Chapter #13

BIG DATA IN AGRICULTURAL AND FOOD RESEARCH: CHALLENGES AND OPPORTUNITIES OF AN INTEGRATED BIG DATA E-INFRASTRUCTURE

Pythagoras Karampiperis, Rob Lokers, Pascal Neveu, Odile Hologne, George Kakaletris, Leonardo Candela, Matthias Filter, Nikos Manouselis Maritina Stavrakaki, Panagiotis Zervas[1]

Abstract: Agricultural and food research are increasingly becoming fields where data acquisition, processing, and analytics play a major role in the provision and application of novel methods in the general context of agri-food practices. The chapter focuses on the presentation of an innovative, holistic e-infrastructure solution that aims to enable researches for distinct but interconnected domains to share data, algorithms and results in a scalable and efficient fashion. It furthermore discusses on the potentially significant impact that such infrastructures can have on agriculture and food management and policy making, by applying the proposed solution in variegating agri-food related domains.

Key words: e-infrastructure, agro-climatic, economic modelling, food security, plant phenotyping, food safety, risk assessment

1. INTRODUCTION

Over the past years, big data in agricultural and food research has received increased attention. Farming has been empirically driven for over a century but the data collected was not digital. Agriculture Canada's family of research centres (circa 1920s) meticulously accounted for wheat yields across farms and weather patterns in order to increase efficiency in

[1] Corresponding author

production. Big Data is different from this historic information gathering in terms of the volume and the analytical potential embedded in contemporary digital technologies. Big Data proponents promise a level of precision, information storage, processing and analysing that was previously impossible due to technological limitations (Bronson and Knezevic, 2016).

Farmers have access to many data-intensive technologies to help them monitor and control weeds and pests, for example. In this sense, data collection, data modelling and analysis, and data sharing have become core challenges in weed control and crop protection (van Evert et al., 2017). As smart machines and sensors crop up on farms and farm data grow in quantity and scope, farming processes will become increasingly data-driven and data-enabled (Wolfert et al., 2017). This is due to the fact that vast amounts of data are produced by digital and connected objects such as farm equipment, sensors in the fields or biochips in animals. Moreover, robots are becoming more and more popular, as it is well illustrated in dairy production. Alongside with the continuous monitoring that is producing well-structured data, other sources of data are produced and used. Interconnections of information systems and interoperability among them are also increasingly important. This is leading to new management modes of agricultural and food production, new services offered and new relationships along the supply chains regarding data sharing and re-use.

Current developments in ICT and big data science potentially provide innovative and more effective ways to support agricultural and food researchers to work with extremely large data sets and handle use cases involving big data. This new paradigm raises new research questions, new methods and approaches to perform research in agriculture and food. To support this paradigm, research facilities and e-infrastructures need to be revisited and new partnerships among academic institutions and private companies should be emerged. This will possibly lead to the involvement of scientists from areas of science and technology, who are not involved to the agricultural and food sector, to boost the innovation process in this sector. This also creates new challenges and opportunities for informed decision making from micro scale (i.e. precision agriculture) to macro scale (i.e. policy making).

Within this context, the main aim of this book chapter is twofold: (a) to present the design of AGINFRA+, an integrated Big Data e-Infrastructure for supporting Agricultural and Food Research, which will be based on existing generic e-Infrastructures such as OpenAIRE (for publication and data set aggregation, indexing, mining and disambiguation), EUDAT (for cloud-hosted preservation and storage), EGI.eu (for cloud and grid resources for intensive computational applications), and D4Science (for data analytics); (b) to present three use cases for performing research in

agriculture and food with the use of the proposed Big Data e-Infrastructure. More specifically, the first use case will be related to global agricultural modelling, inter-comparison, and improvement of the research community that studies short and long-term food production under environmental and climate change conditions. The second use case will be related to addressing specific problems on food security, namely the need to efficiently analyze data produced by plant phenotyping and its correlation with crop yield, resource usage and local climates. The third use case will be related to the design of new high-performance food safety data processing workflows facilitating (i) the efficient extraction of data and models from the rich corpus of scientific literature, and (ii) the issue of generating easy-to-maintain open food safety model repositories.

This book chapter begins by presenting the suggested integrated big data infrastructure for agricultural and food research. It will then go on to provide an overview of the existing initiatives and e-infrastructures as well as details of the above-mentioned use cases. Finally, it will analyze the new challenges and directions that will arise for agriculture and food management and policing.

2. AN INTEGRATED BIG DATA INFRASTRUCTURE FOR AGRICULTURAL AND FOOD RESEARCH

AGINFRA+ aims to exploit core e-infrastructures such as EGI.eu, OpenAIRE, EUDAT and D4Science, towards the evolution of the AGINFRA data infrastructure, so as to provide a sustainable channel addressing adjacent but not fully connected user communities around Agriculture and Food.

