The roadmap towards an overarching data sharing infrastructure at Wageningen University & Research

Smart infrastructures for farm-generated data towards circular agriculture

Jandirk Bulens, Marc-Jeroen Bogaardt, Rob Lokers, Claudia Kamphuis, Fedde Sijbrandij

WAGENINGEN
UNIVERSITY & RESEARCH

The roadmap towards an overarching data sharing infrastructure at Wageningen University & Research

Smart infrastructures for farm-generated data towards circular agriculture

Jandirk Bulens[1], Marc-Jeroen Bogaardt[2], Rob Lokers[1], Claudia Kamphuis[3], Fedde Sijbrandij[4]

1 Wageningen Environmental Research
2 Wageningen Economic Research
3 Wageningen Livestock Research
4 Wageningen Plant Research

This research was (partly) subsidised by the Dutch Ministry of Agriculture, Nature and Food Quality (project number KB 38-001-004 and KB-49.02-001-003).

Wageningen Environmental Research
Wageningen, May 2024

Reviewed by:
Willem Jan Knibbe, Director Wageningen Data Competence Center (WDCC)
Jene van der Heijden, Coordinator Research, Programme Leader (WDCC)
Bas van der Velden, Head of Data Science (WFSR)
Wies Vullings, Team leader Applied Spatial Research (WENR)

Approved for publication:
Wies Vullings, Team leader Applied Spatial Research (WENR)

Report 3357
ISSN 1566-7197

WAGENINGEN
UNIVERSITY & RESEARCH

Bulens, J.D., M.J. Bogaardt, R.M. Lokers, C. Kamphuis, F.D. Sijbrandij, 2024. *The roadmap towards an overarching data sharing infrastructure at Wageningen University & Research; Smart infrastructures for farm-generated data towards circular agriculture.* Wageningen, Wageningen Environmental Research, Report 3357. 70 pp.; 16 fig.; 14 tab.; 10 ref.

Dit rapport beschrijft de bevindingen van de verkenning van een aantal bestaande onderzoeksdata-infrastructuren bij WUR. Het hoofddoel is het ontwikkelen van een strategie gericht op een gemeenschappelijke data-infrastructuur voor de hele WUR waarbij tegelijkertijd de huidige state-of-art data-initiatieven worden behouden en geïntegreerd. Voor de verkenning zijn deze infrastructuren in een bestaand raamwerk voor het delen van data geplaatst, en zijn ze met elkaar verbonden in specifieke use-cases. Daarnaast hebben interviews plaatsgevonden met WUR management, onderzoekers en data stewards om het belang van het delen van data, en de uitdagingen daarvan, te achterhalen. Alle resultaten samen hebben geleid tot het opstellen van een roadmap en aanbevelingen voor WUR om naar de gemeenschappelijke infrastructuur toe te werken.

This report describes the findings of an exploration of several research data infrastructures at WUR. The main goal is to develop a strategy for a common data infrastructure for all of WUR whilst also reusing and applying current state of the art data initiatives. To explore these infrastructures, we placed them against an framework for data sharing and connected them in specific use-cases. Additionally, interviews with WUR management, researchers, and data stewards were held to clarify the importance of sharing of data and the challenges involved in sharing it. All the findings were used to prepare a roadmap and recommendations for WUR to work towards a common data sharing infrastructure.

The pdf file is free of charge and can be downloaded at https://doi.org/10.18174/659257 or via the website www.wur.nl/environmental-research (scroll down to Publications – Wageningen Environmental Research reports). Wageningen Environmental Research does not deliver printed versions of the Wageningen Environmental Research reports.

Wageningen Environmental Research Report 3357 | ISSN 1566-7197

Photo cover: Shutterstock

Contents

Verification

Report: 3357
Project number: KB 38-001-004 and KB-49.02-001-003

Wageningen Environmental Research (WENR) values the quality of our end products greatly. A review of the reports on scientific quality by a reviewer is a standard part of our quality policy.

Approved reviewers who stated the appraisal,

name: Willem Jan Knibbe, Director Wageningen Data Competence Center (WDCC)
Jene van der Heijden, Coordinator Research, Programme Leader (WDCC)
Bas van der Velden, Head of Data Science (WFSR)
Wies Vullings, Team leader Applied Spatial Research (WENR)
Mieke Weegels, Team leader Expertise Groups/Other (WECR)
Quirijn van der Goes, Head Data Science and Innovation (WECR)
Roel Veerkamp, Researcher, Professor (WLR)
Bert Lotz, Team leader Applied Ecology (WPR)
Shauna Ní Fhlaithearta, Researcher (WDCC)

date: 6 April 2024

Approved team leader responsible for the contents,

name: Wies Vullings, Team leader Applied Spatial Research (WENR)

date: 6 April 2024

Executive Summary

This report describes the work performed to establish a strategy towards a shared data infrastructure for Wageningen University & Research (WUR). We define a data infrastructure as the complete set of components (including hardware, software, networking, services and policies) that enable data sharing. This work was part of the 'Smart Infrastructures for Farm Generated Data towards Circular Agriculture' project of the Data Driven & High Tech Knowledge Base programme. It evolved from the project's broader activities on exploring and advancing a range of data infrastructures that exist within WR and prototyping connections between these infrastructures. It was a joint effort between the participating research institutes, combining the knowledge and experiences of researchers, data scientists and IT specialists with various infrastructures.

Many data infrastructures have been developed within WUR in recent years. These have occasionally evolved from specific projects or programmes with targeted aims and requirements. However, over time, they have also proved to be sustainable and usable for broader purposes. Their application domains and the used architectural concepts and technologies differ because they often focus on the targets of a specific project, institute or client, rather than a more holistic WUR perspective on data architecture as a driving force.

Assessment of WUR data infrastructures

A variety of data infrastructures were analysed to obtain an overview of the different infrastructure solutions and their practices. The seven assessed data infrastructures were all part of the 'Smart Infrastructures for Farm Generated Data towards Circular Agriculture' project. All these data infrastructures are currently used in various WUR research institutes. We used the Innopay conceptual framework for data sharing to analyse and assess them. This framework distinguishes nine 'soft building blocks' that need to be considered when setting up a data sharing infrastructure (see **Figure 1**).

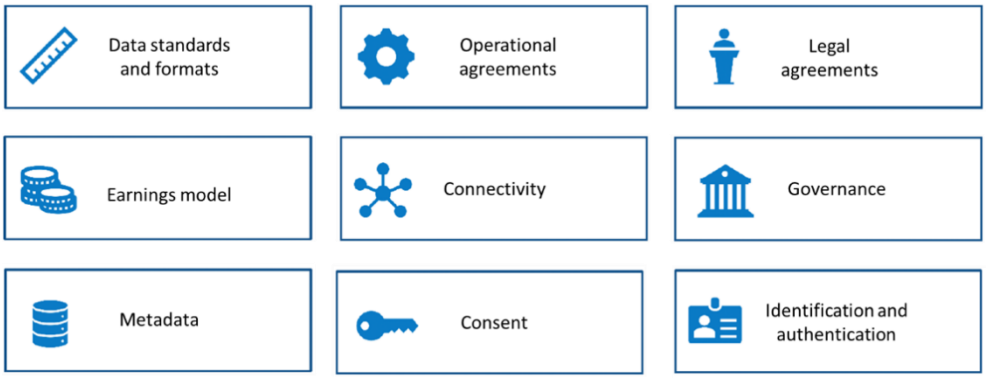

Figure 1 *A soft data sharing infrastructure based on nine building blocks according to the Innopay framework*

As expected, the outcomes of this assessment show that the implementation of the building blocks was diverse (**Table 1**). Overall, many of the building blocks are already addressed to varying extents, with red indicating not covered, orange indicating partly covered, green indicating sufficiently covered, although further improvements may be required. This indicates that there is an awareness of their importance across the research institutes, but also that their implementation is a learning process. At the same time it should be noted that none of the data infrastructures fully covered all building blocks.

Table 1 *Assessment of the current status of the used data sources per building block*

Data source / Building block	Initial /experimental phase	Data standards and formats	Earnings model	Metadata	Operational agreements	Connectivity	Consent	Legal agreements	Governance	Identification and authentication
Farm Information Net (BIN)										
Farmmaps										
AgroDataCube										
Farm Data Safe	yes									
Methane Data Lake (MDL)	yes									
Federated Food Fraud Data (RASFF/EMI)										
Freshwater Fish										

Earnings models were missing in some cases, although this can be explained by the specific goals of these data infrastructures. Notably, most of the evaluated data infrastructures did not fully cover data standards, often implementing technical standards but lacking harmonised semantics. This is also true for the metadata standards, as no standard metadata template and exchange protocols are used in many of the evaluated cases. The building blocks for identification and authorisation, and for consent were implemented by all infrastructures, although ad-hoc, often without considering future broader integration.

Obviously, the chosen implementation of the building blocks is geared towards the targeted application area of the infrastructure. Within these demarcated 'application' environments, the relevant components covering the building blocks appear to operate well. To reveal how the infrastructures 'behave' in a broader context, we also explored prototypes of different connections between these infrastructure to discover their strengths and weaknesses regarding connectivity, interoperability and their ability to add value outside of their specific contexts. In the establishment of these connections, surprisingly no major issues were encountered. This is partly because of the controlled and experimental character of these prototype connections, where lack of clarity regarding earnings models, legal agreements and governance could be easily overcome or ignored. What can also be concluded is that the lack of shared semantics and the heterogeneity of the available metadata and provenance required substantial efforts to bilaterally align these connections. Such lack of documentation (as in metadata, and provenance) and semantic interoperability will cause serious interoperability hurdles if larger and more business-critical cross-infrastructure applications need to be developed. In such cases, the alignment over other building blocks will probably also need to be considered.

Broader WUR perspectives on data infrastructure

To get a broader view on the organisational perceptions regarding the future needs for a WUR shared data infrastructure, a series of interviews were conducted with representatives of WUR management, researchers and data stewards. Interviews with WUR management revealed that the importance of data sharing is widely recognised as a critical asset. Broader vision, knowledge, capacity and support for data sharing within the organisation is needed, particularly to support the full data life cycle even beyond the duration of the project. The role of data management plans, which is less embedded in the research institutes, is crucial in improving this situation. More clarity is needed around issues like data rights, control and ownership and the available opportunities to fund sustainable infrastructures. Better cooperation and integration between institutes, for example through dedicated projects, could catalyse further professionalisation of WUR's infrastructure. WUR

researchers and data stewards have mentioned important barriers for data sharing, like the lack of data storage protocols, the fear for loss of control over 'their' data and the fact that there are no personal benefits for data sharing. Lack of knowledge and time to manage data were also mentioned, as well as the fact that data are often hard to find. Motivation to share data could increase if it were more actively promoted and rewarded, if sufficient resources (time, budget) were allocated and if clearer guidelines were available.

A roadmap towards a future infrastructure for data sharing
Using the learnings from the practical work, data source assessments and conducted surveys, we propose the adoption and implementation of a roadmap to further develop WUR research practices regarding data management. This includes the integration of FAIR data management into a WUR corporate digital strategy, the establishment of a sustainable and future proof data infrastructure and the improvement of data management at all levels. This report provides important recommendations to support the many challenges involved in realising this in Chapter 10. These recommendations can be summarised as follows:

Taking corporate responsibility for a digital strategy aimed at FAIR data management, including the sustainability of the required data infrastructure in a project-oriented environment like WUR.

This should take a broader perspective that includes the essential existing infrastructure components that are currently overlooked. It should aim at financially and organisationally supporting important challenges that transcend the individual science and research groups, like improving semantic interoperability and developing collaborations aimed at connectivity in a federated environment. It is advised to use a soft infrastructure building block framework as reference, preferably the one used in the development of the European Common Data Spaces. Driven by a sound digital strategy supported by the WUR Board of Directors, with main leading and supporting roles for WDCC and corporate IT.

Further adoption of good data management practices (metadata, persistent storage and adoption of semantics) needs to be encouraged on the level of science and research groups.

More and broader multidisciplinary collaborations between different groups are needed to be able to share knowledge and experiences, to learn and to move to a more interoperable and interconnected WUR data ecosystem. In this respect, it is important to take future sustainability into account, for example by establishing earnings models and more generally, by considering the different infrastructure building blocks. This can be initiated between research groups, following a corporate strategy and supporting programmes.

WUR should initiate the movement towards an approach that exposes data in a federated ecosystem.

This leads to a Data Mesh (
Figure 2), in which data are separated from applications and accessible as 'data as a product'. In such constellations, data exist completely independently and is 'self-servicing', and use can be guaranteed without further negotiations. A Data Mesh can also accommodate many of the domain-specific infrastructures that currently exist at different organisational levels. It is then extremely important to separate data and applications to ensure that we prepare our organisational infrastructure for a data mesh. This could be established as a co-development between FB-IT and the groups that maintain independent data infrastructures at WUR, supported by WDCC.

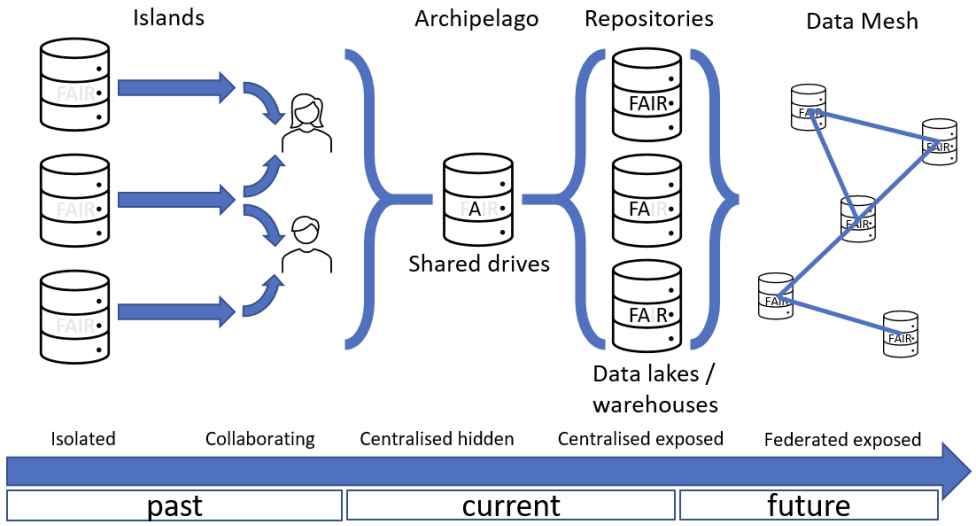

Figure 2 *A roadmap towards a 'common data infrastructure'*

Next steps and Roadmap maintenance

We can see increasing interest in data sharing at WUR, and the demand for a single shared infrastructure gradually leads to initiatives to support researchers with knowledge, capacity and tools to bring data sharing into practice. This includes steps to improve on some of the challenges and recommendations presented in this report. At the same time, we see that many of these efforts are targeted at supporting individual researchers and research projects. Truly interdisciplinary data intensive research, which requires combining data from different domains, is still very challenging. Moreover, as revealed by this report, there has been little or no attention (nor interest) in bringing together the many well-developed data infrastructures at WUR up till now. While there is great added value in linking these systems, they still largely operate in silos. Implementation of the roadmap should, therefore, specifically onboard the deployment of this broader perspective on a single data sharing infrastructure for existing and future data infrastructures to align with each other.

The presented roadmap is not meant as a definitive endpoint. While we believe its foundations are solid, it is also obvious that many initiatives around data infrastructures are evolving on a national, European and global level. The emergence of the Common European Data Spaces as part of the European Strategy for Data is a good example of an important development that will also affect Life Sciences research at WUR. In the coming years, we need to keep close track of such ongoing developments and adapt the implementation of this roadmap where needed.

1 Introduction and background

Wageningen University & Research (WUR) consists of multiple research groups within the domain of life sciences that use and produce data for their research. The data may originate from different sources, from farmers to industries, academia, research institutes and governmental agencies. The data are collected and generated for specific, often knowledge-based, research purposes. In the era of Big Data, in which huge volumes of data are generated at high speed, and increasing computational power, there is growing interest in data-driven research to gain new insights. However, one of the biggest challenges in data science is the accessibility and reusability of data that were originally collected for a specific goal. On top of this, it may be unclear whether these data can be used for other research purposes because they may contain sensitive information and/or come from external parties. These uncertainties make data sharing in and between projects and reusing data in new projects difficult, which negatively affects data-driven research.

The 'Smart infrastructures for farm generated data towards circular agriculture' Knowledge Base Data Driven High Tech (KB DDHT) project ran from January 2019 until December 2022. The objective was to develop an innovative and privacy conserving data infrastructure across WUR, while reusing and benefiting from current state-of-art data initiatives. This infrastructure should allow for a safe, efficient and controlled sharing and reusing of data that are generated or relevant at the farm level. This infrastructure should facilitate data-driven research while respecting data rights and data privacy regulations. It would stimulate activities to reuse, combine, process and integrate data within WUR, and outside WUR ultimately, and support scientific data-driven research where data control and sensitivity is preserved.

This report details the results of the exploration of such an innovative and privacy conserving research data infrastructure at WUR. The requirements and state-of-art data initiatives are explored through several use cases and the findings are placed in an existing framework for sharing data, which is made up out of nine building blocks developed by Innopay and referred to by the Netherlands AI Coalition[1].

In Chapter 2, we detail two concepts for sharing and reusing data: federated data and FAIR data principles. This chapter describes the nine building blocks of the 'soft data infrastructure' Innopay framework that can be used to smoothen the implementation process. Chapter 3 describes the data sources within WUR that were used in this project to explore options for data sharing and reuse. It also assesses how each of these data sources link to the nine building blocks listed in Chapter 2.3. Chapter 4 describes the findings from different use cases in which connections to share and reuse data between the data sources described in Chapter 3 were explored. Chapter 5 integrates findings from the previous chapters to report current state-of-the-play at WUR in relation to the building blocks. Chapter 6 summarises interviews held on data infrastructures at a managerial level and with data stewards of all WUR research groups. Based on this, we report on what we think is needed on different levels to come to a shared data infrastructure at WUR in Chapter 6. Chapter 7 focusses on current developments, starting with an overview of WUR's strategic visions relating to data infrastructures. In Chapter 8, we look at global developments relating to data infrastructures, looking more specifically to European contexts, since these strategies provide direction for the design of the data sharing infrastructure at WUR. Finally, Chapter 9 provides recommendations and a roadmap towards a shared infrastructure at WUR.

[1] https://nlaic.com/wp-content/uploads/2020/08/NL-AIC-Naar-First-time-Engineering-en-Operationalisatie.pdf (accessed 12 -02-2024)

2 Federated data, FAIR data principles and a conceptual framework data sharing

This chapter describes two main concepts, the federated data and FAIR data principles, that allow data to be shared and reused. Building on these principles, it describes a conceptual framework, also referred to as a soft infrastructure, which offers practical guidelines to smoothen the process of implementing such an infrastructure.

2.1 Federated data

McLeod and Heimbigner (Heimbigner & McLeod, 1985) were among the first to define a federated database system in the mid-1980s. In principle, data federation is a data management strategy that can help improve data quality and accessibility. Data federation is the process of querying data from different sources. It is based on the concept of leaving data at their source and using multiple sources. This is extremely helpful if the data are either sensitive, as is the case with personal data, or if there are large volumes of data. This principle is especially used in AI models, where training data are decentralised and results are transferred back to the central AI model. This is called 'Federated Learning' (Abdulrahman et al., 2021).

