The challenges of big data for societies in a changing world

1st SC6 Workshop Report

In the framework of the BigDataEurope (BDE) project, CESSDA (Consortium of European Social Science Data Archives) and SWC (Semantic Web Company) organised a workshop on "The challenges of big data for societies in a changing world" on 18 November 2015, which took place at the Eurostat BECH building in Luxembourg.

Background

Role of CESSDA as networking partner in the BDE consortium is to coordinate the Societal Challenge 6 Interest Group: "Europe in a changing world - Inclusive, innovative and reflective societies", and potential users of big data in the fields of social sciences and humanities (SSH). Apart from building the interest group and collecting its requirements, CESSDA should assists the building of the ICT big data infrastructure access point for social sciences and humanities, explore and evaluate the input data, and discover the implications for the future of big data in social sciences and humanities. It is done in collaboration with the SC6 technical partner, SWC from Austria.

Being a consortium itself with a large network of members, observers and engaged non-members, CESSDA aims to ensure that all relevant stakeholders in this domain, in Europe and beyond, are reached out to and engaged with the BDE project using internal communication channels, as well as the W3C Community infrastructure. As a recognised stakeholders in this field, participants in the project will have the opportunity to influence the design, and ultimately benefit from the platform that the project aims to deliver.

The first of three annual workshops envisaged within the Work Package 2 "Community Building & Requirements" took place on 18 November 2015 and introduced the background, covered the main challenges, and sought real examples of the potential, challenges and complexities of using big data in our societies. Workshop should provide additional information for the social sciences and humanities' requirements elicitation process, their definition of and prioritization regarding the specific nature of the SC6.

General information

42 participants attended the workshop: total of 63 had registered with 28 on a waiting list. Participants were relatively evenly split up between domain experts (academics), strategic decision makers (EU/EC bodies representatives) and technical experts , with few participants from the industry sector. The workshop hosted speakers with a policy and data management/providing background (European Commission, Eurostat) and a technical background from within the BDE project (Fraunhofer IAIS, SWC). The workshop consisted of an opening joint session, followed by two parallel sessions and a final joint wrapping up session. Agenda available on this link.

Opening Session

Michail Skaliotis, Head of Unit of the Eurostat Big Data Task Force, started off by presenting the notion of "smart statistics", which would be owned by the Statistical Office of the future. The world of official statistics is submitted to a swiftly changing environment. The most recent transition is complying with the variety of big data sources and requirements leading inevitably to the 'smart statistics' – representing on the one hand large-scale data analytics, predictive modeling, and visualization required for big data, and meaningful signals

and patterns that have been extracted by intelligent algorithms turning thousands of numbers into smart data on the other. Collecting large quantities of numbers has no added value unless there is a meaningful overview. One visionary approach explains the National Statistical Institutes (NSIs) of the future with official statistics among the main users of such data which is produced automatically[1]. Finally, Michail Skaliotis once again stressed the importance of moving forward from 'official statistics' to 'smart statistics'.

Kimmo Rossi, BDE project officer at DG Connect, informed the participants that DG CONNECT is involved in 100 research and innovation projects on big data, open data, and language technologies, covering all possible application areas with the "common denominator" being big data. Rossi drew attention to the fact that, together with the "Internet of things" and the cloud, big data is a "hot topic" and DG CONNECT's role is to liaise between actors in the area and coordinate activities. He pointed out that transfer technologies and methodologies were available for domains that require the use of big data.

Sören Auer, Fraunhofer IAIS, coordinator of BDE, mentioned both negative and positive aspects of big data, stating that BDE aims to maximise the societal value of it. He presented several examples of companies which are changing their business models in the aerospace and automotive fields, pointing to proactive maintenance at Rolls Royce in airplane turbines with 2000 sensors inbuilt into each, then to the rolling smartphone and data value chains (e.g. car windscreen wipers used as rain sensors for micro-weather prognosis & then used to inform farmers), and finally to predictive analytics whereby technology can help humans to make smarter decisions (e.g. driving). The underlying purpose of the BDE project is to enhance the usage of big data in all areas under Horizon 2020: it aims to increase the potential and advance the tools for using big data technologies by developing common ICT infrastructure for all seven societal challenges of Horizon 2020, to increase data science skills, and to develop needed technical components and architecture that will foster various data types and sources. Auer explained the measures & results of BigDataEurope as a Horizon 2020 Coordination & Support Action (CSA): coordination regarding big data and efficient data management in and across the H2020 societal challenges by, beside others, the creation of societal interest groups as well as support realisation of a big data aggregator platform to be used by different stakeholder groups.

