D8.3 – FUTURE SCIENCE RECOMMENDATIONS

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D8.3
DELIVERABLE TITLE	Future Science Recommendations
RESPONSIBLE AUTHOR	Nikos Manouselis & Eva Bozou (Agroknow)

GRANT AGREEMENT N.	731001
PROJECT ACRONYM	AGINFRA PLUS
PROJECT FULL NAME	Accelerating user-driven e-infrastructure innovation in Food & Agriculture
STARTING DATE (DUR.)	01/01/2017 (36 months)
ENDING DATE	31/12/2019
PROJECT WEBSITE	plus.aginfra.eu
COORDINATOR	Nikos Manouselis
ADDRESS	110 Pentelis Str., Marousi GR15126, Greece
REPLY TO	nikosm@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mr Christophe Doin
WORKPACKAGE N. | TITLE	WP8 | Dissemination & Exploitation
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D8.3 | AGINFRA Future Science Recommendations
RESPONSIBLE AUTHOR	Nikos Manouselis (Agroknow)
REPLY TO	nikosm@agroknow.com
DOCUMENT URL	http://www.plus.aginfra.eu/sites/plus_deliverables/D8.3.pdf
DATE OF DELIVERY (CONTRACTUAL)	31 December 2018 (M24)
DATE OF DELIVERY (SUBMITTED)	15 January 2019 (M25), 31 January 2020 (M37, updated)
VERSION | STATUS	V2.0 | Final
NATURE	Report (R)
DISSEMINATION LEVEL	Public (PU)
AUTHORS (PARTNER)	Nikos Manouselis (Agroknow)
CONTRIBUTORS	Rob Lokers (DLO), Rob Knapen (DLO), Alice Boizet (INRA), Matthias Filter (BfR), Irene Kouriantaki (Agroknow), Alina Petri (Agroknow), Francesca Tsaropoulou (Agroknow), Giannis Stoitsis (Agroknow)
REVIEWER	Athanassios Ballis (UoA)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Table of contents	07-12-2018	Panagiotis Zervas (Agroknow)
0.2	Chapter 2	10-12-2018	Panagiotis Zervas (Agroknow)
0.7	Chapter 3	07-01-2019	Rob Lokers (DLO), Rob Knapen (DLO) Alice Boizet (INRA) Matthias Filter (BfR)
0.8	Chapters 1 and 4, overall formatting	09-01-2019	Panagiotis Zervas (Agroknow)
1.0	Final version after internal review	15-01-2019	Panagiotis Zervas (Agroknow)
1.5	Editing of the "Digital Science Recommendations for Food & Agriculture" discussion paper	31-12-2019	Nikos Manouselis (Agroknow), Giannis Stoitsis (Agroknow), Irene Kouriantaki (Agroknow), Eva Bozou (Agroknow)
2.0	Final Version	31-01-2020	Nikos Manouselis (Agroknow)

PARTICIPANTS		CONTACT
Agroknow IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com
Stichting Wageningen Research (DLO, The Netherlands)		Rob Lokers Email: rob.lokers@wur.nl
Institut National de la Recherché Agronomique (INRA, France)		Pascal Neveu Email: pascal.neveu@inra.fr
Bundesinstitut für Risikobewertung (BFR, Germany)		Matthias Filter Email: matthias.filter@bfr.bund.de
Consiglio Nazionale Delle Richerche (CNR, Italy)		Leonardo Candela Email: leonardo.candela@isti.cnr.it
University of Athens (UoA, Greece)		George Kakaletris Email: gkakas@di.uoa.gr
Stichting EGI (EGI.eu, The Netherlands)		Tiziana Ferrari Email: tiziana.ferrari@egi.eu
Pensoft Publishers Ltd (PENSOFT, Bulgaria)		Lyubomir Penev Email: penev@pensoft.net

EXECUTIVE SUMMARY

This document serves as a white paper, which describes the AGINFRA PLUS vision of a next-generation community driven web-based research infrastructure, as well as present and future application scenarios that can be envisaged from experiences with the piloted AGINFRA PLUS user communities and provides recommendations that aim to inform the roadmap for developing AGINFRA PLUS infrastructure further.

The initial version of this white paper has been prepared with contributions from all partners and was submitted to the EC on M25. In order to further enhance the positioning of the project in the digital science ecosystem, Agroknow worked on a next version of this deliverable that has been organised as an edited volume. A variety of stakeholders, including all project partners, were invited to contribute to this volume titled “Digital Science Recommendations for Food & Agriculture”. External contributors included strategic digital infrastructure initiatives (such as OpenAIRE, the FNH-RI Research Infrastructure for Food, Nutrition and Health, and the METROFOOD Research Infrastructure for promoting metrology in food and nutrition), as well as international stakeholders (such as the University of Guelph, Canada; and the Chinese Academy of Agricultural Sciences). The full text of the volume can be found in Annex A of the present document.

TABLE OF CONTENTS

1 INTRODUCTION

AGINFRA PLUS 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. The challenges that these communities have to face require our immediate contribution to be resolved. To this end, the project develops and provides 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 PLUS implemented 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.

This document demonstrates all the needs that each of the three user communities had and is expected to have in the future, as well as the challenges they face. Particular mention is made of how the cloud-based infrastructure(s) have managed to solve some of these challenges. As the challenges remain, some examples of them are listed as well as their possible solution by using the cloud-based infrastructure(s).

2 AGINFRA PLUS VISION

AGINFRA PLUS 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 uses, adapts and evolves 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.

AGINFRA PLUS particularly aims to demonstrate how core e-infrastructure services and resources may be used to support future science scenarios in agriculture and food. In this sense, AGINFRA PLUS aspires to be part of a wider strategy that the participating key stakeholders have agreed upon. In this context, it works in full complementarity to the INFRASUPP-3-2016 project "*e-ROSA: Towards an e-infrastructure Roadmap for Open Science in Agriculture*". e-ROSA formulated the context in which the various scientific and infrastructure stakeholders in agri-food will come together in order to work together on a roadmap for the e-infrastructures of the next 10 years, in sync with the developments at a broader scale. As a next step, AGINFRA PLUS constitutes an ideal testbed for assessing the viability, effectiveness and sustainability of the e-ROSA set principles and respective roadmap.

This project builds upon the extensive experience and work of its partners, who are key stakeholders in the e-infrastructures ecosystem. It also implements part of a strategic vision shared between Agroknow, the National Agronomic Research Institute of France (INRA), Wageningen Environmental Research (ALTERRA), the German Federal Institute for Risk Assessment (BfR), and the Food and Agriculture Organization (FAO) of the United Nations - the latter one, not participating as a funded beneficiary, but supporting the project and its activities. These stakeholders are part of a core group of internationally recognised players (including the Chinese Academy of Agricultural Sciences) aiming to put in place a global data infrastructure for research and innovation in agriculture, food and environmental science. This data infrastructure will become an incubator of the large infrastructure investments that global donors (including the European Commission) make in the field of agricultural research around the world.

In accordance with the main pillars for an e-infrastructure for open science in agriculture and food defined by e-ROSA project[1], AGINFRA PLUS focuses on the integration and harmonisation of multiple data assets (publications, datasets, models, etc.) under a standardised semantically rich framework that will allow the discovery and reuse of assets and results from researchers within a community, researchers of adjacent communities and to an extend citizen scientists. Towards this, AGINFRA PLUS evolves and enriches resources and services available as open-source software and/or results of previously successful EU-funded projects and initiatives.

New work will be steered towards the evolution of existing resources and services in order to seamlessly integrate with existing open e-infrastructure resources such as OpenAIRE, EUDAT, EGI.eu and D4Science. It tries to demonstrate how scientific communities working on agriculture and food topics may carry out rapid and intuitive development and deployment of innovative applications and workflows, powered by

[1] https://zenodo.org/record/1479659#.XDyuP81S-Ul

open e-infrastructures. It also aims to therefore strengthen and illustrate the value and potential of AGINFRA PLUS as a virtual research environment for the domain of agriculture and food.

In this context, the AGINFRA PLUS project is exploiting the Virtual Research Environments (VREs) paradigm for the three (3) prominent research communities. VREs are a prominent existing cloud-based solution provided by the D4Science Initiative (see Figure 1). VREs are web-based, community-oriented, collaborative, user-friendly, open-science-compliant working environments for scientists and practitioners working together on a research task. These research communities are (a) the Agro-climatic and economic modelling research community (b) The Food safety risk assessment research community and (c) the Food security research community.

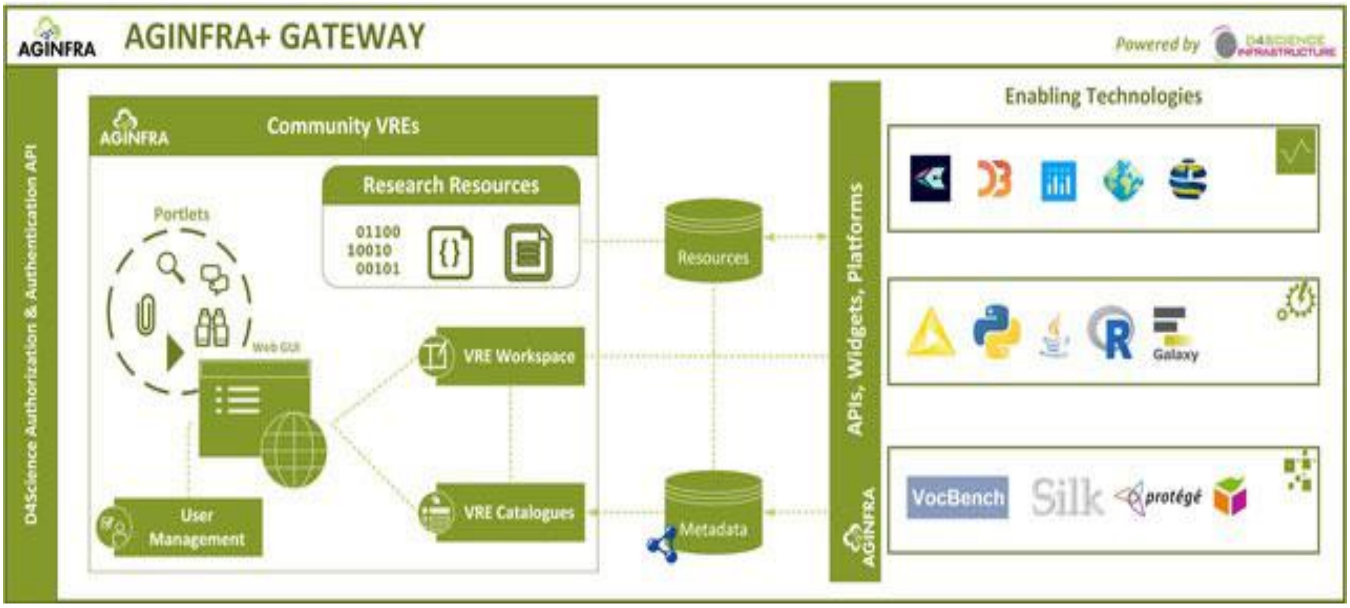


Figure 1: AGINFRA PLUS Technical Infrastructure based on research community VREs

In order to support the aforementioned research communities, the AGINFRA PLUS project has collected requirements from these communities and appropriate VREs have been set-up. These VREs encapsulate the technical solutions serving the community requirements within a collaborative environment that allows the setup, execution, monitoring and sharing of research activities and their results. More specifically, as presented in Figure 1, the VREs provide researchers access to research resources such as data, models data analysis pipelines, data analysis tools and publications, in order to design and execute their research. The resources are findable via the VREs' Catalogues, which are informed by semantically rich metadata organized and generated using the AGINFRA PLUS Data & Semantics Layer technologies. Experiments are carried out via the execution of the available models and algorithms over the services provided by the Analytics & Processing Layer. Finally, results are visualized, organized and shared using the technologies incorporated in the AGINFRA PLUS Visualization & Publishing Layer.

3 AGINFRA PLUS USE CASES

3.1 AGROCLIMATIC AND ECONOMIC MODELLING COMMUNITY

3.1.1 Community overview and their needs

The Agro-climatic and Economic Modelling Community uses advanced scientific models and algorithms to, among others, assess the short- and long-term impact of climate variations and future climate change on agricultural production, practices and markets. Many of the used models have been developed over decades, and their architecture does not always fit the requirements of current technologies and infrastructure. Also, many scientists in the community are most familiar with working in smaller communities of peers, based on longer-term trusted relationships. Their ways of working are not always open, which is partly cultural and partly because of the rather localized work processes and technical infrastructures that have been developed over the years. On the other hand, the character of scientific research these days changes dramatically. The interdisciplinarity of research in the nexus of grand societal challenges has grown substantially, and there is a drastic increase of the size and amounts, velocity and heterogeneity of the data being processed and analysed. Thus, it is inevitable that the community adapts its way of working and moves towards practices of better (re)use and sharing of resources to be better equipped for these developments.

To respond to these challenges, the community use cases (crop modelling, and crop phenology estimation) that were selected by AGINFRA PLUS specifically focus on opportunities to bring researchers from their current usually local, single computer and mostly peer network-based work space to a scalable computing and cloud based collaborative work environment. Providing the community with such advanced options for virtual research will increase the buy-in to use such frameworks and support the community in making the transition to effective collaborative, cloud-based research. The use cases are relevant for a range of different stakeholders' groups in this research domain: researchers, intermediaries and business analysts working on crop modelling and yield forecasting and related activities in the area of policy and decision support in food security, farm management advice and related activities.

3.1.2 Community challenges

The agro-climatic and agro-economic modelling community aims to assess grand challenges like food security, food safety and climate change impacts in an integrated manner. 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, eco-physiological, and meteorological data. The community working on the implementation of such use cases is a diverse network of agricultural, climate and economic researchers, practitioners and service providers in the science, policy and business domains.

The following are the main challenges for the agro-climatic and agro-economic modelling community:

- The amounts and velocity of data in the community's domain are increasing day by day. Examples are the increasing amounts of weather and climate data and remote sensing data that are generated with increasingly higher frequencies and resolutions. Scientists will need to find ways to adapt their ways of working and scaling up their infrastructure, to be able to cope with these developments;

- The growing interdisciplinarity of research will require knowledge, technology and skills to be able to effectively combine data and information from an increasing range of domains. In the nexus of agri-food research there are for instance clear linkages to the domains of water management, energy, ecology and ecosystems. Besides being able to scientifically and technically cope with this heterogeneous data, it will force different communities to work with larger and less familiar research groups and to share data and information in a trusted manner;
- Consequently, there's an urgent need for the community to advance their data science using high-performance, cloud-based environments. This will, however, require drastic technical and cultural changes. Many (legacy) models and algorithms will need to be fit into new infrastructure, and the community should shift to collaborating in less protected environments that share a culture of sharing and reuse.

3.1.3 Present science scenarios

1st Science scenario: Crop modelling

Typical users of this science scenario are:

- Researchers in various scientific domains (e.g. agronomy, agro-economy, climate change) that use agro-climatic modelling as part of their research;
- Information intermediary & service providers (e.g. extension services, ICT service providers) that either use crop models or the output of crop models as a resource to provide added value services for farm advisory and farm management support.

Context and Challenges

The first science scenario focuses on the work of an agronomic modeller in a scientific or commercial environment. Important elements for two different applications within the scenario have been deployed and tested on the VRE. They include discovery and download of (raw) datasets, pre-processing of datasets (harmonization, integration), running a crop growth model and analysis and visualization of model outputs.

A. Regional yield forecasting

Finding correlations between historical yield (statistics) and indicators derived from remote sensing and climate data time series, using the strongest correlations for yield forecasting. An example and description of the methodology can be found in this article: http://www.sciencedirect.com/science/article/pii/S0168192315000702. It can be extended by looking at other indicators, e.g. precipitation and temperature sums, and find correlations with the NDVI time series (or the fitted growth curves).

This application focuses on the generation, analysis and visualization of modelled regional crop growth indicators that can be used to perform crop yield statistics and to predict regional and local crop yields for short term seasonal yield predictions on future projections of yields under climate change.

The challenges in this application are:

- Collecting, pre-processing and organizing model input datasets;
- Refactoring and deploying an analytical crop Simulation model for execution on a VRE;
- Developing and executing algorithms to derive advanced indicators from raw model output data;
- Developing and performing statistics and analytics on large amounts of historical and model data time series;
- Visualisation of spatio-temporal model outputs in different forms (e.g. time-series, maps, combined spatio-temporal visualisations);
- Linking model and statistical output to NDVI time series or growth curves (thus, linkage to the 2nd scenario)

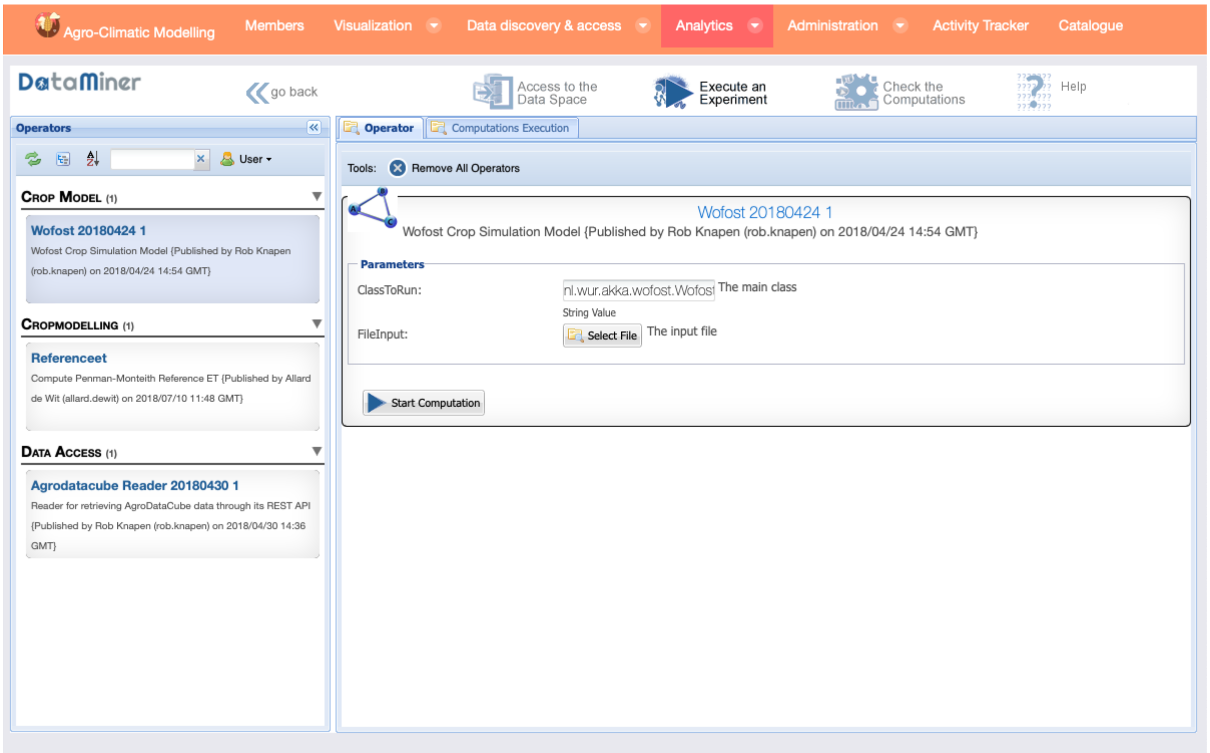


Figure 2: Example of running a crop simulation model in the VRE

<u>B. Large scale regional crop model simulations</u>

Most crop models are point based simulation models. With the proper input data (crop type, soil, climate, field management) available, the model can be run in parallel for many regional locations, several crops, and input variations, and results can be combined to an integrated dataset. This can only be done in a limited way using traditional hardware. Cloud solutions, e.g. on Amazon Cloud Services are an option, but issues of costs, lack of skills, and lack of trust prevent many research groups to start using such infrastructures. This application develops the software and data architecture to be able to run a crop model (WOFOST) in parallel on the D4Science community VRE.

The challenges in this application are partly similar to the one's stated above for the regional yield forecasting application. The following additional challenges exist:

- Setting up a data infrastructure that can efficiently feed the parallel modelling process
- Data wrangling and combining data to feed (parallel) model runs.
- Setting up the software infrastructure for efficient parallelization of model runs, including its data feeds and output streams

Solution

To implement the applications and address the associated challenges, a full crop modelling workflow was developed and deployed on a virtual research environment which had to be able to be performed expeditiously and operationally.

The following are functional and technical features have been realized and are parts of the pilot VRE to support this science scenario:

- Refactoring of a version of the WOFOST (World Food Studies) crop model suitable to run on the VRE. From several implementations of WOFOST, the recently developed Java implementation of this model was most suitable;
- Deployment of modelling workflows on the VRE so that less experienced users can operate crop model configurations;
- Performance – the option to access high performance hardware through the VRE so that crop model simulation runs can be executed quicker compared to regular workstations and laptops used by researchers;
- Scalability – the option to distribute crop modelling runs effortlessly over a variable number of computing nodes through the VRE, e.g. running model simulations for a certain crop for all relevant parcels in the Netherlands for multiple years in parallel and combining the results;
- Findability and accessibility of required (raw) input data through the VRE. This includes access to data from the AgroDataCube (a PostGIS spatial/relational database), and sets of (crop) parameter files (currently in YAML format);
- Development of data integration functions to process the mentioned (raw) input data to usable data formats for the WOFOST crop model;
- Data analytics and data visualization options (both for input and output data), particularly focused on handling and analyzing spatial datasets, e.g. display simulated yields per crop per parcel as a geographic map;
- FAIR publication of models and analytic components and generated output files through the VRE for reuse by third parties and for data analysis.

