D8.3 - AGINFRA Future Science Recommendations

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

DELIVERABLE NUMBER	D8.3
DELIVERABLE TITLE	AGINFRA Future Science Recommendations
RESPONSIBLE AUTHOR	Panagiotis Zervas (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	Mrs. Georgia Tzenou
WORKPACKAGE N. | TITLE	WP8 | Dissemination & Exploitation
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D8.3 | AGINFRA Future Science Recommendations
RESPONSIBLE AUTHOR	Nikos Manouselis (Agroknow), Panagiotis Zervas (Agroknow)
REPLY TO	pzervas@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)
VERSION | STATUS	V1.0 | Final
NATURE	Report (R)
DISSEMINATION LEVEL	Public (PU)
AUTHORS (PARTNER)	Rob Lokers (DLO), Rob Knapen (DLO), Alice Boizet (INRA), Matthias Filter (BfR)
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)

PARTICIPANTS		CONTACT
Agro-Know IKE (Agroknow, Greece)	Agroknow	Nikos Manouselis Email: nikosm@agroknow.com
Stichting Wageningen Research (DLO, The Netherlands)	ALTERRA WAGENINGEN UR	Rob Lokers Email: rob.lokers@wur.nl
Institut National de la Recherché Agronomique (INRA, France)	INRA SCIENCE & IMPACT	Pascal Neveu Email: pascal.neveu@inra.fr
Bundesinstitut für Risikobewertung (BfR, Germany)	BfR Bundesinstitut für Risikobewertung	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)	National and Kapodistrian University of Athens	George Kakaletris Email: gkakas@di.uoa.gr
Stichting EGI (EGI.eu, The Netherlands)	egi	Tiziana Ferrari Email: tiziana.ferrari@egi.eu
Pensoft Publishers Ltd (PENSOFT, Bulgaria)	PENSOFT	Lyubomir Penev Email: penev@pensoft.net

EXECUTIVE SUMMARY

This document serves as a white paper, which describes the AGINFRA+ 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+ user communities and provides recommendations that aim to inform the roadmap for developing AGINFRA+ infrastructure further.

The document provides an overview of the involved user communities and their different user profiles, as well as the way to involve each community in the projects' activities. It describes the objectives and the obstacles that each community had to overcome and provides all the necessary details regarding the special and indispensable aim that has been provided through the use of the AGINFRA+ Virtual Research Environments (VREs). Specific emphasis is provided to risks that the communities might face in the near future and to recommendations regarding how they should address these challenges based on the existing VREs.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

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

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

In this context, the AGINFRA+ project is exploiting the Virtual Research Environments (VREs) paradigm for the three (3) prominent research communities. VREs are a prominent existing cloud-based solution

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

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.

