refreshing datasets on a regular basis. But similar principles could in theory be applied to mappings between other heterogeneous datasets, for example patient data.

12:15 – 12:45 Lunch and Networking

12:45 – 13:15 Invited Keynote: The MIDAS Project

Michaela Black, University of Ulster (Slides)

Project Overview

Dr Michaela Black was invited to present the MIDAS project, which aims to use big data to support public health policy. The project's aim is to connect as many different kinds as possible, including data volunteered by individuals. The project is then creating an integration layer and visualisation tools to help policymakers and citizens understand connections between data, with the ultimate goal of informing policy decisions and enhancing health outcomes for individuals in policy regions.

A key need for policymakers is to be able to connect as many kinds of data relevant to healthcare as possible – not only from health agencies, insurance companies, and care providers, but also data about local infrastructure and schools, and even information from social media data in order to understand public perceptions of new policies. These data can potentially be used to evaluate new and existing policies, providing evidence to review them. However there is a lack of tools to "close the loop" and help policymakers and citizens understand and make decisions based on these data. Furthermore, policymakers do not tend to have a data science or technical background, and rarely have access to data scientists. MIDAS has found that there is a need not only for linking data and creating visualisation tools, but also for education among end users about how data can be visualised, and the risks and limitations of how it can be interpreted.

Discussion

The MIDAS project is also working with personal data volunteered by citizens. In our last SC1 workshop it was clear that working legally and ethically with personal data will be a major challenge to realising the benefits of big data in healthcare, and as with our last workshop, the challenges of working with personal data became a key point of discussion among workshop attendees.

One such challenge is anonymisation. The first question that arose was whether standards of anonymisation exist, and the inevitable trade-off between data integrity, granularity and the risk of reverse identification. Dr Black gave the example of Finland, where many individuals have been happy to volunteer their personalised data for enhanced healthcare; Finland have existing standards under which a lot of healthcare data has been anonymised and released to researchers. However linking datasets is a risk to anonymity, and there is some question of how future-proof existing standards are. And even anonymised datasets can face ethical challenges; Dr Black gave the example of a case where anonymised data was used to build a preventative process to prevent children with Down's syndrome being born. Although legal under GDPR rules, parents of children whose data was used in this process were very unhappy with the way it was used.

Attendees discussed alternative options to anonymisation, such as creating synthetic data to use as a representative replica for real data. The MIDAS project is looking into proof-testing this ideas, as ultimately policymakers are not interested in individuals' data, just in the underlying metadata that is relevant to policy-level decisions.

Another potential solution raised was for honest brokers to act as intermediaries. Dr Black gave examples of Northern Ireland and the Basque region where the honest broker solution is being considered: researchers would identify the data they want, and subject to approval, honest brokers could gather and connect data from different sources, remove identifiable information, anonymise the data, and allow it to be analysed in a restricted in-house environment.

Attendees were curious whether MIDAS had made any interesting discoveries through linking data so far. Dr Black said that a few interesting trends had been observed – for example that Northern Ireland appears to be on a trajectory similar to the USA's in drug usage and obesity – but that at the moment data is bringing in more questions than answers. An important part of MIDAS's work will focus on helping policymakers understand what kinds of questions they can ask of datasets, and how to effectively use visualisation tools to ask the right questions and obtain useful and reliable answers.

13:15 – 13:45 Invited Keynote: BigMedilytics

Supriyo Chatterjea, Philips Research Europe (Slides)

Project Overview

Dr Supriyo Chatterjea was invited to present the BigMedilytics project, which aims to take a holistic view of healthcare and improving outcomes in the healthcare industry. The project is taking a very broad perspective of healthcare, with 35 partners from a wide range of backgrounds, and ultimately aims to increase productivity in the healthcare sector by 20% by applying and adapting state-of-the-art big data techniques and algorithms.

The project has identified the main disease groups that will be responsible for the greatest burdens on society in the near future, currently responsible for 78% of deaths in Europe. These have been split into two themes: oncology, and population health and chronic disease management. A third project theme will focus on the industrialisation of healthcare during treatment phases in hospital. The project is addressing these three themes by working with key experienced partners on a series of 12 pilots (e.g. Comorbidities, Prostate Cancer), each of which will connect datasets through a flexible architecture to address that pilot's specific challenge.

Discussion

Again the first question raised by attendees was around the ethics of personal data – specifically, the risk that one or more partners may not be able to resolve the ethics issues around data needed for their pilot. Dr Chatterjea explained that in fact, this is not the first time BigMedilytics has been submitted as a project – and one of the reasons it failed previously was that they could not guarantee all the data they wanted would be accessible and usable. To address this issue BigMedilytics refocussed on partners who have already worked with similar datasets, and planned out ethics procedures in advance; the project has actually been delayed to carry out ethical considerations and ensure as much as possible that data will be available from day one of the project.