Martin Kaltenböck, Semantic Web Company, lead of requirements engineering in the BDE project & technical lead of SC6, explained the motivation and the need to make use of big data principles and technologies because of the enormous amount of data in place that data comes from: sensors, social media posts, digital pictures and videos, transactions, mobile phones, etc. He explained the four Vs of big data to the audience: volume, velocity, variety, veracity, not least pointing out the importance of the 5th V: value of the data, saying that we need to "see what to do with the data". He explained that the BDE project follows the data value chain in requirements engineering as well as the 7 Societal Challenges and look for and evaluate patterns regarding the requirements. Martin Kaltenböck informed participants of the happenings in the project in its first year of operation regarding requirements

[1] Svein Nordbotten (2010), The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, Journal of official statistics, Official Statistics in Honour of Daniel Thorburn, pp. 205–223

engineering: that a core question matrix had been built as a basic tool for the whole requirements engineering process, and that on top of this matrix several methods of requirements elicitation has been managed, as A) an online survey and B) face to face interviews (around 100 interviews), as well as C) seven workshops and finally D) seven BDE pilot use case descriptions. All this information helped to find the requirements for building the BDE aggregator platform in the next stage of the project. Kaltenböck furthermore presented details of the online survey. He pointed out that there were about 400 participants in the survey, but only 131 completed the survey, and that he was therefore well aware that it was not representative. He added that survey respondents were mainly from academia and public administration filled in the survey as well as some large companies.

As a result he presented that variety is the most important topic for the social sciences, mostly economic and social data and that social sciences were a bit behind in terms of investments in the area of big data. Currently, big companies and also some SMEs are starting to invest. Data provided in the social sciences is often mainly open data in comparison to other societal challenges. Kaltenböck added that following the survey results Societal Challenge 6 has the infrastructure in place for long term preservation of data and that there is a need to process larger data and especially in big companies.

Bert Van Nuffelen from TenForce (BDE team) then presented the architecture of the BDE aggregator platform and furthermore informed the audience that the intention is to build a platform together but to keep in mind that we need flexibility and the ability to grow elastically in the first year of the data aggregator platform.

Fernando Reis, member of Eurostat big data task-force, presented his views taken from the task force which also involves NSI personnel. He explained that they are trying to work with big data themselves in the task force. Reis presented the three main elements for statisticians of big data: data deluge, analytics and a data-driven economy (see the presentation). Reis tried to answer to the question: what does big data mean for official statistics? He underlined that it was a change of paradigm. "We are going from finite population sampling to additional statistical modelling and machine learning. We are moving from designers of data collection processes to designers of statistical products." He explained that in official statistics, people are obliged to submit their data (no need for consent from individuals) but this means that the data subject knows his data will be used and that the data is kept secure (e.g. survey). However, using mobile phone data for instance, raises some issues on whether people actually agree with this. He raised the issue of whether or not to produce statistics on the Roma people, point out that there are now have multi source statistics where several data sources are combined (mobile phone data and a survey). He drew attention to the ESS big data action plan and roadmap, developed to prepare the European Statistical System for integration of big data sources into the production of official statistics across the ESS.

Reis presented some of the challenges for data management, for example, the lack of control of data sources, pointing out that when the data source is owned by a private corporation, it is unclear how official statistics can ensure that the figure was not changed, which is of

crucial importance considering that it will have an impact on policy making. Another concern is volatility: will these data sources be there in the future? The example was given of MySpace and now Facebook. Next was the issue of data integration due to the variety of the sources of big data, as well as open data, and the level of detail of the data. Reis insisted that: "The big question in data management is do you store everything? Two terabytes of data a day produced by the Lufthansa fleet."

Furthermore, a technological challenge also exists as the big data tools themselves change every year, meaning that the technical architecture needs to be adaptable over time. Lastly, privacy was also deemed important, as some anonymization tools today are no longer up to date and metadata standards are not yet up-to-standard either. In research, you need to make sure data is repeatable, which is a challenge with very large data sets.

