eROSA

e-infrastructure Roadmap
for Open Science in Agriculture

Community Building and Fine-Mapping Workshop

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

DELIVERABLE NUMBER	D3.1
DELIVERABLE TITLE	Community Building and Fine-Mapping Workshop
LEAD BENEFICIARY	INRA

GRANT AGREEMENT N.	730988
PROJECT ACRONYM	e-ROSA
PROJECT FULL NAME	Towards an e-infrastructure Roadmap for Open Science in Agriculture
STARTING DATE (DUR.)	01/01/2017 (18 months)
ENDING DATE	30/06/2018
PROJECT WEBSITE	erosa.aginfra.eu
COORDINATOR	Odile Hologne
ADDRESS	Route de Saint-Cyr RD 10, Versailles, 78026, France
REPLY TO	odile.hologne@inra.fr
PHONE	+33 1 30 83 33 92
EU PROJECT OFFICER	Mrs. Pilar Ocon-Garces

WORKPACKAGE N. | TITLE	WP3 | Roadmap co-Design & Uptake
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D3.1 | Community Building and Fine-Mapping Workshop
RESPONSIBLE AUTHOR	Madeleine HUBER (INRA)
REPLY TO	madeleine.huber@inra.fr
DOCUMENT URL	http://www.erosa.aginfra.eu/sites/erosa_deliverables/D3.1.pdf
DATE OF DELIVERY (CONTRACTUAL)	30 June 2017 (M6)
DATE OF DELIVERY	6-7 July 2017 (M7)
DATE OF REPORT SUBMISSION	31 August 2017 (M8)
VERSION | STATUS	V1 | Final
NATURE	O(Other)
DISSEMINATION LEVEL	PU(Public)
AUTHORS (PARTNER)	Madeleine Huber (INRA)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
1	Final version	31 August 2017	Madeleine HUBER (INRA)

PARTICIPANTS		CONTACT
Institut National de la Recherche Agronomique (INRA, France)		Odile Hologne Email: odile.hologne@inra.fr
Stichting Wageningen Research (Alt-DLO, The Netherlands)		Sander Janssen Email: sander.janssen@wur.nl
Agro-Know IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com

EXECUTIVE SUMMARY

This document presents the report of the 1st e-ROSA Stakeholder Workshop "Community building and mapping of the research infrastructures for agro-food research" that was held on 6-7 July 2017 at Montpellier SupAgro. The goal of this workshop was to bring together scientific communities and existing data-related infrastructures and initiatives that can support researchers throughout the data life cycle in the context of their research activities.

It gathered various stakeholders such as IT specialists, Knowledge managers, semantic experts and researchers from disciplinary fields related to the agri-food domain. In particular, the workshop sought to:

- Initiate a community building process around the issue of Open & Big Data in agri-food;
- Improve our knowledge on the current landscape of existing research infrastructures and e-infrastructures, networks, initiatives and policies;
- Justify the need for a common e-infrastructure in agri-food by identifying gaps and key collaboration actions to implement;
- Link the envisioned e-infrastructure for agri-food with the EOSC vision and architecture;
- Prepare for the next e-ROSA Stakeholder Workshops during which we will elaborate our vision for the future and identify related needs (2nd workshop), and develop a common roadmap on how we can achieve this vision (3rd workshop).

It was organised around three main parallel sessions: 1) FAIRifying data, 2) Data repositories and discovery services, and 3) E-infrastructures and data services.

The workshop in itself corresponds to the Deliverable 3.1 under Work Package 3 "Roadmap co-Design & Uptake". e-ROSA Stakeholder Workshops consist in a collaborative mechanism that allows to bring together the e-ROSA Stakeholder Community in view of envisioning the future e-infrastructure for Open Science in agri-food and co-elaborating the common roadmap that will support the implementation of this e-infrastructure.

Community building and mapping of the research infrastructures for agriculture & food

First e-ROSA Stakeholder Workshop - 6 & 7 July 2017

Montpellier SupAgro, 6 & 7 July 2017

Agenda

Day 1

09:00 – 09:30	*Arrivals & coffee*	
09:30 - 09:45	**Welcome & Introductions**	Odile Hologne (INRA)
9:45-9:55	Vision and the state-of-play of the European Open Science Cloud	Wim Haentjens (EC DG RTD)
Opening plenary: The data challenges in agriculture & food		
09:50- 11:00	The data challenges in agricultural sciences	Michael Chelle (INRA)
	Data-driven innovation in the agri-business sector	Graham Mullier (Syngenta)
	The data challenges in the food supply chain	Christopher Brewster (TNO)
11:00 - 11:30	***Coffee break***	
Plenary II: E-infrastructures as an opportunity		
11:30 – 12:30	Defining an e-infrastructure for Open Science in agriculture & food (e-ROSA) and mapping the e-infra landscape	Johannes Keizer (GODAN)
	EOSC pilot : governance, architecture	Simon Lambert (STFC) Gianpaolo Coro (CNR ISTI)

<table>
<tr><td>12:30-13:00</td><td>Discussion</td><td>all</td></tr>
<tr><td>13:00 - 14:00</td><td colspan="2">Working lunch break</td></tr>
<tr><td colspan="3">Interactive Sessions: Today’s EU e-infrastructure landscape for agriculture & food (1/2)</td></tr>
<tr><td rowspan="2">14:00 – 16:30</td><td colspan="2">Parallel session 1: FAIRifying data

Chair: Johannes Keizer, rapporteurs: Elizabeth Arnaud

Input: From an end user point of view, an e-infrastructure should help discover and integrate data coming from different domains (omics, observation, social data, etc.) and data sources in a transparent, easy way. It means than the e-infrastructure should rely on trusted data repositories dealing with agricultural data issues (time, space, ...) to make the data FAIR. Most of the time we are able to provide FAIR metadata and less often FAIR data.

Feedback from practical use cases: The talks (10 minutes) will focus on encountered issues (data types, scales, geolocation, ...), how they were overcome, the "return on investment" (real or expected), the perspectives : what could make it easier in terms of process, tools, practices; what is reasonable to do; how do you go from data lake to linked data; etc.

<table><tr><td>WUR</td><td>Rob Knapen (Alt-DLO)</td></tr><tr><td>INRA</td><td>Anne Françoise Adam-Blondon (URGI)</td></tr><tr><td>CGIAR</td><td>Leroy Mwanzia (CIAT)</td></tr><tr><td>CGIAR</td><td>Elizabeth Arnaud (Bioversity)</td></tr><tr><td>AgGateway</td><td>Andres Ferreyra (Syngenta)</td></tr></table>
Collaborative work:
a. What recommendations to make DATA FAIR? When is data FAIR for the agriculture and nutrition domain?
• What are specific problems in making data Findable?
• What are specific problems in making data Accessible?
• What are specific problems in making data Interoperable?
• What are specific problems in making data Reusable?
b. What is reasonable to do?
c. The role of semantics
d. Our commons?</td></tr>
<tr><td colspan="2">Parallel session 2: Data repositories and discovery services

Chair: Nikos Manouselis, rapporteur: Martin Parr

Input: The analysis of the data landscape shows a huge diversity of data repositories but it is not easy to identify them, nor to assess the relevance of their content.

The data landscape: CIARD RING, Vest registry
<table><tr><td>GFAR</td><td>Valeria Pesce</td></tr></table>
Feedback from data repositories
<table><tr><td>Dataverse</td><td>CGIAR</td><td>Indira Yerramareddy (IFPRI)</td></tr><tr><td>Dataverse</td><td>CIRAD</td><td>Sophie Fortuno</td></tr><tr><td>EUDAT</td><td>CINES</td><td>Marion Massol</td></tr><tr><td>AgroPortal</td><td>University of Montpellier</td><td>Clément Jonquet</td></tr></table></td></tr>
</table>

<table>
<tr><td></td><td>
Collaborative work:
a. What kind of discovery services? How to implement them? ...
b. Trusted data repositories: what does it mean in terms of technical requirements, business models, policy?
c. Quality certification? seal of approval ?
d. Data culture vs. API culture
e. Role of the data repository in the value chain
</td></tr>
<tr><td></td><td>
Parallel session 3: E-infrastructure and data services

Chair: Rob Lokers, rapporteur: Sophie Aubin

Input: E-infrastructures should offer data discovery services, data interoperability and integration services and the ability to build a virtual research environment in a big data context. The goal of this session is to establish a dialogue between “generic e-infrastructures and services” and the “application domains”. An important objective is to examine how to achieve virtual research environments that provide the required functions and workflows to support interdisciplinary research with FAIR data, big data analytics (e.g. workflows and code) over distributed heterogeneous resources.

