eROSA

e-infrastructure Roadmap
for Open Science in Agriculture

# Impact Chains and Science Solutions

## Requirements and solutions for an e-infrastructure supporting Open Science with impact in the agriculture and food domain

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

| | |
|---|---|
| **DELIVERABLE NUMBER** | D2.2 |
| **DELIVERABLE TITLE** | Impact Chains and Science Solutions |
| **RESPONSIBLE AUTHOR** | Rob Lokers, Sander Janssen, Wageningen UR |

| 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 | Ms Georgia Tzenou |

| | |
|---|---|
| WORKPACKAGE N. \| TITLE | WP2 \| Challenges & Ambitions |
| WORKPACKAGE LEADER | WUR |
| DELIVERABLE N. \| TITLE | D2.2 \| Impact Chains and Science Solutions |
| RESPONSIBLE AUTHOR | Sander Janssen, Wageningen UR |
| REPLY TO | Sander.janssen@wur.nl |
| DOCUMENT URL | http://www.erosa.aginfra.eu/deliverables |
| DATE OF DELIVERY (CONTRACTUAL) | 30 April 2018 (M16) |
| DATE OF DELIVERY (SUBMITTED) | 16 July 2018 (M18) |
| VERSION \| STATUS | V1.0 \| Final |
| NATURE | R (Report) |
| DISSEMINATION LEVEL | PU (Public) |
| AUTHORS (PARTNER) | Rob Lokers (WUR), Sander Janssen (WUR), Jandirk Bulens (WUR), Rob Knapen (WUR), Michael Chelle (INRA) |

| VERSION | MODIFICATION(S) | DATE | AUTHOR(S) |
|---|---|---|---|
| 0.1 | Initial version | May 2018 | Rob Lokers (WUR), Sander Janssen (WUR), Jandirk Bulens (WUR), Rob Knapen (WUR), Michael Chelle (INRA) |
| 0.9 | Final draft for review | 13 July 2018 | Rob Lokers (WUR), Sander Janssen (WUR) |
| 0.91 | Review | 13 July 2018 | Odile Hologne (INRA) |
| 0.92 | Review procesed | 14 July 2018 | Rob Lokers (WUR) |
| 1 | Final version (after internal review) | 16 July 2018 | Rob Lokers (WUR), Sander Janssen (WUR), andirk Bulens (WUR), Rob Knapen (WUR), Michael Chelle (INRA) |

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

## EXECUTIVE SUMMARY

E-Infrastructures for research have to help researchers to be more effective and efficient in contributing to the achievement of the societal or grand challenges as captured by the Sustainable Development Goals, or the Europe 2020 strategy. Even if this sounds obvious, this is not self-evident how to link challenges as comprised by the societal goals and impacts of research, to the role of research and the supporting role of e-infrastructures and associated ICT and data solutions.

To analyse the mechanisms of endeavours that aim to contribute to the European and global societal challenges, e-ROSA has followed an approach based on impact chains. The rationale was that deconstructing the path towards impact via the outputs of initiatives and the subsequent short-term and long-term outcomes would provide a better view on the logic behind this. As such, it provides better insight in the role of research and specifically the advances that need to be made in the area of Open Science and the development, broad adoption and effective use of an e-infrastructure for agriculture and food. The approach has been used to collect information from stakeholders in the field in various ways. A stakeholder workshop was held, that aimed at collecting views on achieving impact through research in various areas of agri-food. Additionally, stakeholders were asked to describe case studies of data-intensive research areas. Both, activities were directed by an impact chain analysis and from there on linked through to the role of research in general and the associated challenges and requirements with regard to Open Science and the role that an e-infrastructure for agriculture and food could play.

This deliverable provides an analysis of the workshop and case study results and the underlying impact chains. It considers the linkages of research with the impact chains of a diverse range of subdomains and cases, and states the relevant roles of research in achieving impact. It also provides a shortlist of requirements and solutions in the areas of sharing of resources, connecting resources and collaboration. We consider these as important inputs for the roadmap towards an e-infrastructure for agriculture and food. As such, they need to be included as part of the subsequent strategy to develop this infrastructure and its organisation, its adoption by the broader scientific communities and its governance.

# 1 CONTENTS

**TABLE OF FIGURES**

# 1 INTRODUCTION

As part of the development of a roadmap towards an e-Infrastructure for Open Science in Agri-Food Sciences, the e-ROSA project has identified the “Grand Challenges” and links to the role of research and e-infrastructures for Open Science in the agriculture and food domain.

To identify the relevant requirements and solutions that are needed to develop a roadmap towards an e-infrastructure for agriculture and food, impact chains of agri-food cases have been examined, deconstructing the paths towards impact, tracking back from impacts to the output of (research) initiatives though short-term and long-term outcomes. To support this approach, information from stakeholders in the domain was gathered in various ways. A set of representative use cases for Open Science and the use of e-infrastructures was collected through the agri-food networks. The impact chains of these use cases were analysed, as well as the challenges and solutions that experts thought that existed for these use cases and their trajectory towards Open Science. Moreover, a stakeholder workshop was organized to identify challenges for the pathway to using e-infrastructures for research in agriculture and food.

This document describes the process followed to collect information from stakeholders through use cases and interactive sessions. It provides an analysis of the workshop and case study results, considers the linkages of research with the impact chains of a diverse range of subdomains and cases, and states the relevant roles of research in achieving impact. It also provides a shortlist of requirements and solutions in the areas of Open Science: sharing of resources, connecting resources and collaboration. We consider the outcomes as an important input for the roadmap towards an e-infrastructure for agriculture and food and as important elements of the subsequent strategy to that is needed to develop this infrastructure, its organisation, the adoption by the broader scientific communities and its overall governance.

# 2 APPROACH

## 2.1 OBJECTIVES AND ACTIVITIES

The operational objectives of the work that underlies this document were (1) to link research for impact towards the societal challenges to data and IT challenges that could potentially be solved by e-Infrastructures and (2) raise awareness in the community of researchers working on research for societal impacts of potential e-services they could benefit from.

e-ROSA examined Open Science use cases from the domain of agriculture and food to analyse a range of research impact chains and to identify the challenges and solutions with regard to future Open Science and the use of e-infrastructures. To accomplish that the following activities were performed:

- A campaign was set up to collect representative scientific use cases that are expected to be relevant in the face of the further development of Open Science and use of e-infrastructures, and that are thus expected to be able to link their path towards impact to demands with regard to the desired development of an e-infrastructure for the agriculture and food domain.
    - Base these storylines on the Impact Chain approach (see background below at Section end)
    - Incorporate an awareness of end-users and beneficiaries of the research
    - Reflect on the role of research in tackling societal challenges, also in relation to supporting decision making by private and public sector or by civil society
- The e-ROSA project organised a “Challenges and Solutions Envisioning” workshop, discussing Open Science and e-infrastrucure related challenges and potential solutions through the perspective of different research subdomains of the agri-food sector.
- The results of the use cases exercise and the outcomes of the workshop were analysed and were translated into a set of requirements and potential solutions that can be incorporated into the roadmap developed by e-ROSA and that should drive the development of a future e-infrastructure for agriculture and food research.

To apply this approach effectively, to be able to delineate the problem space and to be able to track back from high level societal challenges to research activities, and subsequently link to Open Science and the future use of e-infrastructures, the following general consideration have to be taken into account:

- The societal challenges as covered in for example the SDGs, Food2030 or policy agendas (such as gender inclusive development, resilient agriculture) are generally on a very high level of abstraction, therefore potentially a massive amount of research could fit under it. In many specific research projects a link is made between these abstract societal challenges, and the concrete day-to-day research objectives. Thus, it is required to reason through the link between the abstract societal challenges and the concrete research priorities and activities, to get to a link with IT and data as comprised in e-infrastructures.
- There is a challenge in delineating the problem space. For example, a societal goal as SDG 2 on Ending Hunger links to malnutrition, undernutrition, obesity, and sustainable agriculture, thus combining health aspects to nutritional intake to agricultural production of those nutritional foods and finally, environmental conditions like soil health and climate change. For the purpose of this project, it seems relevant to include the nutritional aspects of agricultural production, but not include the full link to health. Especially a value chain perspective from production to consumption of foods (i.e. a healthy diet in local context) seems relevant as here some of the data and ICT challenges will be addressed in combining different types of data.

## 2.2 IMPACT CHAIN APPROACH

For a broad theme like agriculture and nutrition, it is a complex task to properly address the key features mentioned in the ToR. To support that, the Impact Chain Approach provides a useful logic model and framework for the general project design in a well-structured and transparent interaction with all partners involved.

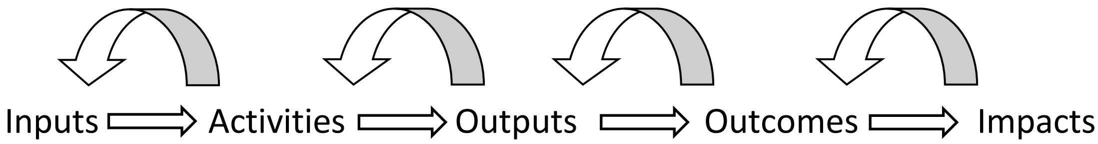



**Figure 1 the theoretical concepts behind an impact chain approach**

**Inputs** are the materials and resources that are used in an activity. This level of detail is not relevant for the project design at this early stage, but will come back later, when a more detailed project planning will be made.
**Activities** are what you actually do to initiate the change you want to achieve.
**Outputs** are the most immediate results of your activities. They create the potential for outcomes and impact to occur.
**Outcomes** describe the true changes that occur as a result of the activities. In many cases, they are divided into shorter and longer term outcomes to create an easier understandable link between output and impacts.
**Impact** represents the vision of a preferred future that underlines why the action is important. It refers to the longer-term change that you hope your action will help create.

