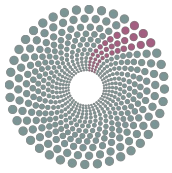

BIG DATA EUROPE

Big Data Europe for

Food and Agriculture

1st Workshop Report

INRA headquarters, Paris, 22 September 2015

Agenda

- 14:00 - Welcome & Introduction, 30 mins
  - Tour de Table
  - Pascal Neveu (INRA), introductory presentation: “"Introduction to INRA's big data perspective and implementation challenges“”
- 14:30 - Data Session
  - Lightning Talks, 30 mins
    - Tim Verhaart (Wageningen UR, LEI) "Big data opportunities for marketing of horticultural products”
    - Elisabeth Arnaud (CGIAR Bioversity International) "Big data analytics in the CGIAR research portfolio and the Bioversity perspective"
  - Interactive Session, 45 mins - split in 3 groups with 1 facilitator each
  - Report to Plenary, 15mins
- 16:00 - Technologies Session
  - Lightning Talks, 30 mins
    - Rob Lokers (Wageningen UR, Alterra) “Big data challenges and solutions in agricultural and environmental research”
    - Valeria Pesce (UN FAO & GFAR) "A global linked and open data infrastructure for agricultural development"
  - Interactive Session, 45 mins
  - Report to Plenary, 15mins
- 17:30 - Wrap up

- Sören Auer (FhG) on BDE big data technologies
- Stefano Bertolo (European Commission), position statement on the big data research agenda and portfolio of the EC
- Q&A session
- Closing

Workshop summary posted on the BDE website 7 October 2015; slides available in the BDE area in slideshare and also linked to this report.

Expectations and Background

FAO and Agro-Know, the BDE partners responsible for bringing in stakeholders and eliciting requirements from the food and agriculture community, have already convened stakeholders in the past to discuss challenges and opportunities around data and therefore had a rough idea of the level of awareness and the expectations around "big data" in the food and agriculture community.

However, the Big Data Europe project needed far more precise indications and requirements than the general scenario that past consultations could provide.

Normally, Big Data in agriculture are associated with information collected by sensors, satellites or drones combined with genomic information or climate data, which can all help farmers to optimize their farms' operations. In addition, challenges and opportunities have been identified by existing communities of data managers in this area also around the heterogeneity of the data that need to be combined and integrated for both fostering new research and innovation and providing meaningful information for decision making.

As an example of generic needs expressed by actors in the broad domain of "agriculture" (which includes food, forestry, fisheries and biotechnologies), a much envisaged big-data-empowered scenario would be the generic ability to deliver better added-value integrated services that can answer the needs of different types of actors. Examples of such integrated systems go from alert systems (pests, disasters) to image-recognition-based plant / pest identification systems to food tracking systems to global research information systems to any conceivable advanced decision making platform combining climate, soil, crop, pest, price, political and social data…

The objective of convening a workshop on big data in this area was that of checking if this scenario still reflected the perspectives of key actors and above all that of asking them more precise questions in order to elicit more specific requirements.

This was the background for the BigDataEurope workshop on "Big data for food, agriculture and forestry: opportunities and challenges", held in Paris on 22 September 2015. The workshop was organized by Agro-Know, FAO, GFAR and the Big Data Europe project and hosted by the Institut National de la Recherche Agronomique (INRA).

The workshop was intentionally co-located with another event that was going to bring together a good number of key actors in food and agriculture-related research: the pre-meeting of the Research Data Alliance Agricultural Data Interest Group. The Research Data Alliance (RDA, https://rd-alliance.org/) is an international initiative started in 2013 by a core of group of agencies (the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation). The Agricultural Data Interest Group (ADIG) is a domain oriented interest group in RDA, co-chaired by FAO, to work on all issues related to data that are relevant to agriculture and related domains.

This ensured the participation of more than 40 experts from around the world, representing universities, research institutions, private companies, international organizations, international projects and the European Commission (see participants' list in the Annexes).

Setting the scene

The workshop was organized around a short introductory session and two discussion sessions, one on data and one on technologies.

While the main results of the workshop are to be found in the summary of the discussion groups, where interesting requirements and challenges for BDE came out, a few presentations in the three sessions had the objective of setting the scene, describing current experiences and plans on big data in institutions working in agricultural research.

The workshop opened with Pascal Neuveu, senior research engineer at the French National Agronomic Institute (INRA) and director of the MISTEA Laboratory. He gave the big data perspective and implementation challenges that such a large and distributed agronomic research organisation has (INRA has 18 different centers spread across over 40 geographical locations, and more than 10,000 people from which about 4,000 are researchers). Pascal also presented a specific case study, the one of high throughput phenotyping at five large, open-air experimental fields, two greenhouses with controlled environments, and two fully equipped laboratories for carrying different types of omics analyses. Variety and Velocity were identified as the most challenging dimensions (velocity especially for streaming images).