This is in high contrast with current research practices, where data is often collected from different sources and combined and integrated into a large database for a specific research purpose. Because the data stays at its source, the data owners remain in control of their own data, which may create a willingness to share data while they are preserved and protected at the source. When working with data, one may question if data users really need all the available data or if the computed results suffice. Despite the advantage of data owners staying in control, there are downsides for users. This includes uncertainty of the data quality of and uncertainty of how long the data remains accessible at this specific data source.

Examples of federated data concepts are the FAIR Data Train and Federated Learning (Appendix 1). Examples of a FAIR Data Train are the Personal Health Train[2] and the Farm Data Train.[3] In Federated Machine Learning, the data are not brought to the AI model, but the AI model is brought to the federated data sources. The AI model is retrained on each data set that includes the data of that data source, which results in more accurate AI models. The retrained models are then used further.[4] General examples of federated learning in real-life are face recognition and voice recognition.[5]

In this context, it is useful to note that, while federated learning has been a focus area in WUR data science, there are several other mechanisms that allow working with federated data. Compute-to-Data and Secure Multi-Party Computing are other examples of often used constructs that offer opportunities to work on federated data sets.

2.2 FAIR data principles

When reusing data, it is essential that these adhere to the FAIR principles (Wilkinson et al., 2016). These data need to be 1) Findable: both human and computers should be able to find data easily. Published metadata and machine-readable metadata is important to finding data automatically; 2) Accessible: once found, it is essential that the data can be accessed. This does not imply that data are always freely accessible. Sometimes authentication, authorisation and/or payments are required before gaining access; 3) Interoperable: data needs to interoperate with other data, applications and workflows for analysis and

[2] https://www.dtls.nl/fair-data/personal-health-train/ (accessed 21-12-2023)
[3] https://www.dtls.nl/fair-data/farm-data-train/ (accessed 21-12-2023)
[4] https://pht.health-ri.nl/use-cases/health-research/use-case-amicus-ai-medical-imaging-novel-cancer-user-support (accessed 21-12-2023)
[5] https://www.v7labs.com/blog/federated-learning-guide (accessed 12 -02-2024)

processing. It is essential to exactly know the structure and definitions of all data elements, both semantically and technically; and 4) Reusable: to finally be able to reuse the data, metadata should describe the data well, so that they can be replicated, combined and extended to new settings. These four principles make data FAIR. They appear to be simple, straightforward and easy, but it is a complex task to realise them in real-life and to implement them simultaneously. Ignoring one or more of these four principles will result in a failed data sharing infrastructure.

2.3 The Innopay framework

A conceptual framework offers a foundation and practical guidelines to smoothen the process of realising and implementing a FAIR data infrastructure. An example is the framework developed by Innopay, which consists of nine building blocks to describe the accessibility of a data sharing network in the logistics sector.[6] This framework is a collaborative effort to improve conditions of data-sharing for organisations that seek collaboration in the data ecosystem. It was initially developed for the Netherlands' Logistics Top Sector, a collaboration between public-sector and private-sector partners within the Dutch transport and logistics sector. They collaborated on the development of a uniform set of agreements to connect with each other on the basis of mutual trust. Of the Innopay framework the iSHARE Trust framework[7][8] is more known. It focuses on one of the building blocks of the Innopay framework, namely the identification and authentication building block.

We used the Innopay framework (**Figure 3**) as the principal elements for creating an overarching WUR research data infrastructure. Describing the data sources at WUR according to these nine building blocks will provide insight into the strengths and weaknesses of data sources in relation to data sharing. Note that this framework describes a soft infrastructure on a conceptual level, providing insights into what needs to be considered when implementing a certain data initiative. Solutions implemented for particular use cases for an aspect belonging to a specific block may reveal what is required for a successful WUR data sharing infrastructure. The nine building blocks of the Innopay framework are described in more detail in the subsequent nine sub-sections. By analysing different data sources within the use cases (Chapters 3 and 4), we determined which blocks are relevant in specific research situations, what is omitted and how this reflects in their implementation.

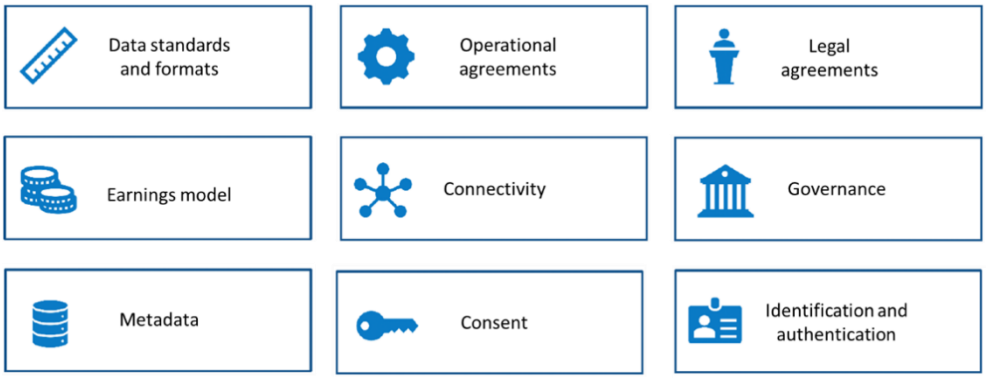

Figure 3 *A soft data sharing infrastructure based on nine building blocks according to the Innopay framework*

[6] https://www.Innopay.com/en/publications/how-are-soft-infrastructures-ishare-promoting-data-sovereignty-create-new-value (accessed 21-12-2023)
[7] https://ishare.eu/ (accessed 21-12-2023)
[8] The iSHARE Trust framework is also referred to by the Dutch Artificial Intelligence Coalition (NL AIC) in their 'Blueprint NL AIC Data Sharing System Architecture'https://nlaic.com/wp-content/uploads/2020/10/Responsible-data-sharing-in-AI.pdf (accessed 21-12-2023)

2.3.1 Data standards and formats

Data standards and formats are relevant for any sharing data infrastructure on both a technical and semantic level. A data standard is an assembled collection of data components that uniformly describe data according to the expectations and understanding of all data users. Technical standards for data formats refer to values of properties and variables that are collected and exchanged, if the data type of these values are an identifier (ID), date, text, integer and whether the values of these variables are expected to meet a certain pattern. Many technical data exchange standards are available, such as JSON, CSV, RDF and XML, and standards to exchange data are typically developed for certain applications, domains or disciplines. In principle, this does not have to be a problem to data sharing as long as it is well-documented according to the used standard or format. Semantically, data can be described in an ontology that expresses the objects and the relationships between them, either as class diagrams or by linked data. In practice, existing ontologies should be adhered to as much as possible, as there are already many domain-specific standards.

2.3.2 Earnings model

Earnings models for data sharing are of key importance for allowing the creation of data sources and for maintaining data sharing infrastructures. The foreseen benefits of these earnings models can be economical or societal, or both. For example, when using data and data infrastructures to support a policy cycle, which follows the flow of preparation, decision making, implementation and evaluation, societal benefits are the driving forces for a data sharing infrastructure, rather than monetary benefits.

2.3.3 Metadata

Metadata is the documentation of data, also known as 'data about data'. Metadata documentation describes details about the contents, formats and internal relationships of data, and it enables others to find, use and properly cite the data. It should provide clear information on the quality of the data that is exchanged, what the provided information was originally intended for and the expected application domain. Metadata should also provide information on the data licence, which contains the rights to use the data. This could be a standard license or, if no standard licence is used, access rules with usage limitations and restrictions could be provided. This links closely with the Legal Agreements building block.

There are many metadata standards available and these can vary between applications, domains and disciplines, as illustrated in **Figure 4**. Some metadata standards are designed for the purpose of documenting the contents or technical characteristics of files, and others for expressing the relationships between files within a data set. Metadata standards are based on a number of agreed elements. The 15 elements from the Dublin Core Metadata Initiative (DCMI)[9] are a well-known example.

[9] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (accessed 21-12-2023)

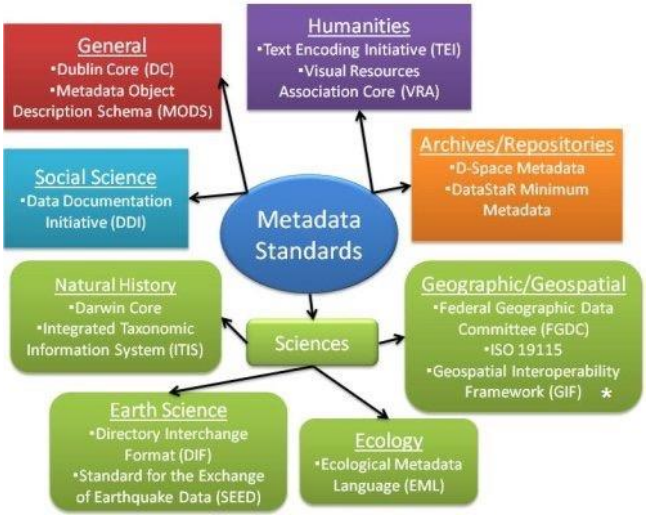

Figure 4 *Some examples of metadata standards, varying between applications, domains and disciplines*

2.3.4 Operational agreements

Operational agreements define the common arrangements that are established regarding the organisational roles and responsibilities and the technical specificities that need to be in place to setup, manage and maintain data exchange processes. These are relevant in case a workflow consists of a chain of activities that are distributed over different partners, organisations and, in the case of computational and/or data storage, even nodes. This is even more important when processes are distributed or federated, and requires more detailed agreements, including data structures, specifications and formats. Operational agreements should cover organisational, executable software and hardware and data aspects of the entire chain or workflow to guarantee a correct and smooth completion of the process. This links with the Connectivity building block, which concerns the technical implementation of the operational agreements in the used environments.

2.3.5 Connectivity

The established operational agreements between partners when sharing data need to be implemented in the technical provisions so that it enables the connections of all components in a data flow as well as their technical communication. Application Programming Interfaces (API) are a common technical mechanism to implement such provisions. In contrast to a user interface, which connects a computer to a person, an API connects computers or pieces of software to each other, allowing applications to talk to each other. It is a programming-based software interface that acts like a bridge between two or more applications and/or systems, enabling them to seamlessly communicate with each other without making any change in the original application or system.

2.3.6 Consent

Consent mechanisms allow data owners to control the access to and reuse of their data outside of a regular license. Permission can be given to access and reuse data that is not declared as open data. Non-open data are data that are non-public or otherwise subject to restrictions. This consent usually determines which user can access and reuse the data and for which purpose. It can also contain the type of data and the time period for which data sharing (processing) and use is granted. The conditions that apply to consent can be detailed in a Data Use Agreement (DUA).[10] These DUAs, sometimes called 'data exchange agreements', are contracts between data owners and data recipients, usually initiated by the data owner. Composing a DUA may be particularly relevant if the shared data involves sensitive data. A template DUA is available for WUR.[11] When composing a DUA, the following aspects are to be considered: legal aspects, including the General Data Protection Legislation Act; technical aspects, for example aspects describing how the data

[10] Alex Ball (DCC), How to License Research Data [https://www.dcc.ac.uk/guidance/how-guides/license-research-data] (accessed 21-12-2023)

[11] https://intranet.wur.nl/umbraco/en/frequently-asked-questions/what-are-the-frequently-used-templates/ (accessed 21-12-2023)

should be stored and protected; data rights; privacy, pseudonymisation or anonymisation of human data; use of data by third parties; embargo period; requirements for citations and/or co-authorship of publications; and the specific purposes of using the shared data.

2.3.7 Legal agreements

Legal agreements, providing the right to use data, are usually described in a licence. A data licence is a legal instrument that specifies a standard set of terms and conditions regarding the sharing and reuse of data. It defines what others may or may not do with your data and is an important aspect in making sure your data meets the R (Reusable) principle in FAIR data management. Some specific regulatory aspects regarding licensing are particularly relevant for WUR Research.

A legal framework is in place for situations where data have been collected using public funding. At EU level, it is composed of several directives and regulations that are legally bounding in every member state (Appendix 2). Such legal acts provide specific exceptions on when not to share, disclose and disseminate data to the public. Personal data, for instance, are protected by the General Data Protection Regulation (GDPR). All EU legal provisions share a general guideline: all data must be shared as open data with one of the open data licences when financed with public money.[12] In the Netherlands, we have an open data policy in line with this EU policy. It means that all data collected by or for the government should use the Public Domain Mark or Creative Commons 0 (CC0) license.

2.3.8 Governance

Governance is the process of managing the availability, usability, integrity and security of data. Governance is firstly an organisational topic, where an organisation is considered as the 'whole' in which collaborative work is taking place. This is irrespective of whether the collaborative work is distributed or federated amongst parties. The Governance building block relates especially to information technology (IT) governance, which can be defined as: "the process used to monitor and control key information technology capability decisions - in an attempt - to ensure the delivery of value to key stakeholders in an organization."[13] According to Louis & Associates (L&A),[14] IT governance is a process with the objective to deliver business results. The IT governance process, therefore, monitors and control key IT decisions that might have a positive or negative impact on business results. The objective of IT governance is not just the delivery of risk optimised business value, but also to engender the trust of the key stakeholders in the people who they have entrusted their funding. Data governance refers specifically to the management of data to support adding value to the data products and is very similar with IT governance. Appointing a chief data officer improves aligned data within the organisation.

2.3.9 Identification and authentication

To technically implement legal agreements, for example implemented through consent or authorisation mechanisms, identification and authentication are essential to uniquely and genuinely identify users when exchanging data. In a WUR data infrastructure, provisions should be made to implement identification and authentication within WUR and externally, working towards a more uniform, straightforward and controlled way of exchanging data on a larger scale than is possible right now. Identification and authentication should be 1) Uniform: one way of working which is compatible with all modalities, big and small organisations, public and private organisations, suppliers and receivers of data as well as their software partners; 2) Straightforward: easy to connect with new and existing, internal and third-party partners throughout the sector, providing more certainty on trustworthiness of parties you exchange data with; 3) Controlled: based on the principle that the owner of the data stays in control at all times; the owner decides with whom what data is exchanged and on what terms.

[12] An exception on this that Member States may never limit access to information on emissions into the environment
[13] From the Wikipedia of CIO-Index; https://cio-wiki.org/wiki/IT_Governance#google_vignette (accessed 21-12-2023)
[14] https://www.l-a.lu/it-governance/ (accessed 21-12-2023)

3 Data sources at Wageningen Research

Within WUR, large volumes of various kinds of data are collected, generated and reused as part of research activities. These data are historically stored in many ways, ranging from individual storage on users' personal devices, shared project folders and institutional shares, to well-structured and governed databases and repositories. As a result of all these different approaches to storing data, it is a challenge to find, acquire and reuse data in new research. In recent years, with the adoption of Open Science, the FAIR principles, the rise of big data technologies and other advances, the importance of organising and connecting data to make them reusable and to store them for longer has grown.

This project assessed the current state of play of the WUR data infrastructure by exploring some data sources and by experimenting with use cases that connect them. This chapter introduces the data sources that were part of this assessment, their main characteristics and how they are currently managed and being used as part of research. Each data source is described briefly, explaining the context in which it is used and the types of data that are stored. An assessment was made on how the individual data sources implemented the nine building blocks described in Chapter 2.3. This assessment provides an overview of the different concepts, methods and technologies that are used within each data source to link to data providers, data users and other components in their environment, as well as how these data sources are currently managed and sustained. This overview helps to understand how WUR and its research institutes respond to challenges like:

- How do we make sure that data can be shared and be understood and reused by others, for example, sharing and reusing hardware, software, standards, metadata and data models. How do we make our data FAIR? This relates to the Data standards and formats, Connectivity, Metadata, and Operational agreements building blocks;
- How do we financially sustain our data sources? How do we organise the data sharing business model? Who pays, and for what? What costs are covered by these fees? How do we cover the total cost for sharing data? What are the societal benefits? This relates to the Earnings model and Governance building blocks;
- How do we organise our data governance and data management? What agreements and/or licences do we use? Who owns and maintains the product and the agreements? What is the organisational structure? What roles and responsibilities are there? This relates to the Operational agreements, Legal agreements, Governance, Consent, and Identification and authentication building blocks;
- How do scientific, social and cultural aspects of data affect how we work with data? How do we disseminate, share and exploit the knowledge gained? This relates to the Earnings model, Governance, and Consent building blocks.

All of the data sources used in this project are introduced in separate paragraphs outlined below.

3.1 Farm Information Net (BIN) – Wageningen Economic Research

The Wageningen Economic Research institute (WECR) 'Bedrijveninformatienet' (Farm Information Net, BIN) has been collecting and administrating the financial and sustainability data of 1,500 agricultural and horticultural companies, 100 fishing companies and 150 private forestry companies in the Netherlands since 1965. This is a legal task undertaken on behalf of the Ministry of Agriculture, Nature and Food Quality. Aggregated tabular data and analytics derived from this data source are published on the www.agrimatie.nl website. Under strict conditions, it is also possible to get access and use anonymised individual company data for scientific research.

WECR has been using the tailor-made system Artis to manage and process the data collected and stored in BIN for more than 20 years. In the WECR projects, the BIN data are used and combined with many other

data sources, like Comext Trade Data, RVO Agricultural Census Data and World Bank Data. To facilitate FAIR data management for these data sources, WECR has developed the Adagio data management solution in cooperation with FB-IT. Adagio contributes to the availability, quality, reusability and enrichment of data and the integration of WUR-internal and external data sources. Adagio is a knowledge-driven data warehouse that facilitates data management in every phase of the research cycle. Data stewards and data engineers can register their knowledge about data in a broad sense, including its use, concepts, validation, heuristics and logic rules. This is done to the individual attribute level to make it interoperable with other data sets. It helps data stewards to convert data from different sources under one query engine and enables them to link information from different sources through classification management, so called 'lightweight ontologies'. Furthermore, it improves data quality by avoiding manual errors and by applying validation rules like checks on data format and validation rules specified by domain experts. Moreover, Adagio stores history and supports traceability and reproducibility. This enables researchers to become less dependent of IT systems and to focus on their core business: their research. This way, researchers' valuable time will be used more efficiently.

Presently, access to and the use of BIN data is facilitated by the Artis and Adagio systems, depending on what level of detail is needed for the research. In the future, the BIN data management will be transferred to Adagio to fully profit from the aforementioned capabilities.

We have analysed the BIN data management system according to the nine building blocks of the Innopay framework (Table 2).