Introduction: Analysis of Big Data technologies for use in agro-food science
<table><tr><td>Wageningen UR</td><td>Rob Lokers</td></tr></table>
E-infrastructure services or components (5-8 minutes talk to illustrate the topic, problems encountered, solutions find or wishes for a better solution)
<table>
<tr><td>ELIXIR</td><td>Paul Kersey (EBI)</td></tr>
<tr><td>EUDAT</td><td>Alexis Jean-Laurent (CINES)</td></tr>
<tr><td>Agrisemantics</td><td>Sophie Aubin (INRA)</td></tr>
<tr><td>EGI</td><td>Enol Fernandez (EGI)</td></tr>
<tr><td>OpenMinted</td><td>Robert Bossy (INRA)</td></tr>
</table>
Collaborative work:
a. Inventory and assessment of existing services, Kiss method: what should be Kept, Improved, Started (what is missing), Stopped?
b. Interaction between e-infrastructures (generic vs. thematic)
c. What services should be generic? How to ensure affordable and sustainable access to data analytics and computation to researchers, regardless of their disciplines or where they are located?
d. What are the specific service requirements for the Agri-food sciences?
e. What are our “common goods”?
</td></tr>
<tr><td>16:30 - 17:00</td><td>Coffee break</td></tr>
<tr><td>17:00 – 18:00</td><td>Synthesis: wrap-up in each parallel session</td></tr>
<tr><td></td><td>Dinner</td></tr>
</table>

Day 2

08:30 – 09:00	*Arrivals & coffee*
09:00 - 10:00	**Report of the 3 parallel sessions** • FAIRifying data • Data repositories • e-infrastructures and services
Interactive Sessions: Today's EU e-infrastructure landscape for agriculture & food (2/2)	
10:00 – 11:30	**Final group discussions** (3 groups) According to the landscape depicted during the collaborative work sessions, identify: • the pitfalls & gaps • the key actions to launch and the partnerships to build
11:30 - 12:00	***Coffee break***
12:00 – 13:00	Synthesis: wrap-up of the tree final group discussions
13:00 – 13:30	**Closing plenary: Outcomes & next steps**
13:30 - 14:30	***Working lunch break***

Scope and objectives

The European project e-ROSA (Towards an e-infrastructure Roadmap for Open Science in Agriculture) seeks to build a shared vision of a future sustainable e-infrastructure for Open Science in agriculture & food. It aims at facilitating the co-development of a common roadmap by and for involved research communities and key stakeholders related to scientific data and research infrastructures, in line with the EOSC vision, agenda and architecture. e-ROSA organised its first Stakeholder Workshop in order to bring together scientific communities and existing data-related infrastructures and initiatives that can support researchers throughout the data life cycle in the context of their research activities.

The workshop took place at Montpellier SupAgro on 6 & 7 July 2017. It gathered various stakeholders (see profile of participants in Annex) such as IT specialists, Knowledge managers, semantic experts and researchers from disciplinary fields related to the agri-food domain. Key objectives of the workshop were presented by Odile Hologne (INRA), coordinator of the e-ROSA project:

- To initiate a community building process around the issue of Open & Big Data in agri-food;
- To improve the knowledge on the current landscape of existing research infrastructures and e-infrastructures, networks, initiatives and policies;
- To justify the need for a common e-infrastructure in agri-food by identifying gaps and key collaboration actions to implement;
- To link the envisioned e-infrastructure for agri-food with the EOSC vision and architecture;
- To prepare for the next e-ROSA Stakeholder Workshops during which we will elaborate our vision for the future and identify related needs (2nd workshop), and develop a common roadmap on how we can achieve this vision (3rd workshop).

The opening plenary provided the opportunity to present the overarching data challenges in the agri-food sector from various perspectives, i.e. from the research point of view, from the private sector's perspective, as well as throughout the food supply chain. It focused on how e-infrastructures can benefit to agri-food science, especially in the context of the implementation of EOSC.

The workshop was organised around three main parallel sessions as follows:

Session 1. FAIRifying data. The first session focused on experiences and difficulties encountered in making data FAIR. In particular, it highlighted the role of semantics, accepted standards and efficient tools as well as adapted mechanisms to support and incentivise researchers to make their data FAIR.

Session 2. Data repositories and discovery services. This session focused on the role of (i) local data repositories in making data and metadata FAIR, (ii) federated dataset catalogues and standardised semantic resources in providing efficient discovery services, and (iii) policy support and sustainable business models in creating an effective change of culture for data sharing.

Session 3. E-infrastructures and data services. The goal of this session was to give a first approach on how to connect EU generic e-infrastructures (e.g. EUDAT, EGI, OpenMinTeD) to domain-specific infrastructures and services (e.g. ELIXIR, Agrisemantics) in the context of EOSC.

During these parallel sessions, dedicated presentations provided food for thought for discussions in each of the three breakout groups. All three breakout groups reported to the rest of the workshop participants in plenary in order to initiate further discussions. Final group discussions took place in order to identify trends, gaps and disruptions, as well as to discuss high-impact demonstrators and use cases that we should focus on in the near future.

The data challenges in agriculture & food

The European Open Science Cloud: Vision & state of play

Wim Haentjens (DG RTD) presented the progress towards the EOSC's establishment in view of supporting the agri-food domain. This first presentation provided the general European context of the envisioning of a future e-infrastructure for agri-food, i.e. the European Open Science Cloud (EOSC).

The European Open Science Cloud (EOSC) is a valuable instrument for promoting data sharing as well as supporting and federating existing structures. As a follow-up to the EOSC Summit on 12 June 2017, the EOSC Declaration will list the key actions participants have agreed on. The EOSC Governance Roadmap will be established by the end of 2017.

EOSC has adopted a specific approach in order to support thematic areas thanks to the development of Thematic Clouds. In particular, one will focus on Agriculture, Food and Nutrition, relying on the overall concept of Food Systems.

DG RTD can facilitate and invest in the process of advancing EOSC in this field. The specific needs and potential opportunities of the agri-food sector in line with the EOSC implementation need to be identified.

The (big) data challenge in agri-food science

Michaël Chelle (INRA) described how agricultural research can address emerging challenges by taking advantage of the data opportunity and the development of infrastructure services.

Agriculture consists in a complex science from a data science point of view, with different disciplines (from genomics to social sciences), different scales (from genes to ecosystems) and geolocalisation. The ability to integrate these heterogeneous data is a key issue to tackle new societal challenges (e.g. food security in the face of climate change, sustainable food value chains, digital agriculture and foodtech).

In addition, the automation of data collection, new techniques in omics as well as the development of new types of data sources (e.g. Internet of Things, crowd-sourcing, text mining) has allowed to collect an exponentially increasing amount of data. Thus, there is a need for an "entity" (i.e. e-infrastructure)

that connects data, infrastructures, resources and people and that allows to share efforts and expertise, and support innovation. In particular, such an entity can develop Virtual Research Environments (VREs) that provide tailored solutions for specific communities.

Many questions arise when envisioning a new e-infrastructure: centralised vs. distributed e-infrastructure, evaluation of an e-infrastructure; articulation between general services and VREs, identification of needs, etc.

The envisioned e-infrastructure can facilitate the required changes in research practices and support the potential shifts in paradigm (e.g. hypothesis-driven vs. data-driven research) related to the way we do research thanks to Open and Big Data.

Discussion

- We first need to identify examples of use cases for which connecting and sharing the data is critical: these examples will make the case for the need for e-infrastructures. We also need to show concrete examples to external stakeholders of what we can already do now thanks to data. In particular, the infrastructural need we seek to address is not in developing new data infrastructures, but in reducing their fragmentation by connecting them and integrating their data. As we aim at accessing large amounts of various data, we need to demonstrate the value of doing so in agriculture.
- Developing skills for the various stakeholders involved in research and data management is a major issue for the envisioned e-infrastructure.
- We need to identify required investments in the short/medium-term & expected impacts in order to pave the way for a long-term roadmap.

Data-driven innovation in the Ag sector

Graham Mullier (Syngenta) explained how the private sector can contribute to the R&D chain in agri-food thanks to the data opportunity.

Highly significant investments in data management are made by private companies to support their R&D systems. In order to justify these investments, we need to connect the data to the problems that need solving, otherwise the data has no meaning. Collaboration is becoming crucial for private R&D as discovery is becoming an increasingly difficult process and costs are rising. There is a strong need for pre-competitive cooperation amongst all players globally. In particular, we need common standards that support interchange. In order to achieve this, we should not get caught in an idealistic exercise. Indeed, *de facto* standards that are being used today can set the official standards of the future for efficient Linked Open Data in agri-food.

In the case of Syngenta, the Good Growth Plan is the company's programme to reach six strategic commitments around resource efficiency, soil and biodiversity conservation practices, and rural prosperity including smallholders. It supports a global network of farms where technology adoption and efficiency performance are monitored each year by an independent company. Data is collected and shared in order to track the progress towards Syngenta's commitments, but also to inform on innovative, sustainable farming practices and allow other stakeholders and organisations to analyse it, which can feed into Syngenta's work. Another example is the sharing of Syngenta's RNA-based biocontrol research results as open data in order to be transparent about potential innovation and speed up the innovation process (i.e. others can analyse the data to provide recommendations and statements).