2nd Science scenario: Crop phenology estimation

These days, the state of crops and over time can be monitored through remote sensing. Remote sensing-based indicators, e.g. Normalized Difference Vegetation Index (NDVI) or Weighted Difference Vegetation Index (WDVI), can provide a good indicator for the development of a crop at a certain point in time. However, the frequency of satellite images becoming available is quite low, and depending on the location, cloud coverage can decrease the amount of usable measurements over time drastically. Statistical algorithms that allow to derive reliable crop phenology curves from a set of scarce and irregular measurements over time, possibly using additional auxiliary data from other sources, can be a great support in many applications, e.g. for crop yield forecasting and monitoring of land use and agricultural practices.

Typical users of this science scenario can be:

- Researchers that perform explorative modelling to develop and test crop phenology estimation or similar models and algorithms;

- Data analysts that aim at combining various data sources like satellite data, statistics etc. to generate policy advice and decision support;
- Intermediaries that either use crop models or the output of crop models as a resource to provide added value services for farm advisory and farm management support.

Context and Challenges

This scenario is relevant to modellers and data scientists in the science, policy as well as the business domain. Understanding crop development and being able to do short term predictions based on reliable information serves many objectives. Scientists can use VRE solutions in this scenario to experiment with different algorithms and statistics to improve methodologies. Policy analysts can use it to monitor agricultural land use, e.g. to control eligibility of provided (CAP) subsidies. Business can exploit the knowledge to inform agricultural insurance and financing policies.

The scenario focuses on data science and data analytics methods to cope with data gaps and uncertainty in remote sensing time series. It includes data wrangling and data integration, explorative design and deployment of statistical algorithms, visualization and FAIR publishing of results. The emphasis of the scenario is on explorative and collaborative modelling, allowing to work in an agile manner and experiment with different datasets and algorithms using commonly used open environments (e.g. Python/Jupyter, R/RStudio) on a VRE.

There are two applications within this scenario that are feasible for VRE deployment and execution:

A. Estimating growth curves

Remote sensing imagery, e.g. from satellites or drones, can be used to calculate variables such as the Normalized Difference Vegetation Index (NDVI). Typically, this gives an irregular time series of weekly values. By fitting a curve (e.g. a double sigmoid curve) through these data points an estimation can be made of crop phenology and hence expected yields. However, such curve fitting is not trivial and can be time consuming, e.g. http://www2.geog.ucl.ac.uk/~plewis/geogg124/phenology.html gives an example of the process. Using VRE compute capabilities calculating NDVI for large areas can be done in shorter time, and perhaps refined with machine learning algorithms that take more data into account.

Figure 3: Work on NDVI data analysis and curve fitting for crop phenology estimations in the VRE

The challenges in this application are:

- Providing easy access for data scientists to remote sensing data streams, tackling the volume and velocity issues and using on-line services, e.g. provided by Copernicus and added value services;
- To access and merge data from different sources (local and remote) that are provided through different technologies using collaborative modelling environments;
- To visualize spatio-temporal data in several ways (e.g. time-series, maps, combined spatio-temporal visualisations);
- To publish output data and algorithms in a FAIR way, allowing communities to reuse.

B. Crop yield risk assessment - Detrending models

Detrending is a widely used technique for obtaining stationary time series data in residual analysis and risk assessment. The technique is frequently applied in crop yield risk assessment and insurance ratings, e.g. see http://link.springer.com/article/10.1007/s00477-014-0871-x. It allows to separate the (non-insured) yield deviations caused by long-term technological developments and environmental impact from the (insured) residuals, caused by abnormal weather conditions and disasters. Since this is a very common practice for researchers and analysts it would be good if the algorithms required are supported by the VRE.

Additional to the challenges for crop growth curve estimation, the following challenges need to be tackled:

- Developing, deploying and testing statistical methods for detrending, including the integration of various existing algorithms and scripts that compose such detrending methods;
- Protecting the part of data and algorithms that is sensible and/or falls under IPR. This is probably a prerequisite for many commercial organizations, and in specific cases possibly also policy makers and scientific organizations.

Solution

To address the challenges, a VRE had to be prepared to support explorative modelling, specifically to develop and test data science algorithms to determine crop phenology development and associated agronomic statistics. It also included the deployment of remote sensing data streams (from the AgroDataCube for the Netherlands and from Copernicus for Europe/world wide applications) on the VRE, making its data accessible and reusable for such explorative modelling exercises. The following functional and technical features have been implemented as part of the pilot VRE to support this science scenario:

- Using the VRE for the explorative development and testing of algorithms (e.g. using a Notebook environment such as Jupyter) for assessing crop phenology development. It includes integrating data access, data processing, statistics and visualization into documented workflows;
- Making the content of the AgroDataCube (a PostGIS spatial/relational database, containing agro-climatic data for the Netherlands) available (findable and accessible) on the VRE for this specific application and for further explorative research;
- Making on-line Sentinel data streams available on the VRE for this specific application and for further explorative research;
- Using the pilot VRE to define workflows for assessing larger scale crop phenology development based on data coming from remote sensing data streams, e.g. for different crops and all agricultural parcels in the Netherlands (through the AgroDataCube) or elsewhere in Europe (through Sentinel data);
- Using the pilot VRE to train machine-learning algorithms using crop phenology development and other input data, and subsequently use the trained model to predict yield per crop per parcel, based on forecasted weather;
- Performing data analytics and visualization for e.g. (spatio-temporal) input data, the crop phenology, derived crop growth curves, yield forecasting. Preferably in an easy to use and customizable dashboard;
- Performance improvements – the option to access high performance hardware through the VRE so that algorithms can be executed quicker compared to regular workstations and laptops used by researchers;
- Scalability improvements – the option to distribute algorithm calculations effortlessly over a variable number of computing nodes through the VRE. E.g., running machine-learning algorithms for yield prediction for a certain crop for all relevant parcels in the Netherlands in parallel and combining the results.

3rd Science scenario: AgroDataCube

Running agro-climatic models and deriving crop phenology data require range of agronomic and climatic data, including remote sensing data. The AgroDataCube is a Dutch initiative that seeks to open up agronomic data for data science and service development. Currently it contains among other services to access agricultural parcel geometries, weather data, crop data, soil data and satellite derived crop development indicators. As these services collectively provide most of the required input data for the previously described scenarios, the AgroDataCube was added as a third, auxiliary scenario, supporting the applications to be developed under the other scenarios.

Typical users of this science scenario are:

- The user groups mentioned in the previously described scenarios, using the AgroDataCube as datasource;
- Analysts that aim at combining various data sources like satellite data, statistics etc. to generate policy advice and decision support;

- Intermediaries that use the data services that the AgroDataCube offers as a resource to provide added value services for farm advisory and farm management support.

Context and Challenges

The AgroDataCube is a valuable data source for developing a broad range of services related to (Dutch) agriculture. The AgroDataCube collects and integrates agronomic data and provides services to retrieve it. Semantics and metadata (including provenance) plays an important role (the European Open Science Cloud initiative increases the focus on data stewardship). Much of the agro-climatic data is spatio-temporal in nature, and there has been a long history in the development of Geo Information Systems (GIS) and Spatial Data Infrastructures (SDI) for working with these complex and large datasets. However, the standards developed within Geo-Information Science are academic and heavy based on closed world assumptions and need to be bridged to lighter community web and semantics standards.

Most of the challenges related to the AgroDataCube concern the processing of the data. These are in fact challenges of the specific application that uses the data. The following challenges relate directly to the AgriDataCube:

- Scalability: the AgroDataCube, as an RDMBS based data service, does not support the most advanced technologies to scale up. Massive access to the AgroDataCube could result in performance loss or even failure of the service;
- Interoperability: most of the data provided through the AgroDataCube is spatial or spatio-temporal data. Provision in standardized formats, that can be processed by geo-analytical algorithms is key to ensure reuse;
- Findability: Retrieving (spatial) data often includes selecting only a specific part of the data, the region of interest, to avoid downloading large datasets, and re-projecting the data so that in the end all datasets share a common spatial projection. Preferably such subselections and reprojections should be supported.

Solution

To deal the aforementioned challenges, the use of a resource like the AgroDataCube from a VRE had to be enabled and to make its data findable, accessible, interoperable and reusable (FAIR). The scenario was integrated in the previously described science scenarios on crop modelling and crop phenology estimation, as these use the AgroDataCube as a main data provider.

The following are functional and technical features that have been realized that are parts of the pilot VRE to support this science scenario:

- Making the content of the AgroDataCube (currently a PostGIS spatial/relational database, containing agro-climatic data for the Netherlands) available (findable and accessible) on the VRE for further processing;
- Using the VRE and its components to implement a range of pre-processing, data analytics, visualisation and similar jobs that support the other scenario applications and that use (among others) the AgroDataCube data services.

3.1.4 Future science scenarios

Context and Challenges

There is a multitude of future science scenarios for AGINFRA that relate to the work of the agro climatic and economic modelling community. First of all, there is the challenge of extending the currently implemented scenarios. Several relevant crop models exist that are still not suited to run in collaborative, cloud-based environments. Also, many alternative and useful data sources and streams exist that could be valuable resources to extend and improve the scenarios.

With regard to new scenarios, it is expected that many of them will have an emphasis on improved interdisciplinarity. Tackling the grand societal challenges, requires integration of data science and modelling components and workflows from different domains, like agriculture, environment, economics, energy. Examples are scenarios that link agronomic and economic models to include feedback loops to the economy and integrate the effects of overall socio-economic developments on agriculture and vice-versa. Also, links with land use models can be essential to cross-link demand and availability of resources for food security, energy security and biodiversity conservation.

It cannot be expected that there's one conceptual, architectural or technical solution that fits all of these scenarios. Some scenarios will require models and algorithms to be operationally coupled, while others might use pre-processed model outputs as an input, or re-use and adapt specific analytical components.

Recommendations

In the face of the range of scenarios sketched above, the following recommendations can be prioritized to support the further development of AGINFRA PLUS:

- To support highly interdisciplinary research, AGINFRA should be recognized as the agri-food hub for data science. It should be operationally connected to the wider network of European and global science hubs and allow its users to operate over these hubs to collaborate and share its data with communities in other domains, and vice versa (re)use data from other domains;
- Consequently, organising interdisciplinary communities and collaboration over domains would require that a collaboration model is adopted that allows research groups over different domains to create a common workspace;
- A well-thought out security model for authorisation and authentication, as well as options for licensing and (optional) protection of access to resources is required. This should create the required trust among users and communities that they are in control of their mechanisms for sharing and reuse, respecting IPR and ownership where needed;
- Interoperability will need to be improved in different ways. AGINFRA should promote data standards, not only based on the common practice in agri-food, but also considering standardisation in adjacent domains. Besides scientific algorithms, methods to convert or merge (data) standards could facilitate linking of domains;
- Consequently, the use of semantics, and in this case not the definition of semantics or the annotation of resources, but the actual use of semantics for (semi)automatic data linking and merging, should be improved;

- Meta-data and documentation of data, algorithms and models should be encouraged (or even enforced). It must be possible for researchers to understand the backgrounds of available resources and trace back how they have been derived. This will allow them to decide if resources are suited to be reused, without having to get in direct contact with the creator.

3.2 FOOD SAFETY RISK ASSESSMENT COMMUNITY

3.2.1 Community overview and their needs

For the domain of food safety modelling, the general objective is to support researchers, food business operators and governmental agencies in the multidisciplinary field of food safety risk assessment and emerging risk identification. There is an increasing demand for software-based solutions that can support decision making in these domains. In order to develop such services and tools the scientific knowledge (e.g. data, mathematical models, data analysis pipelines) generated in this domain needs to be integrated and exchanged in a more efficient way, as it is currently done via classical research papers. Domain-specific, cloud-based research environments are a promising solution to overcome current limitations and frustrations in these domains. Specifically, features that facilitate scientific collaboration as well as features to store and share data and knowledge are important. Furthermore, the usage of shared infrastructural and technical solutions can facilitate the creation and adoption of standards which will increase efficiency along all knowledge generation and decision processes.

3.2.2 Community challenges

Food safety as a global challenge requires efficient knowledge transfer between academia, business operators and governmental agencies. In Europe, a rich variety of useful resources like predictive models, software tools and data repositories relevant for food safety risk assessment exists, but exchange and practical application of this information between different tools used by different stakeholders is currently difficult and time consuming, e.g. many tools used by researchers are far to complex for end users from governmental agencies or FBO. This is a significant challenge as it is widely accepted within the community that the application of mathematical models, new data mining technologies and simulation tools is vital to cope with the numerous existing and emerging food safety risks and challenges.

A specific area that becomes increasingly relevant for many stakeholders is the task of identifying so called "emerging risks" in the food (and feed) chain. This is an important governmental and business activity that is undertaken to protect the consumer with timely and effective preventive measures as increased global trade is making food chains more complex, both in terms of geographical spread and the rapid distribution of goods. Integration of this complexity into the risk assessment and risk identification work is however still a very new field of research that require a high degree of interdisciplinary scientific exchange and expertise. The provisioning of a cloud-based research environment to do fast prototyping and share best practice solutions (specifically those linked to data analysis and modelling tasks) is therefore a promising approach that could help all stakeholders interested in identifying emerging food safety issues at an early stage.

3.2.3 Present science scenarios

1st Science scenario: Determination and Metrics of Emerging Risk - Demeter

Typical users of this science scenario are:

- Risk Assessor;

- Data Scientist.

Context and Challenges

The early identification of emerging risks in the food (and feed) chain is an important governmental and business activity carried out to protect the consumer with timely and effective preventive measures. Identification, integration and analysis of information collected from public and non-public sources is considered essential to support decision making of public and private sector stakeholders. For this data and text mining solutions needs to be developed, applied and continuously improved to identify emerging food safety issues at an early stage. Currently there are only very few open-source data-science based solutions availables that could be used by interested stakeholders. A VRE-based infrastructure could therefore be a very good resource to share existing knowledge within the community and to jointly develop or improve community solutions.

Solution

With a view to resolve the aforementioned problems, a special VRE has been developed for the Food Safety Risk Assessment community namely, DEMETER.

The aim of this pilot VRE was to answer the question, if the new web-based resource for the Emerging Risk Identification community (DEMETER VRE) has the potential to serve as an Open Science resource in the future. To accomplish this, the generated resource had to go beyond an infrastructure for sharing text documents as it is the currently applied practice. Specifically, it has been assessed if the generated pilots (second and third version) could be used to share information, data and data-analysis pipelines developed by different stakeholders, e.g. members of the EREN network. Furthermore, it has been assessed if new opportunities to visualize results of calculations performed by domain-specific data mining and data analysis operations could be provided.

The main objective of the DEMETER community was the creation of an open, transparent, modular web-based system, which will support processes established by EFSA and other food safety authorities for emerging risks identification. Specifically, it provides resources that allow sharing and executing emerging risk data retrieval and data mining pipelines as well as sharing data and knowledge in a broader sense. The envisaged system has a user-management system in place and supports the integration of non-public (e.g. governmental or supplier audits) information in a privacy protecting manner. A strong focus has been laid on a service-based system architecture that provides the necessary flexibility for fast integration of new internal or external services (if desired by the community).

The following needs of the DEMETER community were identified and resolved by exploiting DEMETER VRE:

1) Data and Relevant Semantics Needs

    1. Easy access to open scientific literature and other free online information sources on the WWW;

2. Easy access to social media data (Twitter, Facebook) via dedicated KNIME data analysis pipelines;
3. Easy access to RSS feeds from community information providers: e.g. MediSys, via dedicated KNIME data analysis pipelines;
4. Access to ontology management services that could support automated knowledge generation and extraction in the future.

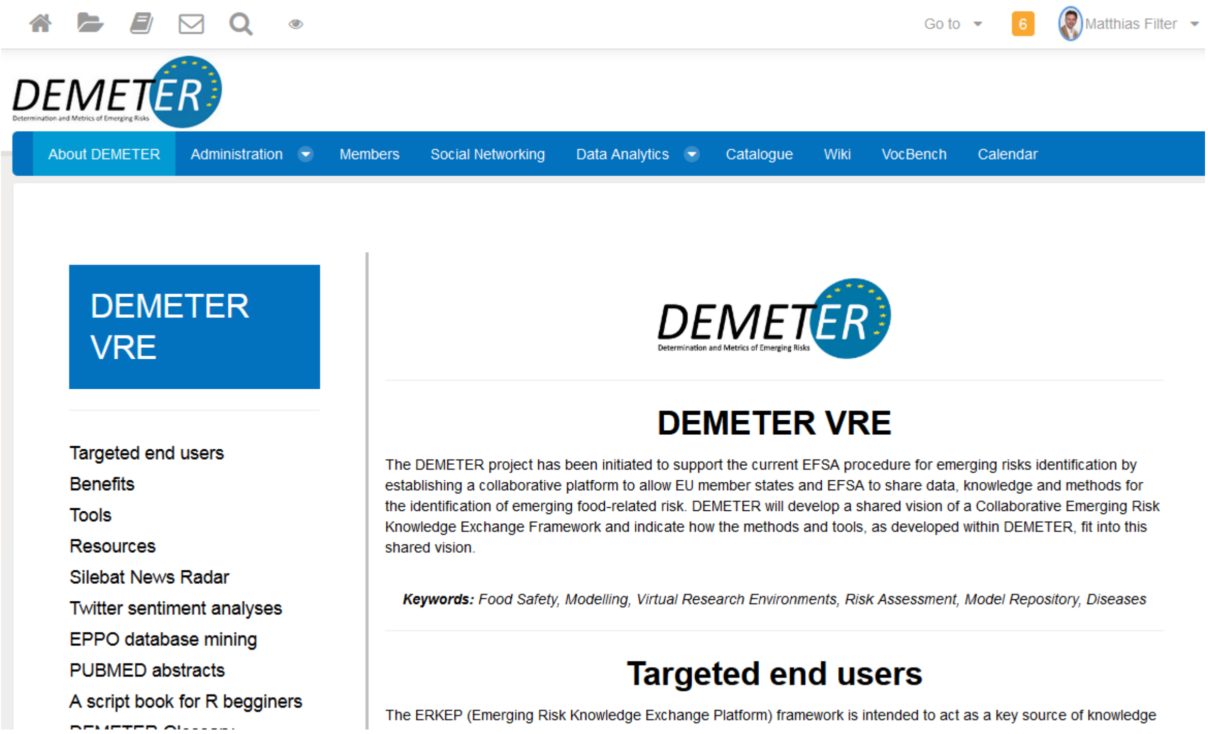


Figure 4: DEMETER VRE Main Page

2) <u>Data Analytics and Processing Needs</u>

1. KNIME workflow execution including support for R and Python extensions;
2. API access to integrated emerging risk identification services;
3. Exploitation of high-performance computing infrastructure (where necessary) to execution DEMETER data mining workflows;

3) <u>Data Visualization and Publishing Needs</u>

1. A service to streamline the publishing of models/ data mining workflows to the scientific community (Publish in the sense of “creating a citable scientific publication” - ideally with a DOI);

2. A public service to search/ filter for emerging risk identification models/ workflows in a model/ workflow catalogue

3. Interactive online data and knowledge visualisation feature via MindMaps

4) <u>Other Needs</u>

1. User management system, i.e a component that allows managing the rights and accounts of community members, e.g. passwords, permissions, etc.;
2. Data inventory/ workspace, i.e. a storage space for electronic files for each community members. This component also needs to provide options for each user to share or publish files with others;
3. Tracing of documents, i.e. a component that allows creating a document history or so-called provenance reports.

2nd Science scenario: Knowledge Integration Platform - RAKIP

Typical users of this science scenario are:

- Risk Modeller;
- Data Scientist;
- Risk Assessor;
- SME

Context and Challenges

Food safety as a global challenge requires efficient knowledge transfer between academia, business operators and governmental agencies. In Europe, a rich variety of useful models, software tools and databases for food safety risk assessment exists, but exchange of these kinds of information between different stakeholders is currently extremely difficult and time consuming. However, integration of mathematical models and modelling tools is vital to cope with the numerous existing and emerging food safety risk and challenges.

The RAKIP Initiative has been initiated by three European institutions specialized in food safety modelling and risk assessment (ANSES, BfR, DTU). These institutes collaborated in a joint effort on the establishment of a Risk Assessment Modelling and Knowledge Integration Platform (hereinafter referred to as RAKIP) where the term “knowledge” specifically refers to data and models relevant for risk assessment tasks. The development of a RAKIP portal aimed at improving transparency in the data- or model-based risk assessments work and should facilitate the exchange of “knowledge” between different software tools that are already available in each of the three institutions. Further it was a joint wish to develop this portal into an open, community-driven Food Safety Model Exchange Portal, that allows to upload, review, execute, search and download risk assessment models and modules also by other interested stakeholders. An implicit requirement for the development of this resource was the development of a harmonized information exchange format for risk assessment models (called Food Safety Knowledge Markup Language - FSK - ML).