Table 2 *Assessment of BIN according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS BIN
Data standards and formats	• Data web services API to retrieve data from the system. • Broad variety of technical connections with external data sources - Web service API; - All generally used file formats, for example xlsx, csv, xml.
Earnings model	• Enable researchers to easily cooperate by sharing and using data, knowledge and practices in every phase of interdisciplinary research. • Relieve researchers of time-consuming complexities in data usage and data sharing with regard to operational efficiency and the requirements of security, privacy and FAIR. • Enable researchers to acquire and execute projects to answer research questions that require an interdisciplinary approach (that could not be fulfilled earlier). • Enable customers and partners to view and explore data and information, compliant with privacy and security guidelines. • Reduction of efforts and costs by increased efficiency and transparency in every phase of research.
Metadata	• The solution is metadata driven. The data in research domains are described in terms of common concepts that are defined by domain experts. For the technical implementation, the logical level is automatically translated to physical databases, data models, validation procedures and tabular models. Validation rules to check and ensure high quality of research data, for example on the consistency of data (e.g. a specific number of animals fits only to a specific size of a farm). • A metadata and knowledge service provides an overview of the available data.
Operational agreements	• Access to detailed BIN data is restricted by contractual agreements and can only be done in a microlab environment where no data can be extracted. • Agreements on frequency of updates have to be made depending on needs, frequency of suppliers and agreements between WECR and data suppliers.
Connectivity	• Commonly used APIs for data retrieval out of the system and for data import into the system are available (data as a service). • Specific selections of data (queries) and graphical representations can be retrieved through these APIs. • Interactive Dashboards (e.g. PowerBI) are easily integrated
Consent	• The participants (farmers) in BIN, as well as other data source suppliers (e.g. RVO landbouwtelling) consented to the use of their data in research, but only with results published on a sufficiently aggregated level, not traceable to individual farmers. • Model owners (domain experts) are in control of the granted access and the level of access to their data in their software models. They are responsible for the configuration of the corresponding level of authentication and authorisation.
Legal agreements	• Legal agreements have been made with data suppliers. • Legal agreements also have to, or can be made with data users, like for using microlabs.

BUILDING BLOCK	KEY ASPECTS BIN
Governance	• A data steward or data management expert role sets up and harmonises procedures and guidelines for data management. They will be an ambassador towards researchers, data engineers, project managers and the general management. • A group of data stewards or engineers is operationally responsible for the governance of the current available data. • An information security officer and privacy officer provide support and governance advice.
Identification and authentication	• Identification and authentication is implemented using the standards as used by FB-IT, for example Active Directory Federation Services (AD FS), a software component that provides users with single sign-on access to systems and applications located across organisational boundaries.

3.2 Farmmaps – Wageningen Plant Research

Farmmaps is a real-time farm data and service platform created to develop and deploy field data, WUR models and digital twins for research and practical applications (Been et al., 2023). It is a geo-based platform, which means that the data are collected and stored based on plot level. Various data sources are available on Farmmaps, such as worldwide weather and satellite imagery and crop registration data, which may be imported from a Farm Management Information System (FMIS). Farmmaps also has user accounts and a billing system. The platform was created in 2010 as a geo application for the farmers decision support system NemaDecide[15] to generate management advice regarding different nematodes in potato production. Akkerweb was launched in 2016 as a geodata platform for arable farmers. In 2020, Akkerweb was succeeded by Farmmaps. Farmmaps uses well-known and broadly used technical standards like REST and (Geo)JSON for the implementation of API and formatting. API documentation provided through Git[16] with sample code serves as an example on how to use the Farmmaps APIs (e.g. to run a blight advice or to make a nitrogen topdress taskmap based on satellite data). A reference of the API can be found on Farmmaps'[17] swagger page.

The platform runs worldwide and hosts over 20 decision support modules, which it refers to as 'apps'. With these apps, farmers and farm advisors can directly apply the results of scientific research in education programmes, study groups, trial stations and transition projects, like Farm of the Future (Boerderij van de Toekomst, BvdT) and the Digital Future Farm. The available apps can be divided in three types: 1) Scenario studies and predictions; 2) Crop management advice, and; 3) Benchmarking and accountability. Scenario studies and predictions apps assess the impact of certain actions or measures farmers can take and focus mostly on the long term. Crop management advice apps provide in-season advice, which can be used in daily farm operations. The benchmarking and accountability apps use historic data to score farm performance. Farmmaps complies with the Code of Conduct for data use in arable farming, which originates from BO Akkerbouw, the umbrella organisation for arable farming (Gedragscode Datagebruik Akkerbouw van BO Akkerbouw).[18] Users are in full control of their data on Farmmaps, and an authorisation system has been developed to share data for a specific research questions and with specific organisations. In Farmmaps, users can indicate if specific data can be shared in their account, as well as for what purpose and with whom. This permission can be withdrawn at any time. Given the rights authorisation, users can easily import farm operational data form their FMIS, such as fields and crop recordings. **Figure 5** provides a schematic representation of Farmmaps. The assessment of Farmmaps in relation to the nine building blocks of the Innopay framework are summarised in Table 3.

[15] http://www.nemadecide.com/english/home.html (accessed 21-12-2023)
[16] https://git.akkerweb.nl/FarmMaps/Documentatie
[17] https://farmmaps.eu/swagger/index.html
[18] https://www.farmmaps.net/nl/Over-FarmMaps/Gedragscode-Data-Delen (accessed 21-12-2023)

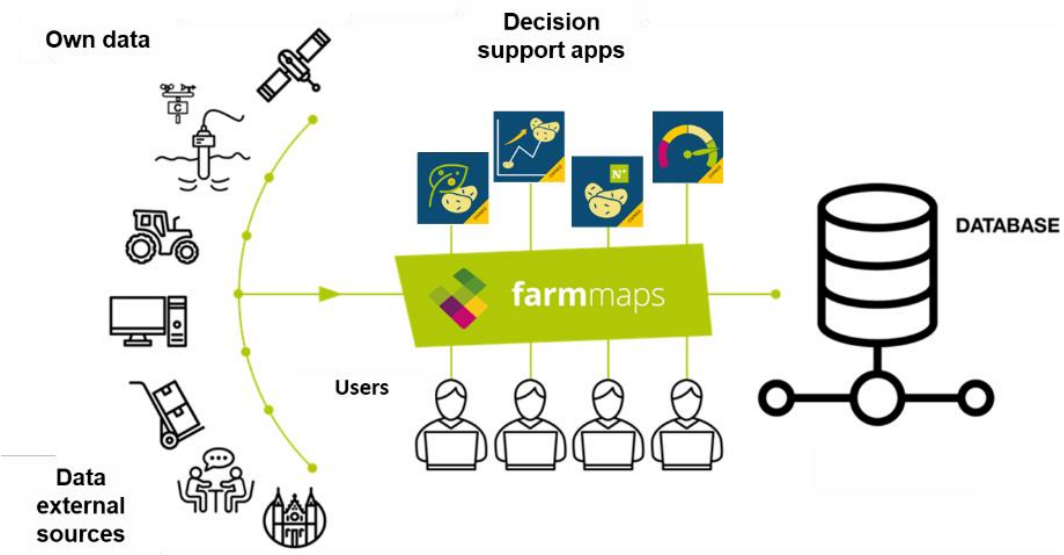

Figure 5 *Schematic representation of Farmmaps; the different data sources are depicted on the left hand side; the cloud storage of data on the right hand side, with a space for each user. Different apps are available on the platform, shown in the middle*

Table 3 *Assessment of Farmmaps according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS FARMMAPS
Data standards and formats	• (Geo)JSON – used for data provision towards other applications. • EDICrop – used for exchange with Farm Management Systems. • (Geo)TIFF – used for geodata in image files, such as satellite and sensor data. • Shape and ISOXML – used for taskmaps and as applied machine data.
Earnings model	• App owners pay a fee for being hosted on the Farmmaps platform. They can decide how they offer their services. All earnings made through the app are for the app owner. • The apps can be accessed via the Farmmaps website or via the Farmmaps API. This last route is used to provide external parties access to the WUR models and algorithms. The app owner and Farmmaps make arrangements with external parties for this access.
Metadata	• Data are stored in a relational database. Metadata is used to describe the data. Information concerning the unit and the dimension can be stored in the metadata. • Farmmaps relies on the metadata made available by these sources for external data sources. • Metadata is stored in JSON format. There are no standards.
Operational agreements	• Users create an account on Farmmaps. All data is under control of the user. Data can be shared with other users after authorisation. This authorisation can be withdrawn at any moment. • Data from third party data sources like RVO, FMIS and Eurofins are established using authorisation systems. The user authorises the third party to share data from their own account with their account on Farmmaps. The implementation differs per data source.
Connectivity	• An API is provided for external parties to access the models and algorithms running on Farmmaps.[19,20] • Commonly used standards like REST-full API's are used for data connections with third parties.
Consent	• Users maintain ownership of own data uploaded to their account. • Users can share data with other users on the platform. This can vary from single files to complete cropping schemes with all their data, and everything in between. • An authorisation system has been developed to share data for specific research questions and with specific organisations. In Farmmaps, users can indicate in their account if specific data can be shared, for which purpose and with whom. This permission can also be withdrawn at any time.
Legal agreements	• Farmmaps is exploited by the Akkerweb Foundation, under license of WPR.

[19] https://farmmaps.eu/swagger/index.html (accessed 21-12-2023)
[20] https://git.akkerweb.nl/FarmMaps (accessed 21-12-2023)

BUILDING BLOCK	KEY ASPECTS FARMMAPS
Governance	• Farmmaps complies with the Code of Conduct for data use in arable farming of the umbrella organisation for arable farming (Gedragscode Datagebruik Akkerbouw van BO Akkerbouw). • Farmmaps runs on the Irish Amazon servers and are subject to European legislation, for example concerning data privacy. The data security is following the ISO/IEC 27001:2013 security management standard.[21] • A Data Protection Officer has been appointed for Farmmaps to conduct supervision that governance and data handling comply with the aforementioned Code of Conduct.
Identification and authentication	• Farmmaps implements identification and authentication on two different levels: on a platform level and on the level of I&A for external systems for data exchange. • On the platform level, a username and password-based authentication is used to log in to the platform. • Open ID Connect is used to provide user authentication and authorisation services for access to the Farmmaps API. Open ID Connect uses the OAuth protocol. • On the level of I&A for external systems for data exchange, Farmmaps uses several acknowledged sectoral mechanisms implemented by these systems, such as eHerkenning[22] for RVO data and RESTful APIs and AgriRouter connection for data exchange with agricultural machines.

3.3 AgroDataCube – Wageningen Environmental Research

AgroDataCube (Janssen et al., 2018) is an open data repository in which a range of available Dutch open data sources related to agriculture are stored in a connected and harmonised data model. AgroDataCube's main objective of is to bring together, structure and provide open source data on agriculture for data science and analytics in a harmonised way. All data is implicitly connected to agricultural parcels to provide the necessary resolution and to take decisions at a farm and parcel level and on aggregated scales. Its parcel-centred setup makes it a valuable data source, specifically for field-based and farm-based applications and data analytics. It prevents data scientists and application providers from having to collect, organise and curate the data themselves from the many dispersed, heterogeneous and disconnected data sets. Access to the data repository is provided through a REST API.

AgroDataCube provides spatiotemporal data on agricultural parcels derived from the Dutch RVO parcel registration, crops, crop history of individual parcels (IACS), soil types and properties (Dutch soil registry), observed weather (KNMI), vegetation indexes (NDVI, WDVI) derived from open satellite data (Copernicus/Sentinel), farm management practices derived from open satellite data (mowing, ploughing, sowing, harvesting) and altitude (Dutch altitude map). AgroDataCube is a good example of an element of the WUR data sharing infrastructure that can be relatively easily reused for research purposes because of its open nature. In the past years, AgroDataCube has become an important provider of harmonised open data for agriculture. It is frequently used in WUR research as well as in broader national and EU research and in policy and commercial domains, such as by water boards, agricultural contractors and farm management systems.

Nevertheless, improvements on additional operational and legal agreements beyond the non-commercial cases, formalised through the Creative Commons BY-NC-SA license, are required in the future. Although the used combination of governance, legal and operational agreements allows for an earnings model with some monetary returns, it remains hard to maintain and sustain an open data repository like AgroDataCube without additional funding from projects and other funds and subsidies. AgroDataCube strives to reach agreements with parties that retrieve large amounts of data from the system. The main aim is to be able to at least partially cover the costs of operations, maintenance and infrastructure improvements. The assessment of AgroDataCube in relation to the nine building blocks of the Innopay framework are summarised in Table 4.

[21] https://aws.amazon.com/compliance/iso-27001-faqs/ (accessed 21-12-2023)
[22] https://www.digidentity.eu/nl/services/digital-identity/eherkenning (accessed 21-12-2023)

Table 4 *Assessment of AgroDataCube according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS AGRODATACUBE
Data standards and formats	• Implemented standards are primarily aimed at software developers and developing machine to machine interfaces. • (Geo)JSON – primarily aimed at developers and machine to machine interfaces. • WKT – used to define geometries. AgroDataCube does not use a standardised data model or semantics, but 'reuses' code lists and taxonomies from the open data sources that are integrated, such soil characteristics from the Dutch soil inventory and crop codes from RVO.
Earnings model	• Monetary: • Using data from AgroDataCube is free. However, large data users and commercial users are asked for a monetary contribution, following the Creative Commons BY-NC-SA non-commercial use license. • Additional earnings are established through funding from projects that either contribute to the maintenance costs or add to the data. • Earnings are used to support the maintenance and further development of AgroDataCube. Non-monetary: • Delivers societal benefits because structuring scattered and heterogeneous open data prevents a lot of parties from having to do the same data processing repeatedly. • Free for research purposes so adds to WUR research and indirectly provides scientific and social benefits.
Metadata	• Metadata is provided as an integrated part of the API documentation. However, there are no standards or formalised semantics.
Operational agreements	• There are well-defined operational procedures and agreements to ingest, maintain, update and curate data to ensure the consistency, timeliness and validity of the data. • Standard operational agreements with data users are defined and described. • The setup of specific operational agreements for large data consumers is still in an early phase where specific, dedicated and mostly project-based agreements are made between parties.
Connectivity	• Data is provided through a REST API, providing responses in (Geo)JSON format.
Consent	• Consent has not been an issue so far as it concerns open data, and no data in the AgroDataCube can be linked to individuals or organisations.
Legal agreements	• Open license (Creative Commons BY-NC-SA, only non-commercial use) determines specific legal aspects with regard to access and reuse. • In practice, we see that mixing/exchanging data from different sources might require additional agreements.
Governance	• Data content is partly curated and maintained by external parties and partly by WENR as part of operational and maintenance work. • There is a structured, documented process of data harmonisation and quality control, with continuous, managed updates for the more volatile data (for example, satellite data) and a schedule of planned deployments for less frequently changing data (e.g. agricultural parcels).
Identification and authentication	• Users need to request a token linked to an email address to get access to the AgroDataCube API for downloads. There is no additional user identification and/or authentication.

3.4 Farm Data Safe – Wageningen Environmental Research

Farm Data Safe is an implementation for federated storage and data management of farm data. It focuses on data sovereignty for farmers, which means that farmers keep ownership and control over their data and access of this data for reuse. To make this possible, Farm Data Safe acts as a controlled and optionally decentralised data infrastructure for farm data. It can ingest data from other farmer-owned sources through the integration with existing data access and consent services, such as JoinData for dairy farming and eHerkenning for governmental data. It also offers its own consent mechanism that can be configured to support fine-grained data control access for external parties that want to reuse farm data. This setup allows for the preservation of privacy and data rights while offering farmer-controlled authorisation for data sharing and reuse with the outside world.

As such, Farm Data Safe serves as a framework and demonstrator for how privacy conserving infrastructures could be further upscaled to the entire Dutch agriculture sector, as well as towards the European perspective of Data Spaces to support the European data economy. Through its nature, Farm Data Safe is also becoming a 'hub' for data rich processes, particularly in research, and a testbed for connecting various data streams in the agri-food sector. Currently, its links with internal data infrastructures (WECR data warehouse,

AgroDataCube, Farmmaps) and external sectoral data infrastructures (Kringloopwijzer, Farm management systems DACOM and Cropvision, Milieumeetlat, SCANGIS etc.) are being operationalised. The assessment of Farm Data Safe in relation to the nine building blocks of the Innopay framework are summarised in Table 5.

Table 5 *Assessment of the Farm Data Safe according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS FARM DATA SAFE
Data standards and formats	• Uses (Geo)JSON for data provision towards other applications. • Common sectoral standards are used for data ingestion. Examples are the various XML-schemas implemented by RVO (BRP, I&R, manure transport), ZuivelNL (Kringloopwijzer), EDICrop for exchange with Farm Management Systems and the (Geo)JSON structures used by AgroDataCube. • An Odata interface is implemented to ingest data from internal data sources. It is currently used to ingest data from the WECR data warehouse, see Chapter 3.1.
Earnings model	• Farm Data Safe is still in a development stage and is currently mainly funded through research programmes and applied projects that require farm data.
Metadata	• Farm Data Safe does not yet publish metadata other than the descriptions of the published data elements. More detailed metadata will be developed in the near future. • Regarding data ingested from other systems, Farm Data Safe relies on the metadata made available by data providers.
Operational agreements	• As Farm Data Safe is a farmer-centred asset, operational agreements are established between the farmer and the data provider on the one side and data users on the other side. • Main data ingestion sources for Farm Data Safe are systems that have implemented their own operational agreements with regard to data sharing, such as JoinData and Farm management systems. It is often also linked to a consent mechanism. Farmers can give consent to open these data sources for ingestion in Farm Data Safe. • Data users can apply for a Farm Data Safe registration and will become potential data sharing partners for the farmer. At all times, the farmer can decide which of these parties they want to share what data and for what period. This is done through the implemented consent mechanism.
Connectivity	• An API is provided for external data users to access the data from farm data stores, as far as consent has been provided by the farmer. • Farm Data Safe uses sectoral data exchange interfaces for data ingestion. These are commonly implemented as machine to machine interfaces, often XML-based.
Consent	• Farm Data Safe's main aim is to be a safe data store for farmers. Consent for sharing data with external applications and initiatives is granted through its consent mechanism. Users can decide on a level of high detail (e.g. data element, farm and parcel level, year), what data they want to share with which registered applications and for how long. • Ingestion of data is always done with the farmer's consent, either through the consent mechanisms of the data provider or through direct data sharing agreements with the farmer.
Legal agreements	• In the pre-operational phase, where Farm Data Safe is still mostly used in research environments, specific legal agreements (*data gebruiksovereenkomst*) are established between users and the research initiative. This includes arrangements on data privacy and rights, data sharing, consent, data protection and security measures.
Governance	Users can choose to deploy Farm Data Safe in the cloud or on their own premises, and governance is partly dependent on that choice: • In the first case, part of the governance lies with the party maintaining the cloud infrastructure. This is currently deployed on Kubernetes, a containerised virtual machine, at the WUR IT infrastructure, and technically managed by WUR FB-IT and administratively managed by WENR. Apart from this overall system management, much of the operational data management lies with its users, implicit to the Farm Data Safe philosophy. • In the second case, the infrastructure and data are independently deployed and managed by users of the Farm Data Safe and possibly partly by an independent ICT service provider. In this case, users are themselves responsible for infrastructure governance issues.
Identification and authentication	Farm Data Safe implements identification and authentication on two different levels: • On the application level, a username/password based authentication is used to log in to the application. • On the level of I&A with external systems for data exchange, FDS makes use of several acknowledged sectoral mechanisms implemented by these systems, such as eHerkenning for RVO data and JoinData for dairy sector data.