After this initial presentation Stefano Bortolo from the European Commission raised the type of question that he kept raising at different points during the workshop: precise numbers in order to understand the order of magnitude of these data and the actual need for big data technologies. E.g. what is the largest number of environments in which 1 geno/phenotype has been tested from start till today? Answers from the audience (200 locations in different countries, 50k-60k per location?) tended also to point out difficulties in answering this question (e.g. considering also years of experiments, depending also on data stream rates…). A similar question was asked about the number of queries that can be handled.

Tim Verhaart from the socio-economic research institute of LEI (Wageningen UR) talked about "Big data opportunities for marketing of horticultural products", introducing a very interesting

public-private partnership through which the fruits and flowers industries are calling on big data technology researchers to help them do business better.

Rob Lokers from the inter-disciplinary environmental research institute of Alterra (Wageningen UR) brought forward the high complexity of using big data as input for complex agricultural and environmental modelling, which is then generating new (big) data, information, and knowledge that supports research and policy making – in his talk on “Big data challenges and solutions in agricultural and environmental research“.

Elizabeth Arnaud from Bioversity International provided an excellent insight into the CGIAR Big Data Analytics Platform and the CGIAR big data plans.

Valeria Pesce from the Global Forum on Agricultural Research (GFAR) presented where we stand today with a global linked and open data infrastructure for agricultural development.

At the end of the workshop, Soren Auer described the Big Data Europe project and Stefano Bertolo of the European Commission gave an overview of the big data research agenda and portfolio of the EC, also raising again some issues regarding the more precise identification of quantitative indicators that will justify the use of big data technologies in food and agriculture.

Breakout Groups

Two discussion sessions of 45 minutes each were organized in three breakout groups.

The first session was more around data (identifying the most relevant types of data and the related main challenges) while the second one discussed technologies (currently used solutions, promising technologies, gaps).

1. Breakout session 1: Data

The three groups were asked to answer specific questions on data and all answers were captured on flipcharts.

Questions:

- What are the most important data sources for agriculture and food?
- What is most important for such data in terms of: Volume, Velocity, Variety, Veracity?
- What are the challenges around these data sources (and the data) in terms of: availability, legal issues, policy issues, skills?

1.1. Overall results

Question 1: What are the most important data sources for agriculture and food?

Participants were invited to think especially of data types and data sources that are perceived as “big”.

The main types of data identified across the three groups were very similar:

- Sensor / drone data
- High-rate image streams
- Genomic data
- Phenotypic data
- Combined cross-disciplinary data (climate, economic, social…)

Question 2: What is most important of the 4 Vs for such data?

Across all three groups, the most important V was definitely Variety: even when speaking about an apparently homogeneous type of data (genomic, sensor), the shared opinion among participants was that a) those data were not particularly useful if not combined with other related data; b) even data of the same type don't come in a homogeneous form.

Particular importance was also given to veracity (especially for genomic data).

Volume was of course recognized as a big challenge for sensor data; velocity was mentioned mainly for high-rate image streams.

Question 3: What are the challenges in terms of: availability, legal issues, policy issues, skills?

Most participants complained about the lack of availability of data they would need due to policy or legal issues. In particular, it was noted that even when data are available what is often lacking is easy discoverability (vs. granting of special access).

The lack of skills also came up across all three groups.

1.2. Other findings and discussions

Other relevant types of data identified:

- Food tracking (volume)
- Food prices (volume)
- Multimedia for different purposes
- Public administration data (tax, customs, traffic)
- Historical data
- Model outputs

Other challenges identified:

- Availability: little availability of historical data; field phenotyping data are closed; lack of nutriments database; issue of competitive advantage
- Skills: lack of skills in all sorts of "omics"
- No data type on its own is useful, need to combine

Again the issue came up in discussions of how “big” these data are and what the threshold is beyond which a normal powerful server is not enough.

Regarding some challenges identified by participants in terms of availability of needed data, Stefano Bertolo from the EC highlighted both the existence of public European services providing different types of public utility data for free (e.g. regarding high speed imaging, Copernicus provides terabytes of data daily for open use) and the role of the EC in putting pressure on governments and public research if public data are not made available.

2. Breakout session 2: Technology

The three groups were asked to answer questions with specific reference to technologies for data processing, representation, acquisition and visualization.

Questions:

- What kind of data management solutions are in place at the moment & where are the main issues with those solutions at the moment?
- Are there promising technologies you have identified for the future?
- What are the challenges around these data technologies regarding: data acquisition, data processing, data infrastructure, data publication, security and/or privacy issues?