To deliver research with impact it is important to think backward along the chain, starting with the desired impacts by asking “what do we want to accomplish” and then to ask which outcomes and outputs are required and which activities should be undertaken to achieve this. It is important to note that – in order to generate impact in the real world – the outcomes and impacts should be defined ‘beyond’ the limits of the project. The project should deliver outputs within the project lifetime as specified in the project plan. In a successful project and supported by the right environmental conditions, these outputs contribute to the realisation of outcomes and impacts in the outer world. However, in contrast to outputs/deliverables, the project cannot be held formally responsible for generating outcomes and impacts, since this often depends on conditions and events that are beyond the control of the project.

## 2.3 METHOD AND DATA COLLECTION

The template for the use cases was completed in different ways to achieve a good coverage over the different topics in agri-food sciences, which constitutes the data collection for this deliverable:

1. Initial templates were elaborated by WUR to enable lesson learning, and to fine-tune the questions, required to get a storyline. In these first three templates, it appeared that the topic of monitoring & evaluation and its elaboration into good M&E indicators is challenging to complete. Second, an iterative approach of developing a draft story line, discussing it, then improving it to the final version worked best to get a worked out version, with a well-thought through storyline. This implies that it is not easy for researchers to complete storylines in one go. Most of these results were already captured in D2.1
2. An online campaign for additional use cases was organised: A google-form was set up reflecting the questions as developed in the use case template. The survey was widely advertised through the e-ROSA twitter channel and e-ROSA website, and through the private twitter accounts of e-ROSA researchers. Announcements were made through GODAN and the online campaign was also advertised through presentations at various events, when this was relevant in period Jan-April

2018. Although quite a lot of energy was spent on the Online Campaign, only one use case was received through this process.

3. Individual invitations: The online campaign was supplemented with individual invitations to relevant research networks and researchers. Here the requests could be better explained and introduced, and this lead to three more submissions of use cases.
4. Analysis: Subsequently all the use cases were analysed together. The results are provided through this document and were subsequently compiled to input to the e-ROSA Foresight Roadmap, combined with the outcomes of the second e-ROSA workshop held in Wageningen.

# 3 CASE STUDY DESIGN

## 3.1 INTRODUCTION

As explained in the previous section, to facilitate the reasoning from impacts of research to the role of research and Open Science, a use case approach was adopted. This approach facilitated thinking through storylines, that link to actual societal actors that are active in agri-food systems (e.g. government institutions, private sector, NGO's, farmers, international organizations) that are aiming to solve a particular societal challenge, and then to the role for research and associated data and ICT aspects. The use cases helped to clarify the role of research, as supportive and facilitating towards solving societal challenges, and providing a knowledge system delivering relevant insights to societal actors, but not being the ultimate problem solver.

## 3.2 FORMAT FOR CASE STUDIES 'REQUIREMENTS FOR DATA INFRASTRUCTURES FROM SOCIETAL CHALLENGES'

To structure the development of Open Science use cases that can direct the roadmap for an e-infrastructure for agriculture and food, a use case template was developed with following sections and questions:

**Impact pathways**

- Describe the link from main impacts to outcomes to outputs and activities using the impact chain approach
- Provide a short narrative of the impact path for this particular challenge

**End user groups**

- Which groups are benefiting from a solution developed by science/research based on the more intensive use of data, infrastructure and analytics?
- How do these beneficiaries benefit?

**Role of Research**

- What role does research play in contributing towards a positive impact?
- Describe the use of data and models where and if relevant
- What are scientific challenges that limit progress?

**Challenges in data, infrastructure, processing power and analytics**

- What are current limitations to the use of data, infrastructure, processing power and analytics?
- What needs to be solved to take advantage of current ICT capabilities?

**Solutions for more efficient Research**

- What needs to happen to bring research to the next level?
- What is the role of data, infrastructure, processing power and analytics?

**Monitoring the impact and implementation**

- If these solutions have been developed, how do we notice? How can we monitor the progress?

# 4 SYNTHESIS

## 4.1 IMPACT CHAINS

Research gets more and more aimed at achieving broader impact through focussing at societal challenges and associated societal indicators like the SDG's. This is also shown by the agri-food research cases that were collected by e-ROSA and it clearly revealed from the discussions held in the e-ROSA stakeholder workshops. However, looking at the foreseen (long term) impact of research in agriculture and food does not provide direct leads to the specific challenges and requirements that are related to the role of research, and of future Open Science and the use of e-infrastructures. Still, it is important to identify these leads in order to determine how research can contribute to achieving impact, how e-infrastructures can support this process, and thus, how the roadmap towards an infrastructure should be shaped.

Therefore, in e-ROSA, we have chosen to build our work around the analysis of the impact chains that underlie research initiatives in the agri-food domain. Such impact chains, as also explained in the previous section, explain how research initiatives are seeking to achieve impact by determining the logic of how their outputs lead to the foreseen high level impacts through a chain of short to longer term outcomes. Besides providing insight into the path to impact, the idea behind this approach was that the output and outcome level should be more clearly associated with the scientific process in general and particularly with the aspects of that process that are related to the use of e-infrastructures.

Taking a closer look at the research case studies that were collected through the network of stakeholders we can see how the outputs of initiatives are expected to evolve to shorter and longer term outcomes and finally lead to impact. It reveals that while impacts tend to be defined in terms of societal, economic or other high-level benefits, assets that somehow relate to Open Science and the use of e-infrastructures are clearly recognizable in the path to such impact.

*Long-term outcomes*: These are generally still quite broad and formulated on a high level. Nevertheless, it is obvious that data-intensive research and innovation are required as a component as they are clearly recognizable as results that require interdisciplinary knowledge linked to the data value chain. Some examples are "efficient, adaptive gene-based approaches to ensure food security", "better informed policy and decision making on food security" or "increased understanding of what is healthy for an individual".

*Short-term outcomes*: These outcomes still show a strong relationship with concepts and assets of e-infrastructures, although we can see that short-term outcomes are already more holistically formulated and closer to the process of improving aspects that are related to achieving impact. They often refer to the efficiency and transparency of research workflows and the overall research process, rather than to individual technical components or organisational topics. Examples mentioned in the collected use cases are "more efficient workflows from data acquisition to data publication", "better resourced data scientists" or "more efficient use of resources to control food safety".

*Outputs*: Envisaged outputs for research cases clearly have direct links to Open Science and assets that can be offered by future e-infrastructures for virtual research. They often refer directly to issues like data analytics and data science, methods and models, standards or semantics. Besides, there are also many references to the more process and governance based aspects of Open Science, like improved data curation and stewardship, capacity development on Open Science related topics or strengthening of networks for Open Science.

The following section summarizes the results of the analysis of research case studies and outcomes of the stakeholder workshop. It describes which directions and solutions were identified that should be part of the elaboration of the roadmap.

### 4.2 THE ROLE OF RESEARCH

There is an evident role for research in advancing towards many societal challenges. This was also clearly demonstrated through the cases examined by e-ROSA, which showed clear links from research outputs to impacts. Research has always been and will remain a main driver of innovation in the development and validation of methods and tools. As more and more data becomes available and the field of (big) data analytics is progressing, the research process will need to evolve, using advanced data science as one of its main instruments. Consequently, we are convinced that there will be a growing demand for research environments that are based on e-infrastructures.

From the analysed cases, we can conclude that at least the following roles for research are envisaged with respect to supporting the grand societal challenges:

- Acting as a main driver for knowledge and innovation in the fields of data acquisition, data analytics, modelling and decision support on the various working fields of the agriculture and food domain
- Discovering, developing and transparently publishing scientific resources (e.g. data, data processing, analytics, models, workflows) for reuse by scientists and other users (policy and decision makers, business, citizens)
- Developing the systemic, holistic approaches and required tools and interdisciplinary methods needed to tackle the grand challenges in the complex, multi-disciplinary domain of agriculture and food.
- Acting as a knowledge integrator between different fields in the agriculture and food domain by making data and tools interoperable and advancing interdisciplinary approaches.
- Acting as a process facilitator, bringing the different end user groups together, sharing and exploiting knowledge and jointly designing the interventions required to achieve the envisaged impacts.
- Acting as a sounding board, for example by bouncing ideas and assessing the impact of potential innovations as proposed by the actors (end user groups).

### 4.3 REQUIREMENTS AND SOLUTIONS FOR AN E-INFRASTRUCTURE FOR AGRICULTURE & FOOD

The stakeholder inventory of challenges and solutions for future e-infrastructure for agriculture and food through the previously described stakeholder workshop and the analysis of agri-food case studies resulted in a range of requirements and recommendations for future directions. These directions can be subdivided into requirements for “Sharing”, “Connecting” and “Collaborating”, and are incorporated in the Foresight Roadmap for e-Infrastructures in Open Science in Agri&Food research.

Besides these requirements, that are summarized in the following sections we consider the following general e-infrastructure requirements as being relevant:

- The e-infrastructure is based on and useful for a community
- The e-infrastructure needs governance and a business model for sustainability
- The e-infrastructure needs a technical backbone, that easily can host a variety of resources and services
- The e-infrastructure comprises a variety of resources and services, useful for the community, interoperable, but not necessarily connected or dependent from each other

### 4.3.1 Requirements for sharing

To make European research more efficient and effective, it is essential that scientific resources can be easily shared among research communities. Requirements for sharing imply several challenges, among which the one of (i) professionally managing research data, (ii) support for multi-scale, multi-dimensional data, (iii) sharing other research resources alongside with data, (iv) the socio-cultural challenges, (v) the ethics of sharing and (vi) capacity building.