3.5 Methane Data Lake – Wageningen Livestock Research

In contrast to the other data sources described in this project, Methane Data Lake (MDL) is a new data initiative developed to serve as an experimental environment to explore, apply and develop new technologies to improve the knowledge base of WLR researchers. MDL was set up as a modern data warehouse, and its explicit purpose is to explore and experiment with different proof-of-concepts, such as cloud technologies, as well as to valorise these concepts by implementing them in existing data sources at WLR and/or in new projects. With this specific goal in mind, MDL is a data source with a different objective from the other data sources explored in this project.

MDL was set up in 2019 as part of the KB DDHT programme to experiment with cloud technologies and the real-time collecting, storing, analysing and visualising of farm data. More specifically, it dealt with collecting individual cow methane data generated by sensors installed at five commercial dairy farms across the Netherlands. Furthermore, the goal was to enrich this information with other farm-related data and to provide feedback to farmers through analytical and visualisation tools available in the cloud. MDL was developed together with FB-IT and the knowledge gained was valorised in different ways: it was implemented in a new PPP (Smart Cattle Breeding) that now collects real-time sniffer data from a hundred Dutch dairy farms. The enrichment with other farm data was explored by connecting to Farmmaps for weather information, which is used in the same PPP, and to the Netwerk Praktijkbedrijven[23] to visualise farm data in real-time. Explored methods to monitor data quality and cloud technologies were also implemented in new projects, including Next Level Animal Sciences[24] and Breed4Food.[25]

During the course of the project, MDL kept its original objective as an experimental tool. For this report, we experimented with connecting MDL and AgroDataCube for climatic data and soil data to improve the value of the connection between these two data sources. This also highlighted the necessity for MDL to connect to MyJohnDeere to get access to the polygons that were needed to retrieve soil information from AgroDataCube. Because MDL was set up as an experimental tool, some building blocks had higher priority than others. The assessment of MDL as an experimental tool in relation to the nine building blocks of the Innopay framework is summarised in Table 6.

Table 6 *Assessment of the Methane Data Lake according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS OF METHANE DATA LAKE
Data standards and formats	• Azure conventions and standards described by Microsoft have been used to build MDL during the data engineering certification. Microsoft Data Factory and Microsoft storage were used. • MDL uses JSON or parquet and is primarily aimed at developers and researchers. • A REST-API integration was developed for the integration of MyJohnDeere. This integration works with the OAuth protocol. • Data loaded into MDL is not based on a common harmonised data model and does not use any commonly used semantic standards.
Earnings model	• Since MDL was developed as experimental tool to research proof-of-concepts with new cloud technologies to increase the knowledge base of researchers at WLR, there was no direct need for a monetary earnings model. • However, the MDL did include an indirect earnings model and contributed significantly to innovations and improved data infrastructures within WLR. Implementing proof-of-principles studied with MDL increased data collection efficiency and quality, and knowledge and components were reused in new projects (e.g. PPP Climate Smart Breeding, Breed4Food and Next Level Animal Sciences).
Metadata	• MDL is an experimental tool and does not generate data itself. Instead, it ingests data from different projects where methane data is available. For data that are ingested from other systems, MDL relies on the metadata made available by that data provider. This metadata is therefore available to, but not published by MDL.

[23] https://www.netwerkpraktijkbedrijven.nl/
[24] https://www.wur.nl/en/value-creation-cooperation/next-level-animal-sciences.htm
[25] https://www.breed4food.com/

BUILDING BLOCK	KEY ASPECTS OF METHANE DATA LAKE
Operational agreements	• During development, informal operational agreements were made with AgroDataCube. This informal agreement allowed a quick setup. • With respect to collecting sniffer data from commercial farms, service level agreements were set up with IT. These agreements involved the connections with commercial farms, as well as the service of these connections. This agreement is also used within the PPP Climate Smart Breeding. • Data loaded into MDL is hosted on Azure and standard operational agreements were made. IT made agreements concerning data control, rights and ownership with Azure.
Connectivity	• Data is offered through a standardised interface (SQLdb, AzureSDK, DataBricks) and data formats. • Connections are set up with well-documented API from AgroDataCube. The API for connections between the data lake and researchers is provided through Azure SDK. Access to the data lake is provided by means of SAStokens (shared access key).
Consent	• MDL is an experimental tool and for each of the data sources that were used to load data into it, we were transparent to the stakeholders about the goal of the project. Specific consent included an agreement per email with AgroDataCube to use data within the KB project only and an agreement with project leaders working with methane data to use that data in this KB DDHT programme. The project leaders themselves had their own agreements with the farmers where sniffers were installed. • Consent from the contractor was required to access our own data when integrating the MyJohnDeere data (data from the parcels of Dairy Campus, produced by the DairyCampus contractor).
Legal agreements	• MDL is an experimental tool so legal agreements are not considered at this point.
Governance	• Aspects related to governance were developed to the extent that was suitable for an experimental environment like MDL to operate. No attention was paid to data quality from data providers that were loaded into MDL. During development, informal and minimal agreements were set with AgroDataCube, for example. This was done to quickly develop and work on these proof-of-principles. • Service level agreements with FB-IT are in use for pipelines that are now operational to collect on-farm sniffer data in the PPP. Access to this data by researchers is limited and organised through an Azure account.
Identification and authentication	• Users need to request an Azure account to query the data. Identification and authentication is needed to identify if users are authorised to access this interface. • The OAuth protocol is used to connect with MyJohnDeere. • MyJohnDeere account credentials are required to access data.

3.6 Federated Food Fraud Data – Wageningen Food Safety Research

A federated data infrastructure was developed for the integration of food fraud data from the EU 'Rapid Alert System for Food and Feed' (RASFF) database and the US 'Economic Motivation Adulteration' (EMA)' database. Its main objective is to set up an infrastructure that allows analytics over distributed databases while respecting the strict privacy and security regulations that exist for food fraud data. Personal Health Train and Farm Data Train[26] concepts were used to implement an infrastructure that supports federated learning while preserving all personal and otherwise sensitive data at its original location.

The open source platform Vantage6, which offers many functions required for federated learning, was used to implement the federated infrastructure. Three distributed data sources were connected. A dedicated data station was established for each database and hosted in three different locations: Wageningen, Maastricht and Leiden. The Wageningen data station provides RASFF data from 2008 to 2013, the Maastricht data station provides RASFF data from 2014 to 2018 and the Leiden data station provides EMA data from 2008 to 2017. The assessment of Federated Food Fraud Data in relation to the nine building blocks of the Innopay framework are summarised in Table 7.

[26] https://www.dtls.nl/fair-data/fair-projects-initiatives/

Table 7 *Assessment of Federated Food Fraud Data according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS FOOD FRAUD DATA
Data standards and formats	• The RASFF food fraud data were in CSV-format and in RDF-format. These data were distributed over three databases. Deciphering of data based on ontologies can be carried out using modern data standards like FHIR or OMOP. • Data is offered through a standardised interface and data formats. However, data is not based on a common harmonised data model and does not use any commonly used semantic standards. • The FOODON and AGROVOC ontologies were used to make the data and the model semantically interoperable. These ontologies play a key role in the food safety domain.
Earnings model	• This aspect was not covered in the current setting, as it concerned a research environment.
Metadata	• To realise interoperability between the data sets, ontologies and unique notification identifiers (including the metadata description) were obtained for each data station. The ontologies, identifiers and metadata are available in the project's git repository. • The metadata for each of the data stations data sets were: type of product fraud, category product, year, origin country and report country. The metadata was used by the Bayesian Network model data train.
Operational agreements	• The open source platform Vantage6 was used. This is a privacy preserving federated learning infrastructure for secure insight exchange. It contains all aspects relevant for federated learning, for example policy management, data sharing and data security (with end-to-end encryption). This infrastructure is hosted at WUR.
Connectivity	• First, the model was trained on the three local data station using a BN model. Then, the trained model parameters were transferred to the central server. Afterwards, the parameters for each data station were retrieved via the central server to generate a combined BN model that contains the information of each data station. • The model scripts were published in the form of docker images (for containerised virtual machines) and were used on the data stations. All input parameters were supplied to these docker images. In this way, the model runs on each data station at the data owner's location. The data stations' computational nodes return the results after successful execution of the algorithm. These results are sent to the central server, from where model users can retrieve the results.
Consent	• The permissions and policies around the data use were either fully or semi-automated, respecting the security and privacy arrangements associated with the data.
Legal agreements	• The data owners own the data on their data stations.
Governance	• A so-called 'collaboration network' was established when the three involved organisations agreed to collaborate. The collaboration server was then set up at WUR. WFSR was obliged to comply with the security and privacy of the data (GDPR) regulation. Therefore, the WUR server had a build-in database that stored information on the collaboration for data sharing and privacy policies, defining each collaborator's access to a data station. A system administrator then made the individual organisations part of this collaboration at the central server.
Identification and authentication	• Vantage6 uses token-based authentication and authorisation for each data station to join the collaboration. This central infrastructure component was hosted at the WUR premise.

3.7 Freshwater Fish – Wageningen Marine Research

FRISBE (Fisheries Research Information System Biology & Ecology) is the Wageningen Marine Research (WMR) database that contains all observations about fish-related research projects. FRISBE is an Oracle database in which WMR and its predecessors IMARES and RIVO have been recording fish survey data from 1970.

FRISBE was developed to store biological data collected per station. A station is defined as a location where data is acquired. The station can be placed anywhere. One or more samples can be taken at a station, which means that different types of equipment, also called platforms, can be combined at one station. Biological information such as counts, length measurements and weights can be added for the entire sample or part of it (**Figure 6**).

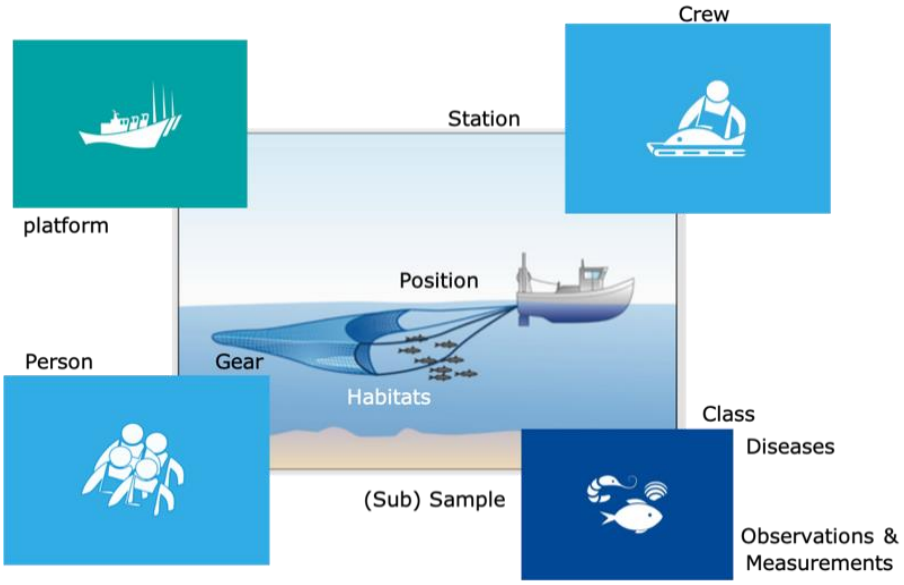

Figure 6 *Schematic overview of data stored in FRISBE*

Aggregated freshwater fish data is supplied by WMR in the Open Data Freshwater Portal. SAS (Statistical Analysis System) code was developed to make data from FRISBE available as open data. SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and also perform statistical analysis on them.

This portal database was set up according to an earlier version of the Aquo Standard (Aquo), which is the Dutch standard for exchanging data within the water sector. Aquo is a semantic standard, which means that it defines the meaning of concepts and data and their mutual relationships. Aquo makes it possible to exchange data between parties involved in water management in a uniform manner and thus contribute to improving the quality of water management.

More recently, a web service was added in the Freshwater Portal, making the data visible by importing the data automatically to a PostgreSQL database. The Digital Delta Ecological standards (DD-ECO API[27]) are employed to connect the RWS transformation to the Aquo standards. The assessment of Fresh Water Fish in relation to the nine building blocks of the Innopay framework are summarised in Table 8.

Table 8 *Assessment of Freshwater Fish according to the nine building blocks*

BUILDING BLOCK	KEY ASPECTS FRESHWATER FISH
Data standards and formats	• The DD-ECO API is used for freshwater fish. The API specification defines a standardised method for providing and retrieving ecological measurements. It is a specialisation of the regular Digital Delta API specification focused on ecological data (biological taxa, chemicals, physical objects) and observations instead of automated measurements.[28] The API can be explored with the DDExplorer.[29] • Uses Aquo and IMMetingen. The Aquo standard is the Dutch standard for exchanging data within the water sector. Aquo is managed by the Water Information House, a cooperation programme of the provinces, water boards and Rijkswaterstaat. IMMetingen is part of Aquo and is used for the exchange of chemical, physical and biological measurement data. The DD-ECO API is built upon the biological measurements of Aquo. • Uses GeoServer, an open source marine server for sharing geospatial data. It is designed for interoperability and publishes data from any major spatial data source using open standards. • GeoServer implements industry standard OGC WMR protocols by, such as Web Feature Service (WFS), Web Map Service (WMS) and Web Coverage Service (WCS). WFS is a streaming and download service and provides multiple exchange formats like xml/gml, geopackage, GeoJSON, KML and GeoTIFF.

[27] https://www.ecosys.nl/digitale-delta (accessed 21-12-2023)
[28] https://github.com/DigitaleDeltaOrg/dd-eco-api-specs/wiki (accessed 21-12-2023)
[29] https://ddexplorer.s3-eu-west-1.amazonaws.com/index.html (accessed 21-12-2023)

BUILDING BLOCK	KEY ASPECTS FRESHWATER FISH
Earnings model	• Compliant with open data policies. The data on fish are publicly available, for some users in an aggregated form. Benefits in its use and reuse are expected to be mostly societal.
Metadata	• No metadata is available in the Nationaal Georegister (NGR). • The Dublin Core (DC) metadata standard is used for metadata (for part of the data) at SeaDataNet.[30]
Operational agreements	• An agreement was made between FB-IT and the IT department of the WUR to provide a Virtual Private Server (VPS) on the Amazon cloud to host the Freshwater Fish Data Wageningen Marine Research[31] and the WMR GeoServer.[32] • The service is based on a Service Level Agreement (SLA) for the maintenance and management of data concerning freshwater fish. Most of the data are collected and maintained based on LNV assignments that are based on legal and supporting tasks (WOT) of LNV. • There is a WMR disclaimer related to the license (see legal agreement below) to guarantee liability. *Disclaimer Wageningen Marine Research ('WMR') institute within Stichting Wageningen Research, is the owner of any and all intellectual property rights in the scrips and datafiles/databases are made available in the WoZEP-repository as part of the KEC4-project. They are released to the public domain under a GNU-GPL-v3-license for scripts and a CC-BY-SA-v4-license for datafiles/databases. Please provide proper references to the source(s) when using the scripts and/or datafiles/databases. The scripts and datafiles/databases made available by WMR via the WoZEP repository ('software') are provided 'as is', without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings in the software. Wageningen Marine Research - January 2022.*
Connectivity	• See the DD-ECO APIs, OGC APIs as described under the section Data standards and formats.
Consent	• Not available because almost all data are open source.
Legal agreements	• Data: shared Creative Commons attributes: CC-BY-SA v4 • Scripts: GNU-GPL v3 • WOT: provides annual data updates (Freshwater fish and Marine data)
Governance	• An open Geodata server has been established at the request of Informatiehuis Marien and there is regular contact about the data and activities related to it. • The responsibilities, activities and costs are arranged with RWS so that this data remains persistently available for them. As such, it is contributing to WMRs expenses and can be considered part of an earnings model. • There are input checks for each data source before entering data so the data, such as names of ships, fish species, code or domain lists are correct. Value margins are specified by percentile. • For other sources, the ex-post checks are executed after entering samples and the digital data. • There are management and control roles in the data entry and disclosure phases. • This includes personal information on who performs certain acts and when. • Scripts are managed in GIT. • Ethics and responsibilities are according to WUR agreements.
Identification and authentication	• The DD-ECO API is only available to authorised users. • Databases themselves are well secured via their own DBMS.

[30] https://www.seadatanet.org/ (accessed 21-12-2023)
[31] https://wmropendata.wur.nl/prod/zoetwatervis/ (accessed 21-12-2023)
[32] https://opengeodata.wmr.wur.nl/geoserver/web/ (accessed 21-12-2023)

4 Use cases to share data between data sources within WUR and the lessons learned

This chapter describes the different use cases that were explored in this project to assess if and how data could be shared and reused within WUR. Table 9 provides an overview of the different use cases explored. Note that most of these use cases involve connections between two data sources described in Chapter 3, but some also explore connections with data sources that contain sensitive information outside WUR. In addition, please note that other connections between data sources, both inside and outside WUR, were established in this project, such as MDL with Farmmaps. The lessons learned were documented in yearly reports and provided to this KB DDHT Topic 4 project.

Table 9 *Overview of use cases explored for this report*

#	Use cases for exchanging and reusing data
1	Connecting BIN to Farmmaps
2	Connecting Farmmaps to BIN
3	Connecting AgroDataCube to Farmmaps
4	Connecting AgroDataCube to the MDL
5	Farm Data Safe piloting
6	Connecting Federated Food Fraud Data to external organisations
7	Connecting Freshwater Fish data in FRISBE to the RWS distribution layer

4.1 Connecting BIN to Farmmaps

4.1.1 Purpose

A connection has been made between BIN and Farmmaps to facilitate yield data transfer and exchange. The purpose of this connection is to offer individual farmers benchmark information by comparing a farmer's crop production performance with that of local comparable farms, for example farms that produce the same crop or under the same conditions. Other publicly available sources of farm yields are BIN's Agrimatie[33] and CBS' WUR and Statline.[34] Agrimatie is a WECR outlet where the results of research activities are integrated to provide an overview of agricultural data. Statline is an open data platform where data gathered by the CBS is published. Both data sources have data on agricultural practices. However, they lack details like data on specific soil conditions, detailed information per region and the different yield percentiles. Those details can be extracted by analysing the raw data from BIN.

4.1.2 Description

Due to privacy and sensitivity constraints, as well as the data protection agreement made with 1,500 farm members, data from BIN can only be exported to Farmmaps in a way that cannot be linked to individual farms. To achieve this aggregated, non-identifiable data, a so-called virtual microlab has been linked to BIN. Data analysis with data of individual farms can be executed within this microlab, and the aggregated results of the analysis can be exported afterwards. It is prohibited to extract or use these raw data outside of the microlab. The applicant must sign a contract before they are granted access to a microlab.

[33] https://agrimatie.nl/ (accessed 21-12-2023)
[34] https://opendata.cbs.nl/statline/#/CBS/en/ (accessed 21-12-2023)

4.1.3 Findings and lessons learned

The availability of crop data depends on the crop type and the year. For example, the number of seed potato growers may vary between 80 to 100 farms within the BIN samples each year. Potato seed yields are known for these farms, as well as the region and the soil types where these farms are located. These yields, the calculated yield percentiles and other information can be relevant for benchmark comparisons. However, for other crops like leeks, the number of farms within the BIN samples may be much lower. This imposes challenges to safeguarding privacy and also makes it difficult to retrieve representative values for benchmarking. 10 farms seems to be the minimum sample size to prevent individual farms from being identified and to provide reliable data analysis for scientific or practical use. More insight is therefore only gained for certain areas, crops and years. Furthermore, data in BIN are linked to a farm, not to a field plot, as is the case with Farmmaps. So, the data of BIN is very useful for specific research questions, but for detailed benchmarking purposes, it is easier to use publicly available data from Agrimatie or Statline. The effort that is required to get access to the data and to analyse them is not worthwhile for benchmarking purposes only, especially considering the limited options to zoom in on regional areas and still have enough data available for reuse.