Parallel Sessions

The opening session was followed by four parallel breakout sessions focusing on A) data, B) risks and challenges of successful data management, as well as C) technological as well as D) legal and policy demands of big data. Each parallel session saw the participation of between 15 and 20 participants.

Each parallel session was moderated by BDE staff and discussions took place in groups of approximately five to six people in the framework of one or more exercises given by the moderators. The outcomes of the discussions were summarised by one spokesperson for each group and summarised by the moderator and appear on the poster attached.

Parallel Session 1: Data in place in the Social Sciences and Humanities

In this session, participants were asked to list the most important 'big' data sources they work with, in an attempt to understand the characteristics and complexity of data sources in the SC6 community. Of all the societal challenges SC6 is perhaps the most open-ended in terms of a core domain. This was reflected by the variety of the participants' background, whose interest lied in making sense of data (in particular figures and statistics) coming from various fields and sectors, e.g. Health, Transport and Governance, to provide information and services to society as a whole. For this purpose, participants were also asked to describe their main use-case for collecting and analysing data from the identified sources. The former ranged from cultural heritage, job market monitoring and policy making to consumer behaviour, trend detection and crime prevention. Sources included national data sources and government data, as well as social media and data coming from other societal domains such as health, transport and mobility.

Following the outlining of the data sources and their principal use-case(s), participants discussed which in their view is/are the most challenging dimension(s) of big data, based on the four V's: Volume, Variety, Velocity, Veracity. In contrast to other SCs, here again there was no clear winner in terms of the most pressing dimension, although velocity emerges as

the least perceived challenge. The former observation could again be attributed to the fact that SC6 has no one primary sector but is instead longitudinal against all sectors. Following the technology session, the last observation was put into question, given that the community representatives (who in their majority came from a statistical background) were not very familiar or involved in technologies dealing with high-velocity data streams, such as social media, in their use-cases.

Parallel Session 2: Risks and challenges of successful data management in the Social Sciences and Humanities

This session was divided in two parts: the first one was devoted to risks and challenges of successful data management in the Social Sciences and Humanities, and the second one in determining the potentials and opportunities from data management. Participants were split into three groups of five to six people and asked to brainstorm and then present the results to the rest of the group.

Risks and Challenges
Among the various risks identified during the session, access to data was emphasized (as well as access to the same data for several usage and that also private companies could provide more data), which can vary depending on data provider, legal restrictions of the respective country, data type, level of access granted, etc. The always-present issue of open vs. privately owned data was also discussed. Another linked issue brought up during the session was the ethical concern when accessing data, which also depends on the level of data anonymisation already carried out on the data. Also the issue of costs of anonymisation was raised. The issue of whether an analysis was replicable was brought up as anonymisation of data in association with stable data sources (along with metadata) make it possible to repeat an analysis and confirm or disprove previous findings, or even add value by adding additional variables in the set. However, if only one of those pre-requisites were to be missing, this would render replicability impossible.

Work in interdisciplinary and multilingual teams has exploded over the past 5-10 years leading to more profound and overarching results, but also confronting data management with several issues: providing multilingual data sets is time and effort consuming, leading to the lack of adequate skills to address it, and at the same time trying to cope with the speed of technological advancement, as well as encouraging people to work together from different areas of expertise (research and academic, data & IT experts, the private sector, etc.). Even with all other requirements fulfilled, data sets are not always completely usable and some degree of "cleaning" is necessary before conducting any analysis. Participants also questioned whether Europe faced a skill gap.

Further remarks concerned data silos (isolated repositories of fixed data, not openly accessible) as barriers to effective operations and a barrier to collaboration, accessibility and efficiency as well as the fact that technologies are changing so rapidly that it is hard to follow the latest trends.

Potentials
Most of the data in social sciences and humanities is publically available at virtually no cost. The huge amounts of data potentially available and the speed at which it can be accessed are assets. The management of such data requires considerable effort as well as new skills, which in turn leads to new job opportunities, generating new ideas and new applications for consumers. Such a propulsive environment enables monitoring and maintaining of data to be done in a completely different way. Hypothesis analysis is easier to conduct in data accessible surroundings, enabling increased reproducibility of outcomes, and finally leading to better forecasting and predictive analytics. With such mechanism in place, data produced can better inform policy & decision making, thus creating easier access to different markets (via e.g. adaptive pricing).