Opening your data provides many benefits for the private sector and in general: e.g. it allows to improve the quality of the published data, give repeatable publishing patterns, influence standards, increase transparency, etc. Furthermore, the Open movement concerns not only data, but also models and innovation in general. Related business models need to build on what is already supported by individual organisations and can facilitate the development of impactful use cases such as identifying weeds with a digital tool (e.g. in the case of weeds that have become herbicide-resistant), identifying diseases, predicting toxicology, etc.

Discussion

- We need to be careful about what we define as "best" standards: if they are not adopted, then they should not be considered as the "best".
- Ownership of farmers' data is a major issue, especially in the case of machine-generated data (e.g. milking robots). The role of the regulator in the governance framework is to stop the monopoly of data. However, there are legal barriers that need to be addressed (e.g. there are regulations for ownership of *databases*, but not for ownership of *data*).

Supply chain data integration – who benefits?

Christopher Brewster (TNO) presented the data-related challenges throughout the supply chain (i.e. from the farm to the customer), especially highlighting the issue of business models that can support clear benefits for data providers (farmers in this case) in sharing their data.

One important question that needs to be addressed is: why should the farmers share their data? Are there benefits for them, especially in terms of related business models? Currently, there are no clear business models that incentivise farmers to share their data. The stakeholders that have a high interest in making data sharing mandatory are governments and policy-makers.

There is a conflict between societal priorities, business interests and privacy concerns:

- Societal priorities imply the maximum of transparency and access to data;
- Business interests can support differing objectives: in some cases it is a good idea to share data, e.g. because customers want access to their data or in the case of a strategic initiative with positive impacts (e.g Syngenta's Good Growth Plan, see above); in other cases, sharing data from a company may reveal too much to competitors (i.e. the natural business position);
- Personal privacy concerns: the issue of personal data needs to be addressed in order to provide access to private data to do research. Indeed, each action of a farmer or consumer reflects personally identifiable activities and falls under the regulations concerning personal privacy (i.e. the EU General Data Protection Regulation).

As a result, there are huge amounts of data we don't have access to along the supply chain (e.g. for the researcher, the farmer, etc.) although there are specific situations where it is of added value to open data from a business perspective, for instance:

- For food security and integrity: i.e. data for solving a food crisis;
- For reputation management;
- To support transparent certification procedures: at the moment, the certification issue implies data sharing in small segments of the supply chain only (e.g. from farmer to certifier or supermarket).

Data sharing can be achieved (i) via power (i.e. regulations) or (ii) economic incentives, which means we need clear business models. In addition, there is a need to make data sharing easy for it to become a habit.

E-infrastructures as an opportunity

Towards an e-infrastructure for open science for agriculture and food: Different views on infrastructure definition and mapping methods

Johannes Keiser (e-ROSA) presented the various concepts and definitions that are at the heart of the e-ROSA project and of the envisioning of the future e-infrastructure for agri-food.

The envisioned e-infrastructure should rely on four overarching components: (i) a community of users, (ii) business models and governance, (iii) a technical backbone and (iv) services (see e-ROSA's definition of an e-infrastructure). Our commons for an e-infrastructure in agriculture can be conceptualised in three logical layers as follows:

1. "Data layer": the data resources and related services for data management;
2. "Interoperability layer": the services allowing access and linking of (i) FAIR data from various sources (1st layer above), and (ii) tools and applications amongst them and with data sources (3rd layer below);
3. "End-user layer": the analytical and valorisation services.

More specifically concerning FAIR data, FAIR compliance can be achieved at a certain degree:

1. UnFAIR
2. Findable, Usable for Humans
3. FAIR metadata
4. FAIR data, restricted access
5. FAIR data, Open Access
6. FAIR data, Open Access, Functionally Linked

A key issue is how to get interoperable data repositories that achieve full FAIR compliance (last degree).

Furthermore, there is a need to integrate existing initiatives into a common framework/e-infrastructure, which is e-ROSA's overall goal. The e-ROSA project is currently undertaking a mapping of key stakeholders (in particular thanks to this workshop) in order to facilitate the development of a common vision for a future e-infrastructure for agri-food.

Discussion

- Importance of always keeping in mind the global dimension
- Need to create not only FAIR *data* but FAIR *knowledge*: we need to develop operational implementation mechanisms for making FAIR data and knowledge.

The European Open Science Cloud for Research Pilot Project

Simon Lambert (STFC) presented the EOSCpilot project and how it supports scientific demonstrators such as ones in the agri-food sector.

EOSC Pilot supports the initial development of EOSC, which means the project is showing it can be done and has value. Overarching goals of EOSCpilot are federation and defragmentation. It is structured around three types of challenges:

1) Scientific challenges through the support of scientific demonstrators;
2) Technical challenges concerning services and interoperability;

3) Cultural challenges, i.e. skills, engagement, policy and governance.

EOSC Pilot is interacting with communities so they can relate to the following issues especially via the EOSC Governance Development Forum:

- Governance
- Policy
- Demonstrators
- Services
- Interoperability
- Community engagement

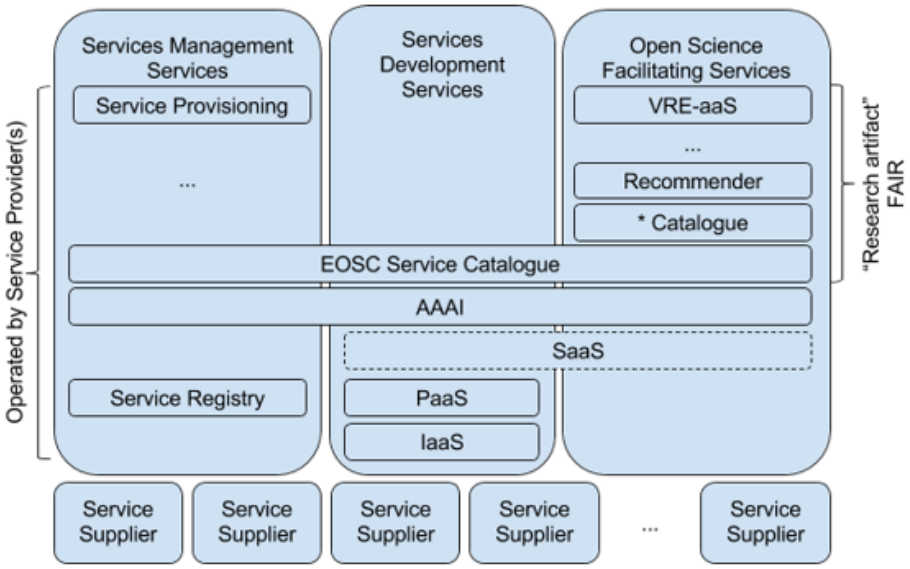
In particular, demonstrators consist in the first mean to interact with specific scientific communities. In addition, specific rules of engagement have been determined for service providers, who need to ensure a minimum level of commitment.

EOSC Architecture

Gianpaolo Coro (ISTI-CNR) described the EOSCpilot proposal for EOSC's architecture.

In line with EOSC's main goals, the EOSC architecture has been defined as a System-of-Systems (same as D4Science's architecture):

- An Open and Evolving System of Systems, connecting existing and forthcoming "systems" (e-Infrastructures, Research Infrastructures, private/public service providers, etc.);
- Constituent systems are service suppliers and/or service providers;
- EOSC will deliver its facilities *as-a-Service*. These services may offer Application-as-a-Service, Data-as-a-Service, Platform-as-a-Service, and Infrastructure-as-a-Service.



The upcoming EOSC meeting on 14-15 September in Pisa will allow to further discuss this architecture proposal.

Discussion

- The issue of interoperability within the architecture was discussed.
- We need to understand how scientists can easily connect with the e-infrastructure.

Today's EU e-infrastructure landscape for agri-food research

Session 1. FAIRifying data

Input

From an end user point of view, an e-infrastructure should help discover and integrate data coming from different domains (omics, observation, social data, etc.) and data sources in a transparent, easy way. It means that the e-infrastructure should rely on trusted data repositories dealing with agricultural data issues (time, space, ...) to make the data FAIR. Most of the time we are able to provide FAIR metadata and less often FAIR data.

Presentations focused on encountered issues (e.g. data types, scales, geolocation), how they were overcome, the "return on investment" (real or expected), the perspectives: what could make it easier in terms of process, tools, practices; what is reasonable to do; how do you go from a data lake to linked data; etc.

Guiding questions:

a) What recommendations to make DATA FAIR? When is data FAIR for the agriculture and nutrition domain?
   - What are specific problems in making data Findable?
   - What are specific problems in making data Accessible?
   - What are specific problems in making data Interoperable?
   - What are specific problems in making data Reusable?

b) What is reasonable to do?

c) The role of semantics

d) Our commons?