2.1. Overall results

An initial answer to questions 1 and 2 was provided by Rob Lokers’s presentation, in which several technologies commonly used in agricultural research were presented:

- RDBMS
- Geo-databases
- Various “old & proven” programming languages (esp. for modelling, data processing)
- Remote sensing: dedicated tools & environments for processing and analysis, ENVI, R, GDAL etc.
- Harmonized information / data models (but still per discipline)
- High Performance clusters / grids
- Still experimental (ICT research for agriculture):
    - RDF databases
    - Vocabularies and ontologies (no alignments)
    - NLP algorithms

Question 1: Data management solutions in place now and main issues?

Most of the technologies mentioned by Rob were also mentioned across all three groups.

Many participants reported that technology solutions are often produced in house and that researchers are still using Excel for managing their data.

All participants seemed to have a clear perception that the current technologies they are using are not enough for certain data processing needs.

Question 2: Are there promising technologies you have identified for the future?

Many participants seemed to agree on a number of promising technologies identified:

- NOSQL
- SPARQL
- Semantic tools
- NLP
- HADOOP
- Elastic search

No participant seemed to be already using a stack of big data technologies.

Question 3: Challenges in terms of acquisition, processing, infrastructure, publication, security / privacy issues?

Acquisition: legacy

Processing: performance; need for adequate processing for decision making

Publication: lack of incentives; no standard tools; preservation

2.2. Other findings and discussions

Other technologies that were mentioned include:

- In use:
  - Dataset sharing platforms: Data Cite, Figshare
  - Global information systems: e.g. GRIN for germplasm
  - Pattern recognition software
- In use, with issues:
  - Spreadsheets
  - In house developed software
- Promising:
  - OpenStreet Map
  - REST services
  - Neural networks
  - Linked Data Fragments
  - Federated search
  - Annotation tools

- ODK for surveys
- GOBII (high density marker data)

Conclusions

Overall, what came out of the discussions was that the special thing about big data in agriculture is its extreme variety.

This is what you get, if you contrast the four V's of big data to the data types and sources that are typically used in agricultural and food research. In most cases we are not talking about an extremely large Volume (other domains have much more voluminous data); it is not that data come with an extremely high Velocity, especially compared to other domains. In many cases, their Veracity is quite high. But in food and agriculture, data Variety matters the most: you need to combine multiple, heterogeneous data types and formats from several sources, trying to solve the information problems and support decision making of the relevant stakeholders.

However, real cases where volume and velocity were high were reported. The discussion indicated that it will be important for the BDE project to more precisely understand the order of magnitude of the data types discussed in this workshop (e.g. streaming data over decades, genomic data…) and the actual need for big data technologies.

Appendices

1. List of participants

Soren Auer	Fraunhofer, Germany
Nikos Manouselis	Agro-Know, Greece
Valeria Pesce	FAO / GFAR, Italy
Martin Kaltenböck	Semantic Web Company, Austria
Timea Turdean	Semantic Web Company, Austria
Stefano Bortolo	European Commission
Rob Lokers	Alterra, Wageningen UR, Netherlands
Marie Angélique Laporte	Bioversity International, Italy
Rosemary Shreshta	CIMMYT, Mexico
Nicolas Saby	INRA, France
Pascal Neveu	INRA, France
Nicolas Tremblay	Agriculture and Agri-Food Canada, Canada
Giovanni L'Abate	Consiglio per la Ricerca e la sperimentazione in Agricoltura (CRA), Italy
Caterina Caracciolo	FAO, Italy
Robert Davey	The Genom Analysis Centre, United Kingdom
Ruth Bastow	Global Plant Council
Nicolas Gengler	University of Liège, Belgium
Elizabeth Arnaud	Bioversity International, Italy
Tim Verweert	LEI, Wageningen UR, Netherlands
Daniel Martini	Association for Technology and Structures in Agriculture (KTBL), Germany
Dana Tomic	IEEE
Anahita Nafissi	Forschungszentrum Juelich, Germany

Hugo Besemer	Wageningen UR, Netherlands
Elf Pavlik	Freelancer
Cyril Pommier	INRA, France
Simon Scerri	Fraunhofer, Germany
Dimitrios Zisis	Institute of Plant Genetics, Polish Academy of Sciences, Poland
Pawel Krajewsky	Institute of Plant Genetics, Polish Academy of Sciences, Poland
Harris Moysiadis	Future Intelligence and QUHOMA project, Greece
Kalin Muldzhawski	Linked Farm
Cristian Vasquez	
Pierre Larmande	IRD, France
Aravind Venkatesan	IBC, France
Erick Antezana	Bayer CropScience NV, Belgium
Karna Wegner	FAO, Italy
Ruthie Musker	UC Davis, US / ETH Zurich, Switzerland

Report by Valeria Pesce (FAO/GFAR), Nikos Manouselis (Agro-Know), Timea Turdean (Semantic Web Company), Martin Kaltenböck (Semantic Web Company), Simon Scerri (Fraunhofer)