Solution

The main objective of the pilot VRE generated for the RAKIP community was to answer the question, if the new web-based resource for the food safety risk assessment community can serve as a community risk assessment model repository. To accomplish this, the RAKIP VRE contains numerous features that go beyond what is currently state-of-the-art, as e.g. in portals like openFSMR (https://sites.google.com/site/openfsmr/) or EFSA’s Knowledge Junction (https://zenodo.org/communities/efsa-kj) or pure listing of QMRA publications, as e.g. in www.foodrisk.org. Specifically, the RAKIP VRE uses the Data Catalogue technology to make risk assessment models FAIR. The VRE further supports the newly developed harmonized information exchange format FSK-ML. It is possible to share data, data-analysis pipelines and to perform user-driven computational-intensive simulations in the VRE. Another important VRE achievement is the support for the open source data analytics platform KNIME through the VRE DataMiner technology.

The following specific needs of the RAKIP community were identified and resolved in the RAKIP VRE:

1) Data and Relevant Semantics Needs

   1. A service to develop and maintain controlled vocabularies / ontologies – see 2.1.4;
   2. Online resource to store/ upload and create new risk assessment models in the FSK-ML format (i.e. as an FSKX file). The FSK-ML format specification is under constant development by members of the RAKIP community. Key feature of FSK-ML is, that this standard allows sharing script-based code that can be executed in the appropriate environment, e.g. KNIME;

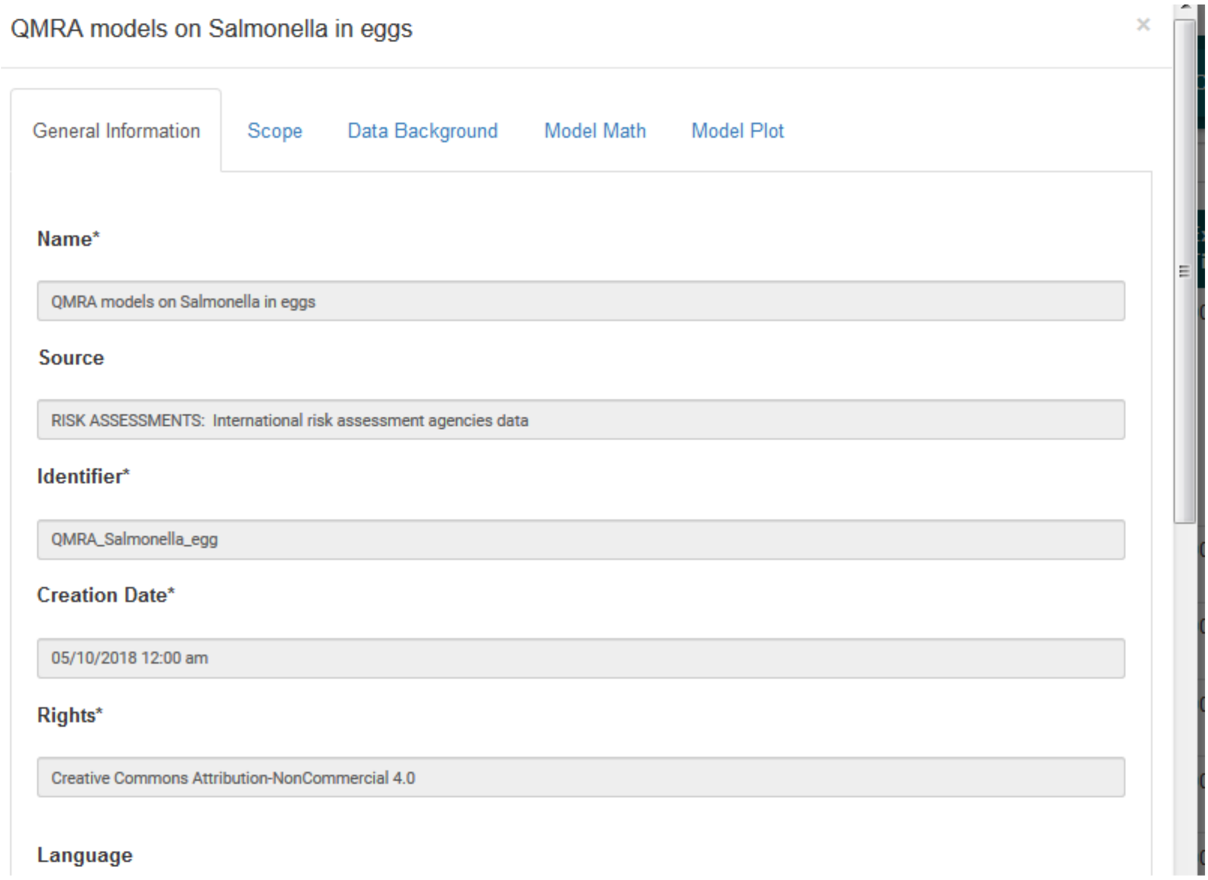


Figure 5: Example of QMRA model

   3. A service that allow the user to visualize metadata or reconfigure FSKX model files from the model repository/ workspace before execution/ simulation.

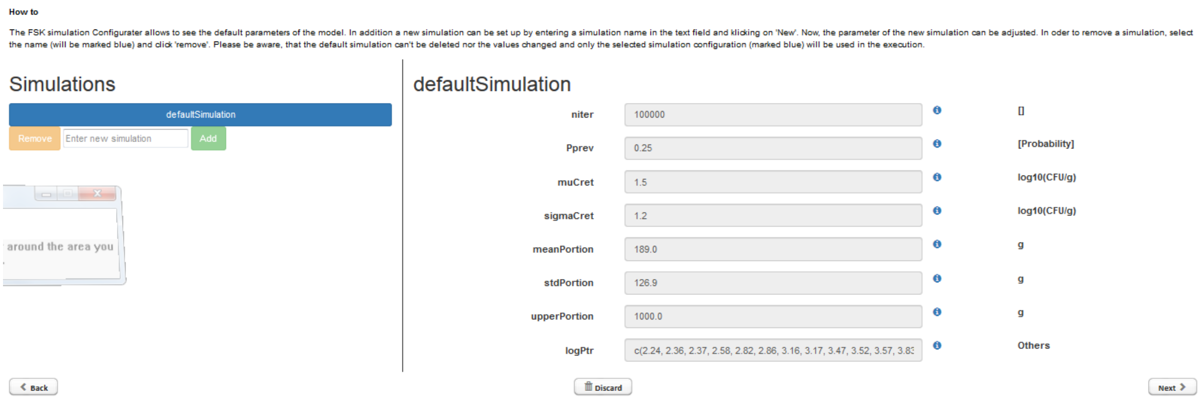


Figure 6: Example of FSK simulation

2) Data Analytics and Processing Needs

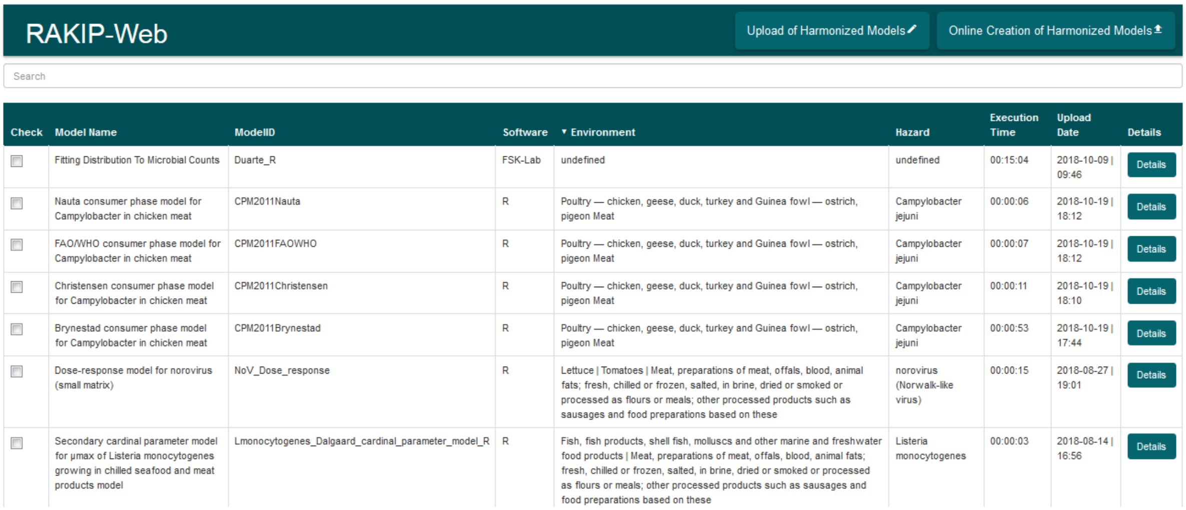
1. Need for KNIME workflow execution inside the VRE. These KNIME workflows will also be able to execute R based models (will be extended to other scripting languages in the future);
2. A service for model execution (simulation);

Figure 7: Example of Fskx Model Runner

3. A service to use high-performance computing infrastructure in case computational expensive simulations need to be performed. The created standardized provenance reports on the simulation process add additional value;
4. API access to all model simulation services;
5. A KNIME-based service that allows to publish data / models directly into EFSA's Knowledge Junction;
6. Creation of URIs for each object published in the VRE DataCatalogue .

3) Data Visualization and Publishing Needs

1. A service that read a FSK-ML formatted file with information on the QMRA model input and output parameters. The visualization service allows users to modify input parameters (in a range defined by the model metadata) and get the simulation results represented in appropriate chart types instantaneously;
2. A service to combine model modules into new models;
3. Interactive and user-friendly GUI of the model repository (including search, browse, sort, filtering functions).



Check	Model Name	ModelID	Software	Environment	Hazard	Execution Time	Upload Date	Details
	Fitting Distribution To Microbial Counts	Duarte_R	FSK-Lab	undefined	undefined	00:15:04	2018-10-09 | 09:46	Details
	Nauta consumer phase model for Campylobacter in chicken meat	CPM2011Nauta	R	Poultry — chicken, geese, duck, turkey and Guinea fowl — ostrich, pigeon Meat	Campylobacter jejuni	00:00:06	2018-10-19 | 18:12	Details
	FAO/WHO consumer phase model for Campylobacter in chicken meat	CPM2011FAOWHO	R	Poultry — chicken, geese, duck, turkey and Guinea fowl — ostrich, pigeon Meat	Campylobacter jejuni	00:00:07	2018-10-19 | 18:12	Details
	Christensen consumer phase model for Campylobacter in chicken meat	CPM2011Christensen	R	Poultry — chicken, geese, duck, turkey and Guinea fowl — ostrich, pigeon Meat	Campylobacter jejuni	00:00:11	2018-10-19 | 18:10	Details
	Brynestad consumer phase model for Campylobacter in chicken meat	CPM2011Brynestad	R	Poultry — chicken, geese, duck, turkey and Guinea fowl — ostrich, pigeon Meat	Campylobacter jejuni	00:00:53	2018-10-19 | 17:44	Details
	Dose-response model for norovirus (small matrix)	NoV_Dose_response	R	Lettuce | Tomatoes | Meat, preparations of meat, offals, blood, animal fats; fresh, chilled or frozen, salted, in brine, dried or smoked or processed as flours or meals; other processed products such as sausages and food preparations based on these	norovirus (Norwalk-like virus)	00:00:15	2018-08-27 | 19:01	Details
	Secondary cardinal parameter model for μmax of Listeria monocytogenes growing in chilled seafood and meat products model	Lmonocytogenes_Dalgaard_cardinal_parameter_model_R	R	Fish, fish products, shell fish, molluscs and other marine and freshwater food products | Meat, preparations of meat, offals, blood, animal fats; fresh, chilled or frozen, salted, in brine, dried or smoked or processed as flours or meals; other processed products such as sausages and food preparations based on these	Listeria monocytogenes	00:00:03	2018-08-14 | 16:56	Details

Figure 8: Example of RAKIP Data Visualization

4) Other Needs

In addition to the list provided for the Other Needs of the DEMETER community, it follows one more need:

1. Data management policy is important: data storage and calculations should not be done on US server.

3.2.4 Future science scenarios

Science scenario: Foodborne Disease Outbreak VRE

Typical users of this science scenario can be:

- Risk managers;
- Risk assessors;
- Epidemiologists;
- Public Health officials;
- (Food Business Operators).

Context and Problems

This VRE would be designed to support foodborne disease outbreak investigations. Such VRE could be instantiated on demand for specific outbreak situations and could immediately provide the required protected web-based platform for information sharing and joined information analysis. At its core this VRE would serve the need for efficient communication between all stakeholders that need to be involved in a given disease outbreak investigation. Another unique selling point of such a VRE-based platform is its capability to provide shared data integration and data analysis functionalities. These kinds of services are nowadays prerequisites for efficient outbreak investigations. For example, such services could provide guidance for targeted and risk-based sampling strategy of suspect food products or provide interactive, online visualization of supply chain information. The VRE could even provide a service to food business operators that allow them to compare anonymously their sales data from their own products against the spatial distribution pattern of an ongoing disease outbreak. The result of such a service could be an important support for the decision if a product should be recalled or withdrawn from the market. In addition, the VRE could also serve as an outbreak specific information source for the general public by using the public VRE "About page".

Recommendations

Based on the experiences with the establishment of VREs for specific communities within the area of food safety it has to be acknowledged, that despite of the service-based architecture there is a significant development effort needed for community-specific customizations. In addition, it has to be acknowledge that nowadays many community members are reluctant to register to new web-based software solutions. Therefore WP6 has the following recommendations towards the VRE service providers:

- In order to attract members to the VREs more resources have to be allocated into general VRE issues like usability, look-and-feel, documentation
- It is strongly recommended to integrate features that have the potential to immediately attract many users in each VRE, as e.g. an integrated web meeting service
- As the customization of scientific computation or visualization services to the needs of a new VRE community is a highly interdisciplinary and demanding development task there is the need for dedicated service providers that are specialized in these tasks, as only VRE experts can have the necessary IT background knowledge to accomplish these development tasks.
- Along these lines - these VRE experts should also develop an understanding on features that were developed based on specific community requests, as e.g. the new integration of KNIME or Galaxy. A in depth knowledge on the potential and application range of such 3rd party technologies would open up the opportunity to create synergies between different communities and give good advices.
- A well-thought out security model for authorization and authentication, as well as options for licensing and (optional) protection of access to resources is required. This should create the required trust among users and communities that they are in control of their mechanisms for sharing and reuse, respecting IPR and ownership where needed.
- Interoperability between the different existing VRE services should be continued to be improved (this is a never-ending task).

3.3 FOOD SECURITY COMMUNITY

3.3.1 Community overview and their needs

The principal objective of the Food Security Community is to select plant species and varieties which are the most adapted to specific environments and to global changes. A way to do that is using High-throughput phenotyping, which is the AGINFRA PLUS Use Case selected as representative for the Food Security Community. High-throughput phenotyping is a good example of Big Data in agriculture because it produces a large amount of data which need to be analysed immediately for decision making. The main goal is to provide a collaborative work environment for phenomics researchers and to assess the effectiveness of this environment to meet the community needs. The user through the use of the collaborative work environment should be able to:

- Discover and access plant datasets at different scales (gene, cell, plant, canopy, crop), at different steps (phenology, food processing), environmental datasets (soil, water, life cycle and sustainability), nutrition and biomass production datasets etc;
- Combine, integrate and (pre-) process these variable and huge datasets;
- Access and edit several ontologies (crop ontology, plant ontology, etc) expressed in different formalisms;
- Access to data produced by phenomics platforms within a web environment;
- Contribute data by uploading the relevant assets (data files, metadata);

- Visualize data (with different plots);
- Import and run data analytics scripts in different languages;
- Import or update and run data analytics workflows;
- Share his results and work with other users.

3.3.2 Community challenges

Plant derived products are at the center of grand challenges posed by increasing requirements for food and feed. 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. Several communities are directly concerned such as breeders, plant science researchers, geneticists, data managers, data scientists, etc. Different platforms produce complex data at different scales and these data require various skills and high-performance computing in order to be managed and valorized.

3.3.3 Present science scenarios

Science scenario: Meta-analysis of phenotyping data

Typical users of this science scenario are:

- Plant phenotyping researchers;
- Agronomists;
- Statisticians from phenotyping community.

Context and Problems

The principal objective of this science scenario is to characterize, in a background of global changes, an environment and see in return which species and varieties are likely to adapt to this specific environment.

To that aim, the user should be able to use a cluster compute - cloud based (VRE-like) collaborative work environment to import, build, and update workflows to:

- Search for, and access, environmental and experimental data which might be large. The main data sources are the different instances of the EMPHASIS information system (currently under development) which will be accessible through web services;
- Choose the variables that characterize a given environment;
- Identify the varieties or species that can be adapted to the characterized environment;
- Share the results with decision makers through a user-friendly interface so they can decide which species or varieties they want to put in a given environment.

According to the Food security use case, the Big Data opportunities should be leveraged in order to sustainably maximize crop performance. That requires determination of which plant species and which varieties are most adapted to climate changes and natural resource preservation. High throughput phenotyping is at the heart of these challenges and produces huge sets of complex data.

The main challenge is to design and implement scientific workflows on massive amounts of complex data produced at different scales (population, plant, organ, cells, etc.), different stages (sowing, phenology, harvest) in interaction with various environment components (soil, climate, agricultural practices, biodiversity, etc.). In order to achieve this aim, we must supply analysis combining various heterogeneous data sources and run scientific workflows on a set of datasets, potentially huge and potentially from different data sources. The given workflow may include some steps of data combination and wrangling.

Solution

The solution to these problems was based on the creation of the Food Security VRE. Through the use of the VRE access control to the data sources and the shared objects (workflows, models and results of workflows) was provided. The following are functional and technical features that are part of the pilot VRE for this science scenario:

- Data access: access and integration of data from various data sources of phenomics platforms with a focus on semantic issues;

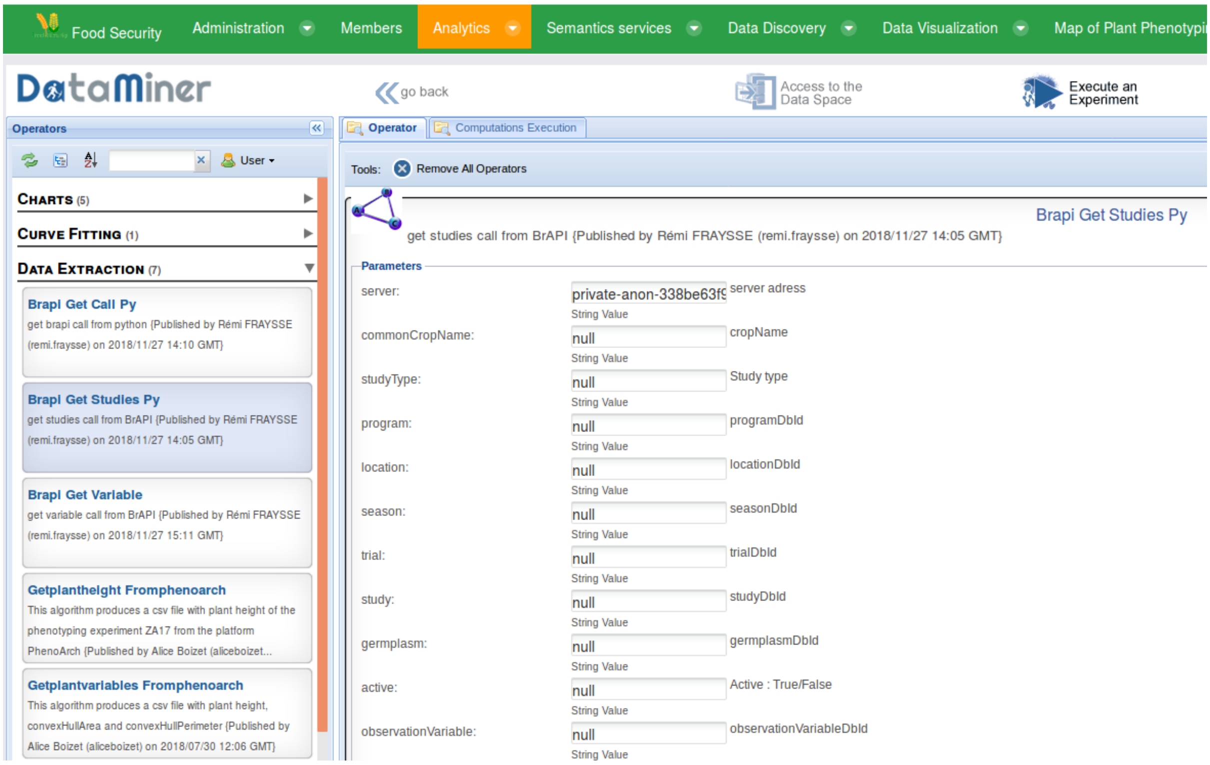


Figure 9: Example of Data Analytics process

- Data exploration and visualization: to provide interactive visualization;