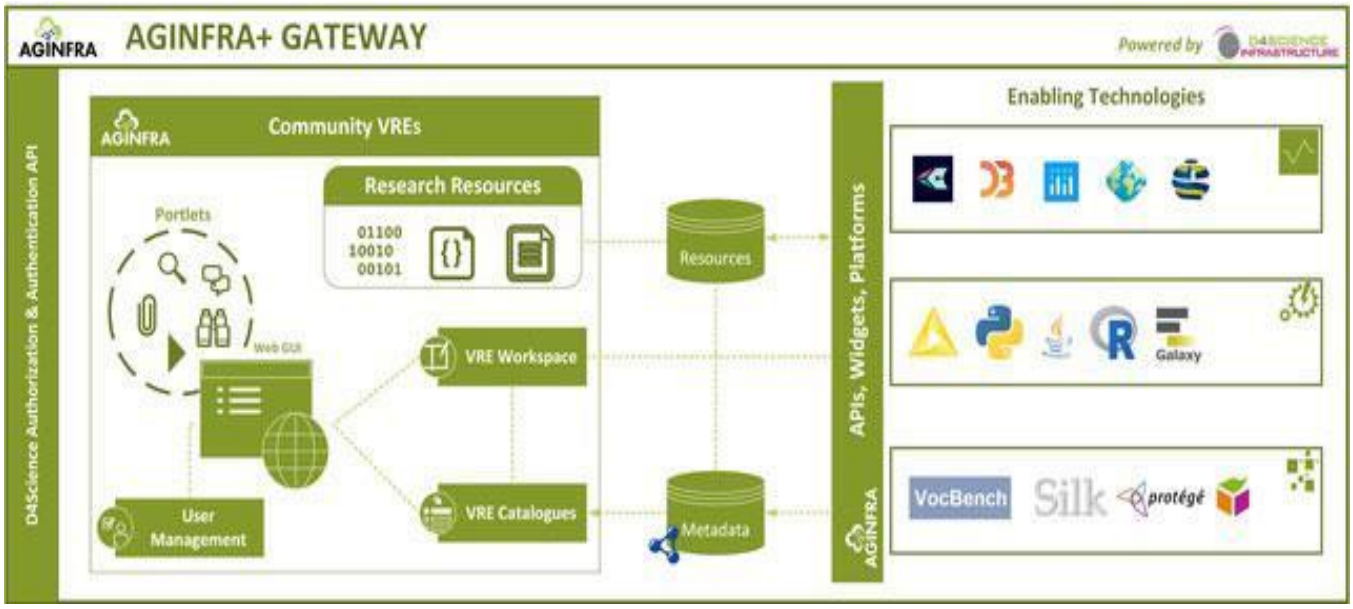


Figure 1: AGINFRA+ Technical Infrastructure based on research community VREs

In order to support the aforementioned research communities, the AGINFRA+ 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+ 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+ Visualization & Publishing Layer.

3 AGINFRA+ 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+ 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)

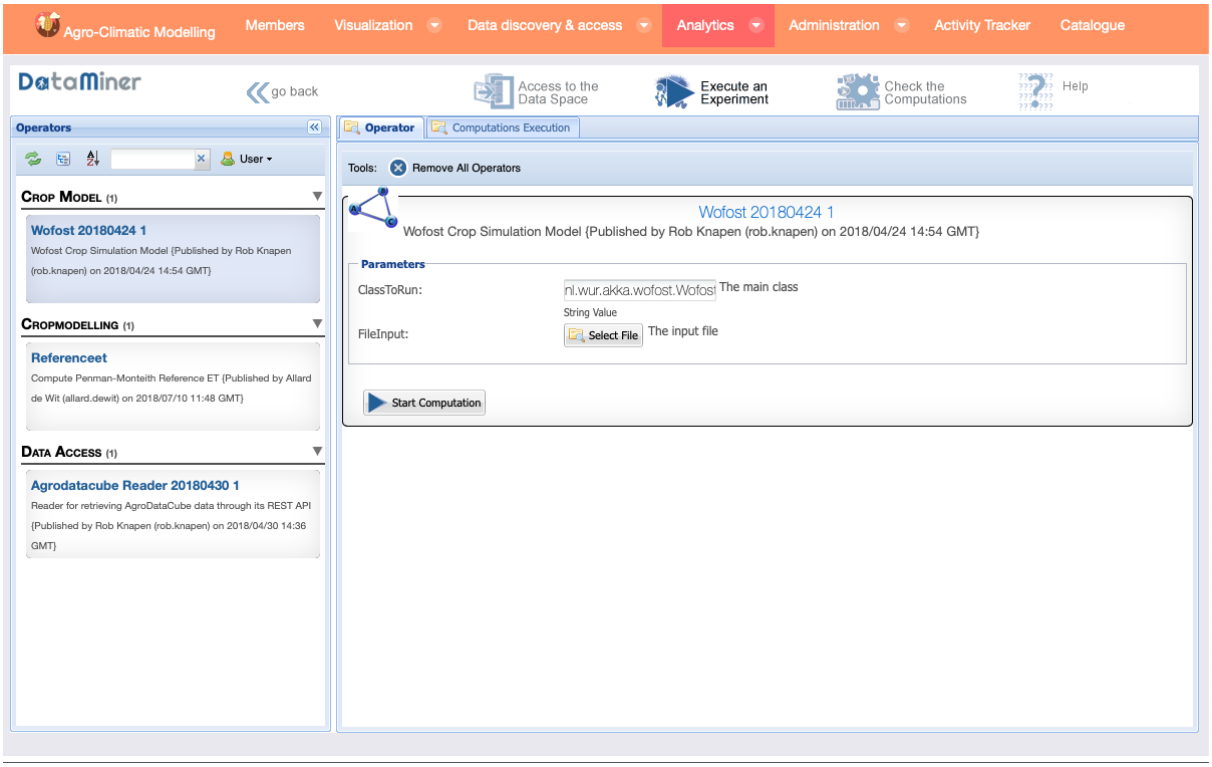


Figure 2: Example of running a crop simulation model in the VRE

B. Large scale regional crop model simulations

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+:

- 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.