To this end, AGINFRA+ entails the development and provision of the necessary specifications and components for allowing the rapid and intuitive development of variegating data analysis workflows, where the functionalities for data storage and indexing, algorithm execution, results visualization and deployment are provided by specialized services utilizing cloud based infrastructure(s). Furthermore, AGINFRA+ aspires to establish a framework facilitating the transparent documentation and exploitation and publication of research assets (datasets, mathematical models, software components results and publications) within AGINFRA, in order to enable their reuse and repurposing from the wider research community.

In a high-level, conceptual view on the AGINFRA architecture (see Figure 1), the AGINFRA+ *Data & Semantics Layer* comprises the

functionalities that facilitate the discovery and consumption of data collections (annotation, description enrichment, linking / mapping, etc.). These datasets and components enable researchers, innovators as well as businesses to compose customized data-driven services via the AGINFRA+ *Presentation Layer*, which builds upon the use of existing open source frameworks such as D3 (an open-source visualization library for data-driven real-time interactivity), Processing.js (a library that sits on top of the Processing visual programming language), Shiny (for interactive web data analysis) etc.

Alternatively, researchers can create (advanced) data analysis and processing workflows which can be packaged as ready-to-deploy components and can be submitted for execution to the AGINFRA+ *Data Analytics & Processing Layer*. The layer is responsible for activating the necessary executable components and monitoring their execution over the available e-infrastructures in accordance with the provided design.

Finally, the evolved AGINFRA architecture entails the necessary tools for making the methodologies and outcomes (algorithms, research results, services) of research activities available via Open Access Infrastructures.

Overall, the services to be developed within AGINFRA+ aim to provide their users with the means to perform rapid prototype production, facilitate the execution of resource-intensive experiments, allow the agile and intuitive

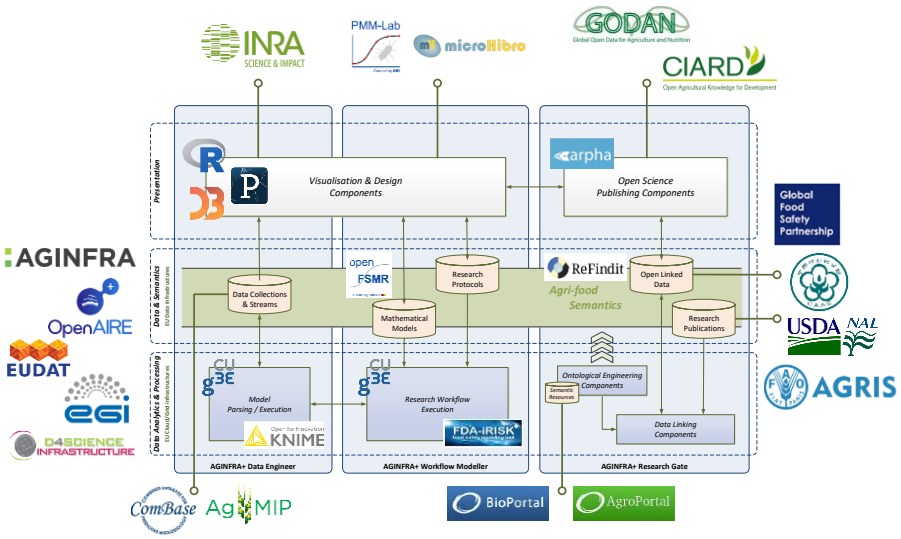

Figure 1: Conceptual Architecture showcasing how different technologies/tools form the vision of AGINFRA+

parameterization and repetition of experiments, and – ultimately – maximize the visibility and reuse of their research outcomes.

3. RELATED WORK: EXISTING E-INFRASTRUCTURES FOR AGRICULTURAL AND FOOD RESEARCH

In this section, we provide an overview of existing initiatives and e-infrastructures that pertain to the vision and objectives of AGINFRA+.

3.1 AGINFRA

AGINFRA is the European research hub and thematic aggregator that catalogues and makes discoverable publications, data sets and software services developed by Horizon 2020 research projects on topics related to agriculture, food and the environment) so that they are included in the European research e-infrastructure "European Open Science Cloud".

3.2 EGI-Engage

The EGI-Engage: Engaging the Research Community towards an Open Science Commons (https://www.egi.eu/about/egi-engage/) started in March 2015, as a collaborative effort involving more than 70 institutions in over 30 countries, coordinated by the European Grid Initiative (EGI) association. EGI-Engage aims to accelerate the implementation of the Open Science Commons by expanding the capabilities of a European backbone of federated services for compute, storage, data, communication, knowledge and expertise, complementing community-specific capabilities.

Agriculture, food and marine sciences are included as use cases in EGI-Engage, providing requirements that shape the new Open Science Commons platform. AGINFRA is positioned as the thematic data infrastructure and is exploring use cases and workflows of storing, preserving and processing data in the generic e-infrastructure.

3.3 OpenAire 2020

OpenAIRE2020 continues and extends OpenAIRE's scholarly communication open access infrastructure to manage and monitor the outcomes of EC-funded research. It combines its substantial networking capacities and technical capabilities to deliver a robust infrastructure offering support for the Open Access policies in Horizon 2020.