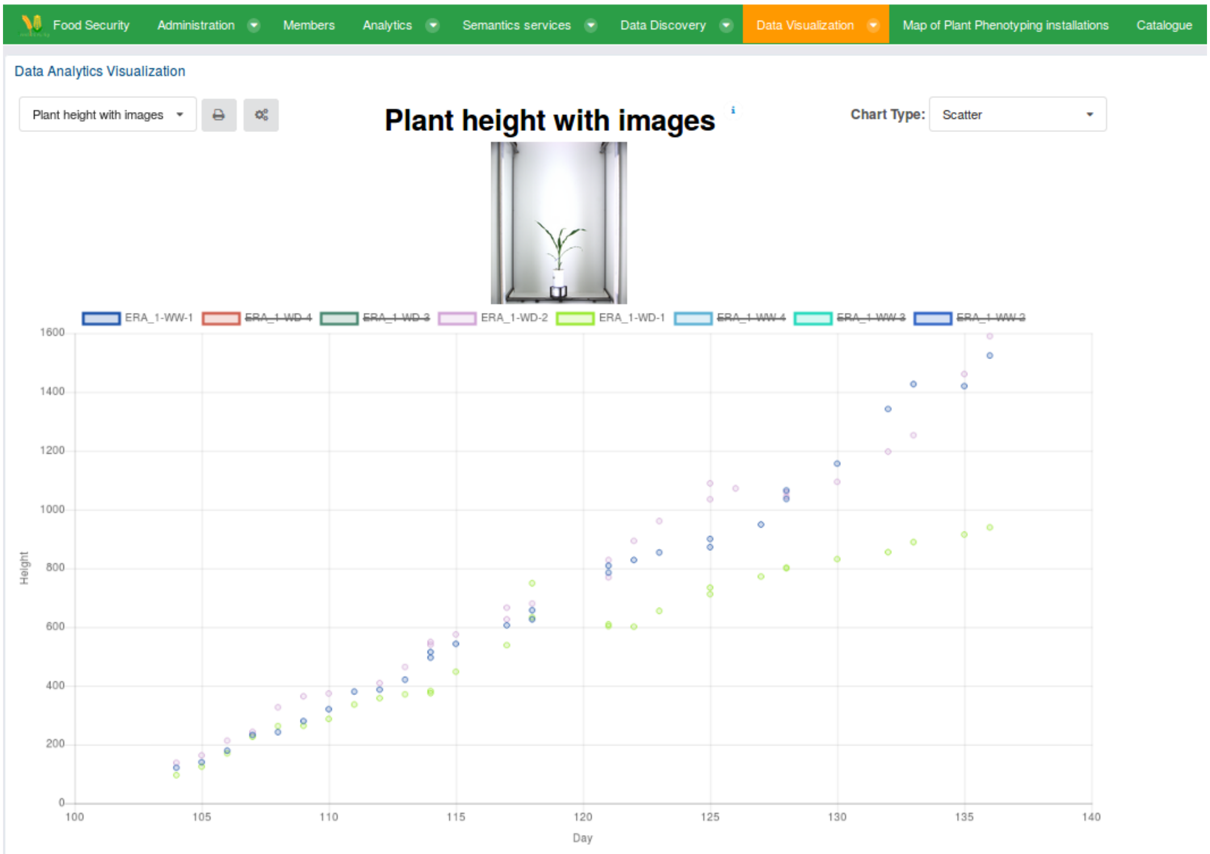


Figure 10: Example of Data Analytics Visualisation

- Workflows management: the VRE provides access to a Galaxy Server and support a visual component for the design of workflows;
- Machine learning: access to machine-learning approaches and also includes support for modern machine-learning approaches like ensemble techniques (boosting, bagging and random forests) and deep learning;
- Flexibility, extensibility and openness: Integration of open-source libraries into the VRE. Work with notebooks should be allowed;
- Semantics tools: ability to collaborate on building ontologies or vocabularies with an ontology management system. Dynamic ontology visualization;

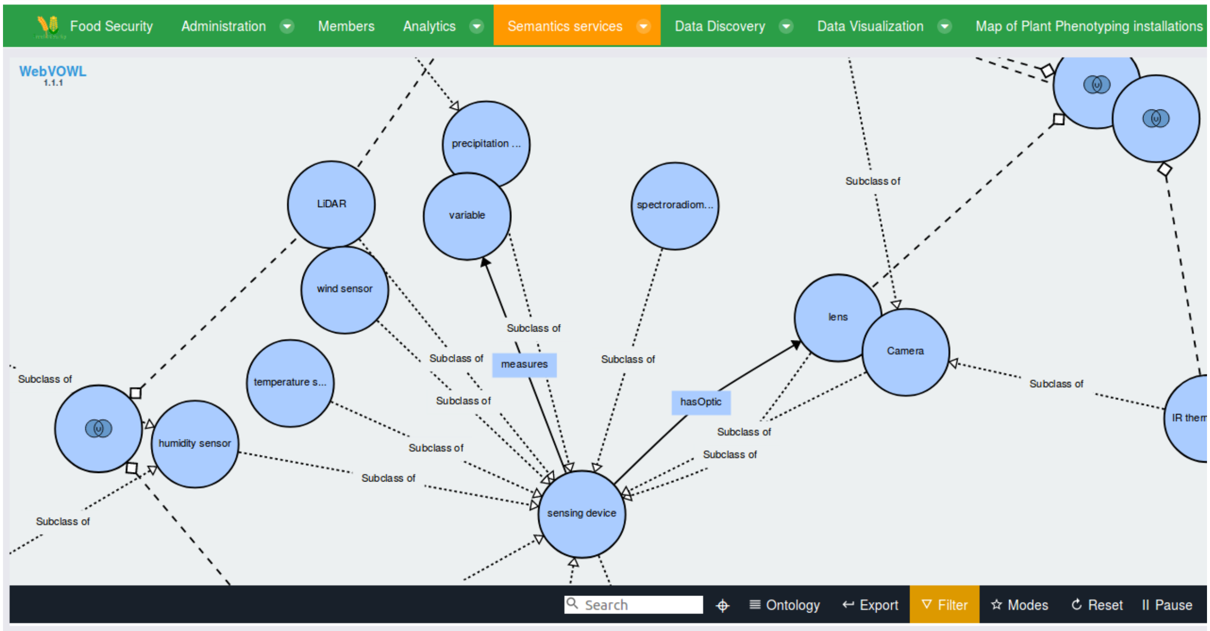


Figure 11: Example of Ontology Visualising

- Delivery: Ability to create APIs or containers (such as code, Predictive Model Markup Language and packaged apps) that can be reused;
- Resource management: management of data and tools. Provides reuse and version management of resources, auditing lineage and reproducibility;
- Collaboration: Makes users with different skills work together;
- Coherence: Provides a seamless end-to-end experience, to make the users more productive across the whole data and analytics pipeline.

3.3.4 Future science scenarios

Science scenario: Linking genotyping and phenotyping data

Typical users of this science scenario can be:

- Plant phenotyping researchers;
- Geneticists;
- Breeding companies

Context and Challenges

Genotyping has always been more developed than phenotyping but today, plant phenotyping is growing really fast. However, these two sectors (genotyping and phenotyping) remains separated. Specific research teams or companies work on genotyping whereas others work on phenotyping. They use different information systems and different vocabularies. Yet you can’t genotype efficiently without phenotyping. Indeed, the need is to link genotyping analysis to plant response to environmental stimuli systems in order to select plants and genotypes adapted to specific environments.

Work is underway to facilitate the exchange of data, such as the Breeding API which specifies web services to enable interoperability between breeding and phenotyping databases. This would enable to target a larger community using the VRE.

Recommendations

It is essential to follow the recommendations below, so that the VRE remains a tool meeting the community needs:

- Keep enriching the VRE with semantics : the semantics are essential to link data from different domains. It is necessary that researcher can easily understand how the traits and variables have been measured.
- Improving the collaboration model to allows research groups over different domains to work on a common workspace. Consequently, the groups management should allow the users to belong to several groups and deal with hierarchy.
- Managing big data : the datasets are less and less stored in files. The VRE should be an interface enabling users to access to data stored on different servers. It should also offer very powerful and up-to-date computing tools.
- Finally, the integration of all services provided should be enhanced in order to improve the user experience.

4 CONCLUSIONS

The present document reports on the AGINFRA PLUS vision regarding a next-generation community driven web-based research infrastructure. An overview of each of the three user communities was provided along with their special needs and challenges they face. The accomplished goal of the AGINFRA PLUS was to manage these challenges by employing the Virtual Research Environments (VREs) paradigm. In the demonstrated present scenarios there were described the challenges managed using the VREs paradigm

In details, the Agroclimatic and Economic Modelling Community demonstrated three present scenarios. The first one, namely *Crop modelling,* focused on the work of an agronomic modeller in a scientific or commercial environment and its challenge was to generate, analyze and visualize modelled regional crop growth indicators that could be used to perform crop yield statistics and to predict regional and local crop yields for short term seasonal yield predictions on future projections of yields under climate change. To address the associated challenge, a full crop modelling workflow was developed and deployed on a VRE which had to be able to be performed expeditiously and operationally.

The second one, namely *Crop phenology estimation,* focused on data science and data analytics methods to cope with data gaps and uncertainty in remote sensing time series. Its challenges were to use VRE compute capabilities calculating NDVI for large areas in shorter time, and perhaps refined them with machine learning algorithms that take more data into account, while in parallel VRE should support the separation of the (non-insured) yield deviations caused by long-term technological developments and environmental impact from the (insured) residuals, caused by abnormal weather conditions and disaster. These challenges managed to be addressed by preparing a VRE that supports explorative modelling, specifically development and testing of data science algorithms to determine crop phenology development and associated agronomic statistics.

The last one, namely *AgroDataCube*, collected and integrated agronomic data and provided services to retrieve it. Most of the challenges related to the AgroDataCube concerned the processing of the data and some of them were scalability, interoperability and findability. The solution of these challenges managed through the use of a resource like the AgroDataCube from a VRE to make its data findable, accessible, interoperable and reusable (FAIR).

The Food Safety Risk Assessment Community concerned two present scenarios. The first one, namely *Determination and Metrics of Emerging Risk – Demeter* focused on the early identification of emerging risks in the food (and feed) chain. The main challenge that this scenario managed to overcome was the identification, integration and analysis of information collected from public and non-public sources which is considered essential to support decision making of public and private sector stakeholders. With a view to resolve this challenge a special VRE has been developed which aims to answer the question, if the new web-based resource for the Emerging Risk Identification community (DEMETER VRE) has the potential to serve as an Open Science resource in the future.

The second present scenario, namely *Knowledge Integration Platform – RAKIP*, focused on efficient knowledge transfer between academia, business operators and governmental agencies. The main challenge was to manage to exchange the existing rich variety of useful models, software tools and databases for food safety risk assessment between different stakeholders. This challenged resolved by the establishment of the special RAKIP VRE that aimed at improving transparency in the data- or model-based risk assessments work.

The Food Security Community displayed one present scenario, namely *Meta-analysis of phenotyping data,* that focused on the characterization, in a background of global changes, of an environment and see in return which species and varieties would likely adapt to this specific environment. The main challenge was to design and implement scientific workflows on massive amounts of complex data produced at different scales (population, plant, organ, cells, etc.), different stages (sowing, phenology, harvest) in interaction with various environment components (soil, climate, agricultural practices, biodiversity, etc.). The solution to these problems was based on the creation of the Food Security VRE. Through the use of the VRE access control to the data sources and the shared objects (workflows, models and results of workflows) was provided.

In the scope of addressing more challenges that each of the communities might face in the near future, some future scenarios were prepared and presented. Along with these scenarios, some special recommendations were provided to ensure that the using tools meet each community needs.

In particular, the Agroclimatic and Economic Modelling Community future scenarios focus on extending the currently implemented scenarios and on improving the interdisciplinarity. Among the recommendations provided, it is worth mentioned that is suggested AGINFRA PLUS to promote data standards, not only based on the common practice in agri-food, but also considering standardisation in adjacent domains. Hence, it is suggested the actual use of semantics for (semi)automatic data linking and merging, to be improved.

The Food Safety Risk Assessment Community future scenario focuses on foodborne disease outbreak investigations. At its core this VRE would serve the need for efficient communication between all stakeholders that need to be involved in a given disease outbreak investigation. The result of such a service could be an important support for the decision if a product should be recalled or withdrawn from the market. For the accomplishment of this scenario it is mentioned among others that there is the need for dedicated service providers that are specialized in these tasks, as only VRE experts can have the necessary IT background knowledge to accomplish these development tasks.

Last but not least, the Food Security Community future scenario focus on linking genotyping and phenotyping data in order to select plants and genotypes adapted to specific environments. The realization of this linkage depends on some special recommendations such as the continuation of enrichment of the VRE with semantics, the improvement of the collaboration model to allows research groups over different domains to work on a common workspace and of the data management since it is has been identified that the datasets are less and less stored in files.

5 ANNEX A: DIGITAL SCIENCE RECOMMENDATIONS FOR FOOD & AGRICULTURE

DIGITAL SCIENCE RECOMMENDATIONS FOR FOOD & AGRICULTURE

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D8.3
DELIVERABLE TITLE	Future Science Recommendations
RESPONSIBLE AUTHOR	Nikos Manouselis (Agroknow)

Table of Contents

FOREWORD

I joined FAO (The Food and Agriculture Organization of the UN) in 1998 and got the responsibility for the AGRIS system. AGRIS was one of the huge bibliographical databases of the time which collected information about scientific and technical publications in agriculture and made them available especially to partners in developing countries. AGRIS already had the two elements about which most of the contributions to this publication are speaking. Community and Technology. AGRIS centers were holding annual meetings at FAO to coordinate their efforts to cover all publications in their area. The AGRIS secretariat initiated the development of specific software which should help them to accomplish this task. CDS-ISIS was developed already in the early 90s. In a way, FAO had a pioneering role in creating collaboration between scientific institutions. WUR and INRA, two contributors to this volume were very important centers of the AGRIS network.

Nearly all of the contributions in this volume emphasize the human factor and the necessity of community building before the technological aspects. This is understandable. Technological questions are straightforward (normally) and resolvable (theoretically). For community building there exists something similar as the 2nd law of thermodynamics. DeltaS>=0. Entropy (non collaboration) in a closed system can only grow. Collaboration is not a given. Every unit has its own business model and even every single person pursues specific goals. Collaboration establishes spontaneously only when the benefit is immediate. If the benefit is long term, a high amount of activation energy is necessary to create a situation in which more energy is set free by collaborative efforts than was consumed to achieve the collaboration. AGRIS was already such a project. Continuously energy input from outside (FAO) was necessary to keep it run. The European Open Science Cloud (EOSC) has the same issue, only dimensions bigger. It is the great advantage of having supra institute and supra country level bodies like the European Union to fuel long term collaboration. The money is the fuel, but without the dedication of people in the participating organization, it would have no positive effect.

When I was entering the area at the end of the 90s, Information was a topic of agricultural librarians and some agroinformatics. Now it is ubiquitous and every Scientist is becoming a data scientist as well. Data, not only experiments become the source of new knowledge. The announcement of 23andme that they have created a new drug out of customer data is only the last example for this. This means that specialists in data processing and evaluation need to work beside the scientist, who own these data and who work with them. There are many successful examples for this. But there are also examples of models and software developed out of other business models and then seeking for clients. This normally does not work out. So, the driving force behind EOSC developments should be more scientists who want to resolve a problem through collaboration than system engineers, who are looking for user of their super efficient virtual research environment. But I admit: sometimes it can be also the genius of system developers, that will inspire scientists to do things they were not even dreaming about.

When I inherited AGRIS in FAO, this was linked to another legacy system called AGROVOC, a big multilingual thesaurus about agricultural science and technology. I got this job not because of my credentials in library and information science (because I had no). I got the job because I had written the Spaghetti Code to publish the first ASP (MS Technology to publish databases on the web) in FAO. I am convinced my management wanted me to kill simply these old fashioned systems. At the same time, Tim Berners Lee made his TED speech about the Semantic Web. I immediately was lost for this concept and decided that AGROVOC would be the starting point for the Semantic Web in Agriculture. Speaking about Ontologies in 2001 had the intrinsic risk to be considered somewhat strange, at least. Now nearly all the contributions to this volume speak about semantics. Graham Mullier from Syngenta, even mentions the GACS (the Global Agricultural Concept Scheme). The GACS was the last "semantics standard" project that I initiated and led. I think I should admit now that it was a failure. One obvious reason is the collaborative aspect. Much more energy had to be invested in it before it would have been possible to see *positive effects. But I see also a problem with all the "semantic systems" we worked on. Too complicated, too much human work needed, too slow in development. I guess there is no big web application outside, which is really Ontology driven. I think we all have to look very carefully to the developments on AI based on big data to exploit quickly these developments for agricultural research and technology.*

I want to close these introductory remarks with emphasizing the role of the European Commission in setting up an ecosystem. It is not only the money that obviously facilitates projects that without these money never would have been undertaken. It is the process of community building that makes the research framework programs so precious. In FP7 NEON, FP7 D4Science, FP7 agINFRA, H2020 eROSA, RDA and other projects I got into contact with all these wonderful people, that at the end formed a community that is existing also outside a specific project. We might have moaned and complained sometimes about the bureaucratic process in setting up and managing the projects, but this is nothing compared to the advantages that we got through the process. Without the EU, its programs and its dedicated officers, this would not have been possible.

Johannes Keizer (dr.rer.nat)

www.johanneskeizer.com

NOTE FROM THE EDITOR

I always wanted to find ways in which my work could support and empower people that work on some of the grand challenges that our planet is facing. In 2005, a professor of viticulture at the Agricultural University of Athens, who had devoted more than 40 years in researching some of the world's most ancient grape varieties, asked me: "We are producing all this information. We are overwhelmed by the volume and variety of all our data. Is there a way in which technology can help us?" This got me into doing a PhD on agricultural information management. And three years later, Agroknow was born. We have been working on collecting, translating, and enriching food and agriculture data from around the world since then.

There are so many aspects of the agri-food sector that require data: improving yield, optimizing the manufacturing process, improving quality and safety, or increasing sales. These days, there is the technology and there are devices to collect a plethora of data from different stages of the supply chain. Starting from the field, where one can collect soil, trial, health, protection and phenotyping data for the plant and crop, genetic data for both seed and plant, environmental data from sensors, and image data collected by drones and satellites. Until the product arrives on the plate of the consumer, a wealth of more data is generated—related to its origin, nutritional info, logistics, certificates, processing, marketing, sales, or others. The data available for each product is abundant; it is also very heterogeneous and dynamic. As new devices are being used throughout the process, the amount of data is also expected to exponentially increase.

A number of questions then arise. How can one make sense of this data? Which are the critical decisions of the supply chain stakeholders that can be supported using this data? How may we extract insights that can inform and empower the people making these decisions?

In Agroknow, we truly share the belief that if we can create an interoperable, shared global data space, we can really propel the food industry forward. Information will become available to all actors producing innovation. Analytical and decision making tools will be able to incorporate a greater abundance of data sources. The digital economy will bloom, with many more online services and applications that will be able to use machine readable, interoperable, and often publicly shared data from all around the world. I have been working on this mission for more than 15 years, first as an individual and then together with our team. This led to the conception of the first FP7 agINFRA project; and to its evolution through the H2020 AGINFRA PLUS.

Today, I think that we still miss necessary infrastructure components, including technology, people, policy, and business ones, that can seamlessly integrate all the data platforms and make them work together. The sector seems to be taking tremendous steps, but for the time being I think that most of our energy is going on increasing the size of our data silos. Our food supply chains are globalized; our data supply chains should also be. This is my vision for the future.

By this volume, we have tried to take stock of where we stand today as a community. I am grateful to the contributions of some very important stakeholders and initiatives. In all these contributions, I read the same message - and the same desire. Let's work more; let's work together.

Nikos Manouselis (PhD, MSc, MEng)

CEO Agroknow & AGINFRA PLUS Coordinator

SECTION A: TOWARDS A DIGITAL ECOSYSTEM FOR SCIENCE

Background & History

The **European Open Science Cloud (EOSC)** is envisioned as a **trusted digital platform** for the scientific community providing **seamless access to data and interoperable services** that address the **whole research data cycle**. It aims to make all European research available to everyone, through a common, shared resource that may fuel innovation from the public and private sector. In a nutshell, the intention is to build **a cloud for all research data in Europe**.

Organisations and people working on **open access and open science related topics in food and agriculture** have been among the early supporters of this vision. The key milestones in this journey have been:

- In 2015, the vision towards EOSC was backed up by the FP7 agINFRA project, the first digital infrastructure initiative for food and agriculture. This continued the pioneer open access work carried out by the CIP PSP VOA3R project.
- International stakeholders came together twice at the Open Harvest events in Crete, to declare their desire to work together so that research data and digital infrastructures can be made available to all.
- A seminal white paper on the need to catalyze the creation of a data ecosystem for agriculture and food has been published in 2016, with endorsement from the Global Open Data for Agriculture and Nutrition initiative (GODAN).
- A number of international projects have been designed and launched to further promote and implement this vision, such as GODAN Action, eROSA and AGINFRA PLUS.

A critical mass of effort and energy has been brought together — to make Europe one of the global leaders of the data revolution in food and agriculture.

Towards a global data ecosystem for agriculture and food

In the GODAN Data Ecosystem White Paper of 2016, it was boldly stated:

"*Agriculture would benefit hugely from a common data ecosystem. Produced and used by diverse stakeholders, from smallholders to multinational conglomerates, a shared global data space would help build the infrastructures that will propel the industry forward.*".