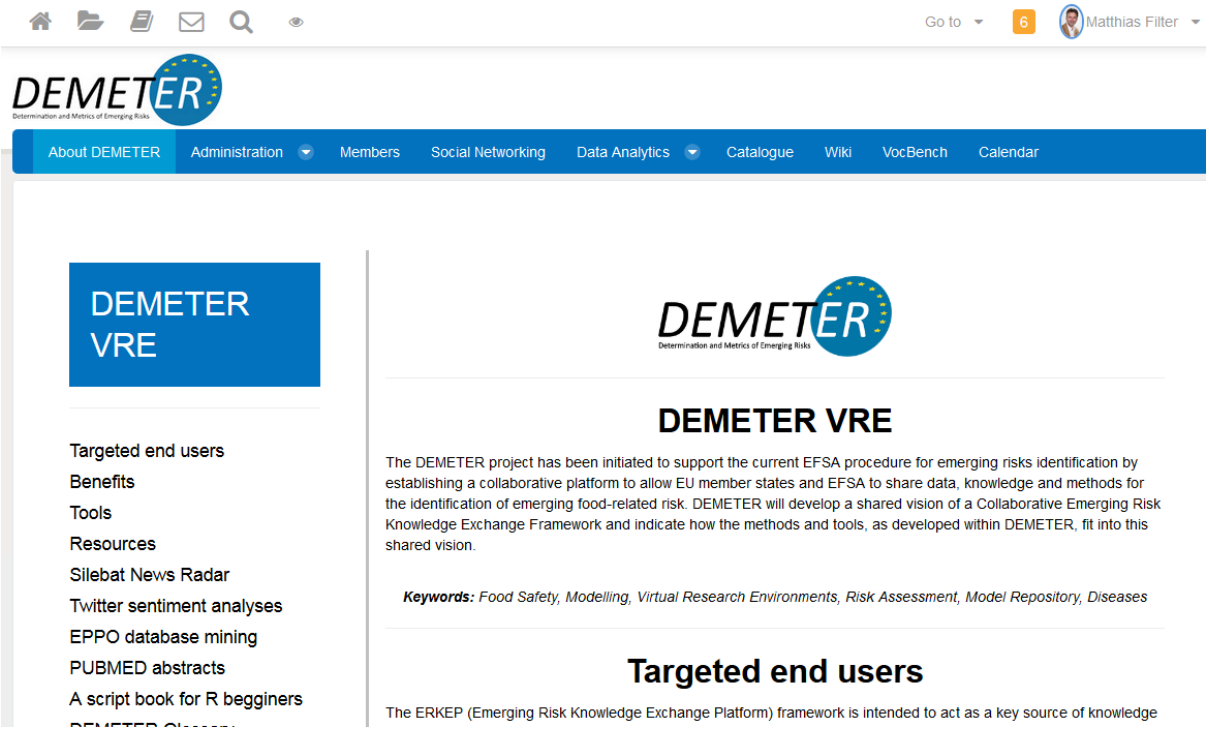


Figure 4: DEMETER VRE Main Page

2) Data Analytics and Processing Needs

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) Data Visualization and Publishing Needs

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) Other Needs

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;