Presentations

1. Challenges in making data FAIR – An Agronomic and Environmental Sciences case study

Rob Knapen (WUR) presented a case study for a Data Interoperability Model and discussed identified challenges in providing FAIR data.

In particular, he highlighted the issue of linked location-based data in tackling the question of how data sets can be geospatially related together. For now, the principles and tools supporting Linked Open Data are not adapted for Geo-Information.

Also, general issues in making data FAIR were discussed:

- Using RDF is demanding in terms of computational and storage resources;
- Existing standards need to be further improved and promoted, there are competing standards;
- Lack of standardized, adopted semantics and variable-types in agronomy (e.g. varieties, units).

2. Making wheat data FAIR

Anne-Françoise Adam-Blondon (INRA) presented the achievements and current challenges of WheatIS (Wheat Information System), the global data repository for wheat.

Several challenges were raised:

- Stronger community management is required in order to promote a data sharing culture.
- The organisation of such an infrastructure in distributed nodes is challenging (e.g. to synchronise technical updates and improvements of the common data model).
- Scientists need to be guided in order to use developed standards and understand what metadata they have to provide in order to deliver interoperable data. Wheat IS has developed Wheat Data Interoperability Guidelines that support this capacity-building objective.
- Issue of accessibility: BrAPI is being developed and deployed across Europe for data used in breeding.
- Issue of sustainability: long-term maintenance is required for repositories of standards as well as tools supporting the automated elaboration of FAIR data.

3. Thoughts on FAIRifying data

Leroy Mwanzia (CGIAR-CIAT) presented CIAT's actions to make its data FAIR.

In particular, several actions focus on incentivising researchers to make their data FAIR and accessible, and on fostering a data sharing culture amongst researchers:

- Producing data is valued at the institutional level in the carrier of CIAT researchers;
- Support is provided to publish data papers, which are then a citable asset for researchers;
- Regular progress reports and the Newsletter which contains a section on FAIR compliance are sent to researchers;
- CIAT monitors the FAIR compliance of project deliverables (including data).

Indeed, CIAT has developed indicators for each letter of FAIR in order to monitor the level of FAIR compliance of its data. The interoperability issue is the most challenging one.

4. Ontological annotations supporting FAIR agricultural data

Elizabeth Arnaud (CGIAR-Bioversity) presented examples of standards and tools used by Bioversity to provide FAIR data.

Standard variables and ontological terms are required for reusability of data: e.g. Crop Ontology and GACS are used for the various repositories, databases and datasets supported by CGIAR in order to ensure data interoperability and promote data reuse.

The issue of smart annotation tools was discussed: up till now, automated annotation tools have never been put into production and there is no systematic evaluation of their efficiency. In addition, there is no fully usable hub of ontologies and vocabularies together with related APIs.

5. AgGateway and FAIR Data

Andres Ferreyra (Ag Connections, LLC) presented the initiative AgGateway.

AgGateway brings together companies mainly and focuses on implementing standards in order to promote interoperability in the fields of supply chain and field operations business processes.

Open data (findable and accessible) is not the way to go for the initiative as farmers generate proprietary and personally identifiable data. However, in many cases growers want to share their data with extension services companies (e.g. to receive recommendations specific to their farm). As the data are not represented in the same way by machine manufacturer, sensors, etc., a harmonization process is required to elaborate interoperable, reusable data and enable data exchange.

In order to make data FAIR, the following points are considered:

- Identifiers and reference data
- Variable-type registries: this is especially an issue for geolocation as taking into account geopolitical-context-dependent data is critical in order to provide relevant data to farmers.
- Link between different terminologies that correspond to the same object (e.g. two different companies will use different names for a same crop or product) and use of controlled, extensible vocabularies

Discussion

Several points were raised regarding the FAIRification of data:

- "RDFization" is a crucial step in "FAIRification". As this process requires significant resources (i.e. computing and storage), it is important to prioritise and select the data that you want to format in RDF. In the case of ELIXIR, there is no strong demand for RDF from users as the use case approach allows to provide semantics in critical fields. RDF has however unique characteristics, such as providing unique global identifiers.
- Making geospatial data interoperable with the outside world for Linked Open Data is a challenging issue (when data is location-based).
- Metadata does not provide enough information about the content of the data, nor systematic access to the data through querying. Global standards can be used to guarantee "Findable & Accessible" data:
    - Standard for Observations and Measurements is ISO 19156
    - Standard for Data Quality is ISO 19157
    - Provenance standard is the W3C PROV standard
- "Interoperable" means that data is linked for query search. There is a lack of standardised, up taken semantics for agri-food (e.g. crops and varieties, units). There is a need for smart annotation tools.

- Need to build a community of practice for FAIRification (e.g. WheatIS) and bring together various stakeholders (private/public, researchers, data scientists, data producers, ontologists,…). Demonstrating integration results can show the value of such communities of practices. Incentives are required at the institutional level in order to encourage researchers to make their data available.
- Operational support to researchers in making data FAIR is especially required when it comes to legal and intellectual property issues, data ownership and related licences.
- Issue of maintaining archives, data repositories and standards in the long-term. This is especially the case for developing countries as FAIRifying data requires increased capacity in knowledge management.

Session 2. Data repositories and discovery services

Input

The analysis of the data landscape shows a huge diversity of data repositories but it is not easy to identify them, nor to assess the relevance of their content.

Guiding questions:

a) What kind of discovery services? How to implement them? ...
b) Trusted data repositories: what does it mean in terms of technical requirements, business models, policy?
c) Quality certification? seal of approval?
d) Data culture vs. API culture
e) Role of the data repository in the value chain

Presentations

1. Data discovery through federated dataset catalogues

Valeria Pesce (GFAR) presented the challenges related to data discovery and highlighted the example of the federated dataset catalogue CIARD RING.

Dataset discovery relies on the following elements:

- The quality of the metadata related to the dataset: local experts that know the data will be more likely to provide high-quality metadata;
- The interoperability of the metadata: machine-readable metadata using shared semantics;
- Secondary catalogues and APIs for dataset discoverability and querying.

In the case of the secondary catalogue CIARD RING, the quality and interoperability of the metadata is very poor. Thus, there is a need to create incentives to provide good, interoperable metadata as well as to clarify the roles for both primary and secondary catalogues in ensuring and/or improving metadata quality and interoperability.

2. IFPRI's Open Data in the Dataverse

Indira Yerramareddy (CGIAR-IFPRI) presented IFPRI's use of Dataverse for publishing and sharing datasets.

IFPRI had initially established an institutional policy to publish datasets in 2000, which was revised in 2010 after the setup of an institutional data repository using Dataverse in 2008 in order to specify

objectives and a related timeline. A general Open Access and Data Management Policy was established in 2013 for all CGIAR centres.

IFPRI has now 250 datasets published in its Dataverse repository, which has increased the findability and reusability of IFPRI datasets. One major difficulty is to achieve interoperability with other research centres (both CGIAR and external research centres).

In addition, dataset publication has occurred more quickly when such efforts have been clearly funded, otherwise there is still a limited buy-in from researchers. There is a need to create incentives and provide support to researchers to make IFPRI datasets available. In particular, linking datasets to publications and vice-versa is an effective mechanism to foster data sharing.

IFRI is involved in the CGIAR Big Data Platform. Metadata standards set under this project help and improve interoperability challenges.

3. Cirad – Dataverse: A platform to manage, work, share and publish data

Sophie Fortuno (CIRAD) described CIRAD's project "Patrimoine Numérique Scientifique" (PNS) and the future use of Dataverse at institutional level.

The PNS project seeks to identify, valorize and preserve data produced through CIRAD's research activities. It has demonstrated the diversity of covered research topics and the large geographic repartition of data. More importantly, it has highlighted the issue of data storage and sharing as current storage practices mainly rely on individual storage resources (e.g. workstation and external device) and data sharing usually occurs amongst coworkers and partners only.

This is why CIRAD had chosen to use Dataverse in order to share and increase the discoverability of the data identified in the PNS project. Two main issues have been encountered so far:

- Scientists know how to describe their data but they need a general data entry framework and common recommendations;
- There is a need for a shared legal framework and guidance.

CIRAD aims at an operational service by beginning of 2018 thanks to training and the development of institutional data management rules. It is reflecting on potential services to implement (e.g. data harvesting, analysis and mining, data visualization, semantics, APIs, etc.)

4. EUDAT – The pan-European data infrastructure

Marion Massol (CINES) presented the EUDAT infrastructure.

EUDAT is a European data infrastructure that offers a complete set of research data services (including personal cloud, data transfer, data repository, data preservation and data catalogue), expertise and technology solutions to all European scientists and researchers.

Pilots are set up in order to deploy EUDAT with specific communities. In particular, long-term digital preservation is a key issue that is addressed by the European Trust Digital Repository (ETDR). In particular:

- In addition to structured and published data, related context information as well as good metadata is key to make the data usable by the next generation of scientists.
- We need to make sure that we generate objects that will be readable in the future.