The GODAN discussion paper aimed to catalyse consensus around what form a global data ecosystem might take, how it could bring value to key players, what cultural changes might be needed to make it a reality and finally what technology might be needed to support it. A number of important recommendations were carved out in this document:

- Finding business models that provide incentives for various entities to collect and share data. If these models provide business value directly to the data providers, the quality of the collected data will be higher.
- Leading by example by providing open data sources. Syngenta has already done this by publishing data about the results of its Good Growth Plan.
- Encouraging data standards that make it easier to produce and share data. In doing so, stakeholders will need to have reasonable expectations of how these standards will be used.
- Automating data collection. Automatically collected data is more likely to be accurate and precise than data collected by hand.

- Annotating datasets. Even automatically collected data cannot be used if it is not described in a consistent and understandable way.
- Following data sharing principles. The five-star maturity model and the FAIR principles provide guidelines for creating and sharing data.
- Using the data. All of the best data sharing efforts have little impact if the data is not used in a productive way. Stakeholders must encourage a cottage industry of data-backed apps that get the most value from datasets.

Bringing an international community together

A number of people involved in open science projects and collaborations from Europe and the world, came together twice at the Open Harvest events of 2016 and 2017 to declare their intention to work more closely together.

This was communicated very clearly in the 2016 Chania Declaration:

"We agree to strengthen our collaboration through building a common roadmap for the infrastructure elements our sector so desperately needs. As we define infrastructure we consider every element that is needed to achieve our goals: physical structures; software; standards; and procedures. First of all we need to support collaboration. This means nurturing relationships that represent truly international perspectives and include both the public and private sector in partnerships."

This intention was made much more specific in the 2017 edition of the Chania Declaration and its resulting Call to Action:

"A shared environment or 'data ecosystem' requires that data, standards and infrastructure be as "open" as possible...This shared environment should be supported by policies and practices that incentivize people to generate social and economic value as well as develop new business models. It should be highly decentralized. It should be a collaborative approach to sharing knowledge, data, analytics and technology across a network of distributed platforms. It should have social, economical and technical dimensions that enhance the impact and efficiency of our work."

The vision of e-ROSA: farmer-inclusive, system-oriented science

Engagement and consultation with the community continued in 2017 and 2018. The EU-funded e-ROSA project organised three stakeholder workshops with participants from the public and private sectors, concluding with the publication of a report titled "Roadmap for a pan-European e-Infrastructure for Open Science in Agricultural and Food Sciences". More than 100 stakeholders were involved in the elaboration of this roadmap to build a common knowledge, foresight and recommendations to achieve this vision. e-ROSA used the "**Share, Connect, Collaborate**" motto as guide through its evolution, as it appears to be the most federative challenges to build an e-infrastructure federating the actual scattered data ecosystem in the agri-food sciences:

- **Sharing**: of the resources of relevance to the scientific process (data, models, papers, etc). Open science is only possible if one is able to share one's research first.
- **Connecting**: available resources need to be connected to allow integration and tackling large scale and more ambitious questions in science.
- **Collaborating**: the research community itself needs to collaborate beyond ad-hoc arrangements to create, maintain and supply domain specific resources for open science in a network of regional or domain nodes.

e-ROSA highlighted that to ensure sustainable and nutritious regional and global food systems, requires a commitment to Open Science. This translates to the need for both research and development institutions, as well as every other stakeholder in the innovation process, to fully embrace a digital transition for each phase of the knowledge production cycle for innovation. For the e-ROSA stakeholders, this extends from research planning and design to collection, analysis and simulation of data and the dissemination of knowledge and the underlying data.

The pilots of AGINFRA PLUS: holistic data-intensive science workflows powered by a shared digital infrastructure

As part of AGINFRA PLUS, this holistic view of end-to-end scientific workflows powered by an open science e-infrastructure has been piloted, tested and illustrated. Three academic institutions (**INRAE** in France, **WUR** in the Netherlands and **BfR** in Germany) and two commercial ones (the academic publishing company **Pensoft Publishers** and the bioinformatics company **BioCos** that won the AGINFRA PLUS data science challenge) piloted the way that a very complex computational model can be executed rapidly over a shared digital infrastructure, to deliver results that support critical decisions. The demonstration scenarios that they have explored were:

- *Distributed simulation of arable crops at farm field scale (WUR)*
- *Sharing and running simulations of harmonized food risk assessment models (BfR)*
- *Management, access & visualisation of plant phenotyping resources (INRAE)*
- *The smart & executable paper for agriculture & food scientists (Pensoft)*
- *Processing DNA genomes for short sequences that are potential food fraud biomarkers (BioCos)*

These selected use cases are representative of a wealth of others that can be found in science and innovation in food and agriculture. Most of them are based on sophisticated mathematical models that are used to simulate or predict how several organisms in interaction will behave under specific conditions. They often use input parameters for training and calibration, e.g. to deploy for a particular crop or geographical region. They also use historical data that serve as the “memory” of the model. They very often use real time measurements (e.g. in situ or from remote sensing) of current environmental conditions to re-calculate a prediction. Their final outcomes need to be visualised and published in ways that will help scientists but also practitioners take critical decisions.

Questions still looking for answers in 2020

As 2020 shed its light, the economic and political environment in which this discussion takes place changed dramatically. All stakeholders working on the digital transformation of food and agriculture are facing important challenges. For the organisations, initiatives but also individuals supporting the role of scientists in both the public and private sector, important questions come forward:

- How close are we to the vision of a common data ecosystem?
- Do we still believe that a shared global data space will propel the industry forward?
- How important, relevant and possible is it to develop digital science infrastructures as collaborative and community-driven resources?

This discussion paper has been envisaged as a volume with invited contributions where important organisations and initiatives from Europe and the world are sharing their position statements. This discussion is facilitated by the AGINFRA PLUS consortium, but it goes much beyond its promised contributions. It rather takes stock on the achievements (but also the failures) of the community so far. To redefine what the future looks like for digital science in food and agriculture. To highlight the important next steps. To ensure that alignment and collaboration between European and international stakeholders continues, stronger.

SECTION B: POSITION STATEMENTS FROM AGINFRA PLUS PARTNERS

B1 Position Statement from INRAe

Authors: Olivier Le Gall, Michaël Chelle, Odile Hologne, Pascal Neveu, Alice Boizet
INRAE, France
olivier.legall@inrae.fr, michael.chelle@inrae.fr, odile.hologne@inrae.fr, Pascal.Neveu@inrae.fr, alice.boizet@inrae.fr

1. Digital science infrastructures as collaborative, community-driven resources

Global change creates new expectations for research in agriculture: adaptation to climate change, food and nutritional security, transition of agriculture, preservation of natural resources, restoration of biodiversity, anticipation and risk management. In addition, there are more territorialized challenges which include living conditions and remuneration of farmers, the economic competitiveness of businesses, access to healthy and diversified food for everyone. Taking all these issues into account requires precise observation and description of various and complex contexts and cases. Agriculture and therefore agricultural research must be redesigned and therefore it requires a very strong collaboration of different stakeholders such as citizens, farmers, researchers, which must share their data and knowledge. This is why INRAE finds very relevant the development of digital science infrastructures which will enable these communities to produce sharable and reusable data as well as models and workflows. Using these infrastructures should change excessively individualistic researcher practices to more collective and transparent ones. And it is collective practices that must drive the development of these e-infrastructures. For these reasons, INRAE, as part of its open science policy, is engaged in e-infrastructure development at national and international levels to produce and reuse FAIR data as well as models and workflows. More broadly, it has led to the creation of a direction for open science (DIpSO), similarly to the Wageningen Data Competence Centre (WDCC), to foster the development of open science, from open access to open data to citizen science.

2. Major achievements towards an open data ecosystem for science

The impact of the lead activities in order to promote and enforce Open Science over Europe are clearly palpable. Growing researcher communities are better and more effective at sharing of digital resources but the re-analysis of other group's data is still in its infancy.

In particular, in the field of food and nutrition security, open resources regarding the environment-and-agriculture segment are lacking in EOSC or lagging behind the nutrition-and-health segment, while they would necessarily be at the same time more diverse, and necessary to assess and design new sustainable food systems.

3. The challenges of the past

Developing Open Science and its supporting methods and operational tools is not so easy to manage. In some cases, this seems to lead to gaps and a mismatch with the research request. This is especially true in the Agri-Env sector (see above).

INRAE is aware of current efforts to link and harmonize existing e-infrastructures. We are still missing success stories. Indeed, even if a lot of progress was made at specific domains, actions have not been enough compelling for the majority of researchers to extensively use Open Science services and resources. Many of us believe that joint efforts should be a top priority to ensure that researchers do not become discouraged and return to their practice of working in closed networks.

There is a high expectation of the agriculture, environment, and food sectors for infrastructural and cross-community support that can cover the full food chain. We can clearly see that this demand comes not only from Science, but also from business, politics, NGOs, and citizens at large. The H2020 eROSA initiative has already sketched in its "e-infrastructure Roadmap for Open Science in Agriculture" the outline and requirements of such an e-infrastructure, defined as the "Food Cloud". While there are several EU-funded initiatives in the domain of agriculture, none of them seems to really provide a fully integration of agriculture (in its environment) and food. There is no initiative in place that has the ambition to develop and deploy such a holistic and overarching infrastructure for the domain. While EOSC could be the frame for such an initiative, this leads to a missed opportunity to link up the involved research communities and for instance close the gap between the production and consumption sides of agri-food science, or even between the environment and nutritional health. However, this perspective is, in our view, required to address the major interdisciplinary challenges needed to move forward on the major societal challenges of our time, especially since these challenges are themselves expressed in a holistic manner.

4. The vision towards 2025

The main objective for INRAE is to enable, facilitate, and stimulate interdisciplinary research because crossing of different research domains is necessary to tackle the new challenges faced in agriculture. Regarding data, the objective is to insure both the capacity of vertical and horizontal integration. The vertical integration consists in connecting data for a given scientific communities, from local to European scale. It relies on the development of community's research infrastructure from INRAE to EU level (ESFRI) e.g. for plant phenotyping, bioinformatics, water and forest ecology. The horizontal integration consists in the ability to connect data from different scientific communities (plant science, animal science, environmental science, humanities and social sciences, etc) to support and foster the development of holistic approach in research and innovation in agriculture, environment, food, and nutrition. At INRAE, it corresponds to the starting project of the development of a federative research e-infrastructure.

5. The most important next step for our community in 2020

The most important next step would probably be to support the engagement of the Agri-Env research community in the EOSC movement by supporting research projects addressing the societal challenges of our food systems and at the same time stabilising the components of a sustainable digital infrastructure to avoid fragmentation and loss of energies and support interdisciplinary research for resilient and sustainable food systems. e-ROSA, AGINFRA PLUS projects were the first steps towards this vision.

B2 Position Statement from *Wageningen Research (part of Wageningen UR)*

Authors: Rob Lokers, Sander Janssen, Rob Knapen
Wageningen Research, The Netherlands.
rob.lokers@wur.nl, sander.janssen@wur.nl, rob.knapen@wur.nl

1. Digital science infrastructures as collaborative, community-driven resources.

Wageningen University & Research (WUR) recognizes the importance of open science as the way forward for future scientific research. Challenges in the life sciences domain are global and highly interdisciplinary, so cooperation and co-development with an open mindset are indispensable to tackle these challenges. As a leading research organization in the agri-food domain, WUR has therefore adopted open science as one of its leading principles. The Wageningen Data Competence Centre (WDCC) and the recently launched Open Science and Education programme aim to facilitate and stimulate our researchers to put open science into practice. We also believe that digital science will play a key role in future research and that collaborative efforts of leading research and technology organizations are required in the European level to stay at the forefront of developments. Therefore, WUR underpins the demand for excellent digital science infrastructures as a collaborative resource and specifically the need for an approach that is, more than now, driven by user communities and their demands.

2. Major achievements towards an open data ecosystem for science

A lot has been accomplished in the past years, through collaborative efforts of European research organizations, to develop and connect research infrastructures and to support the roadmap set out for the European Open Science Cloud (EOSC) and its ongoing implementation. A great number of innovative research infrastructures in the field of life sciences saw the light and are in many cases efficiently used by international research groups. A wealth of newly available generic scientific facilities, like for instance clustering of resources (compute, storage) and the provision of sustainable ways to share research results (catalogues, data journals etc.) are increasingly being used, enforced by the open data and science policies of governments and individual organizations. WUR stimulates that such open science facilities are getting adopted by its own researchers and its research partners and we clearly observe that this is having a positive effect on the effectiveness of research and its impact on society.

3. The challenges of the past

The positive impact of the many efforts to promote and enforce open science over Europe are becoming clearly visible. A vastly growing group of researchers are benefiting from better and more effective sharing and reuse of resources, and more importantly, a culture of sharing is gradually developing. Nevertheless, WUR also notices that the pace of development of Open Science and its supporting mechanisms and tools is sometimes difficult to manage. In cases this seems to lead to shortcomings and a mismatch with research demand.

WUR is fully aware of the current efforts to link and harmonize existing research and e-infrastructures. While a lot of progress has been made at specific domains, we feel that up until now attempts have not been convincing enough to persuade the majority of scientists in our field to extensively use facilities that support open science to get the most out of it for their research. Researchers, especially if they are not directly involved in the area of developing research infrastructures, still experience a high fragmentation. We argue that harmonizing efforts should be a top priority, to prevent researcher from getting discouraged and returning to their practice of working in disciplinary silos and closed peer networks.

We argue that there is a high demand from the agri-food sector for infrastructural and cross-community support that can cover the full food chain, from production to consumption, from seed/feed to fork. We see this demand coming from research, but undoubtedly also from business, policy and NGOs. The H2020 eROSA initiative has already sketched in its roadmap towards open science the outline and requirements of such an infrastructure, defined as the “Food Cloud”. While there are several EU-funded initiatives in the domain of agri-food, none of them seems to cover the whole nexus of agriculture and food. WUR sees it as a missed opportunity that, at this point in time, no initiative is in place that has the ambition to develop and deploy such a holistic and overarching infrastructure for the agri-food domain. This inevitably also leads to a missed opportunity to link up the involved research communities and for instance close the gap between the production and consumption sides of agri-food science. Nevertheless, such a broad perspective is in our view indispensable to tackle the key interdisciplinary challenges needed to move forward on the grand societal challenges of our time.

More from the technical perspective of the evolution of digital science supported by e-infrastructures, WUR realizes that it is difficult to keep up with external, international developments around digital science and digitization in general. A lot of efforts around setting up the EOSC are, logically, based on building on past investments and thus also on currently available infrastructures. As these have been developed in the past, it is inevitable that some of their paradigms and components are overhauled by newly evolved concepts and technologies. As frequent end users of such infrastructures we see in some cases that we are being bound to using what might be considered as an outdated concept, or technologies that are not compatible with what is state-of-the-art in our communities, and that have been replaced by better alternatives. We believe that this might be a serious barrier for especially the forefront of digital scientists to adopt European infrastructures for digital science. We therefore recommend that efforts to build EOSC infrastructure are set up in an iterative way and taking into account the needs of targeted user communities like the agri-food community. This should foster a process of learning and improvement and provide the opportunities to replace or refactor components where deemed necessary to stay upfront.

4. The vision towards 2025

From the perspective of WUR, the European Open Science Cloud and its roadmap are currently still very “data-centric”. There is a large emphasis on aspects like data management, data curation, data stewardship and providing data in a FAIR manner. Consequently, we see more and more data becoming available, but at the same time not a lot of data is actually reused in practice. We argue that the fact that data needs to be reusable, and that thinking about how data can become usable and about supporting the actual reuse through accessible computation architectures and FAIR sharing of data analytics should get more attention. This means that there needs to be a shift from data production and publication to data use, and from data management to computation and data analytics. Besides, it also requires more focus on data quality, next to data availability. This should eventually lead to a situation where users can determine their selves if data is usable. It should for instance promote delivering “fitness-for-use” descriptions that are needed to complement technical and semantic interoperability and to support data users in deciding if, and how data can be reused for their purpose.

In order to mainstream digital science and to deliver its benefits to society, the development of digital science infrastructures, tools and communities of practice should be approached from a more user community faced perspective. While often the development of digital science seems to be predominantly a technical operation, that includes use cases “for the sake of user acceptance”, WUR argues that these initiatives should become truly user driven, with technology being a means of support, rather than the primary objective. We believe that this is on the long term the only way to achieve broad adoption of digital open science and to get the best out of current and future development of e-infrastructures. For future evolution of digital science and digital science infrastructures, we therefore plead for a less technocentric and more community and user-centered approach.

We strongly believe that another important objective for the coming five years should be to make the data science ecosystem more inclusive and democratic. These days, e-infrastructure and facilities for digital open science are still relatively complex and fragmented. Also, many of them are developed and governed and most easily accessible by the larger research organizations. Smaller organizations often lack the knowledge and capacity that is currently still required to become operational. We should therefore strive for equal access to research infrastructures for both larger and smaller institutions in the field of agriculture and food sciences. This includes the integration of data beyond what's coming from research projects, also focusing for instance on operational data collections and access to these data collections as part of research activities. We believe that this would also make the use of and the contribution to digital science more appealing for e.g. business, NGOs and governmental bodies.

In our working field of agri-food and environmental data science, the dimensions of space and time are always present. Therefore, being able to analyze data that has both temporal and spatial dimensions, so called spatiotemporal data, is an essential asset. It is unfortunately a fact that working with such data is often still regarded as a specialist subdomain, and often still requires tools that are not well integrated in regular research workflows. We hope that by 2025, working with spatiotemporal data will be as common as any other data science exercise. WUR therefore advises that specific efforts are made to integrate geo-spatial tooling (e.g. analytics, visualization) in digital science environments. As WUR, being a key user community of such innovations, we are prepared to take our role and responsibility as part of that challenge.

5. The most important next step for our community in 2020

WUR believes that as agri-food community we need to develop demonstrators using the next generation digital science infrastructures, that have a clear link to problems experienced in the agri-food sector. Such demonstrators are indispensable to convince the larger community of the benefits of open science infrastructures and to commit them to contribute and participate. As an example, WUR is nowadays committed to investigate how to establish data infrastructures for managing private farmer data, retaining privacy and data ownership while also keeping it accessible for data use for innovation and research. With institutional partners WUR is developing a Personal Farm Data Space, that allows for the combination of public (open) and private data in seamless manor, giving farmers and agri-food organizations control of access to the data of their business from one location. As another example, WUR with VITO is managing an open infrastructure for crop trial data, based on standardized quality assurance protocols. To successfully develop such demonstrators, as stated before, requires infrastructure and tooling that is more than now based on a community and usage perspective rather than on a technocratic perspective. This again refers back to the previously stated assertion that future development should become truly user driven, instead of driven primarily from technology.

6. Closing Remarks

WUR believes that bringing digital science forward is a cross sectoral challenge that should benefit science and society across sectors for society as a whole. In that respect, and for now constraining this to the agri-food domain, we think that the connection between policy objectives of different DG's (RTD, AGRI, Connect) can be more strongly aligned, with more joint messaging. However, if we consider this from the full working domain of WUR, life sciences, this bridging of sectors and joint messaging would, besides DG-RTD, DG-AGRI and DG-CONNECT, also include the linkages with among others DG-ENV, DG-CLIMA, DG-SANTE and DG-DEVCO.

B3 Position Statement from *Istituto di Scienza e Tecnologie dell'Informazione - National Research Council of Italy (CNR)*

Authors: Leonardo Candela, Donatella Castelli
Istituto di Scienza e Tecnologie dell'Informazione - National Research Council of Italy, Italy, leonardo.candela@isti.cnr.it, donatella.castelli@isti.cnr.it

1. Digital science infrastructures as collaborative, community-driven resources

The National Research Council of Italy (CNR) and its Istituto di Scienza e Tecnologie dell'Informazione (ISTI) is not only believing that the development of digital science infrastructures is important, relevant and possible but also it is investing a lot in activities in demonstrating this. The team has coordinated and coordinates the technical development of two large scale infrastructures for supporting digital science, namely D4Science and OpenAIRE. D4Science is an infrastructure specifically conceived and implemented to provide its users with a rich array of services supporting the various phases of typical research workflows by promoting ***Open Science*** and ***Virtual Research Environments***. OpenAIRE is an infrastructure specifically conceived and implemented to promote ***open access*** to almost any scientific outcome by facilitating innovative ways to ***communicate and monitor science***. Both these infrastructures are a reality nowadays (the serve real communities of practice called to implement actual scientific workflows) and they will contribute to the development of the European Open Science Cloud. The development of such infrastructures and the services characterising them is the result of a long and continuous dialogue among the various actors involved in the development of these typologies of infrastructures where domain scientists, data and information scientists, computer scientists, software developers and system administrators are confronting to develop end to end solutions for scientific workflows. Co-design and co-development are key aspects characterising the development and functioning of both D4Science and OpenAIRE, e.g. the services these infrastructure offer make it possible for diverse communities of practice to simply using and customising them to implement novel workflows.