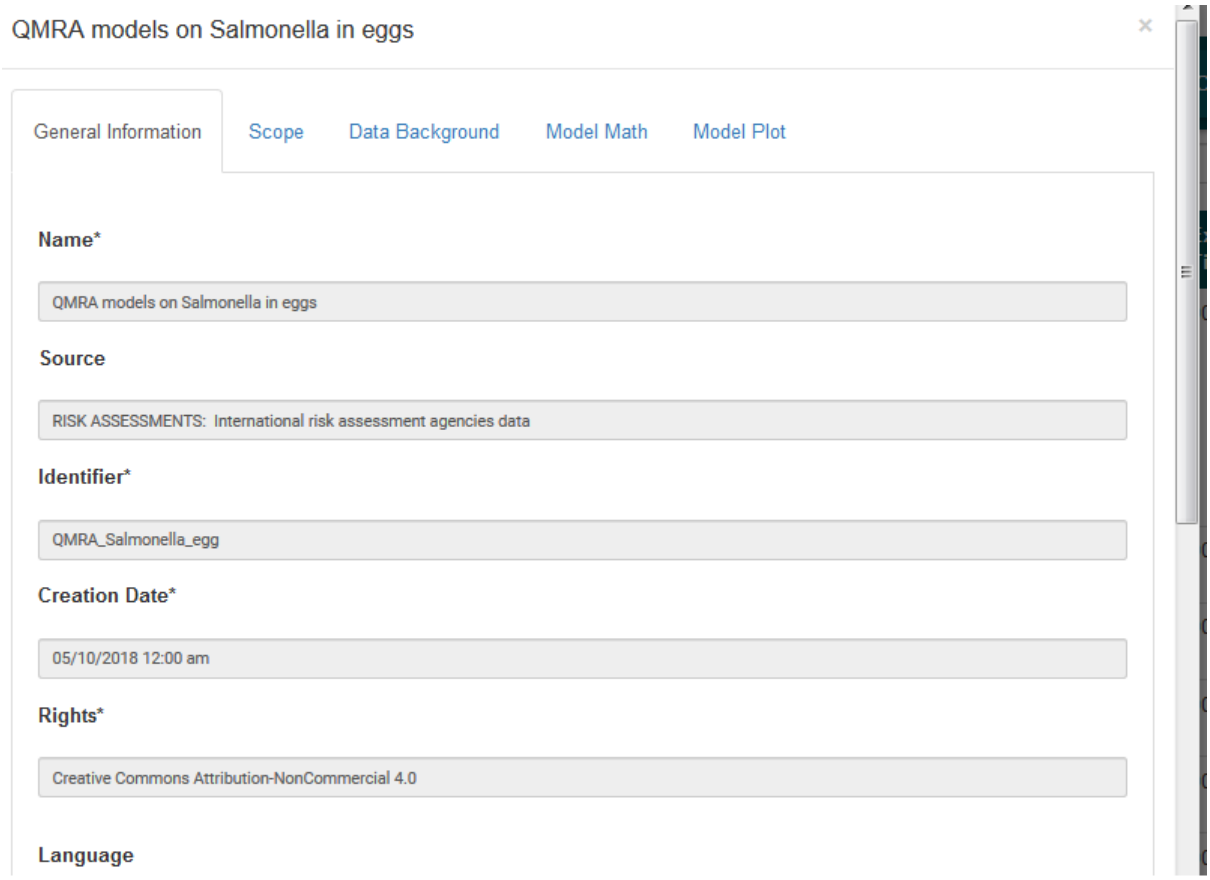


Figure 5: Example of QMRA model

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

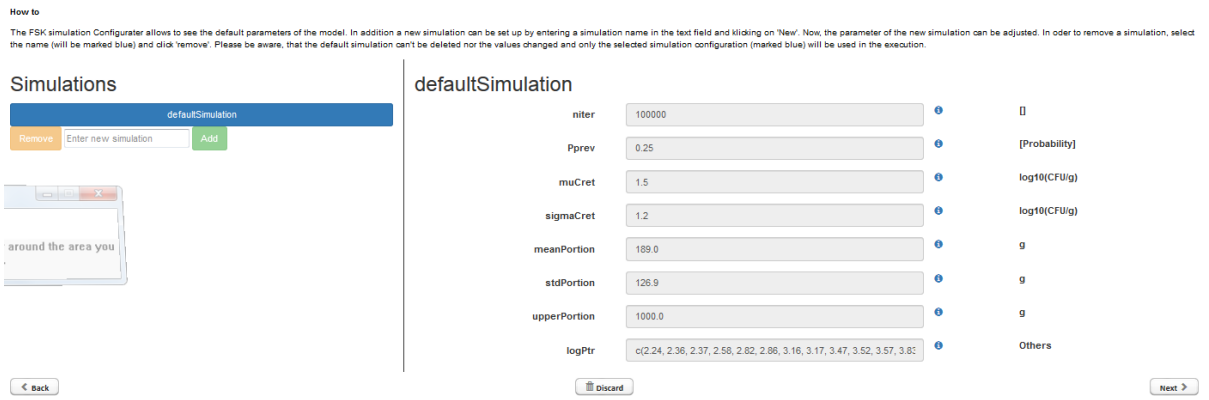


Figure 6: Example of FSK simulation

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);