*Professional research data management*: from the use cases examined by e-ROSA it becomes clear that many research initiatives and communities are struggling with data sharing. First, data is hard to discover and access, because of immature data sharing. It still takes huge efforts to get hold of valuable research data, just because it is not easily findable and accessible. To make the tremendous amounts of existing research data available for future data-intensive research, across use cases in agriculture and food, efforts are required in data management in order to make data available in a proper way. In that respect, it is obvious that data managers have an important role. They should move beyond making data available, and also include the broader aspects of FAIRification as well as data governance aspects. In that respect, there is a clear demand for easy to use tools and interoperability platforms that make data sharing practicable for the broader community of researchers in agriculture and food. While some descriptive metadata (e.g. lineage of datasets, documentation of algorithms, models) needs to be manually provided by the author, support for automatic creation and transfer of additional metadata would greatly help researchers to ensure the quality of scientific resources, starting from data collection. A data curator should be as much as possible unburdened from metadata generation tasks, and be better supported by tools embedding standard vocabularies and methodological tips. This would lead to more availability of metadata, of better quality (e.g., up-to-date, detailed or in consistent formats), and consequently an increased potential for the programmatic access of data.

*Support for multi-scale, multi-dimensional data*: various use cases in the agriculture and food domain report the need to work with multi-scale and multi-dimensional data. Typical for the domain, for example for agricultural monitoring, regional and global impact assessments, climate smart agriculture, is the need to have efficient tools and methods to work with spatio-temporal data, in order to be able to combine different geographical and temporal scales. Another example is the need to be able to work over different resolutions in the use cases of massive scale phenomics in order to better understand gene - environment interactions to be able to anticipate future perturbations. In that case, researchers need to be able to assess the full research domain, “from gene to landscape”.

*Share all scientific resources, not only data*: while there is a focus on e-infrastructures and their capacity to support data sharing, requirements for research infrastructures appear to be a lot broader than the concept of data sharing and the technical and organisational capacities of sharing data. While data is an important fuel for the generation of knowledge within the concept of open science, we need a broader scope with regard to sharing. From the perspective of research, besides data, sharing should also include the sharing of data analytics, models and, at the end, integrated workflows that support the scientific process. As with data, such resources are currently scattered and often not standardized, and sharing them should lead to better reuse of knowledge and methodologies and more harmonized research processes. In short, there is a demand for shared research services, rather than just data services. A good example is the domain of food safety and emerging risks, where currently a lot of models exist, residing in different networks. Similar to data, there is a tendency to use methods that are in scope of a researchers’ network, because it is often hard to access and use external ones. Good sharing practices for such resources would support different networks to perform more efficient research.

*Address the socio-cultural aspects of sharing*: a clear take away, mentioned by many stakeholders in different areas of agriculture and food, is that the social and cultural aspects of sharing knowledge need to be taken into account. A prerequisite for open science to become common practice, is a culture of sharing. This can only be achieved gradually through a combination of advocacy, capacity building and proper guidance and technical support. Moreover, it is clear that such efforts needs to be accompanied by mechanisms to reward sharing, which could provide additional incentives for researchers to share their knowledge and speed up the uptake of sharing practices. An example for this is the establishment of open data journals, like the open data journal for agricultural research (ODJAR), which promote publication of (FAIR) data, motivate researchers to document their data and allow data to be cited. Another important issue to deal with in that respect, is that a large part of the research community is still expecting e-infrastructure offerings to be free, or at least affordable. This indirectly links to the demand for viable business models, e.g. in the form of proven business cases that explain how data value chains can sustainably establish and maintain sharing initiatives. The issue of trust (or veracity) also has to be mentioned here. Many stakeholders indicate that it is often hard to judge the quality and usefulness of datasets that reside outside their own networks. To overcome this barrier for usage of shared data, it is necessary that datasets are well documented and that certification mechanisms for datasets and repositories are in place. Again, recognized, well-managed repositories like open data journals could play an important role in building trust and motivating sharing and reuse of scientific resources.

*Consider the ethics of sharing*: respecting the ethics of sharing, and specifically data sharing is considered essential for open science to succeed on the long term. While open science, including scientific work in the public and private sectors, obviously requires the sharing of data, privacy and also competitive stakes should be respected to ensure broad support and uptake. Moreover, many state that the issue of data ownership needs to be addressed in a proper way. There is a specific reference in that case to data ownership of individuals, like farmers or citizens, e.g. in the case of sharing and reusing farm management data or citizen science based data. Lastly, the issue of having also the less-favoured benefit from open science and e-infrastructures has been frequently mentioned as being desirable. This obviously transcends to a global, rather than a European territory perspective, where for example smallholders and consumers in developing countries should get the opportunity to benefit.

*Invest in capacity building*: in order to evolve from “classical science” to data science and open science, capacity building is crucial. To reach the vast majority of the research community in agriculture and food, and not only the technical front-runners, and support them to make this transition towards sharing and reuse of shared resources, the sharing of knowledge on open science is required. Practical guidance and support tools must be developed to train researchers to become “open scientists”. Examples of topics where capacity building is needed are the FAIR principles, making data FAIR, and the use of tools and methods that allow the reuse of FAIR data. Moreover, there is a need for support on issues such as data stewardship, data curation, data standards, licensing and IPR. As data managers and data stewards will have a central role, training material to educate personnel in these roles is a crucial resource, which needs to be shared across research communities.

### 4.3.2 Requirements for connecting

Better sharing of high quality research data and other resources, like algorithms, models and methods, is an important step towards the envisaged Vision 2030 roadmap for agriculture and food. However, such shared resources will only be effectively re-used if they can be easily harmonized and connected, not only by technicians, but also by the vast majority of European researchers. Only if it becomes possible to easily connect data, methods, models, tools to perform research workflows, e-infrastructures will reveal their full value and only then scientists will be able to conduct open science in a transparent way, maintaining their current research experiences and competences and adding to that the advantages of open science.

Connectivity through standards is required for data intensive research on different levels. It is important to realize that this requires a view on the full data value chain, which is not limited to research, but in many cases includes business and governmental stakeholders. Thus, development of standards and guidance on the use of standards will often need to include the perspective of compliance with legislation and regulation (e.g. INSPIRE) and business standards (e.g. standards and vocabularies in the domains of trade, farm management systems).
There is demand for (1) mapping the existing landscape of standards, be it data formats, metadata, or other resources standards; (2) providing guidance and recommendations on interoperability and (3) finding out and bridging the gaps, developing for instance agreed controlled vocabularies.

*Mapping the existing landscape of standards*: there are several recent initiatives to map the landscape of (meta)data standards in the agriculture and food domain including, but not limited to, the RDA Wheat Data Interoperability and Agrisemantics working groups, and the GODAN Action map of agri-food standards. While the data sharing perspective is important, researchers are in the end aiming at performing full scientific workflows, which also incorporate for instance data analytics, models and visualisation components. So again, if it concerns required connectivity, there's a clear call to not only consider the standardisation and semantic connectivity of (meta)data resources, but to also include the standardisation of analytical models and tools. This is particularly relevant since a lot of the current research is performed using models and tools that are in many cases not fully interoperable with current standards (let alone possible future standards). In some cases, they might even be classified as legacy, and originating from a "pre-standardisation era". One of many examples is the use of numerical models in agricultural monitoring and yield forecasting. These models have often been developed over decades, represent a tremendous amount of knowledge, and have a broad user base. It should be realized that such tools are indispensable in many research areas, and cannot easily be replaced by standard-compliant alternatives. Thus, any open science strategy should incorporate means to adapt (or adapt to) such tools.

*Provide guidance and recommendations*: Already now, there are many developments in the area of agriculture and food that concern the improvement of interoperability, with the aim to improve connectivity within and over research domains. Many standardisation efforts are ongoing, specifically in the area of semantics and the improvement and harmonisation of shared vocabularies and ontologies. Examples are the development of the Global Agricultural Concept Scheme (GACS) and the Crop Ontology.  However, many use cases in different areas of the domain report a proliferation of standards, which makes is difficult to choose which standards to use. On top of that, in cases, standards are still developed aiming at specific subdomains of agriculture and food, which implies the risk that currently existing research silos are maintained instead of being connected. Consequently, there is a strong demand for guidance and for high-level agreement on advised standards to use for different types of research applications. The exemplary approach of the RDA Wheat Data Interoperability working group in breaking the boundaries among the wheat research community could be generalized and expanded in order to clarify the complex landscape of existing standards and deter communities from duplicating efforts.

*Find and bridge the gap*: it is generally agreed by stakeholders in the agri-food domain that the identification and rigorous analysis of some strong interdisciplinary use cases would help in further clarifying the research questions with regard to open science. This could also help to identify in more detail which currently available standards fit such "cross-cutting" research scenarios, which gaps still exist, and how solutions can be developed that have a broad applicability. On the other hand, to ensure short term uptake of open science while advancing on cross-domain interoperability, there's also an immediate demand for guidance and capacity building to support scientists in understanding and effectively using e-infrastructures for research. This might partly be achieved by providing "no-regret scenarios" for open science, that have already proved their value in practice and that are expected to fit into future evolutions of open science. This can be offered through success stories and best practices that are already in place, or by offering learning resources that allow researchers to experiment and work with already available open science services.

### 4.3.3 Requirements for collaborating

Obviously, performing interdisciplinary research in the era of open science, supported by e-infrastructures also implies a different way of collaboration within and between research communities. E-infrastructures can play an important role in facilitating the collaboration and co-development of (international) research groups. To play such a role, the foreseen e-infrastructure for agriculture and food should (i) allow effective collaboration and co-development, (ii) be user-centred, (iii) build on a sustainable business model.

*Allow effective collaboration and co-development*: taking aboard some of the previously stated requirements on sharing and connecting, e-infrastructures should allow research groups to establish shared virtual environments, where they can access and work with trusted, high quality datasets, methods and tools. Moreover, they would then implicitly benefit from the high performance cloud computing and storage facilities that are offered by the underlying e-infrastructure. From a researchers' perspective, they would not be using e-services or e-infrastructures, but rather a virtual research environment that offers transparent access to and seamless use of scientific resources for joint research.