5. AgroPortal: a vocabulary and ontology repository for agronomy, plant sciences, biodiversity and nutrition

Clément Jonquet (University of Montpellier) presented the AgroPortal project.

AgroPortal acts as a repository for agronomic ontologies and offers a wide range of ontology tools (annotation, alignment, etc.). In the long-term, it aspires to become a *Platform-as-a-Service*.

It tracks the use of shared ontologies (who and what) via in-built metrics and relies on five driving use cases that range from the development of specific crop ontologies (e.g. IBC Rice Genomics & AgroLD project) to the collection of ontologies at institutional (LovInra) and global levels (GODAN VEST/AgroPortal MAP of standards).

Within the next two years AgroPortal seeks to map and harvest ontologies and related metadata in order to have a clear view of the ontology landscape in agronomy as well as harmonise the metadata and promote the adoption of semantic web in order to align ontologies (especially with the GODAN Map of Standards and GACS).

Discussion

Expected timelines over a 10-year period were discussed:

f) 2017-2019 Short Term: Focus on building demonstrators and understanding cultural challenges
g) 2020-2025 Medium Term: With an injection of 10M€ deliver working pilots and continue to drive cultural change
h) 2025-2027 Long Term: EOSC is operating; agriculture-focused services are active, operating and sustainable.

Vision

Several long-term objectives were defined (i.e. 10 years from now):

- All research data are preserved and documented;
- A federated ecosystem of data catalogues is achieved, allowing the automation of discovery services;
- Standards for data publishing are developed and used: (i) for data interoperability (metadata, semantics, etc.) and (ii) regarding workflows and tools;
- Avoid duplication and ensure long-term maintenance through efficient and sustainable platforms;
- Foster data reuse through sharing of and open access to data and thanks to strong policy support, especially in order to ensure that enough of the web of data remains in the public/commons domain;
- Create a culture for data sharing through social and business incentives, and by demonstrating value both at community/societal and individual levels (e.g. benefits to share data for data owners);
- Generate and use knowledge relying on shared data.

Proposals for 5-year projects before 2025

[1] Federated dataset catalogue

This catalogue would support efficient dataset discovery through the development of an initial dataset aggregator in agri-food at a global level (including both public and private partners). Hence, it would require rich metadata.

It would rely on a voluntary basis of individual data repositories.

[2] Proofs of Concepts

- For semantics with show cases (e.g. methods, traits, units)
- For applications such as text-mining, data analysis, etc.

[3] Semantics project

This project would seek to increase usage of standardised semantics (e.g. AgroPortal) by helping the community to adopt standards.

Use cases would be developed in order to show the usefulness of standardised semantics.

A coordination and networking mechanism could be established via a "Hub"/portal/… on semantics.

[4] Cultural change project

This project would seek to create awareness amongst communities on open and FAIR data.

It would aim at:

- Convincing communities of the value of sharing data;
- Promoting good practices.

Innovative tools could be used (e.g. competitions).

Contribution to e-ROSA's roadmap elaboration

In order to take this vision exercise further, the latter should facilitate the identification of (i) the steps required to achieve the vision and (ii) the resources that already exist and are available. As such, by crossing those two types of information, you can determine whether a specific step will be quick (already existing resources) or long (no resources available).

Session 3. E-infrastructures and data services

Input

E-infrastructures should offer data discovery services, data interoperability and integration services and the ability to build a virtual research environment in a big data context. The goal of this session is to establish a dialogue between "generic e-infrastructures and services" and the "application domains". An important objective is to examine how to achieve virtual research environments that provide the required functions and workflows to support interdisciplinary research with FAIR data, big data analytics (e.g. workflows and code) over distributed heterogeneous resources.

Guiding questions:

a) Inventory and assessment of existing services, Kiss method: what should be Kept, Improved, Started (what is missing), Stopped?
b) Interaction between e-infrastructures (generic vs. thematic)

c) What services should be generic? How to ensure affordable and sustainable access to data analytics and computation to researchers, regardless of their disciplines or where they are located?
d) What are the specific service requirements for the Agri-food sciences?
e) What are our "common goods"?

Presentations

1. Big Data and Open Science in agricultural and environmental research

Rob Lokers (WUR) gave an introductory presentation to this session.

The agricultural sector heavily relies on integrated research across disciplines (i.e. crop science, animal science, environmental science, economics). While the Volume and Velocity aspects of big data are increasingly being tackled through ICT components, the challenge for multidisciplinary domains like agri-food lies in handling the issues of Variety and Veracity. In addition, unprecedented computational capabilities are now available and significantly increase the potential for data collection, analytics, knowledge generation and decision-making support.

However, despite these high expectations several challenges are to be addressed in order to achieve open, virtual science (e.g. adoption of technologies, semantics and interoperability issue, change of culture, etc.). A large range of infrastructures that seek to address one or several of these challenges already exist and differ in terms of genericity vs. domain-specificity, geographical coverage and services developed.

The key issue is to understand how to connect and effectively use these already existing infrastructures in view of developing a future e-infrastructure for agri-food research and identify missing components and required improvements.

2. Ensembl, ELIXIR and engineering interconnections

Paul Kersey (EMBL-EBI) presented the international organisation for research in life sciences EMBL-EBI and the pan-European data infrastructure ELIXIR.

EMBL-EBI has large data resources and develops services that can serve the agricultural community. This is for example the case of the Ensembl tool that supports genome analysis and visualisation, including for species of agricultural interest.

ELIXIR is the largest ESFRI (European Strategy Forum on Research Infrastructures) initiative. Several challenges are to be addressed in order to implement such an initiative (i.e. agreement on data access and interoperability policies, strong commitment to FAIR and open data, secure sustainable funding for long term resources, agreement on division of labour to maximise total impact). ELIXIR puts a particular emphasis on harvesting FAIR & OPEN data, developing standards for data and metadata (e.g. MIAPPE) and supporting the establishment of interoperable data infrastructures (e.g. TransPLANT).

As for EMBL-EBI, the question of the potential link between ELIXIR and other related initiatives/infrastructures (including the envisioned e-infrastructure under e-ROSA) is key.

3. EUDAT – The pan-European data infrastructure

Alexis Jean-Laurent (CINES) presented the EUDAT infrastructure (see presentation n°4 in Session 2).

EUDAT is a European data infrastructure that offers a complete set of research data services (including personal cloud, data transfer, data repository, data preservation and data catalogue), expertise and technology solutions to all European scientists and researchers.

Pilots are set up in order to deploy EUDAT with specific communities: e.g. the Herbadrop Data Pilot for long-term preservation of specimen in images, including information extraction via Optical Character Recognition (OCR) processing.

4. Agrisemantics – Knowledge organisation for a food secure world

Sophie Aubin (INRA) described the Agrisemantics initiative.

The general issue concerning semantics (all research domains included) is the fragmentation of the landscape of semantics resources: the latter are usually locally elaborated and not reusable, and it is often easier to create new resources rather than extending existing ones. Agrisemantics seeks to support global semantics for agri-food.

In order to do so, FAIR semantic resources relying on Linked Open Data are required in order to create FAIR data and metadata. In particular, Agrisemantics supports the VEST/AgroPortal MAP of standards, AgroPortal and GACS as key components of the global landscape of semantic resources.

5. EGI

Enol Fernandez (EGI Foundation) presented EGI (European Grid Infrastructure).

EGI seeks to support Advanced Computing for Research through a distributed computing e-infrastructure across the world. In particular, EGI provides:

- Common services, which include a service catalogue for researchers,
- Community support services, which focus on thematic services for EGI partners (i.e. scientific application and tools). For instance, EGI provides Cloud Compute services to host Virtual Research Environments such as iMarine.

EGI is directly supporting the EOSC architecture.

6. OpenMinTeD: Overview and challenges

Robert Bossy (INRA) presented the OpenMinTeD initiative.

OpenMinTeD seeks to develop an open Text and Data Mining (TDM) platform and infrastructure for research. It supports the integration of the various existing TDM services and has launched several use cases amongst which one is focused on Agronomy and Biodiversity.

OpenMinTeD builds on three key elements: 1) a registry of tools, workflows and other resources, 2) a workflow engine and 3) a hosting service. At a technical level, workflow reusability is a significant issue as it requires rich metadata.

Discussion

The discussion focused on the services and components that should be Kept, Improved, Started (what is missing), Stopped (KISS method).