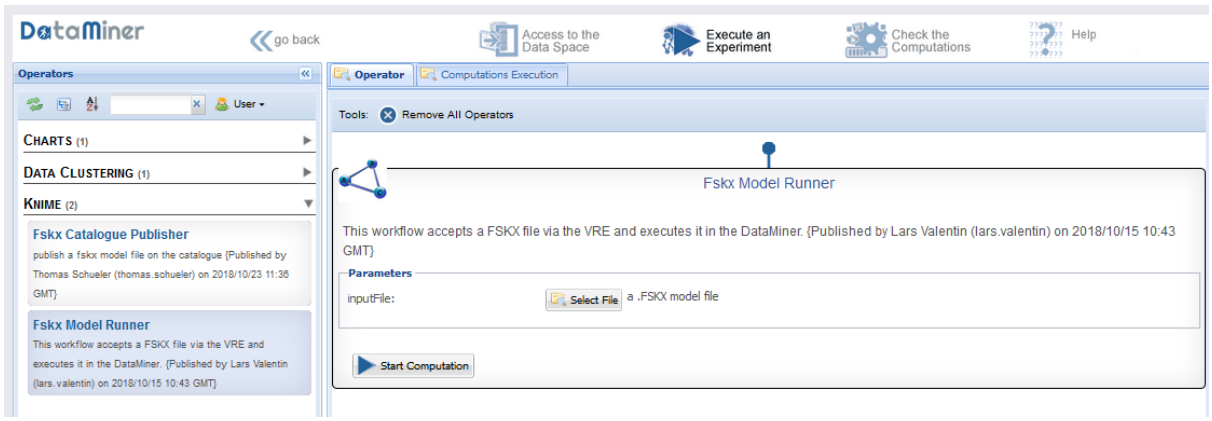
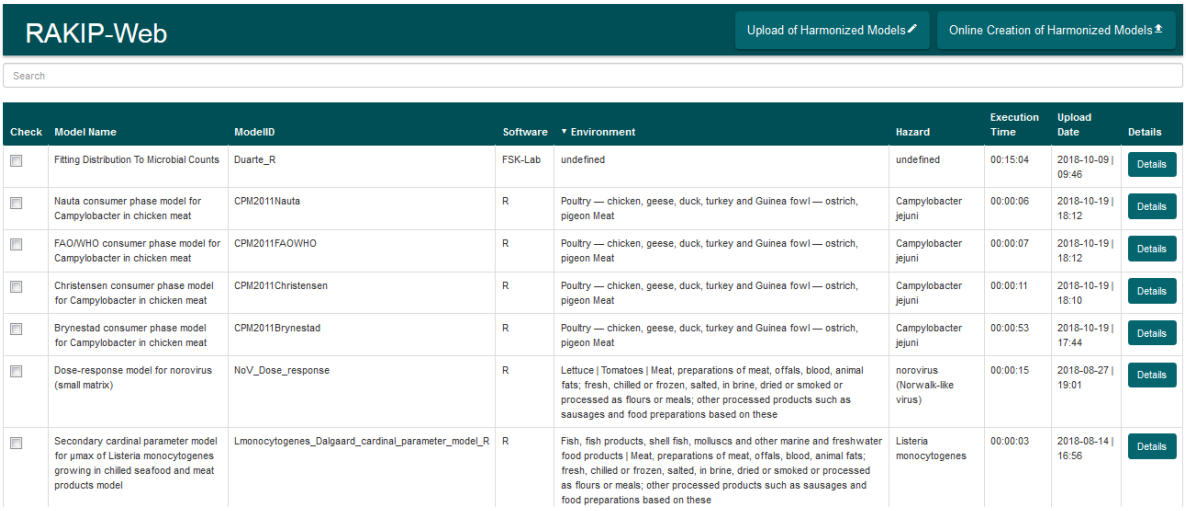


Figure 7: Example of Fskx Model Runner

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+ 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;