*User and research community centred e-infrastructures*: a user centred approach build around some strong use cases could help in steering further directions for the technical development of facilities, and for monitoring progress and steering towards effective e-infrastructure support for European open science. It is considered important that environments for open science, while allowing for effective collaboration and co-development, also respect the individual identity and specifics of different scientific domains and research groups. In other words, virtual research environments should not restrict but rather broaden the opportunities of researchers to perform their work. A "system-of-systems" approach, allowing the establishment of regional and domain specific nodes on top of a shared organizational and technical infrastructure for open science is suggested as a potential model to make this possible. At the same time, this should not impede, but facilitate the cross-nodal connections that are required for interdisciplinary research. Moreover, it should also allow to take aboard individual accomplishments of researchers, and to establish large and small projects to connect to developments in open science. Besides a well thought over, scalable and multi-layered technical architecture, this will require that supporting mechanisms are in place to for example certify open science nodes, and to share resources while respecting IPR, licensing and privacy.

*Build on a sustainable business model*: it is important to always keep the long-term perspective of open science in mind. Stakeholders feel that a "single project funding perspective" will not be adequate to achieve the longer term objectives, and a European Research Infrastructure Consortium was mentioned as a possible mechanism. Moreover, there is a call for strategic initiatives that support the roadmap to open science and underpin the importance of open science to advance the agriculture and food domain. One type of such strategic activities is the development of sustainable business models for open science that cover the whole data value chain, take into account viable assets and cover both the public and the private sector. Thus, open science scenarios should not be limited to public sector research, but should also include the perspective of private sector research and of funders and funding mechanisms. This also implies that there is an important role for public private partnerships in the co-development and sustainable exploitation of infrastructures for open science. The use cases that were adopted and analysed in the e-ROSA project are a good basis in that respect, as besides broadly covering the agriculture and food domain and a multitude of aspects related to data-intensive science, they are also well connected to stakeholders outside the scientific domain. A second strand of strategic activities concern large-scale impact studies, which can underpin these business models as well as the broader societal value of open science. Such studies should clearly outline the Key Performance Indicators (KPI's) that can be achieved through the adoption of open science. They should also, through the monitoring of such indicators, evaluate open science initiatives and seek to advance and show proof of the enabling of innovation through increased sharing in the public, research and private sectors.

## 4.4 MONITORING IMPACT AND IMPLEMENTATION

While some stakeholders have mentioned ways to monitor impact and implementation, we can conclude that e-ROSA activities have not led to a clear vision on this subject. Nevertheless we think that here also the impact chain approach will prove to be valuable. This allows deconstructing high-level societal challenges to tangible outcomes that are more easily linked to the outputs of initiatives and networks. Consequently, it will be easier to define indicators that are easier to influence and measure on the level of individual or groups of projects / initiatives, but are still well connected to impact.

For monitoring impact and implementation, the following aspects are relevant:

- While impact assessment is usually considered as a post-project activity, it is relevant that ex-ante assessments (evaluating and improving design of an initiative, related to the conditions of the environment with regard to adoption, innovation etc.) and intermediate assessment (to monitor and effectively steer initiatives, process evaluation) are also "part of the toolbox"
- Qualitative and quantitative indicators should be available to monitor a broad range of aspects like technology development, adoption/uptake of open science, resource sharing, collaboration and co-development, linkage with private sector science and non-research processes and innovation capacity.
- With/without analysis could be made to evaluate the specific importance of having scientific methods based on data and ICT tools, connected to e-infrastructures across applications with multi-actor processes

# 5 RESEARCH SHOWCASES

## 5.1 EXAMPLE SHOWCASE 1: CROP YIELD FORECASTING

### 5.1.1 Introduction

Crop yield forecasting using combinations of crop growth modelling, remote sensing analytics and yield statistics, can be an important instrument for seasonal predictions of regional and local crop yields. Outputs can be used in multiple ways, among others to forecast regional and global commodity market development, estimate production surplus and shortages and project derived effects on food security like food supply and demand patterns, demand for food aid or financial aid. Moreover, analytics over historical statistics can provide useful insights in relevant patterns and dependencies, which can be exploited by micro-finance and micro-insurance institutions to optimize their schemes.

### 5.1.2 Impact pathways

*Impacts*: Improved food-security; disaster risk reduction; market support
*Long term outcomes*: better informed policy and decision making on food security issues
*Short term outcomes*: more efficient and transparent workflows from data acquisition to data publication; better resourced data scientists in the domain of food security; better semantic linkage between the agronomic and related domains.
*Outputs*: improved methods and models; modelling workflows; data curation and stewardship; targeted semantics.

*Short narrative of the impact path:*
Setup of an e-infrastructure for agriculture and food, including a data-infrastructure, data processing and analytics and modelling facilities and data curation and publication will allow researchers to perform more efficient and effective crop yield forecasting. This will be achieved by the availability of tools and capacity to set up and share modelling workflows and to curate resources. Improved semantic interoperability over the involved domains, combined with more open and better documented resources will allow more efficient data integration. Co-development by researchers, data scientists and end users will result in knowledge that is generated faster, is more transparent and fit-for-use. This will result in better accuracy, more trust and eventually better informed policy and decision makers. This will inform decisions and improve the situation on food related issues like disaster risk reduction and market support and climate smart agriculture.

### 5.1.3 End user groups

*Researchers*: better access to resources (data, analytics, publication of data); support for data curation and data stewardship;
*Policy and decision makers*: timely and more accurate, fit-for-use knowledge to pro-actively act on short-term and long-term food security issues
*Business*: better insight in market development and market potential

### 5.1.4 Role of Research

Research plays an important role in the evolution of crop yield forecasting. Research has always been and will remain a main driver of innovation in the development and validation of methods and tools required in the fields of data acquisition, data analytics, modelling and decision support. While in the near future, operational services in this area might be provided by commercial parties, agronomic science (both public and privately funded) will remain a main source for innovation and broad agronomic expertise will be an important asset to setup viable services. Nevertheless, the domain will need to adapt to new developments in among others fields like "research in the cloud" and big data analytics and will need to be able to liaise with experts in these fields to keep this role.

Some typical research activities that would require support from e-infrastructures are:

- Estimation of length of the growing season and other crop characteristics, getting the most out of available data by (real-time) acquisition and fusion of experimental, crowdsourced and remotely sensed data
- Deriving crop growth sigmoids, based on remote sensing data analytics and statistical interpolation procedures, supported by historical archives.
- Setting up an e-infrastructure supporting automated workflow for crop yield forecasting, using a variety of data acquisition, data analytics, modelling and visualisation modules
- Parallelization of crop growth models (e.g. the WOFOST/WISS model), data processing and analytics to be able to produce faster, higher resolution results

The role of research would be:

- To bring forward scientific knowledge on crop modelling and crop yield forecasting and to improve methods and tools accordingly
- To discover and develop, and transparently publish and re-use resources (data, data processing, analytics, models, workflows)

### 5.1.5 Challenges in data, infrastructure, processing power and analytics

*Improving the availability of research infrastructures*
Infrastructure for high performance computing, analytics etc. is generally not available, not accessible or researchers are not knowledgeable to use it. In many cases, infrastructure that is available is set up on a non-sustainable, local (e.g. within research institute) or temporal (e.g. within projects) basis.

*Improving the availability and access to data and the capacity to work with alternative data*
Access to a lot of data required for crop yield forecasting is limited, either because it is being protected for strategic and commercial reasons (e.g. weather data), because it's scarce (crop calendars), or because it is partly hidden in textual documents or sitting on researchers' laptops. The capacity to overcome this and to work with and combine data sources, including alternative data like crowd sourced data, data from text mining etc. is still lacking.

*Development and testing of big data analytics solutions for geospatial data*
While there are enormous developments in the area of big data storage, processing and analytics, the capacities to work with big spatiotemporal data are still lagging behind. New concepts generally do not seem to scale to the spatiotemporal domain (e.g. lack of storage strategies and performance, lack of spatiotemporal analytics). Currently the domain still seems to rely on classical GIS or at the best hybrid concepts with their specific disadvantages regarding aspects like volume and velocity.

*Development of dedicated semantics*
The domain of crop yield forecasting is cross-cutting domains (agronomy, meteorology, soil science etc.). Yet it does not have developed semantics that can support the required knowledge integration.

### 5.1.6 Other solutions for more efficient Research

*Improving the capacity to work with heterogeneous data*

Development of the capacity to overcome current data gaps and to work with and combine data sources, including alternative data like crowd sourced data, data from text mining etc.

*Improving the capacity to use data science and develop solutions*
To promote and realize co-development of analytics and visualisation for policy and decision makers by agronomists, data scientists and public and private end users.

*Improving the attitude of the scientific community towards sharing of resources*
To realize a culture shift towards open sharing of data, knowledge and tools and the role of data management, curation and stewardship.

### 5.1.7 Monitoring the impact and implementation

Several indicators that can be developed to monitor implementation and impact

Implementation / output:

- The amount of processing, data analytics and modelling components that are available through acknowledged e-infrastructures
- The number of data sources available through acknowledged e-infrastructures that are either potential input for or are generated by crop yield forecasting initiatives
- The amount of crop yield forecasting initiatives that have implemented their knowledge development or operational processes into e-infrastructure supported workflows
- The amount of co-development initiatives between the "classical" research community of agronomists and ICT and the data science community
- The amount of services in the area of food security that are tapping from data, information and knowledge generated by crop yield forecasting initiatives

## 5.2 EXAMPLE SHOW CASE 2: RESEARCH SUPPORTING AGRICULTURAL TRANSFORMATION TOWARDS SDG2 ENDING HUNGER

### 5.2.1 Introduction

As a follow up to the political agreement reached with the Sustainable Development Goals, there is now a commitment of countries/nation states to reach the goals as described in the Sustainable Development Goals, and monitor their progress towards these goals. Looking specifically at SDG goal 2 of Ending Hunger, it is formulated as 'End hunger, achieve food security and improved nutrition and promote sustainable agriculture' which describes the vision of the desired impact by 2030.