In particular, technical services that should be further developed include:

- Strategic disciplinary data sources: genome, literature…; make sure they remain sustainable

- (Findable) registries and federated search engines for data, services, identifiers, concepts… across disciplines
- (Accessible) Authentication and authorization services (until it is completely transparent to the final user)
- (Findable/Interoperable) Develop and organise the semantic layer (resources and services), develop more efficient OWL/RDF technologies, also across disciplines
- (Reusable) long term preservation services should be taken into account in e-infrastructures: technical but also quality issue
- (Reusable) reproducibility of data and workflows (notebooks, workflows…)
- (Reusable) provenance information
- (big) data analytics and visualization
- Virtual Research Environments (VREs)

Also, non-technical aspects were discussed:

- Incentives for use and training to use e-infrastructures: consider user experience, ergonomy, "easy to use" data and services
- Business models: ensure mid/long-term sustaining of tools and services. This is relevant both for data curation and data analytics. What needs public funding and what can be marketed? A business case is required for each use case.
- Cultural change

Lastly, two parallel ways of working need to be considered:

- General infrastructure issues
- Specific use cases/demonstrators that deal with practical problems: they need to be large enough that you can upscale them.

Regarding the latter, we need high impact use cases with demonstrators that address scientific issues or practical needs in order to show that it is possible, quicker, more efficient… using the e-infrastructure. We should not implement use cases that only demonstrate the e-infrastructure usability. It should also aim at being an incentive for people to share their data. Some demonstrators should imply the integration of data from different disciplines that we identify as urgent/valuable to connect to each other.

Use cases could be either functional or more technical. They should engage with the communities that support existing demonstrators (e.g. AgInfra+, use cases on developing countries within the CGIAR Big Data Platform, etc.). Some examples of domain use cases that were mentioned during the session:

(i) Functional/scientific use cases:

- breeding: link molecular, phenotypic data and physical stocks and rely on active/advanced community and technologies
- track food products from farm to fork
- food security modeling: e.g. yield and nutrition gap analysis

(ii) Technical use cases: One important issue concerns the improvement of data 'Veracity' when interconnecting systems (add structured provenance, automatic control systems based on models). From the e-infrastructure point of view, one specific challenge would be "Veracity" throughout disciplines. What services/features are needed to ensure that non experts can safely use data from other disciplines? Possible solutions include:

- More/better provenance information? Information on consequences of using a given dataset
- Automatic control systems based on models: alerts to the user

- Add control on necessary/minimum metadata to run a workflow

Final group discussions

Final group discussions allowed to brainstorm about gaps, disruptions and trends, as well as specific high-impact demonstrators we should focus on.

Trends

- More data produced
- Voice-based apps: this is changing the nature of data entry
- Image-based interaction on mobile devices
- More, smaller autonomous services and robots for collecting data
- Big data in genomics: Crisper-cas9 is a tool for genomic engineering with very high potential
- More Open and/or FAIR data
- Rise of local repositories
- Increasing data catalog interoperability
- Harmonisation of data publication expectations by funders
- Improvement of data management plan
- Using an ontology as a way to support better reasoning
- More complex workflows supporting increasingly complex, integrated reasoning/modeling on data that is heterogeneous in nature (soil, weather, sensors, biological…), semantics, and formats
- More intelligent data processing
- "Platforms" as working environments
- Visualization tools: How can we present data to the end-user?
- Micro-blogging / very short written communication: User behavior is changing
- Citizen scientists as a data source: engage them in order to foster more community- and society-driven services
- Alignment of existing and future e-infrastructures and VREs with the EOSC vision & architecture
- E-infrastructure architectures: development of Systems-of-systems
- Federated nodes systems (ELIXIR, Wheat IS, …)

Gaps

- Standardized way of cataloguing semantic assets and datasets: Bottleneck; A dynamic self-registering system may be a scalable, sustainable way of doing this.
- Automated maintenance of data catalogues
- Regarding semantics:
    - Need for a sustainable, one-stop shop for agriculture-specific semantic assets;
    - Lack of integration between semantic bits and pieces: Recommendation for GACS' mapping
    - Efficient semantic services that are embedded in e-infrastructures
    - Tested smart applications and tools to index, annotate, structure, transform and reason on data
    - Governance mechanism for semantic assets (at global/disciplinary level?)
- Integrated tools methods for data analysis (e.g. VREs)
- Skills to search and transform data, e.g. build and reuse workflows

- Multiplicity of (big) data producers/owners: not only labs but now farms (sensors), shared infrastructures (e.g. satellites), industry & stores, citizens…; technical and governance issues need to be addressed
- Business models: maintain and develop interest for people; sustainable funding (vs. project-based approach) is required to ensure data sources and services' sustainability
- Develop case studies on (i) achieved/potential impact of FAIR and Open data as well as (ii) existing duplication of research and data collection to justify future investments
- Funding to deal with legacy data, models, and tools
- Incentives: engaging communities to encourage them to share and reuse data
- Mechanism (e.g. network, conference) to share experiences
- Legal gaps

Disruptions

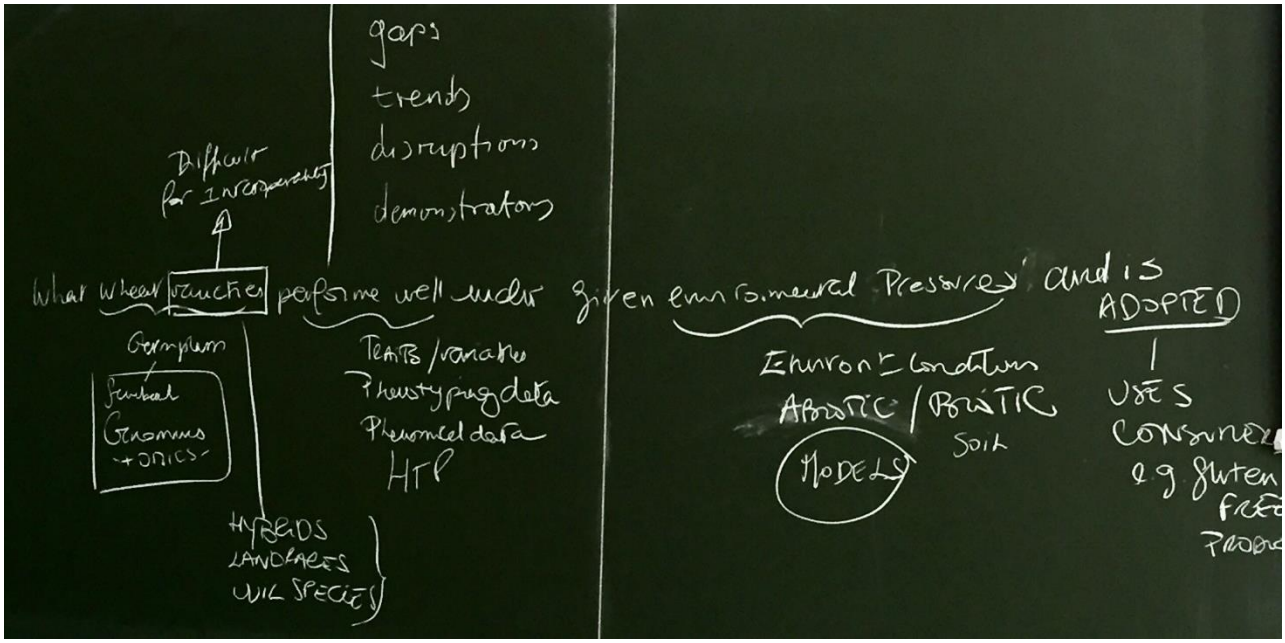
- Blockchain: could be used to implement governance tools for semantic assets, as well as for traceability of sensitive assets
- Artificial Intelligence: data interoperability, processing and analysis *as-a-Service*; in addition, how will this impact our societal models?
- Quantum computing: particularly suited for parallel problems
- 5G: changes for conferences and peer-to-peer communication; possibility of real-time decision-making in the field by integrating machinery and sensors
- Persistent memory: could reduce reliance on databases and introduce new forms for managing data
- Data security: hacking attacks
- Cultural/procedural change for researchers regarding data production: researchers should check data existence, accessibility, reusability, quality and adequacy before producing their own data
- Political and legal changes (e.g. Brexit, economic crisis)

Demonstrators

As explained in Session n°3, use cases/demonstrators can focus both on scientific or technical issues:

- A technical demonstrator identified during the final discussion focused on data visualisation, and more specifically on using semantically-rich ontologies to support visualization, explore datasets and eventually assist discovery.

- A scientific demonstrator was also elaborated during the final discussion and focused on Exploring relationships among wheat varieties and environmental conditions. The latter could be offered to the Wheat IS as an enrichment example. "What wheat varieties perform well under given environmental pressures and are adopted on the market?" was the overall scientific question for this demonstrator (see picture below):
    - Data on wheat varieties especially refers to genomic data;
    - Varieties are a challenging issue for data interoperability;
    - The measure of performance can be achieved thanks to phenomics, phenotyping data and data on traits and variables;
    - Environmental pressures refer to abiotic and biotic conditions, their impact on the performance of a specific variety requires the use of models;
    - The adoption on the market is measured thanks to data on consumers.