The integration of all AGINFRA scientific repositories in OpenAIRE is currently undergoing, aiming to have AGINFRA serving as the OpenAIRE thematic node for agriculture and food.

3.4 EUDAT

EUDAT's vision is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI) conceived as a network of collaborating, cooperating centres, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres. EUDAT offers common data services, supporting multiple research communities as well as individuals, through a geographically distributed, resilient network of 35 European organisations.

AGINFRA+ partners have been early and continuously participating in EUDAT user workshops, providing requirements and use cases on how various EUDAT services may be used by our scientific communities.

3.5 D4Science

D4Science is a Hybrid Data Infrastructure servicing a number of Virtual Research Environments. Its development started in the context of the homonymous project cofounded by the European Commission and has been sustained and expanded by a number of EU-funded projects (ENVRI, EUBrazilOpenBio, iMarine). Currently, it serves as the backbone infrastructure for the BlueBRIDGE and ENVRIPlus projects.

AGINFRA+ partners are technology leaders in D4Science, while also contributing use cases and requirements to D4Science, through their participation of the corresponding projects or project events.

3.6 BlueBRIDGE

BlueBRIDGE aims at, innovating current practices in producing & delivering scientific knowledge advice to competent authorities & enlarges the spectrum of growth opportunities in distinctive Blue Growth areas and to further developing and exploiting the iMarine e-Infrastructure data services for an ecosystem approach to fisheries.

3.7 ENVRIPlus

ENVRIPLUS is a cluster of research infrastructures (RIs) for Environmental and Earth System sciences, built around ESFRI roadmap and associating leading e-infrastructures and Integrating Activities together with technical specialist partners.

Agriculture, food and marine sciences are represented as communities in ENVRIPlus.

3.8 ELIXIR

The goal of ELIXIR is to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments. Some of these datasets are highly specialised and would previously only have been available to researchers within the country in which they were generated.

For the first time, ELIXIR is creating an infrastructure – a kind of highway system – that integrates research data from all corners of Europe and ensures a seamless service provision that is easily accessible to all. In this way, open access to these rapidly expanding and critical datasets will facilitate discoveries that benefit humankind.

In the currently running EXCELERATE project (https://www.elixir-europe.org/excelerate) that develops further the ELIXIR infrastructure, a major use case is on Integrating Genomic and Phenotypic Data for Crop and Forest Plants.

3.9 EOSC

As part of the European Digital Single Market strategy, the European Open Science Cloud (EOSC) will raise research to the next level. It promotes not only scientific excellence and data reuse but also job growth and increased competitiveness in Europe, and drives Europe-wide cost efficiencies in scientific infrastructure through the promotion of interoperability on an unprecedented scale. It aims to bring together and align the existing core e-infrastructure services in order to smoothly support and integrate thematic and domain specific data infrastructures and VREs.

AGINFRA+ partners have already initiated a strategic discussion with the key players trying to shape the EOSC and have participated in the various workshops and brainstorming meetings.

4. USE CASES FOR AGRICULTURAL AND FOOD RESEARCH

4.1 Agro-climatic & Economic Modelling

4.1.1 Landscape

The case of a global agricultural modeling, intercomparison, and improvement research community that studies short and long-term food production under environmental and climate change conditions. In this case, the problem addressed is related to accessing, combining, processing and storing high volume, heterogeneous data related to agriculture/food production projections under different climate change scenarios, so that it becomes possible to assess food security, food safety and climate change impacts in an integrated manner, by a diverse research community of agricultural, climate and economic scientists.

The mission of this research community lies in improving historical analysis and short and long-term forecasts of agricultural production and its effects on food production and economy under dynamic and multi-variable climate change conditions, aggregating extremely large and heterogeneous observations and dynamic streams of agricultural, economical, ecophysiological, and weather data.

Bringing together researchers working on these problems from various perspectives (crop production and farm management methods, climate change monitoring, economic production models, food safety models), and accelerate user-driven innovation is a major challenge. The AGINFRA+ services will enable executable workflows for ecophysiological model intercomparisons driven by historical climate conditions using site-specific data on soils, management, socioeconomic drivers, and crop responses to climate. These intercomparisons are the basis for the future climate impact and adaptation scenarios: instead of relying on single model outputs, model-based uncertainties will be quantified using multi-model ensembles. The close interactions and the linkages between disciplinary models and scenarios, including climate, ecophysiology and socio-economics will allow researchers to prioritize necessary model improvements across the model spectrum.