Workshop attendees discussed the reasons why people may not want their data used by the healthcare industry, and whether it makes sense for consent around personal data to be binary (a blanket 'yes' or 'no'). One attendee suggested that some people may be willing for example to supply data for use by not-for-profit projects and research, but not to help an insurance company refine its tariff schemes.

Although the GDPR requires very specific constraints when asking for consent to use personal data, this raises challenges of its own: for example, one attendee raised the possibility of discovering a new use for data halfway through a project, and not having specific consent to investigate that use. This could limit the ability of projects like BigMedilytics to discover and explore new avenues for improving healthcare outcomes; formulating questions of consent and what happens to derivative data will be an important consideration.

In a more specific use case, one challenge to improving workflows when dealing with stroke patients is that many stroke patients arrive in hospital in no condition to give their consent to having their progress through the healthcare system tracked. Attendees discussed whether in such situations data might be collected, and then only used if and when consent is obtained from the patient, and discarded otherwise.

Finally, Dr Black suggested that BigMedilytics also approach charities and voluntary organisations who 'pick up the pieces' when patients return home from hospital treatment, as they often play a key role in consistency of care, and deliver real face-to-face value for patients.

13:45 – 14:00 Invited Keynote: IASIS

Guillermo Palma, L3S Research Center (Slides)

Project Overview

Guillermo Palma was invited to present iASiS, which aims to integrate data into a big data framework to support personalised medicine. This framework will be based on the BDE framework, with a semantically enriched layer used to populate a knowledge graph which users can then query.

The project is working on two pilots, which will focus on lung cancer and Alzheimer's disease. In each case the goal is to combine a variety of data sources including electronic health records, genomic data, bibliographic data, as well as publicly available pharmacological databases such as GeneOntology and DrugBank. Data will be further enriched by using natural language processing and data mining techniques on clinical notes, and deep learning to analyse medical images, and generate predictive models.

Discussion

Once again, workshop attendees were interested in how iASiS plans to handle the ethics of dealing with medical data at a personal level – in particular whether focussing on two targeted pilots helped to simplify issues. Mr Palma explained that in each pilot, the project is working closely with partners who already own enough data to work with, for example with a hospital specialised in lung cancer. Data also remains under the control of the respective partners for privacy control, with no copies made on iASiS's platform.

Further questions from attendees were around whether iASiS is working with existing ontologies or generating new ones. Mr Palma explained that the project will do both – relying partly on existing ontologies like MeSH, as well as creating their own ontology to build their knowledge graphs.

14:00 – 14:50 The Future of Big Data in Health

Open discussion moderated by BDE

During the course of the afternoon, the presentations given by each invited speaker evolved naturally into discussions of the challenges and opportunities facing big data in health. We allowed these discussions to run over the allotted time for each project presentation as interesting thoughts and ideas were being raised. Although we were left with less time than planned for the open discussion at the end of the afternoon, open discussions had evolved naturally throughout the day and a lot of key points had already been covered.

Simon opened the final discussion by noting that personal data was a topic that had come up throughout the day. Although BDE didn't deal with any personal data in the SC1 pilot, this is clearly a problem affecting a lot of potential big data applications in health, and could become more complicated once the GDPR comes into force.

More generally, we asked attendees for their views on whether stakeholders in the health domain consider using the BDI, or building on BDE as part of their project infrastructure, as iASiS have done. The general view among attendees is that stakeholders must be convinced there is a benefit to switching to a new infrastructure, or adopting a specific infrastructure, and that this requires building up trust. Key factors stakeholders are likely to consider would be whether an infrastructure

can help them solve broader issues beyond just being involved in a specific project, and what costs and risks might be associated with that infrastructure in the future – for example whether there are likely to be expensive licensing fees, and whether the infrastructure will be supported in the case of any bugs or issues.

On the broader question of the way forward for big data in health, there was general agreement that healthcare depends on many factors beyond just health data. Cross-project collaborations involving food, lifestyle, and infrastructure to support healthy living could make health data much richer, especially over the long term. One example given was looking at data about the usage of rental bicycles in major cities, if the owners are willing to make that data public.

This led to discussion of how data owners can be encouraged to share data, in which most agreed there is a need to identify win-win situations. Unless all parties feel like they are benefiting, they may not share their data, in which case everyone loses out in the end. Points were again raised about the complexities of complying with rules around consent to use personal data – especially in the context of consortiums, where citizens may not realise that giving one company access to their data may make it available to others in the consortium. Dr Black gave an example of a Finnish consortium which is looking at an app where users can 'consent-in' or 'consent-out' at any stage of a project, to help reduce uncertainty in cases like these. Ultimately giving power to individuals, and building trust with individuals, seem to be important steps in treating personal health data ethically.

Finally, attendees discussed the fact that more and more projects and solutions are trying to create bench-to-bedside pipelines in the healthcare domain, in a variety of different use cases. If BDE's infrastructure can be adapted for such closed data cases, it could definitely have applications in these areas.

14:50-15:00 Wrap-up