4.2 Connecting Farmmaps to BIN

4.2.1 Purpose

Data collection of BIN requires manual work and interaction with farmers. Using a farmer's data from a FMIS could probably raise the level of efficiency. Therefore, a connection has been made between Farmmaps and BIN. The purpose of this is to reduce manual data gathering and to increase the quality of the data, providing more detailed information on the pesticides application on crops. Instead of a product total per farm, the crop recordings provide more detailed information regarding which crops and in which fields these products have been used.

4.2.2 Description

Crop recordings can be viewed in Farmmaps and used as input for advice models. After having received a grower's permission, these crop recordings can be imported or entered manually from a FMIS, in the case of the Netherlands, from AgroVision or Dacom. This prevents farmers from having to record their data twice. A cultivation report contains, among others, the type and amount of pesticides being used by the arable farmer.

The crop recordings contain data at field level, which provide more detailed insight into the usage of different products on a farm level, crop level and field level, which helps the analysis and validation in BIN. After consent was given, crop recordings from 2020 and 2021 from a selected group of farmers were exported from Farmmaps to BIN. Farmmaps allows data to be shared with other users. However, there was not a system in place yet for sharing data for specific research questions. So, the data was exported manually and sent to BIN. Written permission for sharing these data has been given by the involved farmers.

Crop recording data is transported from Farmmaps to BIN using the EDI-CROP[35] and XML-standard.[36] This standard is set by the Dutch association AgroConnect[37] and is suitable for this type of data exchange. The EDI-CROP messages contain data on a field level, which provides more detailed information for BIN.

[35] Electronic Data Interchange-CROP (EDI-CROP) between companies with regard to cultivation (of crops) https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwi0muyevaCDAxVhhv0HHZSoB9sQFnoECBEQAQ&url=https%3A%2F%2Fwww.rvo.nl%2Fsites%2Fdefault%2Ffiles%2F2023-09%2FBerichtenboek%2520EDI-Crop%2520Bedrijfssituatie.pdf.pdf&usg=AOvVaw1gG3zUKTHCTQBl8swnkv41&opi=89978449 (accessed 21-12-2023). (EDI see: https://en.wikipedia.org/wiki/Electronic_data_interchange, accessed 21-12-2023)

[36] eXtensible Markup Language (XML) is a standard developed by the World Wide Web Consortium (W3C) and is used for exporting data.

[37] See www.agroconnect.nl. The association has a board of nine people that is appointed by the general meeting of the members. The daily management is done by the manager of AgroConnect. And the management of the standards is provided by separate working groups of AgroConnect (accessed 21-12-2023)

In the pilot phase, data governance is based on mutual trust. The data is securely stored on the WUR infrastructure. The connection is managed under supervision of the Farmmaps project manager and the BIN software project manager. The BIN software project manager has made appointments with the overall BIN project manager to organise the pilot and invite participants, and with the software development product owner to implement and maintain the connection. The BIN software development product owner is in direct contact with their Farmmaps counterpart. If the established connection is going to become a permanent data-flow, a formal agreement has to be made with the farmers.

4.2.3 Findings and lessons learned

Based on the experiences and the Code of Conduct for data use in arable farming of the umbrella organisation for arable farming (Gedragscode Datagebruik Akkerbouw van BO Akkerbouw[38]), a consent system should be in place when sharing data for specific research questions and with specific organisations. In Farmmaps, users can indicate in their account for what purpose data may be shared and with whom. Permission can be withdrawn at any time. This permission has to be clear to BIN/WECR as well. So insight into the established permissions is a precondition for further using this kind and origin of data.

Another finding is that the detailed cultivation information can be useful for the validation of data as registered for the WOT-task of CEI.[39] It also turned out that BIN would rather elaborate on connecting directly to a FMIS, which it currently does manually by viewing online reports. The efficiency of this connection is complicated by a dependency of whether BIN farmers are willing to participate in Farmmaps, which limits the number of farmers. But this depends on the relative participation of our BIN farmers in Farmmaps, which could grow in the future.

4.3 Connecting AgroDataCube to Farmmaps

4.3.1 Purpose

The purpose of this connection is to provide data from a WUR data infrastructure component (AgroDataCube) to another component (Farmmaps) to offer added value information to farmers and researchers. This use case focused specifically on the interoperability and reusability of open data.

4.3.2 Description

Farmmaps has implemented a module that allows farmers to view data from AgroDataCube. An interface has been developed that presents data from AgroDataCube using the AgroDataCube API. For example, users can retrieve all field where a specific crop was grown in a specific year and a specific region. Figure 7 shows an example where all the crop fields for a specific region are retrieved from AgroDataCube. More detailed information can be visualised for the individual fields. This includes basic data of crops and soil, but also derived indicators from NDVI satellite data and crop history. The module can compare up to three fields at a time.

Furthermore, the information from about these field is presented when Farmmaps users have the AgroDataCube app installed on their account, as demonstrated in Figure 8.

[38] https://www.bo-akkerbouw.nl/dit-doen-wij/data-intensieve-akkerbouw (accessed 21-12-2023)
[39] CEI = Centrum Economische Informatievoorziening (centre for economic information/data provision)

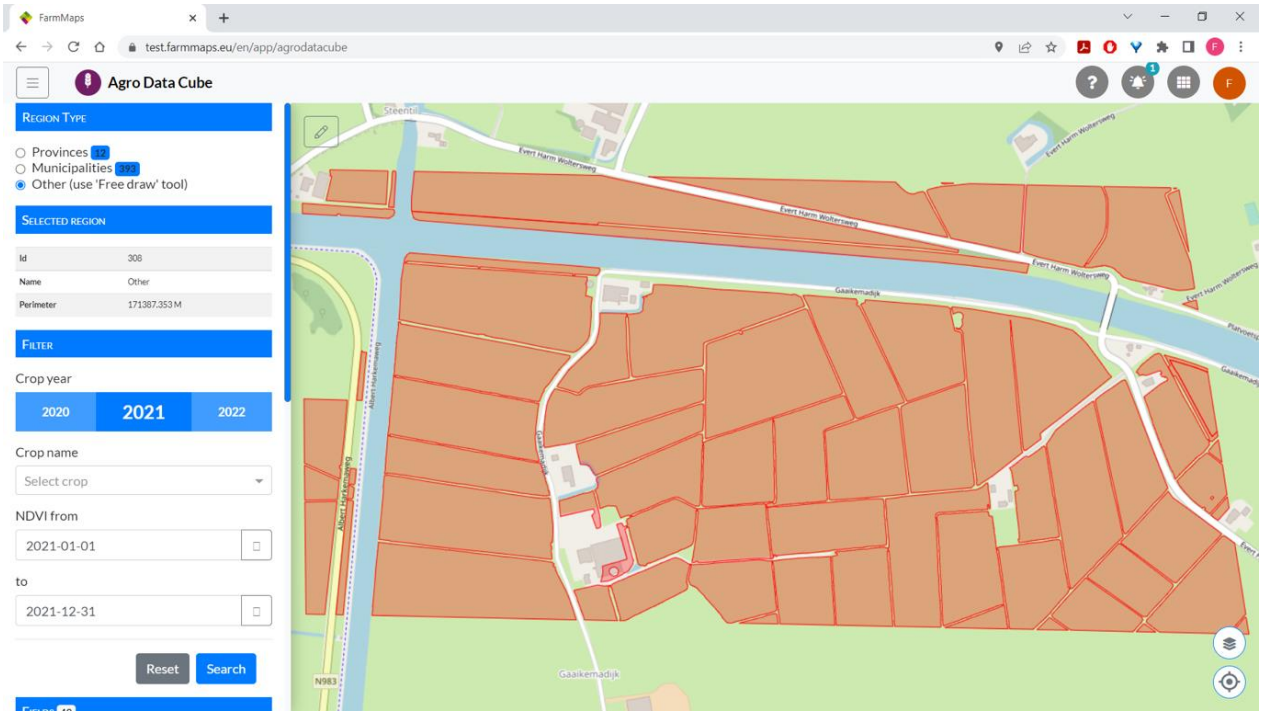

Figure 7 *AgroDataCube app on Farmmaps. All fields from 2021 are shown within an area of interest, as retrieved through the AgroDataCube API*

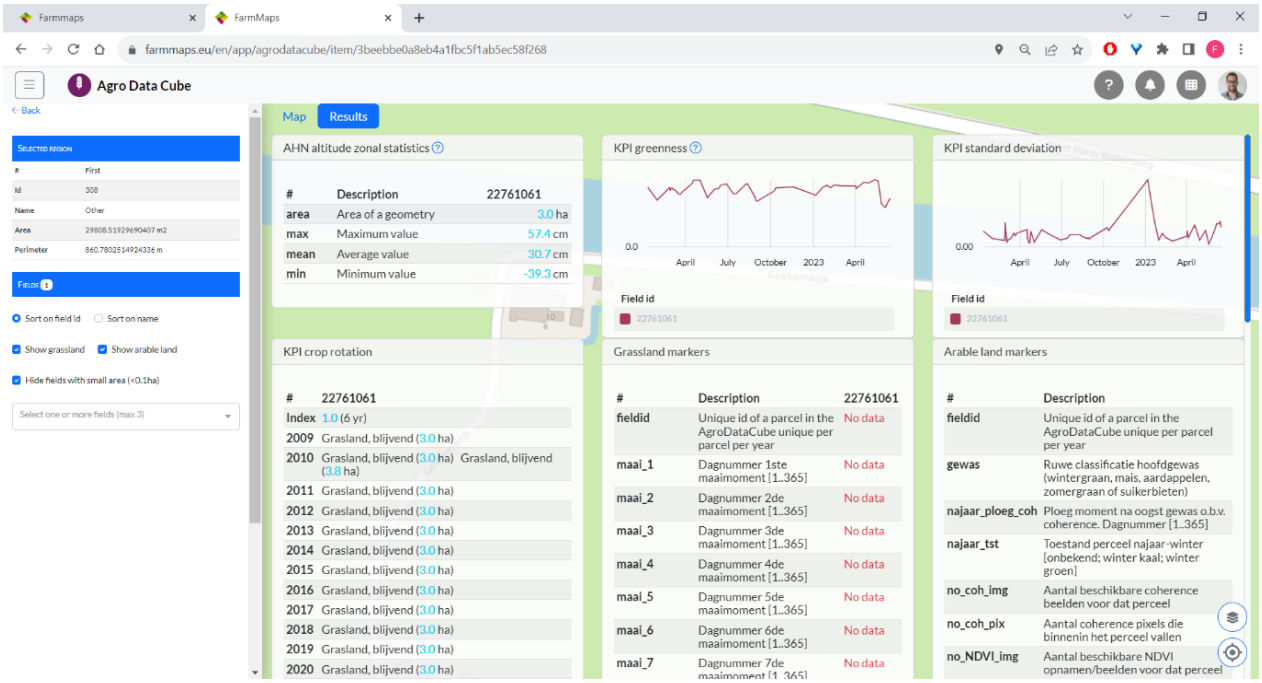

Figure 8 *Information about a user's field, retrieved through the AgroDataCube API*

4.3.3 Findings and lessons learned

This connection tests machine-to-machine data exchange between two infrastructure components, so clear agreements on broadly acknowledged technical standards are required. AgroDataCube uses well-known standards like REST and (Geo)JSON for the API implementation and the formatting and structuring of payloads, which ensures the required syntactic interoperability. Consequently, technical implementation of the connection was relatively simple. Semantic interoperability to ensure that machines can 'understand' the data without human interference, such as by linking data to standardised ontologies, is not yet implemented. Besides, there is a demand for improved documentation on AgroDataCube so its contents and provenance

can be more easily understood by users. This concerns both general metadata according to a standardised schema, and specifics on the semantics and provenance of specific attributes of the AgroDataCube data sets.

Data re-users should be able to learn and understand the meaning of the provided data through its metadata. AgroDataCube provides its metadata primarily through the API documentation provided through Postman.[40] This metadata is actually a mixture of technical documentation and metadata, explaining structure, syntax and semantics. Nevertheless, when developing this data connection, it appeared necessary to provide additional clarifications on the background of the delivered data. This was primarily to provide the connection developers with more detailed insights into the provenance of the data items, to understand the source data and analytics used and to be able determine the value and usability of the data for specific applications.

Through this connection, fields in Farmmaps are linked with those in AgroDataCube. AgroDataCube is based on the BRP (Basisregistratie Gewaspercelen)[41] of the Dutch government and has fixed parcel geometries per year. In practice, the shape and size of fields can change over years and can be imported to Farmmaps fields from other sources, or drawn by users themselves. This means that there can be a mismatch between field boundaries, but more importantly, the Field ID used in AgroDataCube is different from the field ID used in Farmmaps. Fields are being matched based on their geometry, so users can use their existing fields to retrieve information from AgroDataCube. This is a known issue with regard to the differences between the 'agronomic truth' and the 'regulatory truth' that requires the implementation of well-thought over and well-understood operational procedures and the provision of provenance related metadata to the end user.

The API provided by AgroDataCube is aimed at serving structured data on an agricultural parcel level. This works well for the range of applications on that specific scale, such as serving individual farmers in Farmmaps. We have, however, also encountered specific data requests that do not fit well with the API call patterns, such as the request to return all parcels that have a specific crop rotation, for example consecutive years of root crops. For more open queries, which are especially relevant for data science applications, a more flexible way of accessing the data in AgroDataCuve would be useful.

The conditions for access and reuse of AgroDataCube are explained in its documentation and the accompanying license. Nevertheless, additional consultations were needed to clarify and understand what this meant for the connection to and data reuse in Farmmaps in an end-user application. This is partly due to the mix of organisational, operational and technical measures implemented around AgroDataCube to secure the envisaged reuse and foreseen earnings model to cover the costs of operation, maintenance and data curation. Generally speaking, AgroDataCube want to provide this open data free of charge for non-commercial purposes and we require data re-users to share derived products under the same conditions. This is formalised through a Creative Commons BY-NC-SA[42] license. Additionally, an exception is made for massive non-commercial data access that causes excessive use of limited system resources. Therefore, a data access limit is also defined on AgroDataCube accounts. In the case of the Farmmaps connection, where many small users are currently served through one account, we have decided to remove the data limit and make additional operational agreements about data reuse. Modules based on AgroDataCube data should be provided to farmers without charge, and additional agreements are to be made for any large data users where relevant.

4.4 Connecting AgroDataCube to Methane Data Lake

4.4.1 Purpose

At first, this connection focused on retrieving weather data from AgroDataCube. In an earlier proof-of-concept, we connected AgroDataCube with Farmmaps for weather data, but we also wanted to study: (1) the technical connection as such, and (2) whether the connection with AgroDataCube would be more efficient dan collecting data from KNMI directly. At a later stage, we wanted to strengthen the connection with

[40] https://documenter.getpostman.com/view/3284162/TVeqd7aa (accessed 21-12-2023)
[41] https://data.overheid.nl/dataset/10674-basisregistratie-gewaspercelen--brp- (accessed 21-12-2023)
[42] Attribution-NonCommercial-ShareAlike see https://creativecommons.org/

AgroDataCube through a connection with field level information from John Deere to explore the possibilities and value of using GIS locations. Using GIS location would allow the extraction of field-specific information, including crop history, crop rotation index, vegetation and soil information from AgroDataCube, which may be relevant information for improving crop management.

4.4.2 Description

AgroDataCube delivers open weather data from 50 KNMI weather stations throughout the Netherlands. This data can be requested by external users for further reuse through the AgroDataCube API. MDL has used the AgroDataCube API to access and reuse weather data relevant to research locations, in this case, the WLR Dairy Campus research farm. This weather data is provided via a REST API in JSON format. Endpoints were tested using Curl, which is a factory standard, after which the API connections were further developed in Jupiter Notebooks.

At a later stage, when MDL was used to explore the usefulness of retrieving data from AgroDataCube to improve crop management, information on polygons, which is a standard in geospatial systems, were required. These polygons were retrieved from several grassland plots at the Dairy Campus through a connection between the MDL and MyJohnDeere. To avoid delays in building the REST API integration, these polygons were downloaded from MyJohnDeere and were subsequently used to retrieve field-specific information from AgroDataCube.

A REST API integration has been written for the integration of MyJohnDeere. This integration works with the OAuth protocol standard.

4.4.3 Findings and lessons learned

The connection between MDL and AgroDataCube tests the machine-to-machine data exchange between two infrastructure components, so clear agreements on broadly acknowledged technical standards are required. AgroDataCube uses well-known used standards like REST and (Geo)JSON for its API implementation and the formatting and structuring of payloads. This ensures the required syntactic and schematic interoperability. Consequently, the technical implementation of the connection was relatively simple.

The KNMI data is provided by AgroDataCube without any post-processing or data quality control. Users can request weather data from one or more KNMI weather stations for a specific time period. This means that location-specific data are generated at weather station locations. These locations may not be fully representative for the farm location, and thus information from a weather station may not be fully accurate or even incorrect. Moreover, data gaps might occur within the data sets of a single weather station, in which case data from different weather stations might have to be combined to get full coverage over all climatic variables and over time. While the connection for weather data between AgroDataCube and MDL was technically implemented, these limitations have led to the decision to use a paid weather service API delivered by IBM for research within, for example, the PPP Climate Smart Breeding. In the future, interpolation of the current station data to a regular data grid might solve some of the issues encountered in the connection between MDL and the AgroDataCube.

The AgroDataCube API is effective for a lot of field-based and farm-based applications as well as to support aggregations starting from that level. Its focus is on field level. However, the consequential structuring of the API can also be a limitation. There is a clear demand for an additional, more flexible way of retrieving data, especially for data science applications. It is still a challenge how this could be implemented, how the expected increase in (peak) resource demands can be balanced without affecting performance for other users and how to the required additional resources should be financed.

A practical issue is how to link the objects of interest from AgroDataCube to MDL. AgroDataCube focuses on fields, while a lot of research experimented with in the MDL focuses on individual animals. This limits the practical advantages of connecting MDL to AgroDataCube. However, data from MyJohnDeere were used as an additional data source in MDL to strengthen this connection and to provide the parcel geometries and the associated feed quality relevant to crop management. The connection between MDL and MyJohnDeere to

retrieve GIS information of several fields and tillage, sowing and harvest taught us that the technical connection to retrieve data on a field level based on polygons worked well. But there were authorisation issues, ranging from generic logins no longer functioning and changes in the authorisation procedure implemented by John Deere that were not communicated to dairy campus to only granting access to fields that were used for experiments at the Dairy Campus. The lessons learned from this connection are reported in a separate document that was shared with WDCC.[43]

Working with the MDL and data lake technology appeared to be a challenge for researchers that were unfamiliar with the concept and used technologies. Moreover, data lakes are big data infrastructures that are especially useful for working with data too massive to store and process on local resources. To learn all the Azure Standards and conventions, an Azure data engineering certification trajectory has been followed and achieved. Since MDL is designed and exploited as an experimental tool only, which means it only has small amounts of data stored, this raises the question of whether this does not raise more barriers for broad use and if this balances the advantages it offers to experts.