Picture of the elaborated use case on wheat varieties' performance

Conclusions

The agri-food sector is dealing with an increasing amount and variety of data due to:

- The multidisciplinary nature of agri-food science, which is adopting a more and more systemic approach;
- The automation of data collection thanks to robots, sensors, etc., as well as new engineering tools such as in the omics field;
- The development of new types of data sources and providers: e.g. Internet of Things, citizen science, voice- and image-based applications, micro-blogging, etc.

In addition, we have increasing computational capabilities for data collection and analysis, which can support more efficient knowledge generation and decision-making. As such, we today have the opportunity to rely on more complex and integrated reasoning and modelling, requiring the integration of these numerous, dispersed and heterogeneous data.

Many initiatives and infrastructure services already exist and need stronger and coordinated support to promote Open Science for agri-food and implement the European Open Science Cloud (EOSC). There is a pressing need for the development of a common e-infrastructure framework in order to:

- connect data and connect infrastructures,
- integrate existing initiatives into a common framework at a global level,
- share efforts and resources,
- support a collective change of practices through the adoption of shared standards,
- provide a pre-competitive space for sharing data and speeding up innovation processes.

Various issues need to be taken into account when envisioning an e-infrastructure, including:

- The articulation between general e-infrastructure issues (technical and non-technical) with specific scientific data-related needs (e.g. thanks to Virtual Research Environments);
- The easy access to and use of the e-infrastructure by researchers;
- Challenges linked to a distributed organisation (i.e. in nodes):
    - Agree on data access and interoperability policies;
    - Secure sustainable funding for long-term common resources;
    - Agree on an efficient division of labour to maximise total impact;
    - Synchronise technical updates (e.g. at a local level, when implementing improvements of a common data model);
- The identification of needs and services to be covered by the e-infrastructure;
- The evaluation of an e-infrastructure (e.g. assess impact on change in practices).

The future e-infrastructure for agri-food can help address various challenges, which can be categorised into two overall categories: i.e. technical (1) and cultural (2) challenges. In addition, the development of demonstrators (3) is crucial for an efficient development and implementation of e-infrastructure services.

1. Technical challenges

a) FAIR principles

The most challenging issue in making data Findable, Accessible, Interoperable and Reusable is the interoperability issue.

Common standards

Common standards are required to support interchange of data and to achieve interoperability amongst data repositories. Existing standards need to be improved (e.g. OWL/RDF technologies) and the issue of competing standards need to addressed. In addition, it is crucial to understand which standards (e.g. *de facto* standards) have the best chance of being adopted by communities for them to become effective standards.

Adopting RDF formats has been recognised as a crucial step in making data FAIR. However, this technology is demanding in terms of computation and storage resources, which is why there is a need to prioritise and select the data that will be formatted in RDF.

Semantics

Semantics are a major issue in the agri-food sector. Indeed, semantic resources are highly fragmented and there is a lack of standards for semantics in agronomy. There is also a lack of smart annotation tools and of a fully usable hub of ontologies/vocabularies with related APIs. Indeed, existing annotation tools are not really in production at the moment and there is no evaluation of their efficiency.

As such, specific needs have been identified:

- To catalogue semantic assets (i.e. mapping and harvesting) through a standardised procedure;
- To align existing ontologies and vocabularies, and create links between terminologies that refer to the same object (e.g. between different companies) in order to foster open innovation;
- To promote the use of standardised reference semantic resources and of the web of semantic;
- To embed semantic services in e-infrastructures.

Geolocalised data

The proper integration of location-based data is crucial for agri-food research as the latter is location-dependent. Related challenges are the following:

- There are no variable-type registries for geopolitical-context-dependent variables;
- Linked Open Data tools and formats are not adapted for Geo-information.

Reusability

Issues related to data reusability concern:

- The provision of provenance information;
- Long-term preservation: in particular, establishing formats that will still be readable in the future is a key issue;
- Going beyond FAIR *data* by generating FAIR *knowledge*: we need to develop implementation mechanism for this to become a reality.

b) <u>E-infrastructure services</u>

E-infrastructure architecture

The envisioned e-infrastructure should be aligned with the EOSC architecture in order to be easily integrated in a European and transdisciplinary framework of services. The EOSC architecture will rely on a System-of-systems with facilities provided *as-a-Service*. The development of the proposal for the EOSC architecture will directly nurture the one for the e-infrastructure for agri-food. This will ensure an operational articulation between generic services that are developed and provided by pan-

European infrastructures such as EUDAT, OpenAIRE, GEANT, etc., and domain-specific services that will be facilitated by the e-infrastructure for agri-food. In particular, Authentication and Authorisation services need to be improved as a generic service.

Data discovery

Data discovery can be significantly enhanced thanks to:

- High-quality metadata: local experts that collect the data are more likely to provide the required quality of metadata;
- Interoperable metadata: i.e. machine-readable metadata, shared semantics and standards for metadata that guarantee accessible and findable data (ISO, W3C…);
- Federated catalogues that catalogue datasets in a standardised way, and APIs for data discovery and querying.

Data processing

Several points were raised regarding this issue:

- Virtual research environments: VREs can support use cases in providing tailored solutions for specific communities. They should feed into the prototyping, development and implementation of general services provided by the e-infrastructure.
- Workflows:
  - Increasingly complex workflows are required in order to support complex, integrated reasoning and modelling on data that is heterogeneous in nature (i.e. multidisciplinary), semantics and formats;
  - Standards are required to build interoperable workflows and tools (i.e. with rich metadata);
  - Develop skills to search and transform data, i.e. how to build and reuse workflows, is a crucial need.
- Intelligent data processing
- Data visualisation and exploration

c) Disruptive technologies

In the perspective of developing an e-infrastructure for agri-food and related services to support open science and innovation, we need to anticipate to what extent disruptive technologies can positively or negatively impact our field and how we can potentially use (some of) them, including:

- Blockchain
- Artificial intelligence
- Quantum Computing
- 5G
- Persistent memory
- Data Security (hacking attacks)

2. Cultural challenges

a) Community engagement and incentives for cultural change in research practice

Community management is required in order to bring together the various stakeholders that are involved throughout the data management and research cycle: i.e. researchers, data scientists, data producers and ontology specialists.

Dedicated mechanisms should be set up in order to incentivise researchers to share their data. This can for instance be achieved by:

- valuing data sharing and publication at the institutional level (e.g. within the researcher's carrier evaluation procedure);
- providing a technical support to (i) publish data papers (i.e. citable assets that can be valued in the carrier of a researcher) and (ii) share FAIR data (i.e. provide support on which metadata to provide, which standards to use, which licenses to choose, etc.);
- monitoring data sharing/openness and FAIR compliance and disseminating monitoring results to encourage researchers to do so as an institutional strategic objective;
- demonstrating integration results and triggered benefits in order to convince of the value of sharing data for research.

This should eventually support a long-term shift in paradigm regarding the way we do research: researchers should first check data existence, accessibility, reusability, quality and adequacy, before producing their own data.

b) <u>Policies and regulations</u>

The latter play a crucial role in:

- preventing from monopoly of data (e.g. in the case of machine-generated data): legal barriers related to the issue of data ownership need to be addressed;
- supporting societal priorities by keeping enough of the web of data in the public domain to support generation of knowledge of public interest;
- fostering data sharing by harmonising data publication expectations amongst funders.

c) <u>Sustainability & governance</u>

Governance

The envisioned e-infrastructure for agri-food needs to build on a long-term governance model which:

- relies on the EOSC Thematic Cloud on Food systems;
- facilitates the connection of existing e-infrastructures, both generic and domain-specific;
- relies on pilots/demonstrators already supported by existing initiatives;
- builds on what is already supported by individual organisations;
- supports a specific governance mechanism for semantic assets in agri-food (at global/disciplinary level?).

Demonstration of impact

In order to maintain the interest of people in the long-term for a sustainable e-infrastructure, investments need to be justified by making the case for the need for such an e-infrastructure by demonstrating clear, impactful examples of:

- what is already possible thanks to data;
- cases of duplication of data collection or analysis because of a lack of shared data and tools;
- what could be done thanks to the e-infrastructure and shared data (i.e. use cases), for instance:
  - overcoming fragmentation by integrating strategic data sources, across disciplines, across regions;
  - supporting food security;
  - supporting food integrity (e.g. transparent certification).

We need to clearly identify short/medium-term investments required to address potential scientific use cases and link them to expected research and societal impacts.

Business models

Developed business models need to:

- Support long-term maintenance of data repositories, catalogues, standards and tools for FAIR data (e.g. automation of discovery services);
- Take into account societal priorities, business interests and privacy concerns;
- Support clear benefits for individual data owners that share their data in order to further foster data sharing.