Multi-model, multi-crop and multi-location simulation ensembles will be

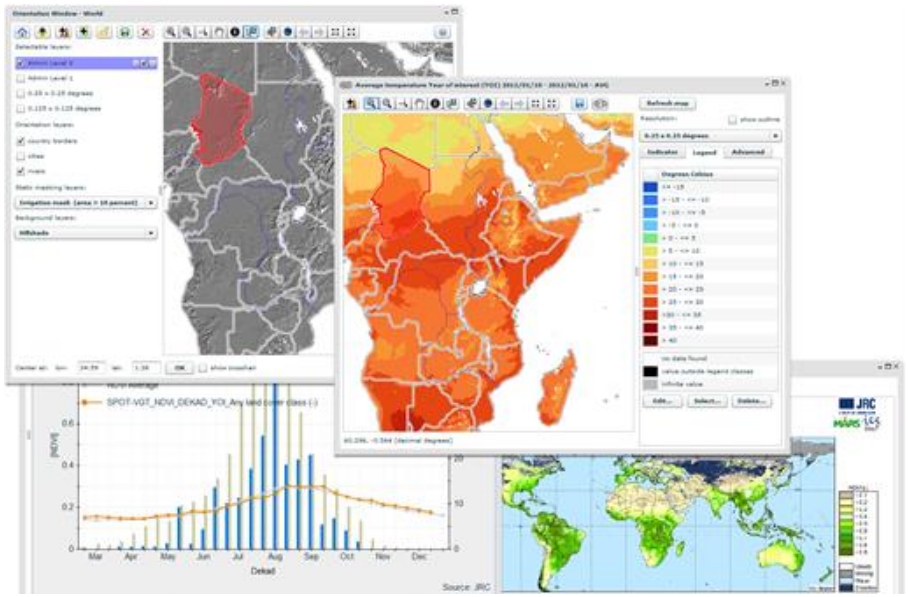

Figure 2:Examples of various data types and sources used to support impact assessment

linked to multi-climate scenarios to perform consistent simulations of future climate change effects on local, regional, national, and global food production, food security, and poverty.

The data sources that can feed such bases required for this work are developed by different community members, are processed using different systems, and are shared among the community members. This creates several challenges that are connected to multiple factors: different platforms, diverse data management activities, distributed data processing and storage, heterogeneous data exchange, etc. and distributed model runs, data storage, scenario analysis, and visualization activities that take place. Thus, AGINFRA+ will also develop a reactive intensive data analysis layer over existing federations that will help the discovery, reuse and exploitation of heterogeneous data sources created in isolation, in very different and unforeseen ways in the rest of the communities' systems.

4.1.2 AGINFRA+ advancement

Improving historical analysis and short and long-term forecasts of agricultural production and its effects on food production and economy under dynamic and multi-variable climate change conditions, is associated with several challenges in the area of Big Data. For short term (e.g. seasonal) forecasting of production timeliness is essential. To provide daily updates of forecasts, taking in the most actual information, real-time processing of up-to-date weather data and remote sensing data is required. For long-term projections, using among others climate ensembles from global or regional climate models, handling the volume of input and output data (combining, processing, storing) is more relevant, together with the availability of interconnected high-performance computing and cloud storage solutions.

As for now, many modelling exercises are performed "locally", mainly because the integrated e-infrastructure (meaning seamlessly interconnecting data-infrastructures, grid computing infrastructures and cloud-based infrastructures) are not available to the modelling community. Organizing such sustainable large-scale integrations and establishing working real-world modelling cases using these combined e-infrastructures would be a great leap forward for the involved research communities.

4.2 Food Security

4.2.1 Landscape

The AGINFRA+ infrastructure will be used to alleviate the big data challenges pertaining to specific problems on food security, namely the need to efficiently analyse data produced by plant phenotyping and its correlation with crop yield, resource usage, local climates etc. It will particularly explore how high throughput phenotyping can be supported in order to:

a) Determine adaptation and tolerance to climate changes, a high priority is to design high-yielding varieties adapted to contrasting environmental conditions including those related to climate change and new agricultural management. It requires identifying of allelic variants with favourable traits amongst the thousands of varieties and natural accessions existing in genebanks. Genotyping (i.e. densely characterizing the genome of breeding lines with markers) has been industrialized and can now be performed at affordable cost and be able to link and analyse phenotyping and genotyping data is strategic for agriculture.
b) Optimize use of natural resources. High throughput plant phenotyping aims to study plant growth and development from the dynamic interactions between the genetic background and the environment which plants develop (soil, water, climate, etc.). These interactions determine plant performance and productivity can be use in order to optimize and preserve natural resources.
c) Maximize crop performance, gathering and analysing data from high throughput plant phenotyping allows a better knowledge of plants and their behaviour in specific resource conditions such as soil conditions and new climates.
d) The tasks generally require intensive big data analysis as they present all the challenges associated with big data (Volume, Velocity, Variety, Validity).
e) Each high throughput plant phenotyping produced several Tbytes of very heterogeneous data (sensor monitoring, images, spectrums). Data are produced at high frequencies and hundreds of thousands of images can be gather and be analysed each day. Such volumes require automatic data validation tools. On of major challenges of plant phenotyping is the Semantic Interoperability.

This AGINFRA+ use case will assess the effectiveness of the proposed framework in data intensive experimental sessions, where the distinct processing steps operate over different datasets, require synchronization of results from various partial computations, and use very large and/or streaming raw or processed data.