Connecting MDL with AgroDataCube was not a challenge from a technical point of view. However, a connection as such is no guarantee for instant value for either parties, or even for a shared infrastructure within WUR. The connection between MDL and AgroDataCube taught us that differences in object interests limits the value of the connecting data sources.

4.5 Farm Data Safe piloting

4.5.1 Purpose

Farm Data Safe is set up as a data infrastructure that allows the safe storing of farmer data in which farmers are able to establish their own data spaces, either in the cloud or at their own premises. The main idea is that this will result in a federation of data spaces where farmers have control over their data and can consent to peer-to-peer data connections.

4.5.2 Description

Several data connections are piloted in Farm Data Safe research projects that use sensitive and sometimes personal farm data. To create and retain trust, these projects rely on safe and secure storage of data provided by farmers solely for the purpose they consented to. In 2022, Farm Data Safe was introduced in the year-long KPI-K project, which aimed to develop and test the KPIs for circular agriculture for 15 to 20 pilots with hundreds of farmers. In this project, farm data from participating farmers were linked with internal WUR and external agri-food sector data infrastructures to develop and implement KPIs and a circular agriculture infrastructure. Farm Data Safe functions as a farmer-owned data safe, allowing farmers to manage their KPIs and to control data sharing with WUR researchers, peers and the broader agri-food sector on a farm level and a parcel level.

Currently, Farm Data Safe connects to the following data sources:
- WECR data warehouse, which is used to retrieve data and KPI values for dairy farms from the Kringloopwijzer.
- AgroDataCube, which is used to retrieve open data and derive specific soil and biodiversity KPIs.
- ScanGIS, which is used to retrieve KPIs on nature and landscape.
- CLM Milieumeetlat, which is used to retrieve data to derive KPIs on crop protection.

In the near future, additional connections are foreseen to directly link to several farm management systems and related agricultural data spaces.

[43] Rikkers R., B. Klandermans, and C. Kamphuis. 2022. KB-DDHT Sharing infrastructures Topic 4 Case Description. internal document

4.5.3 Findings and lessons learned

This pilot worked on connecting both WUR internal and external data sources. Even if a generic WUR infrastructure for data sharing focuses on FAIR access to WUR research data, in many cases there will be direct or indirect links to external, possibly personal or otherwise sensitive data. Therefore, aspects like privacy conservation and data sovereignty have to be an indispensable part of setting up a WUR data infrastructure. During this pilot, it appeared that the currently-available generic WUR infrastructure does not sufficiently support such situations, especially when risks of data loss fall in the disruptive category.

Research participants' trust increases when you allow them to manage their own data and take their own decisions on what data they would like to share for research purposes. Personal data vaults that are sufficiently secured and can be controlled by the data owner are an important component in future data spaces to achieve data sovereignty and implement viable, trusted data value chains.

4.6 Connecting Federated Food Fraud Data to external organisations

4.6.1 Purpose

Food fraud data is considered sensitive, and legislation is in place to ensure that such data is sufficiently protected. Consequently, food fraud data used in research or for other purposes is not allowed to leave the physical location it is stored in. This pilot has set up connections between federated data sources to perform analytics on the available data while retaining security issues by keeping the data at its original location.

4.6.2 Description

This use case has connected food fraud data sources at three locations: WFSR Wageningen, UMC Maastricht and UMC Leiden.

To implement and run a model on these federated data sources, WFSR researchers constructed the Bayesian Network (BN) algorithm, which moves between the three locations and connects them like a data train, allowing them to run the model locally at the data station and integrate local results (see **Figure 9**). Moving algorithms around means the data itself is never exposed in the outcomes only. In this way, you protect sensitive data from being exposed to users.

The algorithm was built by a model developer using R/Python and the routing is facilitated by a central WUR server. The model gathered the parameters by learning from the data at the data station's physical location, implementing a federated learning approach. In this way, the joint BN model provided more information than the BNs trained on local data stations.

A detailed description of this use case can be found in the article by Gavai et al.[44]

[44] https://www.nature.com/articles/s41538-023-00220-3

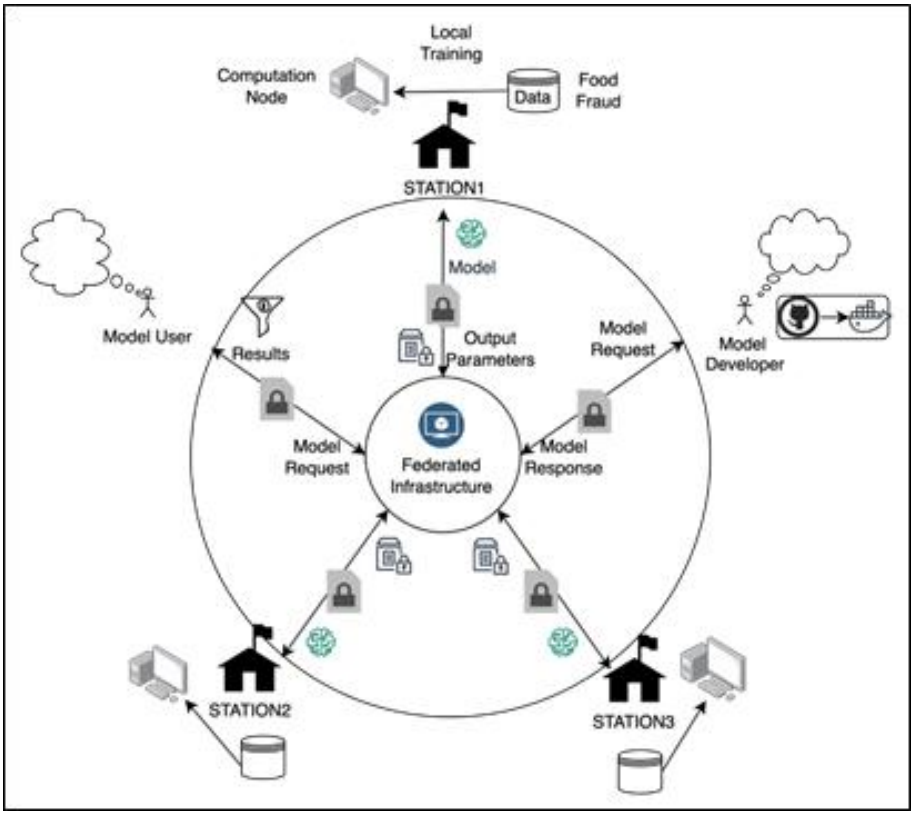

Figure 9 *Illustration of a conceptual framework that connects three data sources (station 1, 2 and 3)*

4.6.3 Lesson learned

The use of federated learning promoted secure collaboration amongst stakeholders in the food supply chain. It kept data within each owner's database and enhanced security while fostering a sense of trust between all actors involved. Furthermore, federated learning proved to be effective in addressing GDPR concerns. It ensured that data did not leave individual databases, so businesses can confidently navigate privacy regulations. Moreover, federated learning encouraged efficient resource use and reduced the costs associated with data collection, especially in critical areas like food safety monitoring. The access to stakeholder data led to streamlined processes and resource optimisation. Federated learning also contributed to enhanced trust levels among actors in the food supply chain. The secure and collaborative nature of this approach established a foundation for stronger relationships and cooperation.

Lessons from this study underscore the potential for federated learning in domains where data privacy is a concern. The findings pave the way for wider adoption of federated learning, especially in the development of data trusts within food supply chains.

4.7 Connecting Freshwater Fish data in FRISBE to the RWS distribution layer

4.7.1 Purpose

Rijkswaterstaat (RWS) wants to make WMR measurement data for freshwater fish and transitional waters digitally accessible via the RWS distribution layer. When connected, the WMR data, together with several water data sets of the Netherlands, becomes available in the RWS WaterInfo-Extra portal.

4.7.2 Description

WMR developed the necessary digital infrastructure for this data exchange under the name 'Aquatic Platform WMR-RWS'. This fits within the larger group of water managers in the Netherlands that coordinate their information flows via the Aquo standard. This is followed by the use of an exchange web service based on Aquo, which is specified in the so-called DD-ECO API. This standard has been implemented by at least one water authority in the Netherlands. The connection to the distribution layer is carried out in a number of steps. An inventory of the aggregated data that need to be included in the connection was agreed with RWS. The first step at WMR involved developing a PostgreSQL database in which the measurement data can be stored in Aquo format and developing Python software with which this data can be automatically imported from FRISBE to the PostgreSQL database. The second step involved the development of API web services based on the DD-ECO API specifications. This allows the desired data and the required metadata to be offered in accordance with the Aquo standard, and more specifically in accordance to the IMMetingen. The DD-ECO API appears to be particularly supportive of filtering, grouping and aggregating data. In the final solution, the RWS infrastructure will call web services to retrieve data from the WMR PostgreSQL database. This data will have to land in the so-called RWS distribution layer. The third step concerns the RWS' acceptance of the API web services developed at WMR. In this phase, RWS tested whether the output of the API web services corresponds with the agreed specifications. Any deviations were addressed. The final step focuses on governance and concerns the preparation of a Service Level Agreement (SLA) between RWS and WMR, and the commissioning of all developed products.

4.7.3 Lessons learned

Aquo and IMMetingen were difficult to implement because of the complex modelling of IMMetingen. The fact that one of the water authorities already implemented the DD-ECO API specifications aided this process. Since APIs collect only the data, this data needs to be presented in a user-friendly portal. Currently, RWS and Deltares are working on a user-friendly interface on their portals[45]. Data aggregation functionality should not be part of an API, but RWS wanted it to be included. Within RWS, it was difficult to connect to the distribution layer in terms of data management.

An SLA is a good instrument to specify tasks and responsibilities. Table 10 shows how this was implemented in the RWS SLA.

Table 10 *SLA for the Freshwater Fish connection of RWS to WMR*

SLA FOR THE FRESHWATER FISH APPLICATION

Application consists of a web service for the disclosure of freshwater fish data, in combination with a separate database, using an Application Program Interface (API). This API complies with the DD-ECO standard required by RWS. This allows RWS to request customised data from the database with targeted questions.
The contractor is responsible for performing the services mentioned below.
These services consists of:
• Functional application management, consisting of:
  - Configuring the API and its database.
  - Resolving malfunctions in the API and/or database.
  - Manage log files.
  - Periodic reports on incident and problem management.
  - Performance management.
• Technical application and server management, consisting of:
  - Managing server + API gateway.
  - Database hosting.
  - Monitoring (availability server, flooding storage) and signalling on the managed servers.
 	
The client is responsible for:
• Functional management, such as authorisation management, reporting and handling user questions.
• Keeping the DD-ECO API + Aquo standard up-to-date in such a way that it conforms to the use of a current version.
• Contacts and management of the contractor.

[45] waterinfo.rws.nl and waterinfo-extra.rws.nl

5 State of play

We introduced the different data sources (Chapter 3) and summarised the lessons learned in different use cases where we connected data sources (Chapter 4). In this chapter, we combine both aforementioned chapters to assess the state of play of the overarching WUR infrastructure to exchange and reuse data.

5.1 Current state

In Chapter 2, the status of the data sources in relation to the nine building blocks of the Innopay framework were described. Table 11 summarises what level of attention is given to each of the nine building blocks for each of these data sources. We distinguish between three levels: missing (red), partly included (orange) and sufficiently covered (green). Please note that 'sufficiently covered' does not mean fully implemented in all cases and further improvements may be relevant. When the initial/experimental phase column (light blue) indicates 'yes', this means this data source is set up in the context of this project or is used as experimental tool. As a consequence, these data sources may comply with fewer building blocks and may pay less attention to them compared to the other data sources.

Table 11 *Assessment of the current status of the used data sources per building block*

Data source / Building block	Initial /experimental phase	Data standards and formats	Earnings model	Metadata	Operational agreements	Connectivity	Consent	Legal agreements	Governance	Identification and authentication
Farm Information Net (BIN)										
Farmmaps										
AgroDataCube										
Farm Data Safe	yes									
Methane Data Lake (MDL)	yes									
Federated Food Fraud Data (RASFF/EMI)										
Freshwater Fish										

Table 11 demonstrates that none of the data sources we explored in this project fully meet all nine requirements. Five out of the seven data sources indicate to have partly included the Data standards and formats building block. For all these use cases, this meant that technical standards were applied but no or very limited mention was made of the use of semantic standards, controlled vocabularies or other ontologies. The Metadata building block was partly covered in five data sources and not considered at all in the case of Farm Data Safe. Earnings models were not fully considered in the cases of Farm Data Safe, MDL and the

Federated Food Fraud Data. Legal agreements were not considered in MDL, which makes sense because MDL was experimental tool set up for researchers to work with new cloud technologies.

Although none of the data sources fully implemented all building blocks, this did not cause any major issues for the use cases (Chapter 4). In the use case connecting BIN to Farmmaps, it was clear that the microlab works well for experimental research and that the data exchange agreements were fine. The same holds for the lack of semantic standard and controlled vocabularies. Setting up a fully automated data flow from BIN to Farmmaps would take work. However, since the added value of this data connection was found to be limited, this automated connection was not explored further. For the use case of connecting Farmmaps to BIN, it was found that the manually signed consent forms work sufficiently for experimental research. However, since a direct link to a farmer's FMIS was preferred, this connection from Farmmaps to BIN was not explored further. In the use case connecting AgroDataCube to Farmmaps, an application where data from AgroDataCube could be retrieved in the Farmmaps user interface was established. This connection made data exchange accessible for people who cannot use an API themselves. Furthermore, the data becomes available for all the user's Farmmaps fields when the AgroDataCube application is added onto the account. Developing this connection went well. The only challenge was that the metadata was not always in sync with the response. In practice, this meant that units were sometimes missing. There were no specific issues with the use case connecting AgroDataCube to MDL. MDL was developed as an experimental tool with the objective to increase the knowledge base of WLR researchers at new cloud based technologies. Consequently, some aspects of certain building blocks like operational agreements were set-up with minimum requirements and effort.

5.2 Summary and findings

The overview in Figure 11 provides a rough estimate on the current state of the used data sources and the explored use cases. Overall, many of the building blocks have been developed to some degree, although not all to the same extent. This indicates that there is a growing awareness of their importance. It also shows that, in general, conformity is a complex task to complete. The data sources mentioned here are mostly designed to be used in different domain-specific applications and can therefore be tailored specifically to their intended use in these applications. Overall, the relevant components of the building blocks operate well within these application environments. Specific findings are given below.

1. We found that the different technical standards are well-defined and applied in use. What is still immature, or even lacking, are semantics standards. Independent descriptions of data structures are missing in most cases. This forms a major barrier to interoperability because clear definitions and relationships between objects are not documented, recorded and published. This is even more of a problem in data and information exchange between parties that are less familiar with each other and when exchanging data across domains.
2. Having a direct monetary earnings model in place is not always the first priority in science, especially in an early phase. However, scientific data from ongoing research can be very valuable for reuse. The overview shows that earnings models are present and well-developed. This is especially valid for scientific benefits and societal benefits, although it still is difficult to monetise this in order to obtain the proper funding.
3. Metadata is often misunderstood. In general, everyone knows the term and acknowledges its importance, but very few are aware how to deal with it. Again, there are different standards to describe metadata, which usually depend on the application domain. A general metadata standards is Dublin Core Metadata Initiative (DCMI). The core comprises 15 elements to describe the resource and includes links to the used terms and the resource itself. Within WUR, PURE, developed by Elsevier, is used as a Research Information Management System, which uses the Common European Research Information Format (CERIF46) as a metadata standard. In the geographic domain, ISO19115 is used as a standard for all spatial related data.
4. With regard to maturity of the building blocks it is good to understand that, in most cases, the framework was not known at the time the data sources were compiled. In that sense, if the building blocks are present, this was often done while being unaware of the existence of this framework. At the

[46] https://rdamsc.bath.ac.uk/msc/m4

same time, there is a strong desire to make data FAIR within WUR, which means that a lot of building blocks were already taken into consideration. Currently, interoperability mainly focusses on supporting technical access. The data structures are almost always lacking standardisation, especially on the conceptual level. This is also the most difficult aspect of 'FAIRifying' data to compile proper metadata. Happily enough, we see that controlled vocabularies and ontologies are used more and more.

5. With respect to standards, in general, the technical standards are followed as much as possible. One could expect that semantic standards, such as ontologies and controlled vocabularies, and web service standards like W3C, ISO or OGC Web Services are also included. They might be used but are not referred to specifically.

6 Views of Science Group directors and researchers on sharing data

6.1 Introduction

To develop an infrastructure that allows research data to be exchanged and reused between the various Wageningen research institutes, it is important to know what both management and researchers of the various science groups think about such an overarching data sharing infrastructure. Their views and opinions are important for how such a data sharing infrastructure should be developed and how it should function. The views of a number of directors and business unit managers of Science Groups are presented in Section 6.2. Subsequently, Section 6.3 describes how researchers and data stewards view data sharing within WUR, in which Sections 6.3.1 and 6.3.2 describe the barriers to data sharing and possible solutions to overcome them.

6.2 Management views

The interviewed directors or managers of each science group are listed in Table 12.

Table 12 *Interviewees for each science group*

SG	Interviewed	Position	Date
SSG	O (Olaf) Hietbrink	Business unit manager	20 June 2022
ESG	Dr JA (Bram) de Vos	Managing Director	29 July 2022
WMR	Dr TP (Tammo) Bult	Director	5 October 2022
PSG	Dr REE (Raymond) Jongschaap	Business unit manager	18 October 2022
WFSR	Dr RFM (Robert) van Gorcom	Managing Director	1 November 2022
ASG	Dr JE (Ernst) van den Ende	Managing Director	8 November 2022

Data sharing
It is widely recognised among the interviewed higher management staff that data sharing within WUR is a must because data should be reused for multiple purposes.

It is considered important that the sharing and reuse of data is driven by societal tasks and issues. Sharing data for reuse must be fed from assignments, especially for interdisciplinary themes. In 2023, the WUR executive board will publish the WUR Digital Strategy, which is part of WUR's vision on digitisation.

For research funded by the national and European governments, data has to be made open and accessible, especially when they are collected or generated using public money.

The added value of data sharing emerges when connections are made between data. Systems biology is a good example of this. In this branch of science, for example soil data and plant data are interlinked together, creating new insights.

Re-using and combining models
In addition to data sharing, the model sharing is also considered important. An example of this is the Taskforce Integrated Area Approach to Nitrogen, in which models are linked to each other. Another example is the Digital Future Farm, part of the digital twins investment theme.

Control, Rights Ownership of research data
More clarity must be created on data control, rights and ownership. According to the directors, "researchers at WUR often think that they own the data they have collected. But the researcher who generates or collects data needs to know that it's not his or her data". Formally, WUR owns the data and researchers must have access to the data they collect. It is often indicated that it must be agreed with researchers for what purpose the data will be (re)used, how the data will be (re)used and what should not be done with the data. It is important to prevent the data from being misinterpreted or even misused, which can be achieved by adding proper metadata.