2. Major achievements towards an open data ecosystem for science

The last three years have been characterised by several developments regarding Open Science and infrastructures for supporting it which culminated in the launch of the ***European Open Science Cloud*** (Nov. 2018) and its advancements. The European Open Science Cloud is promising to benefit from the rich array of experiences and solutions currently existing to realise a comprehensive infrastructure to "*offer 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines …*". EOSC will collect and repurpose the wealth of research outcomes otherwise disperse across several infrastructures into a novel and common space promising to offer new opportunities for their development and exploitation. This will give to the "gems" developed by various initiatives a new showcase.

Among specific achievements pertaining the agri-food domain, the experiences made in AGINFRA PLUS to provide selected communities of practice with suitable solutions resulting from the exploitation of existing and services represent a valuable asset realising - in a small scale - a thematic EOSC. In particular, the services designed and developed to deal with semantic data management, data analytics, data visualization and publishing, packaged together into coherent and community defined working environments promoting collaborative open science practices represent a tangible realization of the opportunities EOSC is promising for the Agri-food section and beyond.

3. The challenges of the past

The challenges characterising initiatives like the European Open Science Cloud as well as AGINFRA PLUS are many. These challenges intertwine technical aspects with organization-, culture- and budget-related aspects. Acquiring a comprehensive and detailed picture on the state of the art solutions in the domains falling under the agri-food umbrella (and beyond) is not a trivial task as well is it is not trivial and immediate to count on these state of the art solutions to serve new needs and scenarios. Placing under scrutiny practices, approaches and solutions is not simple and, very often, it is considered a waste of “resources” leading to from scratch approaches, approaches ending to “reinvent the wheel”. These settings represent one of the major causes of missing opportunities. The developments of large scale infrastructures - be them generic like EOSC or domain specific like those dedicated to agri-food communities - should not escape an almost continuous and unbiased survey of what’s available, with their strengths and limitations.

4. The vision towards 2025

In five years’ time it is auspicable to have in place a comprehensive and feature rich “infrastructure for science” encompassing the boundaries posed by disciplines and organizations, an infrastructure enacting scientists and researchers to focus on their research tasks by seamlessly leveraging the wealth of research supporting resources (services, data, computing). Such an infrastructure should be co-developed by promoting a participatory and open approach recognising contributions and its development process should be driven by equitable practices bringing benefits to our society as a whole. Moreover, such an infrastructure should be recognised as the “one stop shop” for any research activity, no matter the scientific investigation/research discipline it should represent the primary source of research supporting resources (data, services, computing). We do expect that such an infrastructure will make open science the norm in all scientific investigations and we will work to make this happening.

5. The most important next step for our community in 2020

The primary goal - in the short and long period - for the agri-food community is to contribute to the development of any infrastructure like that envisaged above, be it the European Open Science Cloud or a domain specific infrastructure for agri-food. Such contributions need to be organised thus to guarantee that the resulting infrastructure(s) will be an asset (a “minimal viable product”) supporting research in agri-food (and beyond). The expected contribution should result from intra-community activities and inter-community activities. For intra-community activities, it is fundamental to reach consensus on what are the existing solutions for agri-food with their strengths and limitations as well as to carefully identify the open issues characterising the development of an infrastructure suitable for agri-food research. For inter-community activities, the agri-food community should acquire detailed knowledge on developments and initiatives promoted by homologous research domains thus to identify similarities and peculiarities of these with respect to the agri-food research.

From the AGINFRA PLUS perspective it is fundamental to further promote awareness on the solutions we managed to develop thus to maximise their uptake and diffusion. The larger is the user base, the greater the impact of the proposed solution (an indirect indicator of the effectiveness of what has been proposed).

6. Closing remarks

Science is global, scientific investigations are multidisciplinary and crossing the boundaries of specific organizations. Any infrastructure aiming at supporting digital science for food and agriculture should follow the same settings, i.e. it should be global and should be deployed and conceived to cross the boundaries of disciplines and organizations.

B4 Position Statement from *National and Kapodistrian University of Athens (UoA)*

Authors: Georgios Kakaletris, Panagiota Koltsida
National and Kapodistrian University of Athens, Dept. of Informatics and Telecommunications
gkakas@di.uoa.gr, p.koltsida@di.uoa.gr

1. Digital science infrastructures as collaborative, community-driven resources

Open digital science infrastructures can only be realized as collaborative community driven resources, open to inclusion, reuse, adaptation and extension. Under this model, infrastructures will emerge from resources that primarily serve their intended usage and, by being opened up, facilitate their exploitation by other scientists promoting the full exploitation of resources and fostering interdisciplinarity. In this landscape, Open Science will be the major motivator as the growing demand for replicability of research, intensively requires increased access to research resources and transparency. The resolution of common technological challenges and the adoption of common standards will be the catalysator.

However, digital sciences infrastructures shall also facilitate the emergence of new business models that will further empower Open Science including the fusion of commercial, private and public resources.

2. Major achievements towards an open data ecosystem for science

The major achievement towards an open data ecosystem for science is the conceptualisation of EOSC and the establishment of fundamental working groups around its shaping and future evolution strategy. EOSC further strengthens the establishment of Open Science as a culture, further extending the previously formed Open Access paradigm (applied to knowledge and data), into resource sharing that goes beyond the contextual domain (e.g. project, administrative scope, community etc) shaping an open “marketplace” serving researchers, scientists, businesses and the public. Lifting of technological barriers, achieved in recent years and ongoing, is one of the success factors, however building a consensus that is the way forward, independently of the current obstacles, is the biggest one.

Significant other steps are the insistence on Open Access and Data Management Planning which are enablers for Open Science.

3. The challenges of the past

From our perspective the major area where more work is required, is the definition and exploration of new, open business models that will allow industry (including SMEs), citizens, private / public research to take their roles in the landscape of community driven research infrastructures, allowing delivery and use of higher quality, non-locking, open and dependable resources to their audience. In this direction, the collaborative work of all stakeholders in the domain, along with policymakers and the industry is needed, as ethical, regulatory / legal, economic and technological issues arise.

Especially, EOSC and stakeholders around, need to work out the policies and inclusion mechanisms that will allow such development to be feasible and fair, respecting concepts of public funding and business competition etc, which is a multifaceted endeavour.

In order to proceed with this, EU has to define the means for regulating access to shared resources that will both open access to their consumers, yet will restrict abuse especially in the case of restricted resources that do not suffice to cover the need.

4. The vision towards 2025

The following areas achievements will be of utter importance in the sectors where AGINFRA PLUS has been activating:

- Complete research data FAIRness will be the cornerstone of the future in agro-climatic community research.
- Findability and reusability of research, that implies not only data FAIRness, but also all elements of it such as the process followed, the methodology and the “code” behind it, even the errors and uncertainties, expressed all in machine usable packages of information.
- Easy reproducibility of research in agriculture, which will be enabled by the aforementioned achievements, but will be streamlined by the emergence of common protocols for “active” publications where everything related to new knowledge, and the methods to interact with it and explore its internals, will be structured in one machine readable entity.
- Identification of digital game changers in agriculture that will realize the digitalization of it, by pursuing solutions in health and nutrition, food safety, confrontation to climate change etc.

5. The most important next step for our community in 2020

From a technology standpoint, the most important next step is to solidify, standardize and scale up the concepts, methodologies and tools delivered to communities, validating and strengthening those with new ingredients that will further support the goals set for Open Science and the empowering of the agro-climatic sector research. This may be further supported by direct alignment of the offered toolset with the priorities of Sustainable Development Goals (SDGs) and the promotion of this alignment to all relevant stakeholders. Furthermore, the integration of socioeconomic and financial aspects of associated phenomena in the research methodologies of the addressed sectors, exploiting the perspective of different disciplines, will help to shape a more holistic view of the agro-climatic research goals, impact and possibilities.

B5 Position Statement from *Pensoft Publishers*

Authors: Lyubomir Penev, Teodor Georgiev, Pavel Stoev
Pensoft Publishers, Bulgaria

1. Digital science infrastructures as collaborative, community-driven resources

Pensoft maintains several infrastructures (examples: ARPHA, ReFindit, OpenBiodiv and supports several other EU-based or international infrastructures via data delivery, open APIs, etc. We believe that community commitment is a key in the development and adoption of any research infrastructure. At the same time, we are convinced that the community support alone couldn't guarantee the sustainability of the infrastructure without the funding from organisations or private donors, or commercial activities or through mixed models.

2. Major achievements towards an open data ecosystem for science

The major achievement in the last three years were:

1. The implementation of the open data pilot in seven key areas of the EU's Horizon2020 program
2. The EU position towards open science
3. The launch of Plan S by the CoALition_S, a consortium of major EU and USA science funders towards open access to all results of the publicly funded research towards 2024.

3. The challenges of the past

The widely proclaimed understanding that "community ownership", or "community support" of the infrastructures alone can ensure their long-term sustainability often leads to decline and reduced use. Industry and especially SMEs are stronger drivers behind innovations, a well-known fact that is usually well understood by economists and politicians but often underestimated by the researcher's communities.

4. The vision towards 2025

Increasing the number of clients for Pensoft resources and infrastructures, innovations towards open data and open science, developing of new resources and features.

5. The most important next step for our community in 2020

Successfully run and further develop the features and workflows developed within the project, in the first place the two data journals, developed within the AGINFRA PLUS project.

SECTION C: POSITION STATEMENTS FROM EU STAKEHOLDERS

C1 Position Statement from *Research Infrastructure for Food, Nutrition and Health (FNH-RI)*

Authors *(on behalf of the FNH-RI community)*:
Karin Zimmermann, Dept. Consumer and Chains, Wageningen Economic Research, The Netherlands, karin.zimmermann@wur.nl

Sjaak Wolfert, Dept. Data Science, Information Management and Project Management Organisation, Wageningen Economic Research, The Netherlands, sjaak.wolfert@wur.nl

Pieter van 't Veer, Dept. Human Nutrition, Wageningen University, The Netherlands, pieter.vantveer@wur.nl

1. Digital science infrastructures as collaborative, community-driven resources

The present European food system fails to overcome the challenges of personal, public and planetary health. Despite considerable recent research on healthy and sustainable diets, consumers have yet to shift to more plant-based diets which both reduce disease risk and the environmental footprints of our food system. FNH-RI was created out of a need for transdisciplinary research where citizens are actively participating in contributing data and knowledge in an interoperable way, particularly in the fields of (sustainable) food, nutrition, and health. This is necessary to link the health indicators of food to its environmental impact, delivering the research and innovation needed to ensure that a healthy EU population is fed sustainably through the next century. The four core scientific domains concern research on i) food environments, ii) consumer behaviour, iii) diet and eating patterns and iv) personal and planetary health.

FNH-RI improves the quality and impact of nutrition and consumer research via three distinct services: advanced data science services; access to cutting-edge research facilities and tools, including the novel Citizen's Data Platform; and training, education and dissemination services to equip a new generation of food system researchers. Improved research interoperability allows multidisciplinary research on healthy and sustainable eating patterns to be linked to adjacent food system research infrastructures in agriculture and health.

Combining research across these domains will advance core scientific agendas such as the SDGs, the Paris Climate Agreement, DG-Santé, and the European Common Agricultural Policy. FNH-RI builds upon and unites several active research communities, and is a direct result of the EU Joint Programming Initiative healthy diet, healthy life (JPI-HDHL), the Determinants of Diet and Physical Activity (DEDIPAC), the European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI) and Metrics, Models and foresight for SUStainable Food and Nutrition Security (SUSFANS). Closely partnered are emerging EU science programs Fit4Food 2030 and EITFood. Among RIs in the health and food domain, none address the critical intersection of the dual health and environmental impacts of food. However, four RIs are closely related to the FNH-RI: METROFOOD, ECRIN, BBMRI and ELIXIR. Over 100 Universities and Institutes in the food, nutrition and health domain have signed MoUs and intend to collaborate with FNH-RI. Core bodies concerned with FNH-RI, namely the WHO and the Federation of European Nutrition Societies have formally supported FNH-RI.

2. Major achievements towards an open data ecosystem for science

FNH-RI is the result of 8 years (2012-2020) of international and transdisciplinary networking and development activity. The FNH-RI concept development and feasibility studies were funded and reviewed under the H2020 RIA SC and INFRADEV funding instrument in the projects EuroDISH and RICHFIELDS. The PROSPECT FNH-RI consortium has evolved from the needs assessment of our EuroDISH projects (15 partners, 7 countries) and our subsequent ESFRI roadmap 2016 proposal for a DISH-RI (10 partners, 6 countries). The latter was the incentive to extend the DISH-concept (concerning nutrition and behaviour) to the food supply and public health systems. In the subsequent RICHFIELDS project (16 partners, 11 countries), we developed and tested the design for the e-infrastructure presented here. FNH-RI builds on experience in the EuroFIR platform, ENPADASI (JPI-HDHL project), Quisper, and a variety of public-private research initiatives, such as GfK, ICA (food procurement), Statistics Denmark and the City of Göteborg. Feedback from data science experts from CORBEL, ELIXIR, METROFOOD, ECRIN, and EuroFIR have been instrumental in guiding interoperability of FNH-RI with adjacent RIs. Strategic feedback was also obtained via the emerging national nodes, and extensive scientific and policy feedback via symposia and congresses, such as FAO/WHO (Budapest, 2017; Copenhagen 2019), FoodNEXUS (2018), CAPNUTRA (2018), FENS (2015, 2019), EAAE (2015, 2018), UKRI (2019), and OECD (2019). FNH-RI is organized with a central hub located in Wageningen, Netherlands, with national nodes uniting researchers within their respective countries. As of December 2019, 11 national nodes have been established (90+ institutes), and partnerships have been developed with 10 additional EU countries (50+ institutes) and two countries outside Europe (Canada and Australia).

User community needs were assessed via previous design studies EuroDISH and RICHFIELDS, the latter of which performed a user needs survey for an RI in the field. These two projects involved the convention of ca. 100 research community members for symposia to inform the present FNH-RI submission. These groups highlighted the lack of a) data interoperability and standards and the expertise required, and b) data accessibility and access to relevant meta-data, including up-to-date information on research output. Additionally, in-depth qualitative focus groups with young students and researchers were conducted at Wageningen to elicit forward-thinking directions of research and potential innovations in the FNH domain, relevant for the RI's operational phase (2030 and beyond).

As designed in RICHFIELDS, FNH-RI develops a novel e-infrastructure, the Citizen Data Platform (CDP), to harvest real-life data on the determinants of food choice within different food environments, feeding back information to citizens to aid healthy and sustainable choices. A broad panel of EU citizens across regions and socio-economic backgrounds, the CDP will allow the advanced modeling of the relationship between food environment and choice. It will also enable unprecedented testing and modeling of proposed policy or public health interventions via advanced data science techniques (e.g. modeling of complex adaptive systems).

3. The challenges of the past

The ESFRI 2018 strategy report mentions 16 RIs in the Health and Food domain. However, the nutritional and social domain remain disconnected, and none of the existing RIs addresses the food environment, consumer behaviour, eating patterns, personal, public and planetary health in an integrated approach. Yet, this interface is necessary to the food systems transformation.

Three of these RIs address social issues (ESS, CESSDA, SHARE-ERIC), but they provide virtually no data on food, food consumer behaviour or nutrition. Four RIs are close to FNH-RI, i.e. METROFOOD, ELIXIR, BBMRI and ECRIN. Like FNH-RI, METROFOOD is situated between 'food' and 'health', but focuses on metrology and food safety rather than consumers, health and environment. BBMRI (biobanking), ECRIN (clinical trials) and ELIXIR (molecular biology) follow a biomedical paradigm, without a focus on behaviour of apparently healthy consumers nor consumer behaviour as such. FNH-RI will ensure synergy with these RIs to build on their expertise on food supply, clinical and big data. Thus, in the Health and Food domain, there are no multi- or transdisciplinary infrastructures and services that adequately serve the scientific needs for transnational evidence and breakthrough research for the transition to healthy and sustainable diets for the 21st century.

4. The vision towards 2025

FNH-RI unites research fields currently operating in isolation, offering users three unique pillars: Data-, FACT- (research FACilities & Tools) and TED- (Training, Education & Dissemination) services. In the short term, these services:

- Data catalogues – identifying and describing commercial, public and research/academic (meta)data, its provenance, best practice documents and data management protocols
- Research protocols – development, capture and sharing of guidelines and best practice protocols for the use of (linked) big data
- Thesauri/Vocabularies – to minimize semantic ambiguity by ensuring uniformity and consistency in the storage and retrieval of content objects in relation to FNH
- Ontologies/semantic models – encompassing representations, formal naming and definitions of the categories, properties and relations between the concepts, data and entities that substantiate various domains of discourse in FNH (e.g. StandFood, FoodEx2).
- Training & consulting services – to assist both users and data providers to improve data quality, in particular how to make data FAIR
- Catalogue of (lab) facilities – facilitates access to and sharing of labs and physical research facilities
- Micro-lab with remote access to micro-datasets with anonymized personal data of individuals, such as the Citizen Data Platform, which requires stricter access policies
- Software tools for advanced searching, data exploration, dashboards and advanced data analytics (e.g. AI, ML).

FNH-RI will cohere fragmented European research initiatives within the Health and Food domain, funded by its member states and the European Union: cutting edge science by cutting edge data and facilities.

Cutting edge data: In order to design a healthy and sustainable food system, there is an urgent need to combine and make interoperable data concerning food choices and food environments. Critically, data science as applied to nutrition science and consumer behaviour is still in its infancy, representing an unprecedented opportunity for food system innovation. FNH-RI provides a way of accessing currently scattered and non-interoperable data, allowing it to be used for advanced research questions and modelling of interventions. In the initial stage, FNH-RI will organise and allow access (FAIR-ify) to data and modeling tools for researchers and policy-makers alike, facilitating access to groups performing cutting edge data science, integrated with the European Open Science Cloud. In the mature phase, data services within the FNH-RI consortium will incorporate machine learning as applied to food systems data. This will enable significant progress towards the three critical societal challenges: reducing obesity and non-communicable disease, improving the sustainability of diets, and engaging and empowering citizens.

Cutting edge facilities: Through the consortium, FNH-RI will allow access to cutting edge facilities measuring food choice and consumer behaviour. Facilities already present among consortium members include observation/virtual supermarkets and restaurants; sensory labs, mood rooms and brain imaging facilities to assess food choice; and advanced technology facilities for food reformulation.

5. The most important next step for our community in 2020

Extend the FNH-RI community:

- The envisaged final stakeholder community of FNH-RI consists of nodes from each EU member state that are representing the user community, primarily consisting of researchers that study the behaviour of consumers in relation to food, sustainability, lifestyle, nutrition and health domain (see 4.1). Commitment is defined by financial contribution of each nodes, preferably from the designated ministries like the current founding members. Besides, we will reach out to nodes outside the EU to attract and involve excellent research in the FNH domain from outside Europe (e.g. Canada, Australia). Finally, the stakeholder community will also include international public bodies such as WHO, FAO, JPIs and industry associations such as FoodDrinkEurope and GS1.

- The philosophy of FNH-RI is based on an open-access approach to valid, well traceable information, data, services and research infrastructure to any regular member of FNH-RI and joined authorities, with an exception for these, reserved or excluded from this regime due to specific reasons (third party rights protection, etc.). On this basis, the following user strategy is defined:
  - removing/reduction of the barriers in communication, scientific data access and information exchange across Europe;
  - creating national nodes and effective access points for physical/virtual communication and infrastructure sharing;
  - spreading out the philosophy of food, nutrition and health to a wider spectrum of users (specific or general society) via training, education and dissemination activities;
  - relevant real-time information and data sharing on aspects of food quality, safety, life style, human well-being, active ageing to support the target group of users;
  - effective training activities via the shared e-learning platform and/or shared research infrastructure;
  - effective, transparent and efficient networking of the foreground scientific, industrial and educational institutions throughout the European food research, innovation and business area for continuous evaluation and update of the offered data, information and/or research infrastructure.
- The links with RI in the food system (e.g. METROFOOD, E-ROSA) and in the health system (e.g. BBMRI, ECRIN, ELIXIR) as well as the social sciences in general are essential. The Food Nutrition Security (FNS) cloud, currently developed under EOSC, will be an important part of FNH-RI.

Harmonising the research:

- The European research base in nutrition and food science is very well developed but highly fragmented. FNH-RI will collect the relevant data from all the available sources, creating thus an effective platform for top-level pan-European research on the health preventive aspects of food and nutrition as well as sustainable agriculture and bio-economy.
- The methodology of data collection will be based on a voluntary basis of the joined institutions and joined/associated subjects. Analysis tools will work with FAIR data using interoperability approaches to build the interconnections. The data will pass the standardization process (unification of format, harmonisation, etc.). Collaboration with OpenAIRE could be foreseen.

6. Closing remarks

FNH-RI is the necessary and appropriate infrastructure to provide scientific guidance towards realistic, healthy and sustainable diet solutions for 21st century citizens. The added value of FNH-RI lies in (i) unlocking existing research data by connecting data-platforms and exploring big data (DATA). In the initial phase this will focus on ontologies, vocabularies, standardisation and meta-data, being both technically FAIR and GDPR-proof, (ii) developing an ecosystem of research facilities and tools (FACT) and (iii) sharing expertise from top-level bio-psycho-social laboratories to the research community and society (TED). In conclusion, the added value of the FNH-RI lies in the integration of business- and consumer-generated data, the scientific validity of the data and services, transparency of data (data provenance), and a high-level security and protection of personal data in the area of consumer behaviour.