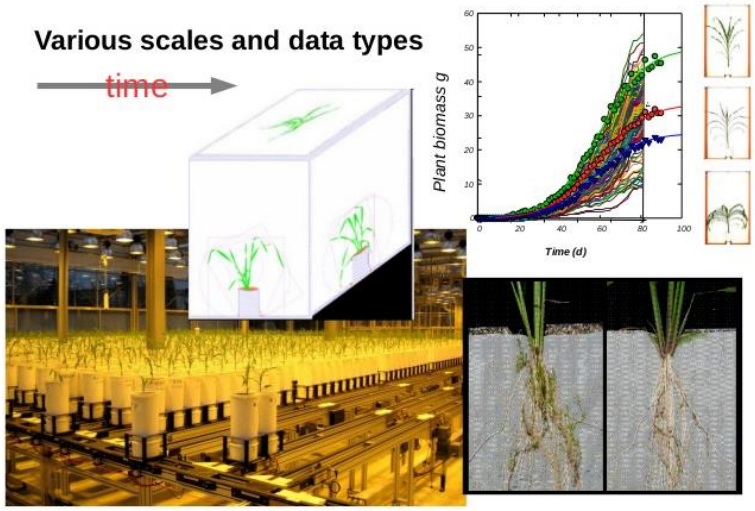

Figure 3:Examples of visualization for high throughput phenotyping

4.2.2 AGINFRA+ advancement

Plant derived products are at the centre of grand challenges posed by increasing requirements for food, feed and raw materials. Integrating approaches across all scales from molecular to field applications are necessary to develop sustainable plant production with higher yield and using limited resources. While significant progress has been made in molecular and genetic approaches in recent years, the quantitative analysis of plant phenotypes - structure and function of plant - has become the major bottleneck.

Plant phenotyping is an emerging science that links genomics with plant ecophysiology and agronomy. The functional plant body (PHENOTYPE) is formed during plant growth and development from the dynamic interaction between the genetic background (GENOTYPE) and the physical world in which plants develop (ENVIRONMENT). These interactions determine plant performance and productivity measured as accumulated biomass and commercial yield and resource use efficiency.

Improving plant productivity is key to address major economic, ecological and societal challenges. A limited number of crops provides the resource for food and feed; reliable estimates indicate that food supplies need to be increased by quantity (50% by 2050) and quality to meet the increasing nutritional demand of the growing human (and animal) population. At the same time, plants are increasingly utilized as renewable energy source and as raw material for a new a generation of products. Climate change and scarcity of arable land constitute additional challenges

for future scenarios of sustainable agricultural production. It is necessary and urgent to increase the impact of plant sciences through practical breeding for varieties with improved performance in agricultural environments.

The understanding of the link between genotype and phenotype is currently hampered by insufficient capacity (both technical and conceptual) of the plant science community to analyze the existing genetic resources for their interaction with the environment. Advances in plant phenotyping are therefore a key factor for success in modern breeding and basic plant research.

To deploy a "big data" strategy for organizing, annotating and storing phenotyping data for plant science in such a way that any scientist in Europe can potentially access the data of several for a common analysis. Which data should be standardized, which not? Standardization with common ontologies is crucial. Elaborate a consensus in the phenotyping community in Europe with high degree of standardization for environmental variables (types of sensors, units) and basic phenotypic data (dates, phenology) and higher flexibility for elaborate traits that need to reach maturity before being standardized. The relation with generalist databases (plant/crop ontologies) need to be established.

How to manage and process increasing data volume with distributed and reproducible data workflows? How to share data / how to access data from various experiments? How to combine large and heterogeneous datasets in an open data perspective? The most recent methods of information technologies will be used for smart data exchanges namely key discovery, matching and alignment methods for accessing and visualizing complex data sets in an integrative way.

4.3 Food Safety Risk Assessment

4.3.1 Landscape

In the context of the Food Safety Risk Assessment use case, the AGINFRA+ project will assess the usefulness of AGINFRA+ components and APIs to support data-intensive applications powered by the FoodRisk-Labs suite of software tools (https://foodrisklabs.bfr.bund.de). This includes the extension of FoodRisk-Labs' capabilities to handle large-scale datasets, to visualize complex data, mathematical models as well as simulation results and to deploy generated data processing workflows as web-based services.