The Dutch government uses the standard ARVODI (General Government Terms and Conditions for public service contracts[47]) for assignments like WUR research. In this agreement, it is determined that the Intellectual Property Rights (IPR) is transferred to the government. During the interviews, it was noted that this is not an option because WUR must retain the IPR, as is also legally determined. At most, the right of use or even the ownership can lie with the government. In turn, the open data policy requires that data remains accessible. Furthermore, it was suggested that an additional condition could be added within ARVODI that all produced data for reuse, if relevant, should be made FAIR.

Data maintenance and (sustainable) storage
According to some directors, "there is no clarity about the aftercare of the data collected in projects. Does responsibility for this lie with WUR Corporate, or should it be arranged within each project?"

Concerning data repositories within WUR, data must be properly stored at the institute premises on a shared drive. In some cases, they also referred to the use of a WUR repository in the future or storing the data for the longer term at an external, international repository like DANS, 4TU or Zenodo.[48]

Stimulate data sharing by means of reward
Some directors stated that "it is also important that researchers, in their role of data generator or data collector, should receive credits for the reuse of their data by others." It is common to look at how often articles have been cited by other researchers when assessing a researcher's value. One director said that "to stimulate data sharing, the citation index could be broadened to include how often a researcher's data has been reused, so to also add a Data-H-Index. In this way, researchers can also gain recognition and appreciation for sharing their research data." Another idea was to give researchers some time before publication to use the data before the data becomes openly accessible. This would entail a temporary restriction or a temporary exclusive right to use the data they have collected to protect the IPR. These two suggestions seem to be especially suitable for WU researchers to publish their results. Additional incentives will have to be devised for researchers working at the WR institutes.

Funding the building of data infrastructure
Currently, each WR institute has to provide funding for building its own data infrastructure. This funding can come from projects by allocating a separate amount of the budget for the institute's data infrastructure. Another way is to use a surcharge of a fixed percentage for each project.

To raise awareness and promote data sharing within WUR, as well as to explore what the requirements for a WUR data-sharing infrastructure are, a new investment theme could be executed for a duration of five to ten years. However, the opinions are divided on this.

The funding for building and managing a WUR-wide internal data sharing infrastructure could be financed by charging a fee to researchers from one institute who reuse data that has already been collected and stored by researchers from another institute. These revenues go into a single collective pot that is intended for the whole of WUR.

Separate development of data infrastructure
Each science group and research institute is developing and exploiting its own data infrastructure, in which one institute is more advanced than the other. Collaborating more intensively prevents every science group

[47] ARVODI: De Algemene Rijksvoorwaarden voor het verstrekken van opdrachten tot het verrichten van diensten
[48] DANS: https://dans.knaw.nl/nl/, 4TU: https://www.4tu.nl/, Zenodo: https://zenodo.org/

from having to reinvent the wheel. The WDCC could play a connecting and guiding role in this collaboration. A setup like the Shared Research Facilities (SFR) at WUR might act as an example.

Type of funding influences possibility of data sharing for re-use
To reuse data within WUR, we have to deal with the differences in the origin of the funding between WU and WR, which may be public and/or private. At WU, all researchers must open up the data used for their publications. But in the privately-funded contract research that WR conducts for commercial companies, the requirement is often set that the data may not be made public. In research projects funded with public money, for example funded by the Ministry of LNV or the EU Commission, the data must be made public. But in case of sensitive data, access to that data will or can be restricted according to legally provided rules.

Fear of losing work due to sharing data
A fear of sharing data with researchers from other Wageningen institutes is that the data will be used to take over new projects or the research work of those who generated the data. We do not know whether this is a realistic concern, but if it is, sharing principles need to be outlined in a clear policy. In addition, sometimes a knowledge field cannot always be strictly assigned to one research institute or one group of researchers. This fear can be overcome by making agreements between the different WUR research groups.

Making data management plans
An important step for making data FAIR is that every project at WUR creates a data management plan (DMP). Currently, WU is more strict in enforcing this, and all PhD candidates are obliged to make a DMP for their research projects. The WR institutes are less strict and structured, but they have also started to implement the DMP requirement. The WUR WDCC provides an online DMP template with guidance.[49] Furthermore, data sharing within WUR could also be promoted by offering training to researchers on how to make their research data FAIR.

To ensure that researchers actually make data findable, accessible and interoperable so that these data can be reused by others, they likely also need to be motivated in other ways.

6.3 Researchers' opinion on data sharing

A new infrastructure for data sharing within WUR is only valuable if the data has added value to other researchers and research domains within WUR. To be of value, three conditions must be met according to the interviewed researchers. The first condition is that researchers can reuse the data effectively. This implies that the data must fulfil the demand of other researchers so that they do not have to collect the data themselves. It also implies that the data collected by other researchers is an equivalent but cheaper alternative to the data that researchers have used until then. And that data can also be suitable for other researchers for reuse if it will provide new insights.

The second condition is that data is allowed to be reused by other researchers by policy, legislation and regulations. The WUR data policy prescribes that the norm is that "research data should be as open as possible, and as closed as needed." Fully open data is not the norm. Several factors determine whether research data should be made open, restricted or closed. This is also visualised in Figure 10.

The third condition concerns the willingness of researchers to share the data they collected. This aspect has to do with, among other things, the organisation culture: unwritten rules and habits.

We do not know to what extent WUR researchers are making or are willing to make their collected data findable, accessible and interoperable to the other WUR researchers for reuse in 2023, nor do we know what obstacles and incentives play a role in this.

[49] https://zenodo.org/records/7801409 (Accessed 21-12-2023)

Figure 10 *The WUR data policy, stating that research data should be as open as possible and as closed as necessary*

We have conducted interviews to get an idea of the main barriers and stimuli experienced by WUR researchers for making their collected data FAIR. WECR researchers conducted 17 qualitative interviews with researchers and/or data stewards from various WR institutes and University departments. These interviews were sometimes carried out in collaboration with the Wageningen Data Competence Centre.

The questions were mostly focused on the cultural and behavioural aspects of research data management, rather than the technological aspects. The interviews were carried out in a semi-structured style.

Table 13 *Departments of the interviewees**

Respondent	Component	Department
1	ASG	Wageningen Bio Veterinary Research
2	SSG	Wageningen Centre for Development Innovation
3	SSG	Wageningen Centre for Development Innovation
4	SSG	Wageningen Centre for Development Innovation
5	SSG	Wageningen Economic Research
6	SSG	Wageningen Economic Research
7	ESG	Wageningen Environmental Research
8	ESG	Wageningen Environmental Research
9	WFSR	Wageningen Food Safety Research
10	WFSR	Wageningen Food Safety Research
11	ASG	Wageningen Livestock Research
12	ASG	Wageningen Marine Research
13	PSG	Wageningen Plant Research
14	PSG	Wageningen Plant Research
15	ESG	Wageningen University
16	PSG	Wageningen University
17	AFSG	Wageningen University

*It was agreed to keep the interviewed employee anonymous

6.3.1 Barriers

Below is a summary of the barriers mentioned by the interviewed employees.

- No standard protocol for data storage. There is no standardised data storing method and no common metadata standard within WUR. Researchers have their own preferences when storing data for their own use, and some researchers are hesitant to change their habits.
- No personal benefit. There are no personal benefits to be gained from sharing your data freely with others. Many researchers feel proud of 'their' data. Sharing data has the risk of not getting the recognition or gratitude they think they deserve.
- Loss of control. There is a strong sentiment among researchers that they have control over who gets to use their data. The interviewees felt that WUR does not provide clear data ownership policy. WUR does not provide a mechanism for researchers to show ownership as the original collector of the data.
- Skills for management of research data. Data sharing might require new skills that are difficult for some researchers to learn. Some researchers might not be comfortable with the IT skills required for proper research data management. The researcher's cultural background can make it difficult to ask for guidance on data storage. Some researchers also prefer to work the way they are familiar with and never get told about new data management practices.
- A lack of data management onboarding means that new employees are highly dependent on their supervisor. If the supervisor is competent at data management and prioritises this aspect, it is more likely that the new employee will learn about data management than if the supervisor is less interested or too busy. Also, long-serving employees are often unaware of proper data management practices. This is closely linked to habits.
- Good data management takes time. Insufficient or improper budget allocation in projects prevents researchers from having the time for proper data management. Data management must take place not only at the end of a project, but also at the start and during the project.
- No online catalogue of the available data sets. WUR researchers are expected to have sufficient networking skills to find information. One of the most frequently heard barriers to data sharing is that the only way to discover the existence of data sets is through your colleagues. There are no other mechanisms in place to facilitate this.

6.3.2 Stimuli

Below is a summary of the stimuli mentioned by the interviewees.

- Easy storage of research data. Researchers should be able to quickly and reliably store their research data. Likewise, finding data should be a simple task. Simple programmes, methods and interfaces may encourage making research data findable, accessible and interoperable.
- Explicit appreciation and recognition. Researchers are more willing to share their data if it is recognised that they were the ones that provided the data sets and if colleagues are appreciative of this fact. For example, adding a 'like' button to someone's professional profile that can be clicked if colleagues appreciate the data, or being cited.
- Creating more awareness of FAIR data principles among students and staff. In addition, it should be easier for researchers to make their data FAIR. Management should encourage researchers to be proactive in making data FAIR.
- Allocate sufficient and separate project budget for data management.
- Reviewing procedures should not merely focus on the final report or article, but also on the data sets.
- Data management should be included in the onboarding process. Some departments have already implemented this. The research institute's management should support researchers in research data management, for example by offering training courses.
- The WUR organisation has to communicate the rules on data rights control and ownership more proactively.

7 Known developments at WUR

7.1 Strategic visions

Three vison documents support the WUR strategy on data sharing (see Table 14).

Table 14 *Vision documents for WUR's strategic plan*

#	Vision	WUR plan period
1	Open Science & Education	2022–2025
2	WUR guidelines on value creation with software and data	2022–2025
3	Digital Strategy	In prep 2025–2028

7.1.1 WUR Open Science & Education plan (OSE)

Today, the Open Science & Education plan (OSE) is well-known among staff and is one of the priorities of the WUR Extension & Update Strategic Plan 2019-2024. An important foundation has been laid by adopting the progressive WUR Open Access and Data Management policy and developing policy for open educational

Open science is a more open way of conducting, publishing and evaluating scientific research. Open science strives for more transparency in the research process, collaboration and the reusability of knowledge among researchers and across disciplines, as well as in society as a whole. Open science principles can also be applied in education, for example by sharing and reusing online learning resources. Adopting the principles of openness and transparency will contribute to a more efficient research environment and strengthen the integrity and reliability of science. 'Finding answers together' is key to increasing the value of science.

WUR employees are encouraged and facilitated to apply OSE principles through practical tools, infrastructure, training courses, networks and help desks, among others. The OSE ambition document for 2022–2025 describes how we will shape OSE in the coming years. The national Recognition and Reward programme has a specific open science component. The local application of open science and education principles in programme evaluation protocols as well as individual academic career paths is expected to play an important role in achieving this cultural shift. resources (**Figure 11**).

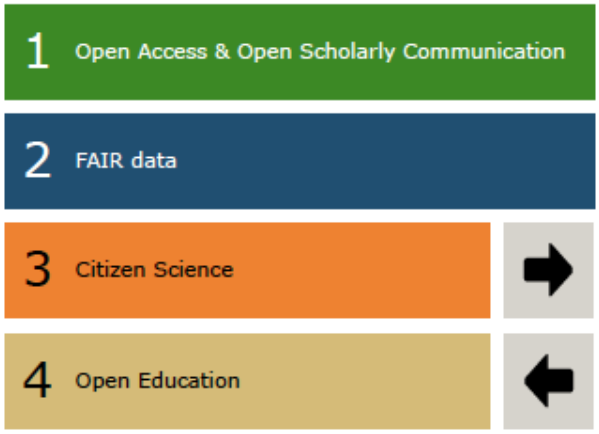

Figure 11 *The approach of the Open Science & Education plan for 2022–2025*

7.1.2 WUR guidelines on value creation with software and data

The WUR guidelines on value creation with software and data offer practical support for increasing our societal impact. They provide a description of creating value with software or data under an open or proprietary license and elaborate on relevant considerations. Open science practices are an integral part of these guidelines. The purpose is to present a clear framework for the many issues concerning value creation that come with software and data. The framework explains the available IPR for software and data protection, such as copyright, patents and database rights, and various models for design tools, products and services that can be used for value creation. These guidelines form part of the WUR policy on Intellectual Property and Value Creation ('WUR Value Creation Policy') and are applicable to WUR employees, PhD candidates and research students.

WUR employees develop or use software and data as part of their research, which influences the way the research results make an impact. Software and data are more widely recognised as stand-alone research outputs, ranging from a single line of code to a complete data set or an entire software package containing multiple interconnected programmes. As a general principle, software and data should be well-documented, and the FAIR principles should be followed whenever possible. WUR encourages open science and FAIR data. These guidelines are predominantly aimed at those who see further potential for societal value creation, commercial exploitation or other utilisation of data or software created during research projects. This is structured around four basic questions in protection and value creation with software and data.

1. What do you want to achieve?
2. What do you have?
3. What can you do?
4. How do you create value?

7.1.3 Digital Strategy (in progress)

The Digital Strategy will be part of the new 2025–2028 strategic plan, which was drafted at the end of 2023. It will state that the WUR digital strategy will help define and manage opportunities through new technologies to improve the quality and impact of our education and research. It is built on four key themes: 1) innovation empowered by digital possibilities; 2) connecting people while developing digital expertise; 3) strengthen an ecosystem that makes use of digitally supported solutions; and 4) an organisation that shares responsibilities for open and transparent digital operations. It will do so by increasing collaboration between our domains and with partners in our ecosystem, mitigating threats and focusing on ethical choices. This vison is strongly supported by the soft infrastructure as is described in the nine building blocks of the Innopay framework presented in this report.

7.2 Wageningen Data Competence Centre

The Wageningen Data Competence Centre (WDCC) was established to organise, facilitate and monitor new applications, opportunities and developments in the field of big data. WDCC brings education, research, value creation, infrastructure and data management together. It facilitates developments in these fields and serves as an internal and external contact point. This way, WUR works on new applications to explore the potential of data to help improve the quality of life.

WDCC shows how relevant data and data science are in our domains. It works the other way around too: we also show how significant WUR's contribution can be to the development of data science. WDCC organises everything that is needed to take advantage of what we can and want to do with data, now and in the future. This builds on the knowledge and data available at WUR.

Research data management is a joint responsibility of different organisations within the facility services: FB-IT is responsible for creating the data infrastructure. The Library is responsible for data archiving and FAIR data services. Legal is responsible for sorting out data rights like ownership and IPR.

7.3 FAIR and Data Stewards

Under the influence of open science developments, the new Code of Conduct for research integrity, global academic discussions on FAIR data, funding requirements and the introduction of a new WUR data policy, researchers are increasingly being requested to perform safe and sustainable management of their research data. Data management plans must be made at the start of projects, and WU PhD candidates and postdocs are obliged to include these plans in their research proposals. Data management prescribes that secure data storage must be used during the research and that the data must be archived and registered once the research is complete.

Data stewards are appointed at all science groups to support employees dealing with research data, departments, teams and in projects. By the end of 2023, approximately 180 data stewards had been appointed. However, the visibility of data stewards is low, and employees do not fully understand their role. Support and coordination of the data stewards is done by the WDCC research data management support team.

7.4 4TU

The four Dutch technical universities (Technische Universiteit Delft, Technische Universiteit Eindhoven, Universiteit Twente and WUR), are united in the 4TU.Federation.[50] This federation aims to boost and pool technical expertise. The universities plan to educate and deliver plenty of excellent engineers and technological designers to realise internationally renowned and relevant research and to promote cooperation between research institutes, businesses and public organisations. Research data is an important element and an international data repository for science, engineering and design has been named 4TU.ResearchData. It offers research data set curation, sharing, long-term access and preservation services to anyone, anywhere. Furthermore, training is offered and community engagement resources are available to research professionals and support professionals working to make their research data findable, accessible, interoperable and reproducible (FAIR).

[50] https://www.4tu.nl/ (accessed 21-12-2023)

8 International developments

Developments in Wageningen cannot be seen separately from what happens elsewhere in the world. The increase in the amount of data, the so-called data explosion, makes adequate measures necessary to enable proper utilisation. Developments are taking place worldwide that unlock data in such a way that use is facilitated. In Europe, the Commission has drawn up a digital strategy that includes data spaces designed to make data available in a well-structured and organised manner for policy, the market and EU citizens. There are also technical developments worldwide, such as data meshes, that are used when data becomes available as a completely independent product. We go into these worldwide developments in more detail in the following two sections.

8.1 EU Common data spaces

A major development with regard to pan-European data sharing and reuse is the announcement of the European Strategy for Data.[51] This strategy aims to create an internal market for efficient and secure data sharing and data exchange between sectors within the EU. The centrepiece of the European Strategy for Data is the concept of data spaces (see xxxx?). A data space can be defined as a federated data ecosystem within a given application domain that is based on shared policies and rules. The users of such data spaces get access to data in a secure, transparent, trusted, convenient and uniform manner.

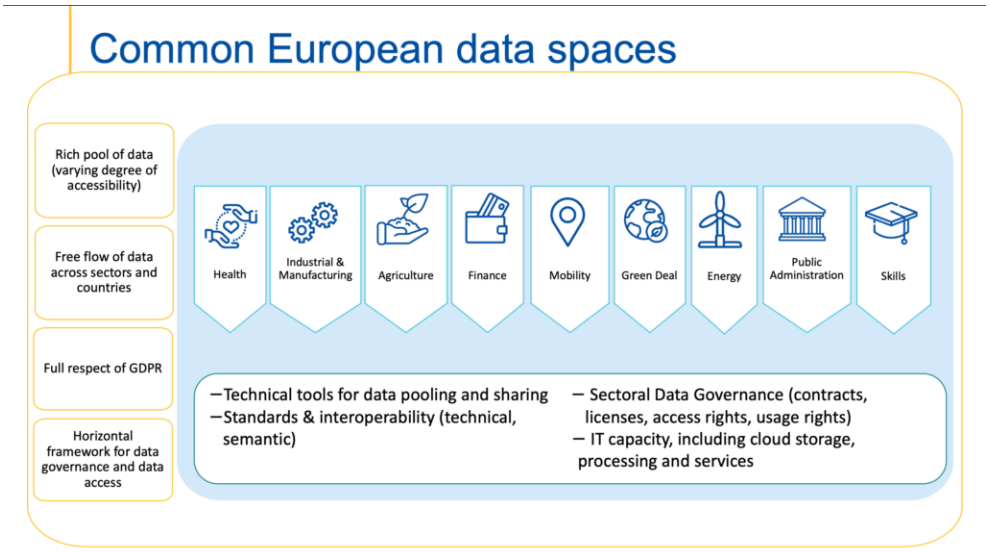

Figure 12 *Illustration of the common European data spaces[52]*

The EU Commission set out the European Strategy for Data as a single market for data (see Figure 13).