### 5.2.2 Impact pathways

This visionary impact translates to different outcomes in different nations, for example, in the Western world, the nutrition aspect is much more important in fighting obesitas and fighting unhealthy eating habits that lead to diseases, while in the developing world, there is much more focus on producing enough food with a good quality and enough variation, and reducing risks in the agricultural production chain. Each nation state needs to define its own actions for reaching the SDG targets, leading to the creation of national networks, called SDG Academies. Statistical agencies have a role to play in measuring the achievement of the goals, via relevant indicators (168 of these indicators have been defined). There are also thematic networks to facilitate this process, for example, the Sustainable Development Solutions Network created by Jeffrey Sachs, and where Wageningen UR is involved in shaping the research efforts towards SDG 2. Thus as outputs and activities, relevant communities need to be formed at the national level, statistical agencies need to establish monitoring networks, and the whole idea of designing packages for agricultural transformation is gaining momentum. The thinking around agricultural transformation (https://www.idrc.ca/en/article/agricultural-transformations) is that a new organisation and way of doing things is needed in the agricultural sector to maintain or increase production, while at the same time providing less burden to the environment and improving the nutritional status of foods (see here for an example of such an approach for Uruguay).

### 5.2.3 End user groups

- National ministries: they need a better understanding from the data on what the specific problems are they need to tackle in their geographies
- National SDG academies: they need data and information to help their national members to agree and design potential interventions around agricultural transformation pathways
- National statistical agencies: they need to monitor the progress towards SDG goals, and many indicators are new, so they need new data sources to adequately quantify these indicators. Also processing capabilities are required.
- Agricultural supply chains: many players (farmer cooperatives, processors, consumer organizations) in the supply chain could be actively involved in designing interventions relevant for their set up and company goals. They need an easy set up to access the results of research and to enable research to access the data sets in their supply chains.
- NGO and citizen organisations: these organizations use the political momentum created by the SDG's to stimulate further concerted action by a range of actors. They could use access to data and research in their arguments to stimulate the right interventions.

### 5.2.4 Role of Research

Research has a role to play in the complex environment around the SDG's as it can act as a facilitator, as sounding board, and as developing new methodologies. With respect to facilitation, as research organizations and individuals often do not have a specific view or opinion how to reach sustainable development, they can organize the fora to bring the different end user groups together and design jointly the interventions. This often starts with first developing a joint understanding of the problems.
With respect to sounding board, here research can help to bounce ideas and check the impact of potential innovations as proposed by the actors (end user groups). Through methods of impact and integrated assessment, different options can be weighed on their potential impacts and their unintended consequences, facilitating the decision process by communities.
Finally, for new methodologies, it is recognised that the traditional disciplinary approaches used by research do not work very well for the complex challenges as put forward by the SDGs and to realise visions. Therefore, new methodologies are needed that incorporate the inclusive, transformative and visionary aspects, and as a first step the Agricultural Pathways have been proposed by the SDSN, which is a method to design back-casting scenarios by visualizing what one wants to achieve, and then reasoning backwards to all the actions needed now to achieve this, while including all relevant actors in the exercise. Given these very recent developments, the role of data and models is not yet clear. However, it is a given that methods need to be as much as possible data and evidence based, that transparency helps in multi-stakeholder set up, and that data sharing and joint analytics are requirements.

### 5.2.5 Challenges in data, infrastructure, processing power and analytics

Given that the implementation of the actions towards SDG 2 is a relatively new process, and the research approaches are also in development, the challenges in data, infrastructure, processing power and analytics are not fully mapped out. The comprehensive nature of the analyses required reinforces potentially all challenges found in the more specific showcases (for example, yield forecasting). Thinking along the DIKW pyramid (figure 3), the challenges for research are sitting mostly at the knowledge level, which is close to the decision making level, and which involves trade-offs on conflicting goals, decision support, participatory models, and building on many specific information sources. Three levels of innovation are required in data, infrastructures, processing power and analytics:

1. Interoperability & data integration to answer comprehensive content questions, linking production to nutrition and nutrition to health, in terms of spatial interactions
2. We need a federation mechanisms over the infrastructures supporting the specific niches/scientific disciplines.
3. Working across scientific disciplines at the knowledge/wisdom level requires different tools and resources for researchers: more focus on decision support, knowledge rules, trade-off analysis tools, participatory models.
4. Have easy-to-use data analytical and presentation tools, which can be run in participatory settings to enable facilitation by scientists.

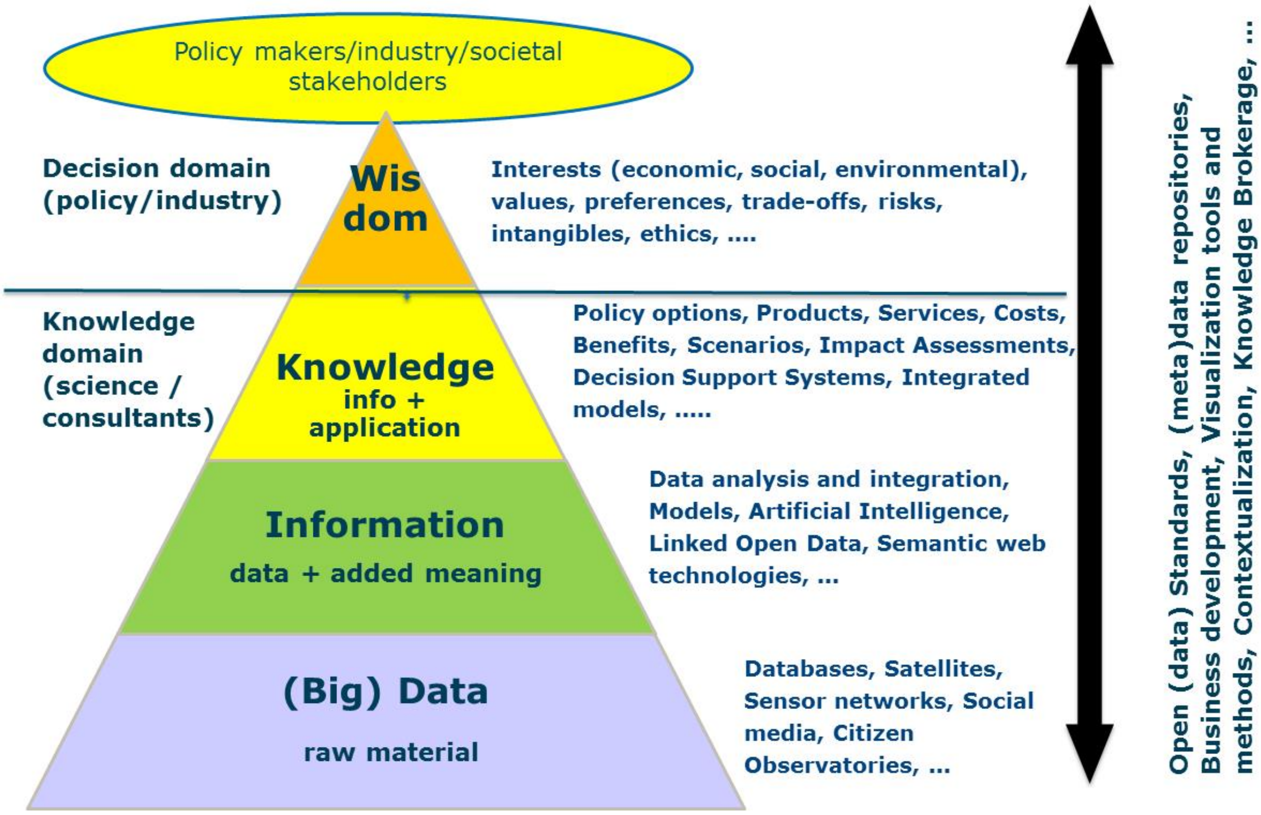



Figure 2 DIKW pyramid of data to information, to knowledge and wisdom

### 5.2.6 Solutions for more efficient Research

For research to become more relevant in the context of SDG 2 End Hunger, it needs to develop new scientific methodologies, while at the same time applying them with lots of different data sources and information points. This combination of development of methodology and the resources to support those methodologies is extremely challenging. This will not work in all applications, and all cases, so accepting that there is room for experimentation required, and that these experiments might not all succeed is crucial in first stages of development. The scientific methodologies need to be tested in many more locations; now some tests have been made and some first examples are available, but many more are required. At the same time, new solutions have to be thought of for reaching impact with the help of research. While a useful contribution can be made from research, the research has to be connected to societal dialogues and multi-actor approaches. Useful solutions required are:

- The development of vocabularies of useful terms, coupled to generally available data sources, to enable first analysis in an efficient way.
- Enabling access to quality controlled data sources processed in a consistent, coherent and quality assured way, to allow maximum use of data available.
- Tools for visualization geared towards participatory discussion to ensure inclusion & joint learning of all stakeholders participating sessions

### 5.2.7 Monitoring the impact and implementation

It is important to monitor the links between the different elements, i.e. the connection between the new scientific methodologies developed, their use and/or dependence on data and visualization tools and the updated applications with real stakeholder involvement. With/without analysis could be made to evaluate the specific importance of having scientific methods based on data and ICT tools, connected to e-infrastructures across applications with multi-actor processes, in other words, comparing an application of the scientific method with the use of common data and ICT infrastructures, to an application without, in the context of agricultural transformation and SDG 2 End Hunger Implementation.

## 5.3 EXAMPLE SHOW CASE 3: RESEARCH SUPPORTING USE OF REMOTE SENSING DATA IN A DEVELOPMENT CONTEXT

### 5.3.1 Introduction

The ongoing project CommonSense is subject of the show case. It focusses on Ethiopia as an example in Africa where e-Infrastructure is still in its infancy. Research to transfer knowledge and experiences form Europe in a developing infrastructure targeting open science in agriculture is challenging. In this show case an outline is given and barriers are identified to indicate its specific nature within this context. It shows research is not only facing 'research' challenges, but also societal, cultural and political aspects. The project is still in progress.

CommonSense is a project carried out in the Dutch G4AW Program targeted at the use of Geodata for Agriculture and Water (G4AW) to improve food security in developing countries by using satellite data. CommonSense is carried out in Ethiopia and uses amongst others data from Remote Sensing to offer services to monitor crop growth, produce local weather information nationwide, localized weather forecasts, monitor yield and market prices. It is supporting the smallholder farmers mainly through Micro Finance Institutes (risk assessment), Coops and Unions, agricultural production networks (sesame business network), Development Agents (Min of Agriculture).