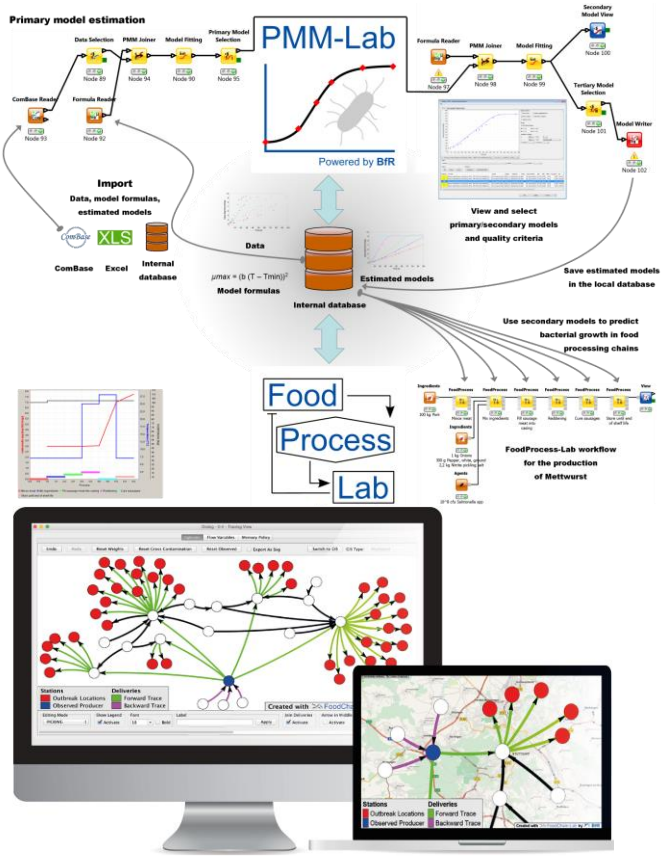

Figure 4: Examples of graphical workflow based data analysis and visualisation features in FoodRisk-Labs

More specifically, FoodRisk-Labs will be extended such that it can use and access AGINFRA Services. This will allow to design new high-performance food safety data processing workflows facilitating e.g. the efficient extraction of data and models from the rich corpus of scientific literature. Another workflow will address the issue of generation of easy-to-maintain open food safety model repositories (openFSMR), which exploit AGINFRA Ontological Engineering and Open Science Publishing Components. Mathematical models published in community driven openFSMR will then be used for large-scale quantitative microbial risk assessment (QMRA) simulations. These simulations incorporate predictive microbial models, models on food processing and transportation, dose response models as well as consumer behaviour models. AGINFRA components supporting the execution of computational intensive simulations as well as those helping to present simulation results will be applied here.

Finally, preconfigured QMRA models will be deployed as easy-to-use web services using specialised AGINFRA components.

4.3.2 AGINFRA+ advancement

The exploitation of food safety and food quality data using data-mining algorithms is a field of increasing relevance to food safety professionals and public authorities. Nowadays, the amount of experimental and analytical food safety data increase as well as information collected from sensors monitoring food production and transportation processes. However, there is a gap in free, easy-to-adapt software solutions that enables food safety professionals and authorities to exploit these data in a consistent manner applying state-of-the art data mining and modelling technologies. In addition, integrated, standardized data and knowledge management is required to establish quality-controlled model repositories that can be used for risk assessment and decision support.

The FoodRisk-Labs collection of software tools has been outlined right from the beginning as a community resource to allow broad application and joint developments. The specific food safety analysis and modelling functionalities were implemented as extensions to the professional open-source data analysis and machine learning platform KNIME (www.knime.org). The KNIME visual workflow composition user interface enables users to apply or adapt preconfigured data analysis workflows or to create new ones from the large number of available data processing modules ("nodes"). The selection of KNIME as the technical implementation framework guarantees important features like modularity, flexibility, scalability and extensibility. All FoodRisk-Labs KNIME extensions also provide preconfigured (empty) databases allowing users to easily manage their domain-specific data and/or establish model-based knowledge repositories.

5. NEW CHALLENGES AND DIRECTIONS FOR AGRICULTURE AND FOOD MANAGEMENT AND POLICING

The variety of stakeholders that are involved in scientific activities that address major societal challenges around agriculture, food and the environment is enormous. They range from researchers working across geographical areas and scientific domains, to policy makers designing development and innovation interventions. These activities have been traditionally informed and powered by a combination of quite heterogeneous

data sources and formats, as well as several research infrastructures and facilities, at a local, national, and regional level. In 2010, a SCAR study tried to give an overview of this picture, which has been documented in the report "*Survey on research infrastructures in agri-food research*" (Survey on Research Infrastructures in Agri-food Research, 2010). As more and more research information and IT systems became available online, the relevance of agricultural knowledge organisation schemes and systems became higher. A recent foresight paper on the topic has been published by the SCAR working group "*Agricultural Knowledge and innovation systems*" (AKIS-3, 2016). The emergence of the open access and data policies has brought forward new challenges and opportunities (as the 2015 *GODAN Discussion Paper* (Open Data Institute, 2015) has revealed), which have to be addressed and supported by future e-infrastructure services and environments. In addition to this, Commissioner Moedas pointed out a clear link between societal challenges and openness as a European priority. He has positioned these challenges across the three dimensions of Open Science, Open Innovation and Openness to the World. AGINFRA+ is best positioned to achieve impact upon all these dimensions.