C2 Position Statement from *METROFOOD - Infrastructure for Promoting Metrology in Food and Nutrition (METROFOOD-RI)*

Authors:

Claudia Zoani, ENEA - Italian National Agency for New Technologies, Energy and Sustainable Economic Development; Dept. for Sustainability – Biotechnology and Agroindustry Division (SSPT-BIOAG), Italy - METROFOOD-RI Coordinator, claudia.zoani@enea.it - claudia.zoani@metrofood.eu

Karl Presser, Premotec GmbH, Switzerland, karl.presser@premotec.ch

1. Digital science infrastructures as collaborative, community-driven resources

The agrifood sector is a strategic asset of all European Countries and one of the largest and most important economic sectors, with particular social relevance: it is vital to ensure employment, preserve rural public goods, supply healthy and quality food, and facilitate the integration of SMEs into the international food chain. High-quality data on the food chain are fundamental to populate the expanding data technologies with useful contents and, according to the FAIR principles, enable advanced research on food and food metrology. METROFOOD-RI addresses these themes through its mission to enhance the quality and reliability of measurement results and make available and share data, information and metrological tools, in order to enhance scientific excellence in the field of food quality and safety and strengthen scientific knowledge, also promoting scientific cooperation and integration.

Based on its trans-disciplinary approach and considering the wide potential impact, METROFOOD-RI, belonging to the Health & Food domain, interconnects with other ESFRI domains: Energy (agriculture and land-use change, food and non-food systems; bioenergy and bio-renewables - trade-offs between food supply, energy supply, biodiversity and ecosystem services), Environment (agriculture and land-use change, food and non-food systems; health, environment and air/water/food pollution), Physical Sciences & Engineering (New devices and technologies to exploit alternative sources of food; advanced digital solutions for food production and waste management), Social & Cultural Innovation (better use of longitudinal studies and cohorts in complex disease studies, e.g. obesity and food demand and consumption; agriculture and land-use change, food and non-food systems), Computing and Digital Research Infrastructures (collection, dissemination and sharing of information about principles, terminology, tools; harmonization and integration of food composition databases; development of new standardized tools for food quality, safety & authenticity).

METROFOOD-RI is a physically and electronically distributed infrastructure with many partners across Europe and beyond. This distributed architecture allows us to develop the infrastructure in a collaborative way where each partner takes care of its part and enlarges the infrastructure with an additional puzzle piece. Our distributed architecture also revealed how little knowledge is available about existing physical and electronic means, data, information and services. METROFOOD-RI emerged from a community-driven need for harmonisation and standardisation in food analysis and to make such data publicly available. Many, if not all of the partners, were already providing laboratory and electronic services to science, public and policy makers. The building of a consortium to bundle forces and reduce redundancy was a logical step. METROFOOD-RI therefore clearly states that the development of a digital science infrastructure is important, relevant and possible as a collaborative and community-driven resource.

2. Major achievements towards an open data ecosystem for science

In the past 3 years many initiatives on national, European and international level contributed to an open data ecosystem. First of all, the EU commission spread the FAIR principle and the notion of open science. This spreading made researcher, policy makers and the public aware that public funded data should be publicly available. The EOSC initiative was a part of the EC to push the electronic part of infrastructures forward and to produce concepts, software and other components. The Research Infrastructures currently supported by ESFRI are other major achievements for the open science, where whole communities collaborate to open facilities, data and information to the research community. Many of these infrastructures like GEANT, ZENODO or EUDAT are operational and help the thematic infrastructure on the electronic part. But also, international, national and institutional initiatives must be mentioned, where many governmental, scientific and academic efforts have been undertaken to make publications, data and information publicly available. Another major achievement is that several open access journals evolved, and institutions started and continued to support them. Open access is a precondition to foster research and innovation by making knowledge and data accessible.

3. The challenges of the past

Despite the tendency – especially inside the EC – to increase the transparency especially for what concern food safety and risk assessment, there are still several barriers. There are still many fears also in making data FAIR, and this is related to the difficult trade-off between transparency and privacy and the reservations of governments about publicity. The concept of “open data” and “data reuse” is closely related to basic questions of ownership, responsibility, and control. Attitudes towards data sharing are generally positive, but open data and data reuse is not yet a reality for most researchers, and research result sharing still mainly occurs through the current publishing system, or direct sharing, or does not get shared at all. Research data management and privacy issues, proprietary aspects, and ethics are the main barriers, to which add ethical and cultural limitations, financial and legal issues. Furthermore, difficulties related to Intellectual Property Rights and data publicity must be taken into account, especially with reference to data from industry or related to research applied to industrial samples. In addition, the last framework programmes supported projects on a temporary basis without helping to maintain systems, software and other components for long-term. Developed database systems and data were mainly created only for the duration of the projects or as deliverables, and no incentives were available to run these systems for long-term. Furthermore, the pressure to publish in the academic community does not support that young researchers take care about long-term preservation of research results or any developed systems.

4. The vision towards 2025

In Food & Agriculture, digital tools built around environmental & food quality & safety continue to progress and be introduced in this system with the advent of the awareness of open science, artificial intelligence and other technological innovations. This implies the need to develop innovative solutions to face these challenges and reflect on the EU Food & Agriculture system of the future, meeting the needs driven by all the different actors, mainly consumers, producers and food industry in general, and public health agencies. The IFPRI Global Food Policy Report 2018 report shows that inequality in access and use of knowledge is increasing and that current actions and commitments are “uneven”. Data management services need to move beyond repositories to enable data-driven research and provide value to users.

In the near future, it would be necessary to foster new schemes to guarantee open access to the available resources (i.e. data, inventories, individual roadmaps, projects outputs, laboratories) by developing models and tools to enable easy and quick access to relevant resources. The increasing focus on open science and open access represents a key to acquire novel knowledge and distribute it among scientists and the public without any limitation. This particularly includes the access to scientific publications deriving from publicly funded research by FPs (H2020 currently and then Horizon Europe). The application and further promotion of the FAIR approach is crucial, together with the promotion of strategies and tools for data management. Data represents in fact a mean for interdisciplinarity, but for its practical realisation common data formats and standards must be ensured. Data and products should be re-used in different scientific contexts and data from different disciplines should be aggregated, integrated and made interoperable: this represents one of the current major challenges. Further opportunities would arise from the development of a new platform enabling the knowledge and data exchange by setting up a systemic approach putting together all the different actors of the agrifood system (researchers, policy makers, food & environmental inspection agencies, Food Business Operators, and consumers), allowing them to use digital tools to facilitate their interaction, and giving them an easy access to data and knowledge related to Food & Agriculture research (e.g. with reference to exposure to multiple chemicals, eutrophication, climate changes, pesticides & pollinators, sustainability and circular economy, the green deal, increasing concentration in the supply chain, changing in diet trends, microplastics, vector-borne diseases, global trade flows and online trading, predictive models, food quality and nutrition, nutraceuticals, authenticity and traceability).

A key aspect is to develop a really integrated system of Research Infrastructures and networks in the area of food and agriculture as well as in the electronic part for data publishing. At the moment, different projects are supported and develop solutions, but their alignment and interconnection are rather limited. What we need to achieve is a long-term and final construction plan for a common data ecosystem, so that every project can contribute to the final big picture similarly like many space missions helped to enlarge the International Space Station (ISS). This is an ambitious goal and a first step could be to support communications and exchange between the different data system partners.

5. The most important next step for our community in 2020

METROFOOD-RI is actively engaged not only in the definition of links at different levels - with European Initiatives (e.g. Eurofir, FNS-Cloud, FNH, LifeWhatch, etc.), Scientific Communities, and Food Business Operators and other actors engaged in data collection, sharing, use & re-use, and interoperability, but also in making these links practical and really useful for the different actors of the agrifood secots, the Health & Food domain and other cross-linked domains.As many projects are on a journey they have never done before, it would be really helpful to have opportunities to exchange experiences and get inspiration from each other. This can also be seen as spreading and exploiting knowledge and is a crowd approach to develop long-term construction plans. In addition, the science & data community could start to publish experiences in providing digital data services. The community can also start to create a landscape for the area of food and agriculture to see which project, research infrastructure and institutions are working in which areas, and have produced what contribution for the open data ecosystem. All that can really support the practical realisation of the FAIR approach and the actual realisation of the above-mentioned integrated system of Research Infrastructures and networks in support to Food, Agriculture & Health.

6. Closing remarks

A really relevant opportunity rising from the development of digital infrastructures applying science and research on different disciplines is the possibility to integrate several scientific communities each other, and with the ICT community, finally allowing a real knowledge exchange and the development of a multicultural interactive system. The realization of an integrated landscape of initiatives/infrastructures/networks should be promoted at any level: global, European, regional and local. Close cooperation with other RIs and with e-Infrastructures within the EOSC will increase the capability of RIs to combine and integrate data and resources in a common environment. In particular, the cooperation of RIs in the frame of the EOSC can enhance their ability to share resources and data, increase flexibility and favour the early monitoring of the budget. Furthermore, it can allow the enhancement of the portfolio of the services they offer to the respective user communities. Therefore, it would be strategic to support – also with dedicated funds – local, national and regional initiatives that build on existing (e-)RIs or aimed at creating new ones where there are gaps, as well as to ensure their interoperability. The links with S3 can further promote the realization of this integrated landscape, especially for what concern the regional and local level; this can be sought both specifically in the regions of the involved countries and with the thematic platforms (the "Traceability and Big Data" one in particular).

The agrifood sector represents one of the most challenging and promising sectors for data sharing and integration, where really great opportunities can arise from the successful development of proper solution for data interoperability. This can be further improved by the relations with data related to the Environment, health, and new materials and technologies, which in turn can be linked to the agrifood. Many progresses have been made in defining the parameters mainly influencing food quality & safety, especially for what concerns industrial processing. However, no system considering in a systematic and holistic way the potential influencing factors all along the food chain, from primary production to final consumption, is available. The influencing variables can vary very widely, concerning sectors even completely different (e.g. the Environment, climate, varieties, agronomic technologies, use of natural resources, processing and storage technologies, etc.). Often, the data are available, but are completely disconnected from one to another and there is no possibility to make them interoperable, as well as there is no awareness of their availability. Instead the possibility to put in relation all these variables with the characteristics of the food (nutritional, functional, etc.) could greatly favour innovation, competitiveness, positioning on the market, etc. Furthermore, the availability of all such types of data also allows the promotion of sustainability, closure of cycles and the reduction of food losses, as well as to promote the integration of technologies and the spreading of knowledge and best practices.

C3 Position Statement from OpenAIRE

Authors: Elli Papadopoulou, Najla Rettberg, Natalia Manola; OpenAIRE

1. Digital science infrastructures as collaborative, community-driven resources

OpenAIRE, the Open Access Infrastructure for Research in Europe, sees the development of collaborative and community-driven digital science infrastructures as **important**, **relevant** and **possible** but, most importantly, as the only way to realize a global data commons where publicly funded resources are available, re-usable and exploitable, among others, through TDM solutions.

Important	Relevant	Possible
In order to more effectively respond to growing demands of both funders and researchers, our efforts in developing infrastructures should be communicated with others and run complementary with existing endeavours to ensure growth and harmonisation of practices. Scholarly Communication should be further realised and promoted as the means to achieve that. These infrastructures should be openly governed, free to use and based on open standards and the collection of open and shareable research data and research output.	The introduction of FAIR principles has led to standardisation of processes being at the top of data infrastructures priorities that are now leaning towards adopting or building new services that, to some extent, ensure FAIRness of data. For open data, in particular, investing in a collaborative, transparent, community-driven ecosystem supports, among others, one of its foundational principles, that of “the free flow of data”.	For years now, OpenAIRE’s socio-technical network for repositories interoperability has heavily relied on collaboration. OpenAIRE facilitates national service providers/ repository managers in complying with its metadata schema, a task that requires mutual effort, based on commonly agreed standards. Also, the National Open Access Desks (NOADs) sometimes undertake initiatives outside their work plan so as to ensure that specific collaborations are feasible and accelerate progress in their countries. The aforementioned work is achieved by utilising own resources thus avoiding applications for new funding. This proves that, in certain occasions, it is possible to develop and/ or adapt to new standards through in-kind contributions. We should all explore these opportunities and contribute to building a flexible and sustainable Open Science framework where collaboration is its merit.

2. Major achievements towards an open data ecosystem for science

The evolution in Open Access has lead to standardisation of practices and policies in Europe and beyond, with remarkable paradigms such as PlanS and AmeliCA. It also led to crucial developments in open and FAIR data management practices (FAIR principles) and policies which has gained momentum under the Open Science paradigm.

- **At European level**

"The European Open Science Cloud is a great and timely opportunity for coordinating our efforts to unleash Europe's potential into making good use and adding value to research data. Collaboration and openness should be in the center to make EOSC work. Open Science - and particularly Open Access to scientific content (publications, data, software) - is now becoming all the more relevant, and we are pleased to see our work in OpenAIRE for the past 10 years fill this gap in EOSC." - *Natalia Manola, OpenAIRE Managing Director*

As already mentioned, 2017 marked the start of the implementation of the European Cloud Initiative (EOSC) which is an ambitious platform to facilitate the access and reuse of research data. This development arose from the culmination of efforts spanning from new policy directions (New Recommendation on access to and preservation of scientific information, General Data Protection Regulation, New Public Sector Information, etc) and has so far funded a number of projects, from the EOSCpilot to a number of follow-up and interdependent projects, namely EOSC-Hub, EOSC Secretariat, and the regional EOSC5 projects. EOSC has provided an exceptional opportunity for a wide range/ diverse set of stakeholders to come together and collaborate towards its realisation and sustainability at all levels to provide a more coherent set of resources to harness public scientific output.

- **Internationally**

EOSC has become a big source of influence for other regions whose interests to invest in Open Science and Open Innovation has significantly grown, with the most recent being the launch of the Australasian infrastructure for Open Science. On the globalisation of Digital and Open Science Commons, a connection with the United Nations has been made to identify the intersection between Open Science and the Sustainable Goals for Development (SDGs).

- **In OpenAIRE**

OpenAIRE has proudly served the European Commission's strategies for Open Access (OA) for 10 years now and is considered to be one of the pillars of EOSC. Among successful efforts of the OpenAIRE team is the FP7 post-grant open access pilot report which was compiled to address the impact of open access publishing during the pilot by providing statistics that highlight adoption and best practices between journals and institutions while revealing perceptions of researchers for complying to the OA pilot. In addition, the Community of Practice (CoP) that was recently set up, is comprised of almost 50 organisations and initiatives with shared objectives to building capacity on data and Open Science in Europe. NOADs contribute highly in aligning national practices with European requirements through awareness-raising and training activities that cultivate data and Open Science literacy and make a smooth transition to the new model of scientific conduct. Moreover, NOADs drive the creation of national bottom-up Open Science communities that tackle major issues concerning possible adoption of Open Science in their countries focusing on their resolution at the policy and infrastructure level. They are also responsible for communicating new developments to key national scientific leaders and funders.

OpenAIRE's work also has a global effect. The launch of the OpenAIRE Research Graph at the end of 2019 was the culmination of years of effort to harness the wealth of scholarly communication resources and aims to identify the impact of Open Science in the global academic and research community by applying semantics and speeding up the process of data exploitation through Scholarly Communication. Information contextualisation enables potentials that may lead to new models on Open Access to be realised, find gaps in the scientific bibliography that were not prominent otherwise, show new ways of collaborating and many more. Additionally, collaborations with La Referencia, KISTI and other core scientific infrastructures and service providers ensure that this activity is not one-sided, moreover it expands to practices overseas.

3. The challenges of the past

Of course the open access movement has come a long way and Open Science is the new narrative however more needs to be done to implement policies within institutions and have leaders who fully support new paradigms of doing Open Science (e.g new ways of evaluation, open peer review, pre-print submission etc). In the last years there have been significant efforts to join up the constituent parts of the scholarly ecosystem, especially metrics and bibliographic information which at present is reliant on publishers and subscription services. A level playing field will occur when we can build truly open and free infrastructures being served metadata directly from libraries not at a price but for free and from trustworthy sources. This could have moved faster.

4. The vision towards 2025

Keeping up with the Open Science evolution, OpenAIRE believes that the next priorities which need to be set up and tackled within the next 5 years are the following:

1. **A global open research commons**, i.e., a global linked interoperable infrastructure for open science, where we can share and discover in context all research results. One of the next steps is to avoid silos at the global scale and avoid the development of fragmented services and repositories. They should all be linked up and searchable. Initiatives such as the EOSC, the African Open Science Platform or the Australasian paradigm, all aim at creating a focal point that assumes no barriers for the use and access to scientific information and research services in their regions.

   Collaboration between those initiatives should start immediately.

2. **An open metrics framework** is essential to help us transparently measure resources (citations, usage and costs) and should be positioned in the heart of delivering open science (enriching reward systems).

   OpenAIRE is committed to monitoring Open Access by making use of and promoting open practices and by establishing collaborations that aim to realize the uptake and impact of Open Science.

3. **A shared Open Science policy and legal framework** which dictates and applies the rules of how data elements are published, shared and re-used.

   Open Science policies need to be consistent, standardised and machine readable. Legal support in copyright clearance, ownership etc should be at the core of infrastructures' and institutions' helpdesks while more centralized support should be provided through common mechanisms and national offices in individual countries'.

4. **A training infrastructure** that backs up the cultural shift to more responsible research and innovation practices and eases the use and exploitation of services and data. Currently, training misses the constituents for capacity building of both professionals who are called to support the new scientific model but also for researchers that are expected to conform to new conditions.

> There is a need to develop a training infrastructure for Open Science that assumes all roles and capacities for data intensive activities at any moment of the research lifecycle or individual's career.

5. The most important next step for our community in 2020

Researchers should be at the core of our activities, which has not been the case until today. Services should be user-driven, developed in order to better researchers' everyday work. Infrastructures should seek for new ways to collaborate with others, especially on providing the skills and training necessary for their services to be widely utilised and for researchers to thrive in Open Science.

C4 Position Statement from the H2020 Project *FNS-Cloud* (Food and Nutrition Security Cloud)

Authors:

Karl Presser, Premotec GmbH, Switzerland, karl.presser@premotec.ch

Paul Finglas, QIB, Quadram Institute, UK, EuroFIR AISBL, Belgium, paul.finglas@quadram.ac.uk

Sian Astley, EuroFIR AISBL, Belgium, sa@eurofir.org

Claudia Zoani, ENEA - Italian National Agency for New Technologies, Energy and Sustainable Economic Development; Dept. for Sustainability – Biotechnology and Agroindustry Division (SSPT-BIOAG), Italy - METROFOOD-RI Coordinator, claudia.zoani@enea.it

1. Digital science infrastructures as collaborative, community-driven resources

FNS-Cloud will launch a first-generation 'food cloud', federating existing and emerging datasets, and it will develop new services to support re-use by researchers. Existing FNS resources (data, knowledge and tools) for health and agri-food sciences are fragmented and they lack critical mass, and access by user communities is 'unevenly' distributed. This means data are not readily found, accessible, interoperable or reusable, and existing services focus on clinical, molecular or biological sciences.

FNS-Cloud is a thematic cloud in the European Open Science Cloud (EOSC) initiative on the topic of food and nutrition security, and it will integrate within the EOSC framework. The project should implement demonstrators to show its usability and feasibility to the research community.

FNS-Cloud considers the development of digital science infrastructures as collaborative and community-driven resources to be important, relevant and possible. It therefore applied and started its project to develop a first-generation digital cloud or infrastructure in the area of food and nutrition security. Many partners of the project are involved in existing and emerging research infrastructures and therefore fully support the idea and concept of digital infrastructures and contribute to building the 'food cloud'. The number of project partners (36 partners) and the estimated number of datasets that these partners bring to the project (72 datasets, exact number is under evaluation) show that collaboration among all partners is necessary to build one digital infrastructure. Our demonstrators were planned with the research communities and the national and European research agendas and built to be useful to the community. The consortium is convinced that a digital infrastructure, as a service provider for the research community, can only survive and reach acceptance if services are developed in a community-driven approach.

2. Major achievements towards an open data ecosystem for science

One of the FNS-Cloud partners is EuroFIR AISBL, a non-profit organisation that arose from the EuroFIR and EuroFIR Nexus projects. EuroFIR was one of many Network of Excellence (NoE) projects that were defined as a network of ‘institutions willing to combine and functionally integrate a substantial part of their activities and capacities in a given field, in order to create a European “virtual research centre” in this field’[1]. Therefore, NoEs must be considered pre-runners or even full equivalents to research infrastructures. The EU framework to support such NoEs was a major achievement to move the community in the direction of digital and physical infrastructures.

The EU initiatives EOSC and ESFRI are two major achievements in the past three years for digital data ecosystems. EOSC aims to provide European researchers and professionals in science with a virtual environment and digital services, such as storage, computational power and making data FAIR.

EOSC helps existing and emerging digital infrastructures with knowledge and solutions that are developed in different projects and that connect experts in the digital area to exchange experiences and inspire each other.

ESFRI directly supports research infrastructures and is likely the biggest achievement in Europe because it is designed for the long-term support of research infrastructures and it requests commitments from countries. Funding goes directly into mature infrastructures, and the network helps existing and new infrastructures to learn, exchange experiences and share parts of the infrastructure. The new master programme to manage and build research infrastructures is another great achievement in ESFRI.