The basic notion is to provide the smallholders adequate farm advice to better secure their yields and improve their livelihood and meanwhile supporting food security in Ethiopia.

### 5.3.2 Impact pathways

Major impacts of informing the smallholders, mainly through farmers associations, is to be expected from strengthening the full chain. We may distinct various actors on the value chain for any crop from field to consumer market. For example for sesame, which is a cash crop in Ethiopia and generating foreign currency from export for Ethiopia, the value chain can be seen as follows is as follows:

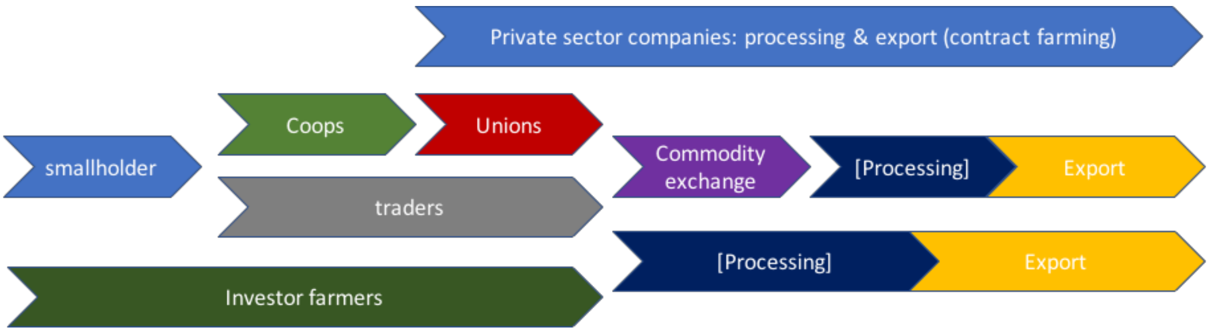



**Figure 3 a graphic representation of value chains in Ethiopia as part of the CommonSense Project**

The services provided by the project CommonSense will support the first part of the chain, by use of intermediate actors like for instance the unions. They provide information for farm management, supply the inputs, store and/or trade the harvest and more.
The challenge is to reach the intended users being the small holders. In Ethiopia there are barriers related to education, language, access to technical infrastructures and the good quality data to provide meaningful information. To overcome these barriers is using intermediaries like Coops and Unions, Extension services offered by the government to directly target at the user group and a second approach is the use the micro finance institutes with adequate information to grant loans with a high pay back rate. Major impact will only appear if the advices produced (farm advice, weather forecasts and support for loan assessments) are sufficient (just the right quality) and well communicated.

### 5.3.3 User groups

The major target group here are the smallholders, but basically any actor in the whole chain can use and benefit from the services delivered. In summary we can distinguish the following users:

- Smallholder farmer
- Coop and union employees
- Loan officers of Micro Finance Institutes
- Development agents (DA) as the extension officers of the ministry of agriculture
- Traders
- Agro industry
- Research and knowledge institutions

### 5.3.4 Role of Research

By its nature the project is a research project. It partly explores ways to reject or to go forward on existing research on crop growth and monitoring, and partly its goal is to implement (intermediate) results as 'viable' products. The services offered are based on existing research, but parametrized, extended and localized to be valid in the context of the selected region/country. It is built on knowledge and experiences showing good results elsewhere in the world, but has to be modified and adapted to fit in the Ethiopian context. The resulting product must be fit for use in Ethiopia and the approach applied is 'user centred', starting to identify the user needs. It follows on headlines the following process.

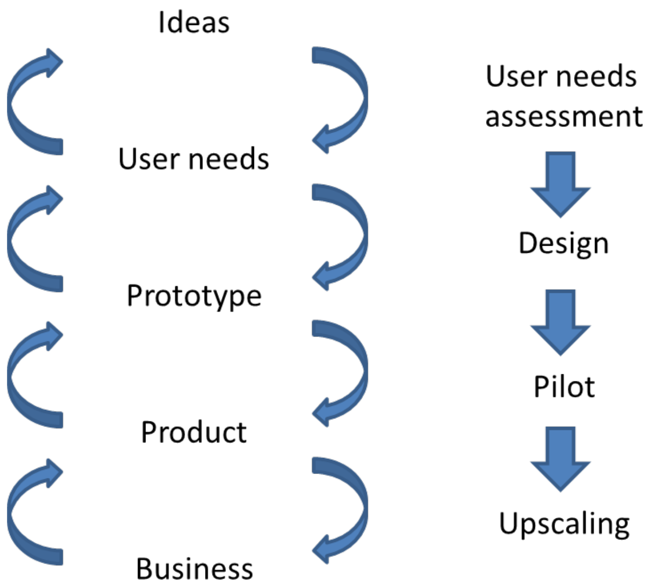



Figure 4 the research-to-product approach tested as part of CommonSense Project

Monitoring and evaluation during the project (and possibly thereafter) will trigger iterations in the process to improve results.

In addition one of the criteria of the donor is that the project has an outreach of more than 200.000 smallholders (direct & indirect) and services and products should be sustained after the project is ended, based on a realistic business proposal. Even that phase is very much supported by the research in the sense that the scientific result should be translated into 'easily' consumable and 'to the point' information to the targeted user groups.

### 5.3.5 Challenges in data, infrastructure, processing power and analytics

The data infrastructure in Ethiopia is weak. Internet is available with reasonable band width in Addis Ababa, the capital of Ethiopia, but more remote areas lack this completely. Instead of data transmission, information exchange is mainly text based and dependent on SMS (short message service) and IVR (interactive voice response) technology.

To be flexible and ready for the future we proposed an SDI set up (Spatial Data Infrastructure). A pre-condition for this is off course that an existing infrastructure is around. Since this is not the case the setup of the architecture is based on principles of a SDI (creating nodes with web based technology), but no full benefit of such an infrastructure can be made yet.

A big challenge was get the support of Ethiopian governmental bodies to what the project would deliver. For example the National Meteorological Agency (NMA) has by law the mandate to be the only one allowed to publish weather related product (forecast, etc.). Since we wanted to provide local weather forecast for farm management advices we needed the approval of the NMA and their support to distribute local weather forecasts.

The Information Network Security Agency (INSA) controls all data related affairs in Ethiopia and also related to the Ethiopian policy all data should be as much as possible provided by Ethiopian organisations. It meant we had to inform this agency to show no threats are generated from the data and information services the project wanted to provide.

Within Ethiopian the ICT capabilities vary a lot. Partly (mainly the business) has very good capabilities, Sill a lot of capacity building is needed especially for governmental organisations to take full advantage of current ICT capabilities.

### 5.3.6 Solutions for more efficient Research

In Ethiopia a big number of donor based research projects are going on. Cooperating more between projects focusing on the same domains could bring a lot of efficiency. Even in the short period of time this project evolved we noticed that much impact and synergy can be reach by cooperating with local networks and to approach research in the country, For Common sense from the start we were working together with the Sesame Business network (SBN) and due course we became more and more connected to the Agriculture Transformation Agency (ATA), the research institute for the ministry of agriculture.

Next finance is especially for poor countries as there are many in Africa is a huge problem. Since research is very donor dependent to sustain good results and/or to follow up on promising research project is crucial to be more efficient and to reach the intended impact.

The role of a good data infrastructure is evident, it will develop well since technology is becoming available more easy and affordable, but there is still a long way to go.

### 5.3.7 Monitoring the impact and implementation

One of the condition for the G4AW program was to monitor and evaluate progress made in the project. This is related very much to the outreach to farmers and creating a sustainable business case. To come up with simple numbers, like the number of small holders reached is the easy part. To assess the impact of our results is much more complicated, especially in this case since the past two years were years of severe drought, which means that assessing the increase in yield (income) is virtually impossible for our project.

Nevertheless indicators to monitor progress are developed so we should be able to draw conclusions once the project will be upscaled.

## 5.4 EXAMPLE SHOW CASE 4: PHENOTYPING

### 5.4.1 Introduction

Phenotyping using combinations of data from omics and phenotyping platforms, but also from field trials, as well as modelling is a key to better understand the gene x environment interactions, crucial to face the overall societal challenge, that is to achieve a sustainable use of the resources by the food system, by a better use of the genetic diversity and new breeding strategies. This means developing efficient breeding to provide genetic solutions to the disruptive changes in food production. Such breeding needs to produce robust and resilient breeds and varieties to outside changes and could thus have to focus on different aspects than the traditional productivity perspective. Breeding thus needs to focus on supporting non-intensive (e.g. low input, organic, small scale) types of farming, targeting efficient resource use and stability of production with a high disease and pest resistance. This means also developing new races or varieties, but also innovative control mechanisms less dependent on chemical solutions and management interventions that increases the natural control of agro-ecosystems and lower the chance of emergence of pest and diseases in the first place or the spread of invasive species. Last, resource use efficiency and circularity in the food system need still to be improved, to ensure an adequate supply of resources to others sectors, meaning the food system has to produce more with less. This also entails the safe-guarding of natural resources, i.e. biodiversity, at the landscape level, which links to a vibrant countryside and urban food landscapes.

To achieve these societal impacts, science needs to better target the needs of stakeholders and help to develop participatory techniques for breeding, trade-off analysis, enabling discussion and tailoring. An unresolved scientific challenge is crossing the scales from genes & cells through organisms to population (field, herd) to landscape. To achieve that, using gene-based approaches is to speed-up the extrapolating innovations from lab to field situations and large scale phenotyping and characterisation of the environmental components, or GxExM interactions (Gene-Environment-Management interactions).

To achieve such goals, multi-scale phenotyping capacities has to be developed in various locations and networked. The ESFRI projects have contributed to this over different targets, e.g. Emphasis for plant species, Embric e.g. for fish breeding or SmartCow for cattle. At that time, plant phenotyping opens the way of such approach, partly due to the range of possibilities that offers plants for breeding compared to animals.