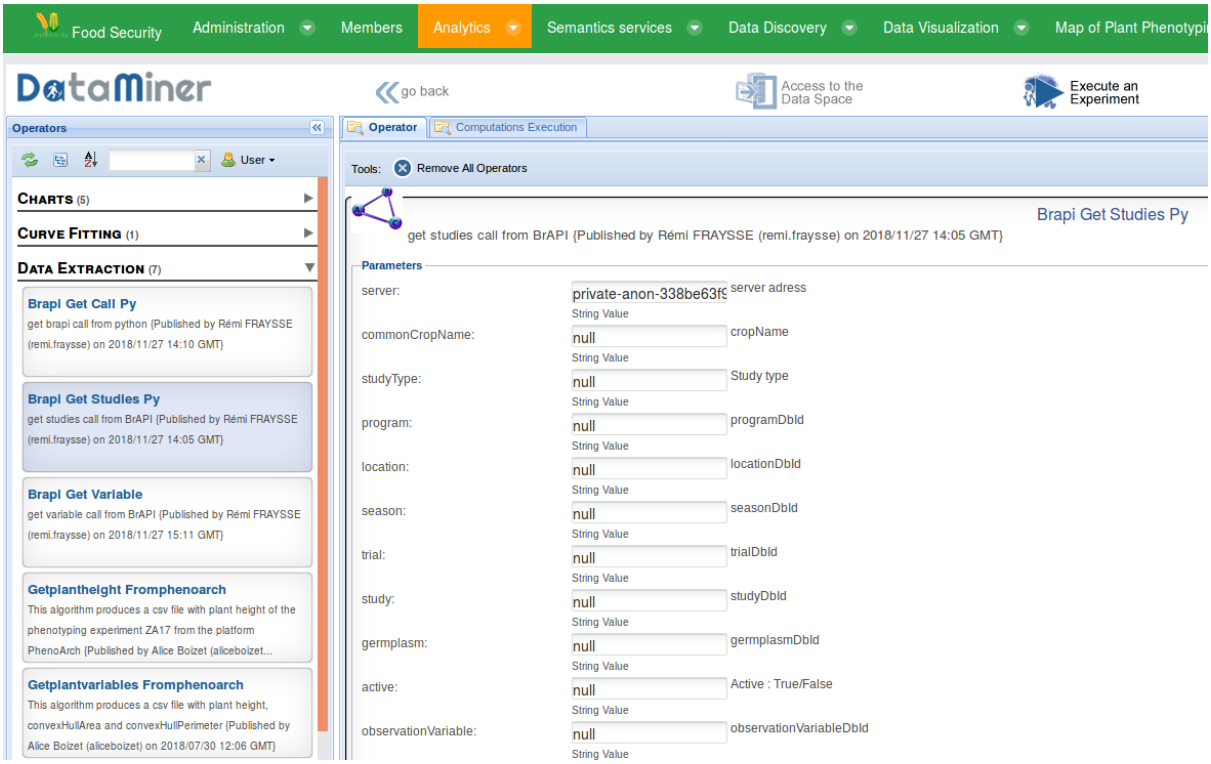


Figure 9: Example of Data Analytics process

- Data exploration and visualization: to provide interactive visualization;