5.1 Open Science

The Belmont Forum[2] is a roundtable of the world's major funding agencies of global environmental change research and international science councils, which collectively work on how they may address the challenges and opportunities associated with global environmental change. In 2013, the Belmont Forum initiated the multi-phased *E-Infrastructures and Data Management* Collaborative Research Action[3]. In August 2015, this initiative published its recommendations on how the Belmont Forum can leverage existing resources and investments to foster global coordination in e-Infrastructures and data management, in a report entitled "A Place to Stand: e-Infrastructures and Data Management for Global Change Research" (Belmont Forum, 2015). The main recommendations of this report included (a) to adopt data principles that establish a global, interoperable e-infrastructure with cost-effective solutions to widen access to data and ensure its proper management and long-term preservation, (b) to promote effective data planning and stewardship in all research funded by Belmont Forum agencies, to enable harmonization of the e-infrastructure data layer, and (c) to determine international and community best practice to inform e-infrastructure policy for all Belmont Forum research, in harmony with

[2] https://www.belmontforum.org
[3] http://www.bfe-inf.org/info/about

evolving research practices and technologies and their interactions, through identification of cross-disciplinary research case studies.

AGINFRA+ is fully aligned with these recommendations and has as its strategic impact goal to ensure that Europe brings forward to the Belmont Forum AGINFRA as a world-class data infrastructure. It aims to demonstrate a number of cost-effective and operational solutions for access, management and preservation of research data that will be based upon (and take advantage of) the core e-infrastructure services at hand. And to introduce, through a number of prototype demonstrators, the three selected research use cases as good practices that may become international and community best practices and drive e-infrastructure policy for all Belmont Forum research.

5.2 Open Innovation

The agricultural business landscape is rapidly changing. Established brands in agriculture such as John Deere, Monsanto, and DuPont are now as much data-technology companies[4] as they are makers of equipment and seeds. Even though agriculture has been slower and more cautious to adopt big data than other industries, Silicon Valley and other investors are taking notice. Startups like *Farmers Business Network*[5], which counts Google Ventures as an investor, have made collecting, aggregating, and analysing data from many farms their primary business. Popular, business and tech press keeps on highlighting the evolution that (big) data brings into the agriculture (Bobkoff, 2015), food (Thusoo, 2014) (Metz, 2014) and water (Fishman, 2016) business sectors – but also into helping feed 9 billion people (Gilpin, 2016). For instance, in the farming sector, data collection, management, aggregation and analytics introduce a wide variety of innovative applications and solutions such as, for example, sensors which can tell how effective certain seed and types of fertilizer are in different sections of a farm. Another example could be a software which may instruct the farmer to plant one hybrid in one corner and a different seed in another for optimum yield or intelligent systems which may adjust nitrogen and potassium levels in the soil in different patches. All this information can be automatically also shared with seed companies to improve hybrids.

This is also creating an investment environment with a tremendous potential for startups and companies that are focusing on data-intensive applications. Last year's investment report from AgFunder[6] states that the

[4] http://techcrunch.com/2013/10/02/monsanto-acquires-weather-big-data-company-climate-corporation-for-930m/

[5] https://www.farmersbusinessnetwork.com/

[6] https://agfunder.com

$4.6 billion that was raised by agricultural technology (AgTech) companies in 2015 from private investors, is nearly double 2014's $2.36 billion total - and outpaced average growth in the broader venture capital market. It also points out that this should not be considered as a new *tech bubble*: apart from the food e-commerce sector that seems overheated, the rest of the agriculture and food market is still facing challenges not seen in other mainstream technology sectors. In comparison to the size of the global agriculture market ($7.8 trillion), AgTech investment (with less than 0.5% of it) seems to be very modest and with amazing prospects.

AGINFRA+ particularly aims to take advantage of this investment trend by targeting and involving agriculture and food data-powered companies (and especially startups and SMEs). It has a dedicated work activity on getting such companies involved, and it will align its efforts with the business outreach (through data challenges, hackathons, incubators and accelerators) of its European (e.g. ODI[7], Big Data Value Association[8]) and global networks (e.g. GODAN[9]).

5.3 Openness to the World

At the 2012 G-8 Summit, G-8 leaders committed to the *New Alliance for Food Security and Nutrition*, the next phase of a shared commitment to achieving global food security. As part of this commitment, they agreed to "*share relevant agricultural data available from G-8 countries with African partners and convene an international conference on Open Data for Agriculture, to develop options for the establishment of a global platform to make reliable agricultural and related information available to African farmers, researchers and policymakers, taking into account existing agricultural data systems*" (White House Office of the Press Secretary, 2012). In April 2013, the prestigious G-8 International Conference on Open Data for Agriculture took place in Washington DC, announcing the G8 Open Data Action plans[10]. The goal of the EC's action plan[11] has been "Open access to publicly funded agriculturally relevant data" and included the flagship projects (such as agINFRA, SemaGrow, TRANSPLANT, OpenAIRE) that some of the key AGINFRA+ partners were coordinating and implementing at that time. To a large extend, when the *Global Open Data for Agriculture and Nutrition (GODAN)* initiative was launched as a

[7] http://theodi.org/
[8] http://www.bdva.eu/
[9] http://godan.info/
[10] https://sites.google.com/site/g8opendataconference/action-plans
[11] https://docs.google.com/file/d/0B4aXVC8hUc3oZlVEdlZ1RVJvZms/edit

result of this conference, these EC-funded projects have been driving the contributions of partners like ALTERRA and Agroknow that made Europe one of the GODAN global leaders.