### 5.4.2 Impact pathways

*Impacts:* Genes are preserved in their diversity and manage from individual to landscape to develop a multifunctional agriculture more resilient in the face of perturbations
*Long term outcomes*: Efficient, adaptive gene-based approaches to ensure food security, farmer quality of life, and the sustainability of our ecosystems
*Short term outcomes*: better understanding of multiscale GxE interactions through the development of more efficient and transparent workflows from data acquisition to data publication and better resourced data scientists in the domain of breeding sensu lato
*Outputs*: connecting data from various origins, academics breeders and from various scales, omics to phenes to farms; improved data mining and models; data curation and stewardship; targeted semantics

***Short narrative of the impact path:***
Setting up e-infrastructures for phenomics, including a data-infrastructure, data processing and analytics and modelling facilities as well as virtual network and training to connect the various phenomics infrastructures will allow researchers to better understand the gene x environment interactions and consequently help the breeders to contribute to develop a multifunctional agriculture more resilient in

the face of perturbations. This will be achieved by first interconnecting the information systems of existing phenomics facilities, that are already working together through various networks such as EFRI ones (*e.g.* Emphasis for plants), with a focus on the multiscale integration. This required not only IT aspects but also changes of practices for data providing, curation, integration, and interpretation. Thus, the networks will be an important actor to make available tools and capacity, and develop shared resources, at various level including semantic interoperability. This has also been done through co-development by researchers, data scientists and breeders. This will result in better accuracy, more trust and eventually better informed policy and decision makers.

**End user groups**

Different groups of end users can be distinguished for the case of phenotyping: researchers, the breeding business and policy and decision makers, all having different ways of benefitting from e-infrastructure use.
*Researchers*: better access to resources ((meta)data, semantics, analytics, models, workflows, publication of data); support for data curation and data stewardship; better support for multi-scale and multi-site integration
*Breeding business*: better knowledge of the G x E integration, better knowledge to extrapolate data from omics to farms. Available data to design new genotypes well adapted for the futures of agriculture.
*Policy and decision makers*: timely and more accurate, consolidated vision to design a political and ethical framework for biotech in breeding *e.g.* GMO. Information for an improved management of biodiversity for landscape managers.

### 5.4.3 Role of Research

Research plays an important role in the evolution of plant and animal breeding as well as at a higher scale the management of biodiversity. The development of phenomics is becoming a main driver of innovation in the development and validation of methods and tools required in the fields of data acquisition, data analytics, modelling for breeding through a better taking in to account for the G x E interaction. To achieve that, the domain will need to adapt to new developments in the domain of cloud computing, big data analytics and HPC simulations and will need to be able collaborate with experts in these fields to keep this role.

E-infrastructures will help scientists in:

- Designing phenomics experiments relying on a better knowledge of infrastructures (lab facilities, phenotyping platforms, farm networks)
- Building models based on shared phenotyping data to develop model-assisted breeding
- Setting up e-infrastructure supported automated workflows for analysing G x E interactions, using a variety of data acquisition, data analytics, modelling and visualisation modules
- Developing HPC access for efficient phenomics model simulations as well as big data processing and analytics to be able to produce faster, higher resolution results to address the complexity of the GxE interactions.

The role of research will be:

- To bring forward scientific knowledge on phenomics and to improve methods and tools accordingly
- To discover and develop, and transparently publish and re-use resources (data, data processing, analytics, models, workflows)

### 5.4.4 Challenges in data, infrastructure, processing power and analytics

*Improving the availability of research infrastructures*
In the case of phenotyping, e-infrastructures offering high performance storage and computing, analytics etc. are generally not available, not accessible or researchers are not knowledgeable to use it. In many cases, if such infrastructures are available, they are set up on a non-sustainable, local (*e.g.* belonging to a research institute) or temporal (*e.g.* part of projects) basis.

*Improving the availability and access to data*
Data for G x E estimation may be difficult to gather due to protection for strategic and commercial reasons, due to their cross-scale and multidisciplinary nature and consequent possible non interoperability, and due to the still common practices of research to keep data within their networks of peers or even to store data on their laptops. The capacity to overcome this and to work with and combine data sources just begins to be developed for some plant species, but are lacking for others.

*Development and testing of big data analytics solutions for phenomics data*
While there are enormous developments in the area of big data storage, processing and analytics, the capacities to work with big multiscale GxE data are still lagging behind. New concepts generally do not seem to scale to the GxE domain (e.g. lack of storage strategies and performance, lack of spatio-temporal analytics). Moreover, gathering big data sets on phenomics may enable the development of machine learning approaches to decipher the complexity of GxE, which produces phenotypes and confers plasticity to individuals to face environmental changes.

*Development of dedicated semantics*
The domain of phenotype forecasting is cross-cutting different domains (genetics, physiology, bioclimatology, etc.). Yet the development of semantics that can support the required knowledge integration is still at its beginning.

Some of these challenges have been identified as common with other use cases, which may mutualise the efforts of development and training of generic and transverse services.

### 5.4.5 Other solutions for more efficient Research

*Improving the attitude of the scientific community towards sharing of resources*
To realize a culture shift towards open sharing of data, knowledge and tools and the role of data management, curation and stewardship.

*Improving the capacity to use data science and develop solutions*
To promote and realize co-development of analytics and visualisation for policy and decision makers by geneticists and ecophysiologists, data scientists and public and private end users.

*Improving the capacity to work with heterogeneous data*
Development of the capacity to overcome current data gaps and to work with and combine data sources, including alternative data like crowd sourced data, data from text mining etc.

Again, these solutions have been identified as common with other use cases, which may mutualise the efforts of development and training of generic and transverse services.

### 5.4.6 Monitoring the impact and implementation

Several indicators have to be developed to monitor implementation and impact. Some of them will rely on technological developments that have been identified as common with other use cases, which may mutualise the efforts of development and capacity building.

Implementation / output:

- The amount of processing, data analytics and modelling components that are available through acknowledged e-infrastructures
- The amount of phenomics initiatives that have implemented their knowledge development or operational processes into e-infrastructure supported workflows
- The amount of co-development initiatives between the “classical” research community of “phenomicists” and ICT and the data science community
- The amount of services in the area of breeding that are tapping from data, information and knowledge generated by phenomics initiatives

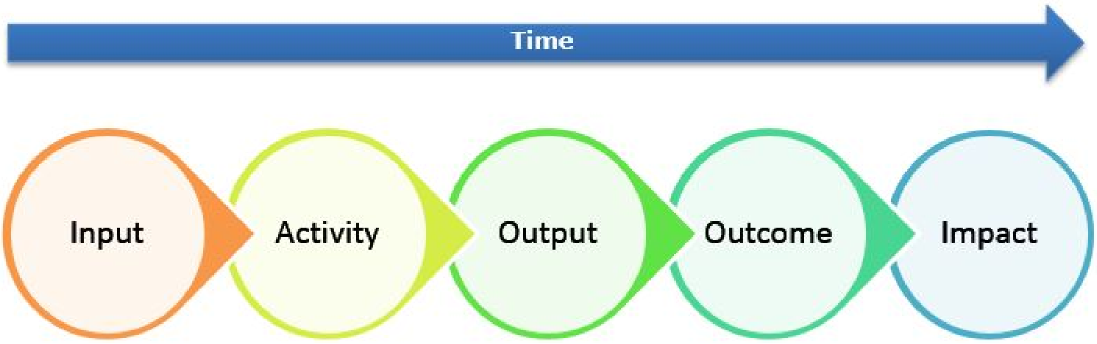
## 5.5 EXAMPLE SHOWCASE 5: FOOD SAFETY

### 5.5.1 Introduction

The growing population (at present time just exceeding 7 billion people, but expected to increase to 8 billion by 2030 and 9 billion by 2050) will require an increase in the total food production. An increase in consumer demand for high quality, high nutritional food (including increased consumer demand for more meat to date, in particular in emerging economies) will accompany this population growth. Food security is therefore of high importance and includes the availability of sufficient and safe nutritious food for all people at all time (FAO, 1996). A global political agreement to achieve zero hunger, food security and improved nutrition and promote sustainable agriculture has been published as Sustainable Development Goal 2 (SDG-2) by the United Nations. Food safety is an integral property of food and agriproducts that must be secured to reach the goals set in SDG-2. The performance of a food supply chain, including food safety and nutrition, is affected directly and/ or indirectly by many factors such as climate, economy, and human behaviour. A systemic approach is needed that takes all of these interactions into account to optimise the food production and to allow mitigation actions when needed (Kendall et. al. 2018). To realise such system approach, data must be combined of different nature and origins. Food safety monitoring data is generated in each EU country by their food safety authority and various information sources are used to direct the activities within the monitoring programme. Data is collected and some are shared via infrastructures set up by European institutions such as the European Food and Feed Safety Authority (EFSA). Furthermore, new technologies are being developed and implemented such as on-site measurements of food safety hazards using mobile phones and the use of citizen science. These developments will move food safety research and control into big data science urging the need in EU and beyond for an infrastructure to share knowledge, data and models related to food safety.