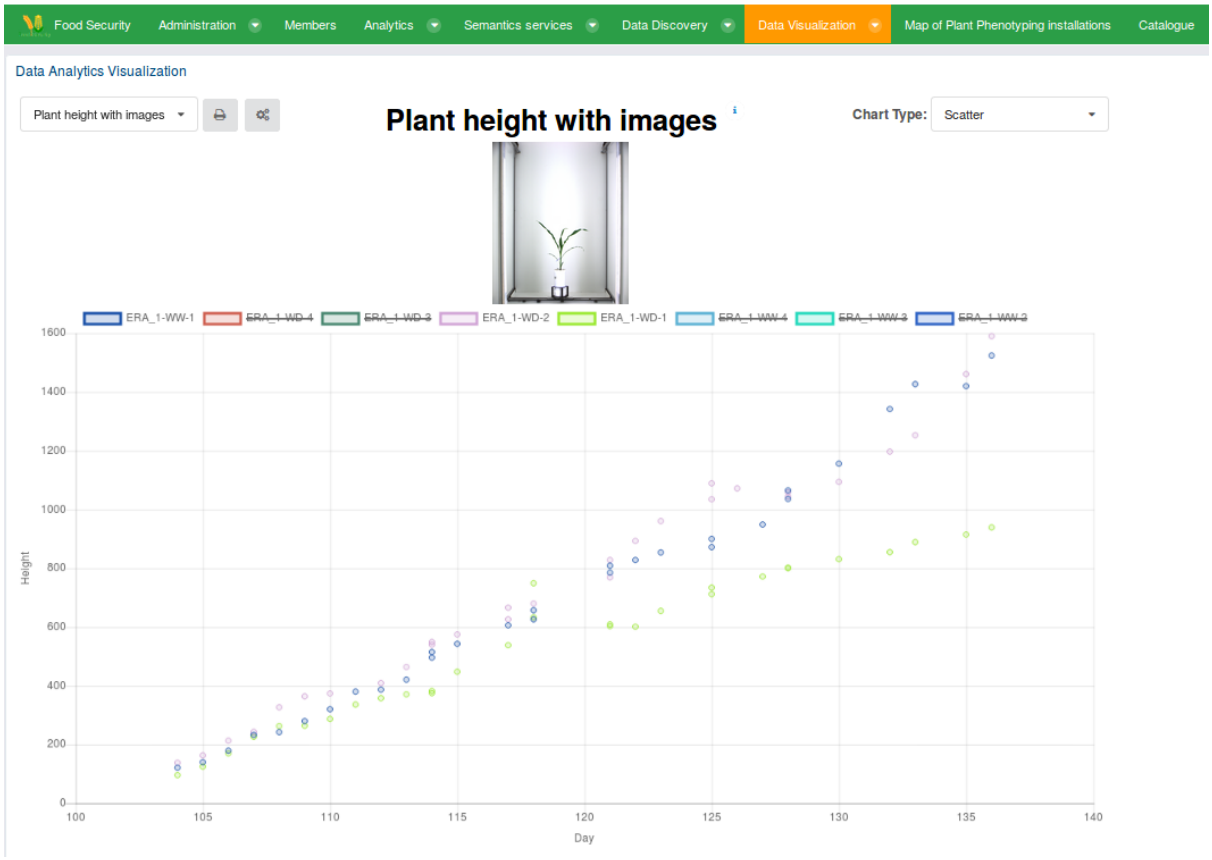


Figure 10: Example of Data Analytics Visualisation

- 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;

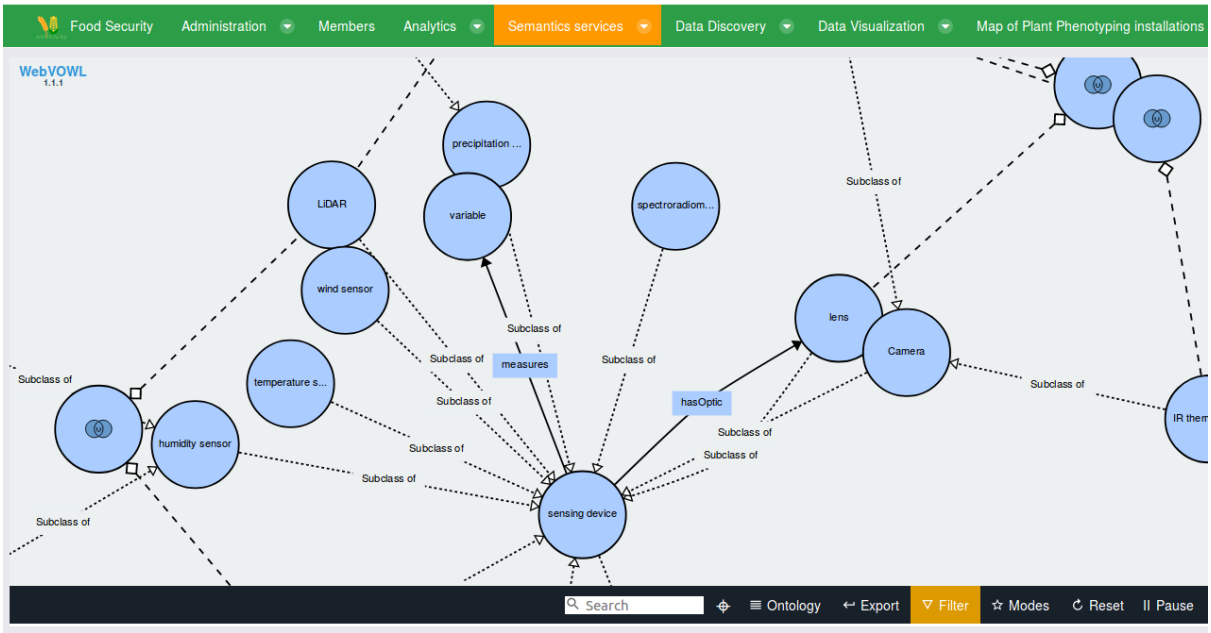


Figure 11: Example of Ontology Visualising

- 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+ 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+ 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+ 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.