In a similar way, and through its representation and active contribution to international networks like *GODAN*, the *Research Data Alliance (RDA)* and the *Coherence in Information for Agricultural Research and Development (CIARD[12])*, AGINFRA+ aims to continue supporting the global outreach and collaboration of European agriculture and food data stakeholders with their international counterparts.

6. CONCLUSIONS AND FUTURE WORK

The proposed e-infrastructure addresses the challenge of supporting user-driven design and prototyping of innovative e-infrastructure services and applications. It particularly tries to meet the needs of the scientific and technological communities that work on the multi-disciplinary and multi-domain problems related to agriculture and food. It will use, adapt and evolve existing open e-infrastructure resources and services, in order to demonstrate how fast prototyping and development of innovative data- and computing-intensive applications can take place.

In order to realize its vision, AGINFRA+ will achieve the following objectives:

- identify the requirements of the specific scientific and technical communities working in the targeted areas, abstracting (wherever possible) to new AGINFRA services that can serve all users;
- design and implement components that serve such requirements, by exploiting, adapting and extending existing open e-infrastructures (namely, OpenAIRE, EUDAT, EGI, and D4Science), where required;
- define or extend standards facilitating interoperability, reuse, and repurposing of components in the wider context of AGINFRA;
- establish mechanisms for documenting and sharing data, mathematical models, methods and components for the selected application areas, in ways that allow their discovery and reuse within and across AGINFRA and served software applications;
- increase the number of stakeholders, innovators and SMEs aware of AGINFRA services through domain-specific demonstration and dissemination activities.

Furthermore, AGINFRA+ will focus on the development of fully defined demonstrator applications in three critical application areas, which will allow

[12] http://ciard.info/

to showcase and evaluate the infrastructure and its components in the context of specific end-user requirements from different scientific areas.

REFERENCES

AKIS-3, S. W. (2016). *Agricultural Knowledge and Information Systems Towards the Future - A Foresight Paper*. Retrieved from http://ec.europa.eu/research/scar/pdf/akis-3_end_report.pdf#view=fit&pagemode=none

Belmont Forum. (2015). *A Place to Stand: e-Infrastructures and Data Management for Global Change Research*. Retrieved from http://www.bfe-inf.org/sites/default/files/A_Place_to_Stand-Belmont_Forum_E-Infrastructures__Data_Management_CSIP.pdf

Bobkoff, D. (2015, September 15). *Seed by seed, acre by acre, big data is taking over the farm.* Retrieved from http://www.businessinsider.com/big-data-and-farming-2015-8

Bronson, K. and Knezevic, I. (2016). Big Data in food and agriculture. Big Data & Society. January-June 2016: 1–5

Fishman, C. (2016, March 17). *Water Is Broken. Data Can Fix It.* Retrieved from https://www.nytimes.com/2016/03/17/opinion/the-water-data-drought.html?_r=0

Gilpin, L. (2016). *How big data is going to help feed nine billion people by 2050.* Retrieved from http://www.techrepublic.com/article/how-big-data-is-going-to-help-feed-9-billion-people-by-2050/

Metz, C. (2014). *Forget GMOs. The Future of Food Is Data - Mountains of It.* Retrieved from https://www.wired.com/2014/09/ex-googler-using-big-data-model-creation-new-foods/

Open Data Institute. (2015). *How Can We Improve Agriculture, Food and Nutrition With Open Data*. Retrieved from http://www.godan.info/sites/default/files/old/2015/04/ODI-GODAN-paper-27-05-20152.pdf

Survey on Research Infrastructures in Agri-food Research. (2010). Retrieved from https://ec.europa.eu/research/scar/pdf/final_scar_survey_report_on_infrastructures.pdf

Thusoo, A. (2014). *How Big Data is Revolutionizing the Food Industry.* Retrieved from https://www.wired.com/insights/2014/02/big-data-revolutionizing-food-industry/

van Evert, F.K., Fountas, S., Jakovetic, D., Crnojevic, V., Travlos, I. and Kempenaar, C. (2017). Big data for weed control and crop protection. Weed Research 57, 218–233.

White House Office of the Press Secretary. (2012, May 18). *Fact Sheet: G-8 Action on Food Security and Nutrition.* Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2012/05/18/fact-sheet-g-8-action-food-security-and-nutrition

Wolfert, S., Ge, L., Verdouw, C. and Bogaardt, M-J. (2017). Big Data in Smart Farming – A review Agricultural Systems 153, 69–80.

ACKNOWLEDGEMENTS

The work presented in this chapter has been partly supported by the AGINFRAPLUS Project that is funded by the European Commission's Horizon 2020 research and innovation programme under grant agreement No 731001.

This is a pre-print of a book chapter published in “Big Data for the Greater Good”. The final authenticated version is available online at: https://doi.org/10.1007/978-3-319-93061-9_6