### 5.5.2 Impact pathways

The main impact this case will have is improvement of the food security by ensuring that the food/ agricultural products are suitable for consumption thereby reducing losses due to waste and minimising hospitalisation of the consumers. Contaminated food is responsible for a high number of persons worldwide being hospitalized and even death and > 30% of the food being produced is lost as waste due to contamination. Policymakers and risk assessors will be able to develop and perform mitigation actions when the capabilities of big data including an infrastructure is available.




| Input | Activity | Output | Outcome | Impact |
|---|---|---|---|---|
| Technology infrastructure<br><br>Standards<br><br>Open data solution<br><br>Data on monitoring and data on drivers of change on the performance of the food supply chain (economic, trade, behaviour, climate, agronomic) | Mapping (data, models and knowledge)<br><br>Stock taking (which data sources are used, location, ownership, accessibility, which level is collaboration, sharing possible, desirable)<br><br>Research, dissemination, training & education<br><br>Technology development (big data science implementation in food safety, development of prediction models, early warning) | More efficient use of resources to control food safety<br><br>Improved collaboration international and support of smooth international trade<br><br>Improved collaboration between stakeholders (citizen, authorities and industry)<br><br>Scientific publications<br><br>New improved technologies | More efficient agricultural production<br><br>Healthier population (less outbreaks, less people dying from eating contaminated food)<br><br>Increased technology and innovation (new devices to measure on-site) | Contribution to food security and sustainable development at national and international level. |

### 5.5.3 End user groups

The main end-user will be persons (risk assessors and risk managers) responsible for the food safety at national and international authorities such as national food safety authorities and European Food and Feed Safety Authority (EFSA), national ministries dealing with food safety and European Commission (DG Sante), FAO.

### 5.5.4 Role of Research

Research will provide the opportunity to utilise the new technologies being developed related to monitoring and the use of citizen science. New models will be developed that enable integration of complex data sets, thus enabling the prediction of food safety problems at an early stage so that proper mitigation action can be developed and implemented. Infrastructures are needed to share data, knowledge and models and research that can realise this will stimulate international collaboration and improve food safety. Furthermore, it will provide opportunities to improve the food safety and citizen health of countries with limited resources (developing countries).

Due to the complexity of the food supply chain a systemic approach is needed, meaning that the food supply chain should be considered as a whole. This notion is now understood, but methodologies and approaches on how to perform such systemic (holistic) analysis are not yet well developed, so dedicated research is needed. Recently, Wageningen University & Research (WUR) demonstrated that by using artificial intelligence, machine learning and data from driver of changes, prediction models can be made that outperform all existing models.

Precondition for further development and implementation are the utilisation of big data analytics in food safety research, the availability of FAIR data (food safety, climate agronomic, etc.) and an ICT infrastructure. The currently available infrastructure is not feasible for such activities (neither technical or on management level).
The scientific challenges are handling and integrating big volumes of data of with different nature and origins automatically and developing prediction models utilising this huge amount of information. Hence, all aspects of the DIKW pyramid of data to information, to knowledge and wisdom have to be addressed (see Figure 1).

There are several scientific networks and initiatives that are working to advance this knowledge field. The European Food and Feed Safety Authority (EFSA) is working on this theme (DEMETER) and WUR is coordinating this initiative. Furthermore, in H2020, initiatives are being developed that addressed many of the scientific challenges (Internet of Things, collaboration EU-China etc.). The WHO has recently built a food safety platform “FOSCOLLAB” to provide integration of different sources from various disciplines. A big data collection system in agriculture has been developed in SemaGrow project to support farmers in decision making. The GODAN project, funded by DFID, focuses on weather data, with special attention to those relevant for use in farm management systems. In the RICHFIELDS project, innovative consumer support tools will be developed to select healthy food. The developed tools will utilize food products data, food intake data, lifestyle and health data, including real time consumer-generate data through the use of mobile apps.

### 5.5.5 Challenges in data, infrastructure, processing power and analytics

What are current limitations to the use of data, infrastructure, processing power and analytics?

There are several barriers and limitations that prevent effective use of data, infrastructure, processing power and analytics in the domain of Food Safety:

- Absence of a suitable infrastructure (for data collection, storage, computing, connectivity and data sharing)
- Lack of food safety models able to handle big data
- Lack of food safety prediction models
- Data availability is limited and available data is scattered
- Sustainability of the developed models and systems

What needs to be solved to take advantage of current ICT capabilities?
Therefore, the following needs to be developed to take advantage of current and future ICT capabilities:

- Models enabling systemic approach of food safety
- Infrastructure (i.e. Web environment) to automatically collect, process, store, compute, visualise and integrate data and information from various origins and natures.
- A food safety data platform to collect and connect, compare and share information about food safety generated across Europe.

### 5.5.6 Other solutions for more efficient Research

To bring food safety research to the next level, improved e-infrastructure is required. A Virtual Research Environment (VRE) should support scientists on their research adventure by facilitating collaboration between them and providing them with more effective means of collaboratively collecting, processing and managing data, as well as collaborative knowledge creation. The resulting network, or system of systems, gathering all actors of the food supply chain, will enable the direct contact between scientists and the food chain actors and enhance the knowledge sharing.

The majority of food safety applications requires dealing with diverse sources and forms of data, ranging from highly structured to unstructured data. Data-driven model development is a key approach to extracting structure and knowledge from Big Data. In Food safety applications, such as application of IoT in agriculture, there is a need to deliver real-time information and to be able to handle the huge increase in the number of interconnected devices. To deliver and process real time data, the infrastructure has to offer enough computing power and connectivity. Cloud computing is required to provide the data storage and massively parallel computing, which means also handling a larger amount of data.

## 5.6 EXAMPLE SHOW CASE 6: FOOD & HEALTH

### 5.6.1 Introduction

The European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI) that ran from December 2014 till June 2017 has brought data of human intervention and observational studies to support the improvement of health together. The main objective of ENPADASI project was to deliver an open access research infrastructure that will contain data from a wide variety of nutritional studies, ranging from mechanistic/interventions to epidemiological studies including a multitude of phenotypic outcomes that will facilitate combined analyses in the future.

### 5.6.2 Impact pathways

*Impacts:* Improvement of health of the human population and reduction of non-communicable diseases
*Long term outcomes*:

- Increased understanding of what is healthy for an individual (focussing on life-style)
- More efficient ways to change life-style behaviour
- New ways to improve health in at-risk subgroups

*Outputs*:

- Better description of risk subgroups
- Systems understanding of the effects of interventions
- Better understanding of the basis of individual needs
- New behavioural chance techniques

***Short narrative of the impact path:***
Health care costs are increasing every year, important to this increase is the growth in non-communicable diseases. These can only be reduced if prevention of disease and maintenance of health is better supported. However, data of human intervention and observational studies to support the improvement of health was nor easily accessible. The ENPADASI project that ran from December 2014 till June 2017 has brought these studies together. The trajectories of the subjects in the different studies can be used to find citizens-like-me to show which life-style is healthy for me. In addition, the different interventions in the studies can lead to a better understanding of the effect of life-style interventions on the human system.

**End user groups**
Different groups of end users can be distinguished for the case of food & health

- citizens,
- insurances companies
- food industry
- researchers
- health companies

### 5.6.3 Role of Research

Currently knowledge is too fragmented to work towards actual improved personal health. Research is needed to tie together all knowledge and explain the outcomes. ICT is key to resolve these issues, by making data interoperable and make models for prediction of health.

Especially for the understanding of the human system and the interaction with its environment, not all knowledge is available.

The scientific networks around ELIXIR, ECRIN, BBMRI and other biomedical networks are working towards similar challenges, however, those are directed towards cure of disease. These challenges are more directed to parts of the human system and to subgroups. For improvement of health personalization and system thinking are key.

### 5.6.4 Challenges in data, infrastructure, processing power and analytics

The limitations with regard to data, infrastructure, processing power and analytics lay mainly in the availability of data. ENPADASI has made several datasets accessible, but still too many datasets are not connected. To have enough data on the human system to work towards a citizen-like-me, more data should be connected. This also will lead to new requirement for analytics. Predictive models are key here.

Many datasets are not yet well interoperable and in general good ontologies are still lacking. ENPADASI has developed a new ontology, but did not succeed to cover all needs. Collection of individual data that can be connected to study data and knowledge (as the personal health train concept) is very important for this use case. It's therefore important to work towards:

- FAIR-ness of "all" nutritional data and knowledge.
- Complex predictive models that well describe the underlying biology.
- Full implementation of the personal health train where the citizen is able to control their data.

# 6 OVERALL REFLECTION ON THE APPROACH

As this is the last deliverable of this Work package, summing up the main outcomes on 'Impact Chains and Science Solutions', it is relevant to reflect on the success of the approach of working through use cases, based on the impact chain approach across the agri-food system, and to draw out lessons for subsequent steps.

Retrospectively, we can conclude that thinking in terms of a structured approach to research case studies for open science and the emphasis on analysing impact chains has effectively facilitated the process of consulting stakeholders in the agri-food domain and analysing their specific science cases. Moreover, it has effectively supported the identification of the main challenges, requirements and solutions that are connected with future research that utilizes e-infrastructures as an empowering mechanism for collaborative research and its mechanisms of sharing, connecting and collaborating.

The strengths of the approach were that:

- A link could be made between societal challenges, scientific challenges and ICT and data challenges connected to the role of research, so that none of these challenges appears in isolation.
- A clear division of roles and relevant actors could be established, to show who has a role to play for which reason. This strengthens the logic of the use cases.
- The approach was able to reach a quite detailed level of descriptions and interactions, thereby going beyond generalizations like 'data need to be FAIR'.
- A synthesis of the most important developments could be made across the use cases that demonstrate the priorities for further development of an e-infrastructure, not as to arrive at a list of standalone issues.
- The use cases could be made across the food system, for different subdomains, so the methodology is generic enough to serve a wide array of purposes.

The weaknesses of the approach were that:

- The method is time intensive: it requires quite some effort for researchers to complete the use case forms, and iterations were needed to mature the thinking and elaborate the storyline. This made it difficult to collect a large number of use cases.
- The approach was not suited for an online consultative process, due to its comprehensive nature and time investment required. Consequently, only one additional use case was retrieved through advertising the online Google Form with the use case template questions.
- The approach did not yield the overall storyline on societal challenges, connecting them together. This required a separate effort to develop such a storyline for the Roadmap.

The approach offers opportunities for future refinement, potentially simplification in some places, and possibly with the use of semi-structured interviews for retrieving storylines as an additional method for data collection. Generic broad-brush dissemination through online campaigns are less suitable in this approach, and should be avoided. The approach could be supplemented with online surveys on more generic issues on scientific, societal or data/ICT challenges.