Parallel Session 3: Technological demands of data in the SSH

Participants in this session were asked to list big data technologies that they are either already using or familiar with, and those that they are aware of and would like to learn more about, or plan to include in the architectures in the near future. The results of this exercise were discussed in an attempt to classify them according to existing big data technology groups or clusters, while keeping them generic enough for the audience.

A distinction between the listed technological components immediately became apparent: a significant amount of participants listed conventional tools that are not necessarily able to handle big data in its true sense. In fact, only the Storage components indicated are designed for or support big data (e.g. Hadoop, Virtuoso). The analysis and processing tools and solutions mentioned to an extent precede the advent of big data, and reflect the amount of technological components available (especially statistical tools like SPSS, Matlab) in this domain that have been around for decades. The question arose as to whether these tools can successfully be integrated in big data architectures. The participants struggled with naming technology that can support the import, export (and representation) of multi-source and multi-format data, as well as its fusion, re-use and sharing (and in particular, provenance).

Following the ensuing discussions, a number of big data bottlenecks and critical issues have been identified. Apart from an identified lack of awareness of velocity-related components by the community, challenges included difficulties of dealing with the increasing volumes of textual data for analysis and geo-locational tagging.

Parallel Session 4: Legal and policy demands of data in the Social Sciences and Humanities

In this parallel session, participants were split into three groups of five to six people and asked to brainstorm and then present the results to the rest of the group.

The reform of the General Data Protection Regulation is currently in “trilogue” phase between the three European institutions based on the document from June 2015, with a

final agreement on the data protection package expected in December 2015. Practical implications are still missing as proposals regarding data transfer and storage are by no means clear and ready to be used. Moreover, the legal framework differs in different countries and largely depends on the type of data collected. For certain types of data (e.g. medical data) it takes too long to get necessary permissions for collecting data. Another issue, especially with official statistics, is the data sources stability and trust (are the resources reliable, will repeated measurements be valid, variations in tested samples, etc.). When it comes to the data ownership, the issues of open data versus enterprise owned data emerges: some countries (e.g. Belgium) started the process of voluntary sharing of data owned by private companies (e.g. mention of the Harmonized Index of Consumer Prices - HICP regulation[2]).

Another acute issue in scientific and official databases is data provenance; it is crucial for the validation of data. The possibility of copying data and transforming it has made it increasingly difficult to trace the origins of a data set. With this mind, participants underlined that the provenance of the information was a clear requirement for big data management. The stability of data sources and data ownership were also raised as important legal issues.

Data profiling is becoming a more and more pressing issue for data protection law to address, especially as it often takes place without the individual's knowledge or consent. Although it is proposed in 2013 that the GDPR should include a definition of 'profiling'[3] together with additional provisions to protect data subjects, the outcome is still uncertain.

Data anonymisation was also addressed. It was suggested that it should be taken into account in the research planning stage. The management of both direct and indirect identifiers before undertaking data collection produces better informed consent and requires a less resource intensive process when doing data anonymisation (e.g. administrative data, although without great research value, represents personal and sensitive information and need for its collection and later anonymisation has to be determined prior to data collection).

Final remarks were regarding the need for the constant update of the regulations since the data research community environment is very dynamic, as well as the need for constant upgrade of tools and skills in anonymised data management.

Q&A, Wrapping up

[2] "HICP Framework Regulation – Council Regulation 2494/1995" of 1995 required that HICPs be produced and published, use a common reference base, provide common coverage of consumer goods and services, and share a common classification.

[3] "Profiling" means any form of automated processing of personal data, intended to analyse or predict the personality or certain personal aspects relating to a natural person, in particular the analysis and prediction of the person's health, economic situation, performance at work, personal preferences or interests, reliability or behaviour, location or movements.

This session brought all participants together and Martin and Simon summarised the main findings of each parallel session to the entirety of the audience (as described in the session sections above and summarised on the flipcharts shared in Flickr).

Links to other material

Workshop Blogpost: http://bit.ly/1QU58Sb
Agenda: link
Slides on slideshare
Photos on Flickr
Social Sciences page: http://www.big-data-europe.eu/social-sciences/
W3C interest group: https://www.w3.org/community/bde-societies/

Report by Eleanor Smith (CESSDA), Ivana Versic (CESSDA), Martin Kaltenböck (SWC), Simon Scerri (Fraunhofer IAIS).