The crowd-sourced initiative, Research Data Alliance (RDA), taking care of many aspects of digital data, is another major achievement. Experts, mostly from ICT but working with data from different research topics, try to harmonise and standardise certain aspects in data management, data exchange and data usage. It is not surprising that the EOSC, ESFRI and RDA began collaborating, as digital infrastructures will become one of the biggest topics in the next years.

3. The challenges of the past

The answer above shows that many things were done correctly to lead the research community in the right direction. However, the balance between exploitation and exploration is a challenge in every scientific area. Exploitation involves the re-use of existing knowledge and experience from NoEs and other infrastructures. Many of the infrastructure projects must solve the same problems, e.g. implement an Authentication and Authorisation Infrastructure (AAI), but general guidance is hard to find. One reason is that implementation work is often not considered research work, resulting in fewer publications, and another reason is that no central documentation or best practise library was created.

Another lost opportunity in many EU projects was that many generated datasets, data management systems and data tools were not designed for long-term use because no incentives were available to make data and tools publicly available after the project finished.

4. The vision towards 2025

Until 2025, we should try to overcome lost opportunities. A central digital library with basic information about digital infrastructures should be created and maintained. Documentation about options for data access policies or distributed infrastructure architectures should be publicly available.

Research should not only get reputational rewards for published papers but also for implemented data systems and datasets made FAIR. This is in line with conceptually separate data producing researchers and data using researchers. While the first group focuses on producing data, the second group focuses on using data to produce knowledge. More focused and experienced researchers would result in better data and knowledge quality. The production of data is often considered minor scientific work because it contains many repetitive actions. If this view can be changed, more investment will be spent on making data FAIR, and a Ph.D. position in digital infrastructures will become normal.

5. The most important next step for our community in 2020

FNS-Cloud sees two steps that can be taken in one year. One is an overview of data types in the agri-food sector and the other is to connect the different infrastructures in the agri-food area better.

The overview should provide a map of the different research areas in the agri-food sector and the possible datasets for each of these areas. In a second step, available datasets should be added to provide an overview of where datasets are available, as well as where datasets are missing. The second step is to organise some common meetings between the infrastructures in the agri-food sector and to discuss specific topics, including architecture or data access policies. Such workshops would strengthen collaboration and support inspiration and exploitation.

6.Closing remarks

Surprisingly, EOSC was planned for only two years, and it should finish this year. Discussions are ongoing about its relation to ESFRI, including how it should continue. Our experiences show that building an infrastructure, running it and maintaining it are not tasks able to be completed in four years. More time is needed, not only because of technical, legal and administrational challenges, but also because the mind-sets of many data owners must change before being able to make data FAIR. We hope there will be more initiatives like EOSC supporting bigger but also smaller infrastructures and that membership constraints can become more flexible so that not only countries can become RI members but also institutes, companies and even individuals.

---

[1] EU Commission FP7 website, **https://ec.europa.eu/research/fp7/understanding/fp7inbrief/funding-schemes_en.html**

SECTION D: POSITION STATEMENTS FROM INTERNATIONAL STAKEHOLDERS

D1 Position Statement from Syngenta

Author: Graham Mullier
IT Lead Business Sustainability and Open Publishing, CP R&D IT, Syngenta; graham.mullier@syngenta.com

1. Digital science infrastructures as collaborative, community-driven resources
Syngenta is a leading company in the agriculture sector, providing growers with tools and solutions they can use to help achieve global food security for a growing population, sustainably. At the heart of our business are innovation, ethical operations and a collaborative mindset. Our innovation enables farmers – from smallholdings to large-scale farms – to safely produce food, feed and other plant-based products as efficiently as possible, without using more natural resources. We take a long-term approach to developing technologies that will not just meet demand in the years to come, but will do so without depleting already overstretched resources.
Climate change, population growth and an increasing demand for more and better food pose structural challenges to global food and agriculture systems. These challenges require a step-change in how we approach global food supply; our growing world population requires more and more efficient food production, with less waste across connected value chains. This grows increasingly difficult in the context of climate change – which is already causing widespread disruption to natural and agricultural ecosystems – requiring agriculture to become more resilient to its effects.

As Syngenta looks at how to respond to these challenges, we see a number of gaps:

- The supply of well-trained hardware and software engineers and ag-tech specialists is of great importance in ensuring the long-term sustainability of Syngenta's research. Digital technology in agriculture is a relatively new area and the industry could find itself with a deficit of skilled people over the coming years, particularly as the agriculture sector is often perceived as a rural or low-tech area with little to attract the next generation of technical specialists.
- Weed and pest control will be heavily influenced by precision delivery, visioning systems and targeted applications but there is little joined-up cross-discipline working to develop commercially viable systems which target these farming problems. Related to this, improved digital connectivity for farms and rural areas is probably a precondition to successfully step up the digitisation of agriculture.
- Agriculture needs to become more sustainable but the agronomic practices that support this are patchy and largely disconnected from sustainable business models. With the right systems in place, agriculture could even exceed the goal of being carbon-neutral, instead serving as a carbon sink by removing more carbon dioxide from the atmosphere than the sector emits.
- Although many organisations play a role in the agriculture and food sector, the strongest links are historical, with poor interconnections in emerging areas relevant to our key challenges. This gap does point to an opportunity: to apply significant resources to problems of industrial relevance and to promote collaboration between research and industrial partners with the potential to provide a step-change in agribusiness, both farmer-facing and downstream in the food value chain ("field to fork").

As part of Syngenta's response to these challenges we are building more capacity wherever we operate, working with both farmers and communities. We are sharing more information about the benefits, safety and risk of our products, as well as gathering and sharing data about their performance on farms. We also collect and share farm data while respecting business ethics and our obligations around data privacy and confidentiality. We see the value of shared data and knowledge, and to that end, we encourage the sharing of open data to make information about agriculture and nutrition more readily available, accessible and usable. We promote multi-stakeholder platforms that mobilize the sharing of knowledge and technology, across the entire agriculture and food chain, to help advance agricultural production systems. We also engage openly and collaboratively with governments and regulators, acting as advocates for innovation in agricultural technology and stressing the adoption of best practices in these regards.

Examples of our approach to collaboration and information sharing can be found at Syngenta.com, particularly the commitment to transparency, the Principles for Sustainable and Responsible Agriculture, and the Good Growth Plan, with outcomes published every year since 2015 at data.syngenta.com.

- www.syngenta.com
- https://www.syngenta.com/site-services/transparency
- https://www.syngenta.com/media/corporate-publications
    - Principles for Sustainable and Responsible Agriculture
- https://www.syngenta.com/what-we-do/the-good-growth-plan
- data.syngenta.com

2. Major achievements towards an open data ecosystem for science

The last three years show what is possible when organisations focus on a challenge and take advantage of developments happening outside of their primary business focus, especially those in computing and data technology. The role of data science is no longer new or unusual in sectors such as agriculture, where it is beginning to be employed on a large scale. Key frameworks for the creation, stewardship, handling and use of data, including the data ethics canvas and the concept of data trusts (developed by the ODI); FAIR data; roadmaps (particularly that from the e-ROSA project); and the beginnings of usable semantic connections between well-known vocabularies and identifiers, as exemplified by the Global Agricultural Concept Scheme (GACS). Also important is the development of an increased range of innovative digital agriculture apps by start-ups and existing organisations, putting new science in the hands of farmers and others in the food supply chain.

3. The challenges of the past

Hindsight is a wonderful thing but rather than pick out examples of “failures” we suggest that the experiments we have seen have given us an opportunity to learn about the complex relationship between digital science and ag-tech, seeing how this could be used for the good of farmers and the food sector. One weak point these highlights is the connection between academic and commercial research, where there is room to make more impactful links between the two to establish a more effective, efficient and sustainable data ecosystem.

4. The vision towards 2025

To address the challenges outlined earlier, we see a need to build a stronger chain of connections across the whole ‘technology readiness level’ scale, from early research to commercial. Together, we need to step beyond niche activities into more complex, interconnected and dynamic ecosystems of players. This will be challenging, but we believe it is vital to help address the needs of farmers in delivering climate-smart sustainable agriculture for the future.

5. The most important next step for our community in 2020

If we are to develop a dynamic, interconnected ecosystem then we should focus on the data we will be exchanging. For 2020, we believe the community needs to learn how to share data more impactfully – addressing the three facets of feasibility (can we do this?), desirability (does anyone want it?) and viability (should we do this and can it be sustained longer-term?). We suggest the community expands well-beyond university research groups by including larger commercial organisations such as Syngenta, together with small companies, start-ups and tech disruptors, as well as initiatives and consortia such as the Global Open

Data for Agriculture and Nutrition (GODAN) initiative and AgGateway and international bodies such as the FAO.

6. Closing remarks

Looking across the range of contributions to the discussion paper, some common themes and perspectives are worth calling out:

- Basic compute infrastructure has improved massively in its availability, scalability and service-oriented nature, such that it is a great enabler for digital science. However, there are still activities which are reinventing some of this infrastructure, perhaps because of the length of some projects in the area.
- Publishing data is useful but of limited value if it becomes an end in itself. Of more value are those activities which connect data together (the 'semantic data management' CNR mentions), and 'pull' data into valuable uses (as covered by WUR, for example).
- Demonstrating value with this connected and meaningful data is very important. Richly-described spatiotemporal datasets, that allow connections to real-world places and moments in time, make the digital science outputs more accessible and valuable to end users who are not specialist data scientists.
- Business model sustainability crops up in several contributions. Being able to do something simply because it is technically possible is unlikely to have a lasting impact. What is needed is a combination of that technical capability with a need (a 'pull'), and the way in which the activity or service can be sustained over time with a business model. Linking that value of the output or outcome back in to a community of digital scientists and others will help the whole data ecosystem of players; it could help the community become more self-sustaining by keeping people out of silos and opening up single-discipline peer networks.

In his introduction, Johannes Keizer is right to point to the early semantic work as being overambitious. However, connecting the meaning of data (semantics!) to challenges and solving them is still what this broader community is doing. Building that community, as he says, has been a very important outcome. Perhaps building a social ecosystem alongside the data ecosystem should be our next ambition.
Open and collaborative communication is crucial to creating shared value for Syngenta and our stakeholders, helping to support and accelerate the development of sustainable and responsible agricultural practices. Our collaboration will enable us to address a range of challenges more effectively, ranging from food security and safety, to climate change and biodiversity loss. A more robust and collaborative open science infrastructure for the agriculture sector could be key to achieving this.

D2 Position Statement from *Agriculture Information Institute, Chinese Academy of Agricultural Sciences (AAI CAAS)*

Authors: Prof. Xuefu Zhang
Agriculture Information Institute, Chinese Academy of Agricultural Sciences, zhangxuefu@caas.cn

1. Digital science infrastructures as collaborative, community-driven resources.

Significance

Digital science infrastructure is a prerequisite for the realization of digital communities. Digital community makes full use of the new generation of information technology, such as mobile Internet, Internet of things, cloud computing, big data and artificial intelligence, to provide a safe, comfortable, convenient and intelligent multi-functional production and living environment for community subjects, so as to form a new form and mode of social management and public service based on informatization and intelligence. It is of great significance to build a digital community.

Relevance
With the development of information infrastructure construction and digital network technology, digital community not only brings new experience to people's work and life, but also changes people's life concept and lifestyle. It can integrate and make full use of all kinds of information resources in the community, comprehensively communicate with residents, communities, e-government and e-commerce, provide all-round community management and services, realize remote and remote distributed management, so as to improve the quality of life of residents and promote the overall progress of society. The digital comprehensive service platform is a platform to deliver happiness to the common people. Like water and air, it is embedded in our daily life, food, housing and transportation, to solve food safety, doctor-patient conflicts, imbalance of educational resources, etc. an App, one button direct access, "good-looking, fun, easy to use", and "saving money, time and effort". Therefore, the digital community has a direct relationship with the improvement of everyone's social living standard.

Possibility
As the smallest unit of social organization, community plays an important role in economic development, social management and community security. Digital community construction is an essential foundation and an important part of digital city construction. The development of China's digital community has characteristics of very short time and fast speed.
Some large cities like Beijing, Guangzhou, Shanghai etc. in China have made gratifying achievements in the digital and smart transformation.

2. Major achievements towards an open data ecosystem for science
In the past three years, large urban communities in China are undergoing digital and intelligent transformation. The concepts of "Digital Earth", "digital city" and "digital community" have been widely recognized. Some cities have carried out pilot construction on the basis of exploration. Beijing, Shanghai, Guangzhou, Hangzhou, Xiamen and other large and medium-sized cities have formulated plans for the construction of "digital city" and "digital community", and others have carried out "intelligent community" Experimental projects such as energy community and smart home. Although the construction of digital community in China is generally later than that in foreign countries, with the rapid development of economy, the acceleration of urbanization, industrialization and informatization, and the popularization of communication, information, digital technology and Internet application, the development of digital community is very fast.

1. The encouragement and support of national policies, the actual needs of the society and digital and more and more matured smart community products and technologies are all promoting the construction and development of smart community.
2. In 2019, with the support of 5G, artificial intelligence, cloud computing, Internet of things and other emerging technologies, China's intelligent industry has developed rapidly. The major enterprises continue to explore the direction of digital transformation and intelligent way for their own development, and the technology innovation and application of intelligent enterprises continue to emerge.
3. According to the Digital China Index report (2019) released by Tencent Research Institute, 31 provinces, autonomous regions, municipalities directly under the Chinese central government and 351 cities in China have a good digital development trend.

3. The challenges of the past

At present, there are still some problems in the construction of digital smart community in China. For example, the community construction in China is generally in the primary stage, the level of infrastructure construction is uneven, and there is no unified planning and intensive comprehensive service platform; the community governance function needs to be improved, and the public service is few to use; and the residents' life is not intelligent enough; the sustainable operation and development model has not yet been developed.

Problem

1. Lack of sufficient intelligent equipment and financial support

At present, communities in China cover nearly 800 million people in the city, which creates the market potential of smart communities. According to the data of smart research consulting, the market scale of smart community in 2018 is about 392 billion yuan. In order to realize the full connection of all parts of the community, it is necessary to have enough intelligent terminal equipment and financial support. However, most enterprises lack sufficient funds except a few. There is lack of fixed capital support for infrastructure construction.

2. Need higher system application integration capability. The construction of the integrated application of the digital community requires a higher level of system application integration capability, which requires the seamless link and collaborative development of the equipment layer, data layer, transmission layer, business layer and user end. Therefore, in the construction of digital community, all parties need to explore together.

3. Lack of community information resources and inadequate service

Although some optical fibers are pulled to the door, the information resources and services are not in place, or the services provided are not complete. As a result, the overall application level of the network is still low.

In any case, smart community is an inevitable trend. It is a manifestation of the development of information society that families move towards intelligence and communities embrace technology. The healthy development of smart community industry in the future is worth looking forward to.

4. The vision towards 2025

The 2019 digital agriculture rural New Technology Outlook Forum proposed that we should accelerate the construction of the digital resource system of sky land integration, build the national and provincial agricultural rural big data center, promote the development and application of agricultural rural big data in the way of combination of blocks and strips, increase the efforts to tackle key technologies, promote the application and transformation of achievements, cultivate and expand the agricultural rural digital economy, and strengthen agricultural credit Interest service, with big data driven rural economic and social development of high quality.

China is taking "meeting the needs of property, government, residents and businesses" as the starting point and foothold, the new generation of information technology as the support, the information infrastructure construction as the guide and starting point, the comprehensive information data service platform as the core and key, the intelligent application as the basis and carrier, building a community ecosystem, and creating a comprehensive solution of "hardware + software + platform + service" Plan to build a real smart community.

(1) To enhance Community information infrastructure construction

The construction scheme of community information infrastructure adheres to the principle of synchronous planning, design and implementation and the concept of green development, adopts the scheme of macro micro integration, inside and outside collaborative integration, three-dimensional and multi scene coverage, integrates the information infrastructure with buildings and environment, meets the diverse communication needs of the community, creates a new experience of intelligent service, and builds "environment-friendly Resource saving "smart community.

(2) To further push Community construction of comprehensive information and data service platform

The service platform shall at least have the following characteristics: intensive, modular and diversified three-dimensional presentation can be realized through the display of large property screen, application system view, mobile terminal, app, wechat, etc. to strengthen the supply of scientific and technological innovation in agricultural and rural areas.

(3) To promote advanced technology application

Mobile Internet, Internet of things, cloud computing, big data, webvr, WebGIS, holographic image and other technologies are applied in the platform.

5. The most important next step for our community in 2020

At present, the new business forms of digital economy, such as online shopping, mobile payment and sharing economy, are developing vigorously, creating an unprecedented mobile information network environment, and the future digital economy will be more open and innovative. Therefore, it is urgent to speed up the construction of digital infrastructure, including 5g communication network, high-speed broadband network and other network infrastructure, speed up the construction of a new generation of high-speed, mobile, safe and ubiquitous information infrastructure, and form a network space integrating everything, human-computer interaction and heaven and earth. At the same time, to speed up the coordination and co--construction of digital resources is the core of the digital economy, which is crucial.

D3 Position Statement from *Food from University of Guelph (UoGuelph)*

Authors: Rozita Dara, Mohammed Amiri-Zarandi, Evan Fraser
University of Guelph, Canada

1. Digital science infrastructures as collaborative, community-driven resources

In our opinion, creating an open data environment is a required precursor of the digital agricultural revolution. Briefly, the same technologies that created the Internet and are transforming medicine are being applied to the food system. These technologies offer us the potential to produce more food on less land with fewer inputs but barriers towards adopting these technologies remain. In particular developing the data infrastructure to ensure privacy, maintain cyber security, and ensure fair governance of data are necessary before the potential of the digital agricultural revolution can be unlocked. There are many challenges to establish a fully open platform for raw data. However, there may be an opportunity to open meta-data , aggregated data, and standards to facilitate interoperability and data sharing.

2. Major achievements towards an open data ecosystem for science

In Canada, there are many barriers to realizing an open data network for agriculture and food. With that said recent investments in an innovation supercluster called “protein industry Canada” shows promise as protein industry Canada has been given the task of developing the data strategy for the country. More locally, the University of Guelph is Canada’s oldest and best established ag food research powerhouse and we are in the process of developing a data strategy that will bring coherence and interoperability to our infrastructure. Realizing the need for agricultural platforms and technology solutions to enable an interoperable ecosystem is a major achievement.

3. The vision towards 2025

We believe to have an open digital framework, it is important to address concerns related to data ownership, privacy, regulation, and fairness.

In a digital agri-food ecosystem, data ownership is a controversial topic. On the one hand, farmers should have the right to access and control generated data related to their farming operations. On the other hand, technology providers such as sensor producers and cloud applications' owners usually reserve their right to store, use or disclose data in accordance to their contracts with farmers. In an open digital framework, the ownership of each data source should be clearly determined. Another challenge that is predictable in an open digital framework is to determine who is beneficiary in the potential added value from data and services. The platform must include clear regulations that illustrate who owns these benefits.

Another point is that many companies produce smart farm tools. The diversity in standards and protocols that are utilized in these tools has resulted in interoperability issues in agri-food applications. The sensors that collect data from the environment use different formats for generated data. These data are not identical in terms of length, fields, and format (CSV, JSON, or XML). In addition, the measurement units that are used in these tools are different. The devices that are used in digital farming utilize diverse technologies for communication including WiFi, Bluetooth SMART, NFC, ZigBee, or 3G /4G cellular communications. APIs and semantic web technologies are the potential solutions to tackle the interoperability issues in a digital framework to ensure standardization.

API is an interface that is provided to let the other applications and software to access data or functions of a system. A cross-platform enable other researchers, developers, and startups to create solutions for the digital platform at a high level regardless of protocols and technologies that are used in the back-end. Semantic is another concept that can reduce the concerns related to the interoperability. It is a paradigm that can be used to provide a common understanding of various tools in smart farming applications. This paradigm can describe standards and agreements using shared schemas and ontologies. To develop a useful digital platform, standardization can be one of the main duties in the coming years.

Privacy is another concern that should be addressed in an open digital framework. Every party in the digital agriculture ecosystem needs reassurance regarding the privacy preservation of sensitive data. The first requirement for this is that sensitive data in the digital framework should be anonymized and encrypted. Anonymizing data means that the stored data by the service providers should be separated from any sensitive personal information including name and location, and data encryption can be used to protect data from unauthorized access during transmission through the network. In addition to these methods, it must be guaranteed that it is impossible to trace back, manipulate or eliminate sensitive data.

Another consideration is that the use of modern technology in digital farming is uneven. Here are the reasons: digital platform for open, interoperable data1) many farmers do not have enough skills to use recent technologies, and 2) an effective factor is the limited budget of local farmers to buy the required modern machines. This means that big companies are able to eliminate small farms' market and play a dominant role in any potential digital framework and consequently make a data monopoly. Any proposed digital framework should provide motivations for farm market actors and encourage them to participate in the framework.

4. The most important next step for our community in 2020

It is important to address procedural, regulatory, and challenges. Also, to raise awareness and educate the community about the benefits of digital platform for open data to understand the needs, requirements, and concerns of the agri-food community in terms of open data platforms. Finally, to build consensus and buy-in among various players to adopt and contribute to such concepts and platforms.