3. Demonstrators

Both scientific and technical demonstrators are required to address specific needs of scientific communities and provide scalable, more generic services that can serve agri-food research as a whole. They allow to address both technical and cultural challenges as described above. Examples of high-impact demonstrators that could be developed are listed below:

a) Scientific use cases

- Food security modeling
- Identifying weeds
- Identifying diseases
- Predicting toxicology
- Breeding data
- Track food products from farm to fork
- Wheat varieties that perform well under given environmental pressures and that are adopted on the market

b) Technical demonstrators

- Semantics
- Federated dataset catalogue
- Linking of location-based data
- Data visualisation
- Data veracity across disciplines

To conclude, this first e-ROSA Stakeholder Workshop succeeded in bringing together targeted stakeholders that could fuel the strategic dialogue on the need for an e-infrastructure for Open Science in agri-food. In particular, the workshop allowed to:

- facilitate a shared analysis of the actual landscape by sharing information on current initiatives that can feed into such a future e-infrastructure, identifying trends, specific gaps and challenges as synthesised above;

- initiate the development of a common framework to build the vision for the e-infrastructure: we discussed the development of a vision for the next ten years as well as a phasing of actions through the identification of potential demonstrators to be implemented in the short-term and of longer-term (i.e. 5-year) projects to be supported afterwards (e.g. federated dataset catalogue, development of semantic resources, support for capacity-building and cultural change, etc.);

- identify and initiate the design of challenging, high-impact use cases that support the implementation of open science throughout the food system value chain.

Overall, the workshop provided valuable outcomes as it allowed to highlight the challenges that need to be addressed in order to support efficient sharing of FAIR data for data-driven research and innovation in the agri-food sector. Overarching issues emerged from the discussions, including:

- The strategic reasons justifying the need for an e-infrastructure in agri-food;
- The technical challenges that this e-infrastructure could help address: this includes the application of the FAIR principles which especially relies on support for interoperability and shared semantic resources, the delivery of data discovery and processing services and the integration of disruptive technologies;
- The cultural challenges that the e-infrastructure can help address: this especially concerns community management and commitment for a change of research practice as well as sustainable governance and business models that provide incentives for data sharing;
- The need to further identify and develop crucial use cases (both scientific and technical) that the e-infrastructure could support.

The outcomes of this first workshop provide strong input:

(i) for the elaboration of a common vision paper that affirms the need for an e-infrastructure for agri-food science;
(ii) to further discussions on the specific needs and vision for semantics in agriculture during the next meetings of the Agrisemantics working group that will take place during the IGAD pre-meeting and the Research Data Alliance plenary in Montreal; and
(iii) to the second e-ROSA Stakeholder Workshop on 27-28 November 2017 that will focus on the needs and use cases identified by scientific communities and the development of the common vision for the e-infrastructure in the context of the Thematic Cloud on Food Systems supported by EOSC.

--

All workshop presentations are available at: https://www.slideshare.net/tag/erosastakeholderws1

Report by Madeleine Huber (INRA)

List of participants

	Name	Affiliation
1	Anne-Françoise Adam-Blondon	INRA
2	Erick Antezana	Bayer
3	Elizabeth Arnaud	Bioversity International
4	Sophie Aubin	INRA
5	Christopher Baker	IPSNP Computing Inc.
6	Robert Bossy	INRA
7	Beert Bredeweg	University of Amsterdam
8	Christopher Brewster	TNO
9	Patrice Buche	INRA
10	Jandirk Bulens	Wageningen Research
11	Michaël Chelle	INRA
12	Gianpaolo Coro	CNR
13	Enol Fernandez	EGI Foundation
14	Andres Ferreyra	Ag Connections, LLC.
15	Sophie Fortuno	CIRAD
16	Gilles Garric	INRA
17	Wim Haentjens	European Commission
18	Odile Hologne	INRA
19	Madeleine Huber	INRA
20	Corina Itcus	INCDSB
21	Alexis Jean-Laurent	CINES
22	Andy Jenkinson	Agrimetrics
23	Allan Leck Jensen	ICROFS
24	Clement Jonquet	University of Montpellier
25	Pythagoras Karampiperis	Agroknow
26	Johannes Keiser	GODAN Secretariat
27	Paul Kersey	EMBL-EBI
28	Rob Knapen	Wageningen Research
29	Simon Lambert	STFC

30	Pierre Larmande	IRD
31	David Legland	INRA
32	Jonathan Levin	INRA
33	Rob Lokers	Wageningen Research
34	Nikolaos Manouselis	Agroknow
35	Daniel Martini	KTBL e. V.
36	Hans Marvin	Wageningen Research
37	Marion Massol	CINES
38	Graham Mullier	Syngenta
39	Leroy Mwanzia (remote)	CGIAR
40	Pascal Neveu	INRA
41	Órlaith Ni Choncubhair	Teagasc
42	Seishi Ninomiya	University of Tokyo
43	Martin Parr	GODAN Secretariat
44	Valeria Pesce	GFAR
45	François Pinet	Irstea
46	Antonio Sanchez-Padial	INIA
47	Anne Toulet	University of Montpellier
48	Xenophon Tsilibaris	GRNET
49	Iris Maria Tusa	INCDSB
50	Indira Yerramareddy (remote)	CGIAR
51	Panagiotis Zervas	Agroknow

Profile of participants

The workshop gathered a variety of stakeholders in terms of location and geographic coverage of the represented organisation or network (see Figure 1 and 2) and type of organisation (see Figure 3). It also succeeded in bringing together a significant number of persons that are external to the e-ROSA project's core team (see Figure 4).

Figure 1. Location of the represented organisations

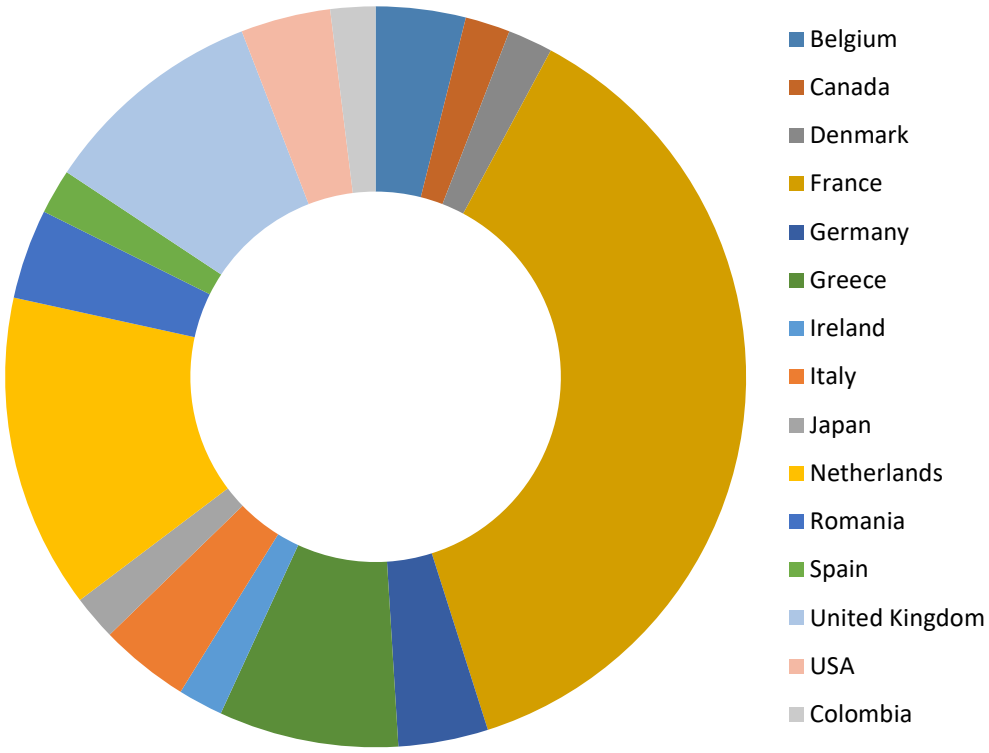


Figure 2. Geographic coverage of the represented organisation or network

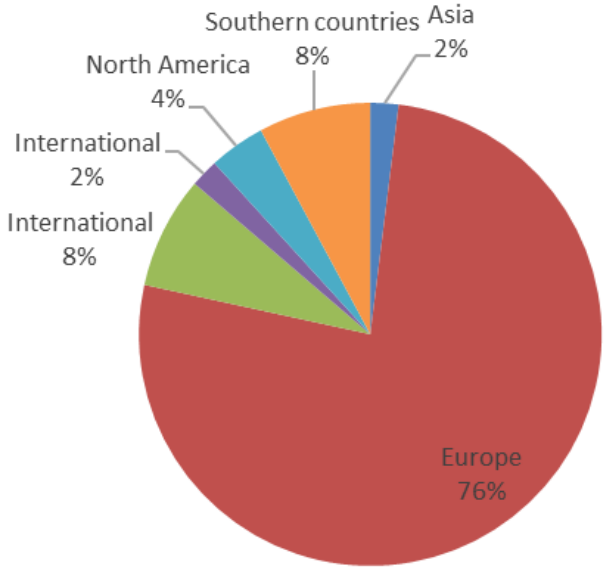


Figure 3. Types of organisations

Figure 4. Proportion of participants that are involved or not in the e-ROSA project team