[51] https://digital-strategy.ec.europa.eu/en/policies/strategy-data (accessed 20-02-2024)
[52] https://digital-strategy.ec.europa.eu/en/policies/strategy-data accessed 20-02-2024)

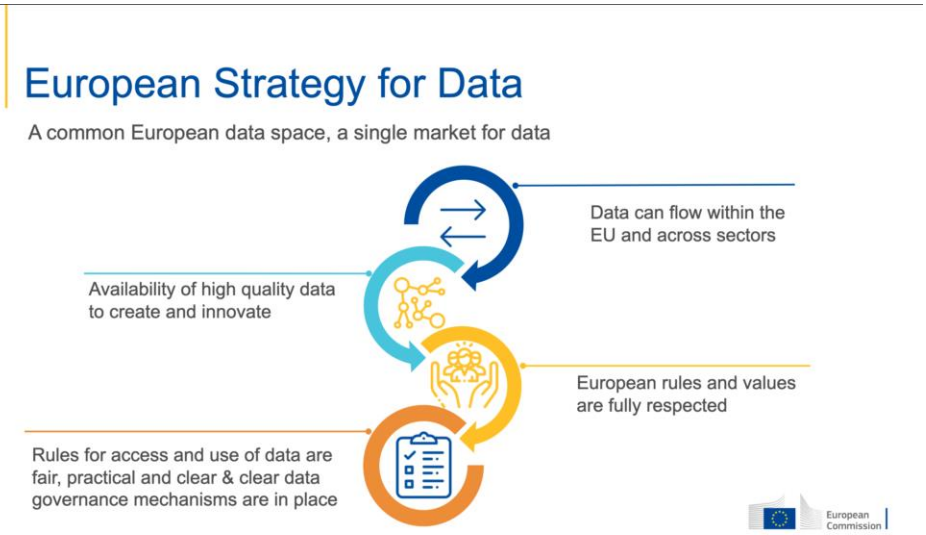

Figure 13 *European Strategy for Data*

Data is an essential resource for economic growth, competitiveness, innovation, job creation and societal progress in general. Common European data spaces will ensure that more data becomes available in the economy and society in member states, while companies and individuals who generate the data stay in control.

Supporting the implementation of the EU Digital Transformation Strategy, the EU-funded OPEN DEI project[53] aimed to encourage synergies, support regional and national cooperation and enhance communication among planned innovation actions. A particular priority is to base the creation of all common data platforms on a unified architecture. Therefore, the OPEN DEI project has defined a European soft infrastructure with 12 building blocks, specifying functional, legal, operational and technical aspects such as security, identity, authentication, protocols and metadata. These building blocks fall into two categories: technical building blocks and governance building blocks. The technical building blocks are differentiated into interoperability, trust and data value (see **Figure 14** EU Data spaces building blocks).

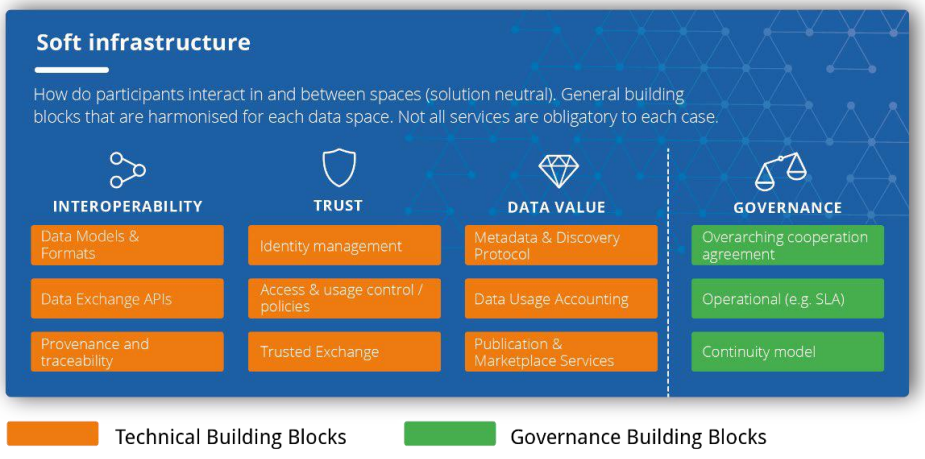

Figure 14 *EU Data spaces building blocks*

The OPEN DEI framework building blocks are comparable with those of the Innopay framework. Innopay was selected by us because the Open DEI project was still ongoing at the time we started this project. However,

[53] https://www.opendei.eu/ (accessed 20-02-2024)

the OPEN DEI framework is the reference point for further EU data space development and will probably be adopted more broadly. Therefore, it is advisable to also adopt this framework for future activities related to WUR data infrastructure development.

8.2 Data Mesh

The term 'data mesh' was first defined by Zhamak Dehghani (Dehghani, 2022). 'Data mesh' is a sociotechnical approach to building a decentralised data architecture by leveraging a domain-oriented, self-serve design. It borrows Eric Evans' theory of domain-driven design (Evans & Fowler, 2019) and Manuel Pais' and Matthew Skelton's theory of team topologies (Skelton et al., 2019). Data mesh is mainly about the data itself, with the data lake and pipelines being of secondary importance. The main proposition is scaling analytical data by domain-oriented decentralisation. With data mesh, the responsibility for analytical data is shifted from the central data team to the domain teams, supported by a data platform team that provides a domain-agnostic data platform.

Data mesh is based on four core principles:
- Domain ownership
- Data as a product
- Self-serve data platform
- Federated computational governance

In addition to these principles, Dehghani emphasises that the data created by each domain team should be discoverable, addressable, trustworthy, possess self-describing semantics and syntax, be interoperable, secure, and governed by global standards and access controls. In other words, the data should be treated as a product that is ready to use and reliable.[54]

Crucial in this approach is that data is considered as a product and, as such, a fully serviced data access point for which the owner remains responsible. The data mesh is built up from data sources as nodes in a network that can be used to create applications.

> *'DATA PRODUCT' VERSUS 'DATA AS A PRODUCT'*[55]
>
> *Any digital product or feature can be considered a "data product" if it uses data to facilitate a goal. For example, the home page of a digital newspaper can be a data product if the news items featured in the home page I see are dynamically selected based on my previous navigation data.*
>
> *"Data as a product" is the result of applying product thinking into data sets, making sure they have a series of capabilities including discoverability, security, explorability, understandability, trustworthiness, etc. What does data as a product look like? A data as a product contains the code, its data and metadata, and the necessary infrastructure components to run it.*

[54] https://en.wikipedia.org/wiki/Data_mesh
[55] [source https://towardsdatascience.com/data-as-a-product-vs-data-products-what-are-the-differences-b43ddbb0f123]

9 Towards a shared infrastructure @WUR

9.1 Introduction

The assessment of the data sources used in this project (summarised in Chapter 3) revealed that there are differences between and within science groups with respect to working with data and models, but it also revealed that a common awareness on the importance of data sharing exists. Science groups acknowledge that if data and models are used, they must be well-documented and made accessible for sharing. WUR as an organisation also realises that a harmonised approach towards data sharing, offering data sharing infrastructure and a policy that defines associated guidelines for WUR research is indispensable. This is one of the reasons why the WDCC was founded and why one of WUR's most important statements is 'Serious about data'. The WDCC promotes the adoption and operationalisation of FAIR principles for data and models, among others supported by data stewards. This also includes introducing data management plans (DMPs) for each research project and proposal. In some situations, models and software are assessed to comply with Status A, which is a certification that guarantees proper documentation, use and a certain level of quality for our research.

The importance of data and data management is widely recognised by WUR management and researchers. At the same time, a clear strategy for data management, data sharing and the required supporting data infrastructure is missing, as well as a guiding infrastructure framework. Consequently, over the past decade, and with the vast amount of new research data, many data infrastructures have been established independently. This has resulted in a fragmented and heterogeneous landscape of data and data infrastructures. Many of these infrastructures have become foundational elements of the research strategy of institutes and research groups.

This chapter provides recommendations to move towards a clearer, shared corporate strategy on data management and data sharing. It considers the corporate level, the research group level and the level of individual researchers. It addresses the aspects reflected in the Innopay building blocks framework, from governance and data management to the more technical aspects of a data sharing infrastructure. It also takes account the importance of existing infrastructures and the need to integrate these into a broader strategic approach.

9.2 Recommendations at WUR corporate level

Corporate responsibility

> WUR should take up responsibility with respect to data governance and the underlying data sharing infrastructure and embed it within the WUR strategy.
>
> It is proposed to assign the role of a Chief Data Officer (CDO) to one of the general directors of WUR.
>
> (Action for: Board of Directors)

A general observation from this research is that data governance requires more attention. Overall, data governance is considered important within WUR, but it is not always fully covered and often set up on an ad-hoc basis, resulting in a fragmented pattern across WUR. To make its vision on data sharing clear and provide a foundation to improve and harmonise data sharing, WUR should define data governance as part of its corporate strategy and take control.

Define a business case on how to fund the creation, maintenance and support of a common data sharing infrastructure. Clarify what costs are covered by the 'initiating' activity, which is usually a project and which costs are covered by the science group (WR/WU) or on a corporate level.

(Action for: WECR, in collaboration with all science groups)

From the interviews with the science group directors, it appears that funding related to data infrastructures is often organised per science and/or research group, and that approaches differ a lot between groups. Most work at WUR is project-oriented. Governance over time in a project-oriented environment is not self-evident. Researchers generally do not have the means (budget and time) to properly maintain the deliverables after a project has ended. This is specifically relevant for high value data sets that are created to be reused. The WUR science groups do not provide common principles on how this data maintenance should take form. More clarity on the WUR data maintenance strategy is needed. To clarify the costs and benefits of data sharing and the required governance and infrastructure, and to support a corporate strategy, a sound business case could help convincing management and researchers to take up a shared approach. The Shared Research Facilities can act as an example of how this can be organised.

Develop a WUR data sharing infrastructure based on a common framework

WUR should adopt a unified soft infrastructure framework as a reference for the future development of the WUR data sharing strategy and infrastructure.

The OPEN DEI framework promoted by the EU DSSC is currently the most logical choice.

(Action for: corporate IT)

The Innopay framework was used for the presented assessment of current data sharing infrastructures and their connectivity and interoperability. This has greatly supported us in our consideration of the current state of play and possible future directions. Adopting a similar framework will also provide a good baseline to establish and maintain a WUR corporate data sharing strategy and the future implementation of a data infrastructure. Considering the current EU developments, the building block framework developed and supported by the Data Space Support Centre (DSSC) is currently the most logical choice. Furthermore, this framework is currently widely accepted in the EU and its implementation is in progress. Another benefit of this framework is that it makes a clearer distinction between technical and governance building blocks, with the former being grouped around interoperability, trust and data value.

Capacity building

On corporate level, WUR needs to raise awareness and build capacity on all aspects of FAIR data sharing, particularly interoperability and achieving reusability. Dedicated staff, particularly data stewards, could support this better.

(Action for: WDCC)

While researchers are generally well-skilled in data science and analytics, awareness of the importance of data sharing and knowledge on how to make data interoperable and reusable is often lacking.

When it comes to interoperability, it is of the utmost importance that the structural and semantic information of the data is given. The structural description, also referred to as the ontology of data, is crucial when it comes to reuse. There are ongoing projects that provide tools to create and use ontologies. A good way to improve skills and make semantics applicable more broadly would be to set up a taskforce to develop a WUR ontology or knowledge graph, specifically aimed at creating better semantic links between WUR domains. While guidelines and tools should be provided through technical support, a strategy for capacity building should be deployed organisation-wide. Support staff for capacity building is available at WUR but this

information does not reach the whole organisation. The pivotal role of data stewards, their visibility and their skills should be addressed.

Introduce corporate standards where needed

> Develop and deploy a WUR metadata standard in collaboration with the WUR Library.
>
> (Action for: WDCC)

Data sharing practices vary between different science groups. Harmonising such practices will simplify future data sharing and exchange. One specific area where standardisation is helpful is metadata. To support the harmonised provision of metadata, a common WUR metadata standard template is a minimum requirement. Preferably, this template will follow a standard that complies with common standards such as PURE (CERIF), which is used by the WUR Library, Dublin Core and ISO 19115.

Broader perspective on data sharing infrastructures

> Take a broad perspective on data sharing infrastructures, aiming at a federated infrastructure (data mesh), respecting and integrating existing infrastructure elements.
>
> (Action for: WDCC, FB-IT)

Much effort is spent supporting regular research and providing policies, procedures and tools that are help individual researchers to organise, manage and share their data. At the same time, we see that data are very heterogeneous and that conditions that apply differ over funders and application areas, among others. This makes it hard to provide a one-solution-fits-all approach. Moreover, there are many independent and often siloed data infrastructures that have been developed over the years. They often have a strong position in their research domain and cannot just be replaced or altered. This unavoidably leads to a federated system of data infrastructures, or a data mesh, also onboarding specific requirements that do right to the characteristics of working with non-research organisations.

9.3 Recommendations on the research group and individual level

The previous section proposes recommendations on the WUR corporate level. These recommendations are fundamental to developing a common data sharing strategy based on an underlying data sharing infrastructure that supports the research process and makes life easier for researchers when it comes to data sharing and data reuse. At the same time, recommendations can be given for the current processes independent from future situations. While a corporate strategy and infrastructure are needed in the future, there are already means to improve the current situation by using tools and methods that are available and already partly embedded in WUR's data management policy and practices. Many of these practices can already be encouraged or even be made mandatory for leaders of research groups, research programmes and projects.

Collaborate more over research teams and science groups

> Science groups and research groups should collaborate more in data sharing to learn from each other and prepare for better interoperability.
>
> (Action for: science groups and research programme leaders)

Researchers are not always fully aware of what colleagues or science groups are doing, and thus interrelated research is often conducted in isolation. Although more and more researchers are collaborating, we are not yet used to working multidisciplinary. This also applies to data sharing practices. In this project, we have

experienced how working with different data infrastructures can improve awareness and catalyse innovations. We therefore advise to encourage such collaborations. We are aware that this process will take time. However, we strongly recommend speeding up this process by sharing good practices in combination with using the established structures in a more advanced way. For example, KB programmes should be used to connect the individual projects within programmes. To ensure this, it is also important to have a specific KB programme for digitisation, including aspects related to data sharing and connecting data-oriented initiatives.

Consider future sustainability and the earnings model

> When setting up a data infrastructure, have a clear earnings model in mind with a clear view of possible future sustainability.
>
> (Action for: project and research programme leaders)

New initiatives are starting with each new funding round or yearly programming cycle. It is important to ensure that these initiatives efficiently make use of and contribute to the data strategy and data infrastructure. This requires a good estimate of what added value can be expected, especially if an infrastructure is used or set up with the aim of preserving and sharing project results, including the data generated. Even more important is to ensure that results can be maintained after the project, if required. We therefore recommend that every proposal contains an assessment of the expected added value for data sharing and reuse, and how the results will be secured after the project ends.

Make use of existing tools

> Researchers should assess existing tools provided and supported by WUR and adopt them in case they fit their purpose.
>
> (Action for: researchers, supported by WDCC and data stewards)

Carrying out a research project or a research is a process that has several phases. Data management plans are a means to describe how data will be organised and managed. For completed projects, there are clear solutions for storing data for reuse. However, managing data during a research project remains cumbersome. Fortunately, there are tools available to support researchers with data management. One of these support tools is iRODS[56]/YODA,[57] which is promoted by the Wageningen Common Data Solutions (WCDS) programme and supported by WDCC.

Pending the results of the WCDS programme, we recommend to take a broader look at methods and tools that help and relieve the operational process related to data management in projects and research. This should take the different ways that data is used and organised (e.g. CSV files, databases, data warehouses, triple stores) and the complexity of data use into account.

Store data in accessible persistent repositories

> Once the project or research ends, store the data in existing, accessible persistent repositories, where feasible.
>
> (Action for: project and research programme leaders)

It is obvious that the data we generate at WUR should be accessible for reuse when relevant. This requires available budget and personnel hours. Think about it beforehand, not as afterthought. There are many

[56] iRODS: integrated Rule-Oriented Data System https://irods.org/ (accessed 21-12-2023)
[57] YODA: YOur Data: https://www.uu.nl/en/research/yoda (accessed 21-12-2023)

existing repositories that provide persistent, long-term data storage. WUR is part of the 4TU[58] consortium that initiated 4TU.ResearchData, an international data repository for science, engineering and design. 4TU.ResearchData offers research data set curation, sharing, long-term access and preservation services to anyone, anywhere. In the European context, there is Zenodo, an open repository for general purposes developed under the European OpenAIRE programme. Even in cases where data is commercially sensitive or contains personal data, there might be ways to share or at least publish the metadata with a reference to a contact point.

Document data with metadata and use semantics

> Researchers should document their research data with metadata using the most appropriate metadata standard and publish it through existing catalogues.
>
> (Action for: project and research programme leads)

Metadata is key to data reuse and provides the necessary information to find, access and use the generated or collected data. As part of the WUR FAIR policy, a lot of attention is paid to the usefulness and necessity of creating good metadata. While WUR agreements on metadata standards are missing, researchers should assess which standard fits best for their objective and how these metadata can be registered, asking advice and support from WDCC when required.

Many data repositories already have a metadata catalogue and provide search functions for data. We recommend publishing the metadata in existing metadata catalogues, especially when the data is suitable for reuse. In the Netherlands, we can recommend the Dutch government Data Register (data.overheid.nl) or, more specifically for research data, 4TU.ResearchData or Zenodo. If a new repository is set up, make sure that the metadata is stored in a catalogue that can easily be harvested by other catalogues.

> Make use of semantic standards, using existing, commonly accepted ontologies in combination with controlled vocabularies or thesauri as much as possible.
>
> (Action for: project and research programme leaders)

Technical standards are well-known and applied in practice, but semantic standards are still underdeveloped at WUR. Nevertheless, using harmonised, agreed upon semantics is a prerequisite to obtaining interoperability as part of FAIR data. Research groups should therefore invest in knowledge and capacity on semantics. Where possible, they should adopt and integrate accepted ontologies and controlled vocabularies to increase data interoperability.

9.4 Roadmap for a common data sharing infrastructure

> As part of a corporate digital strategy, WUR should evolve its data infrastructure towards a data mesh, a federated infrastructure where independent FAIR data components offer the required interoperability for efficient data reuse.
>
> (Action for: FB-IT, in collaboration with the science groups)

Based on the lessons learned from this project and the developments in data sharing within and outside of WUR, this section proposes a roadmap towards a common data infrastructure.

[58] 4TU.ResearchData is led by the 4TU.ResearchData Consortium, which consists of Delft University of Technology, Eindhoven University of Technology, University of Twente and Wageningen University & Research.

We currently see a highly fragmented data ecosystem at WUR. Data, databases, data warehouses, data lakes and data cubes are scattered and integrated within all sorts of information systems and applications. This landscape was established over many years and consists of a multitude of heterogeneous, partly connected elements with quite some legacy from the past. It can be better understood through the maturity model for data sharing environments, shown in Figure 15. It shows how data sharing infrastructures have evolved over the years and is applicable to what has happened at WUR. We have seen a gradual movement from isolated data islands, for example on mainframes and other monoliths, towards initial attempts to connect those silos in what are sometimes called 'archipelagos'. The creation of shared network drives and other data sharing environments has subsequently helped to make data accessible in a far more flexible way but often without a lot of emphasis on reusability. In recent years, the evolution of data infrastructures towards data warehouses, data lakes and similar infrastructures has further improved the access to data, but this time with much more emphasis on reuse, for example by implementing technical and semantic standards, and allowing easier recombination of data.

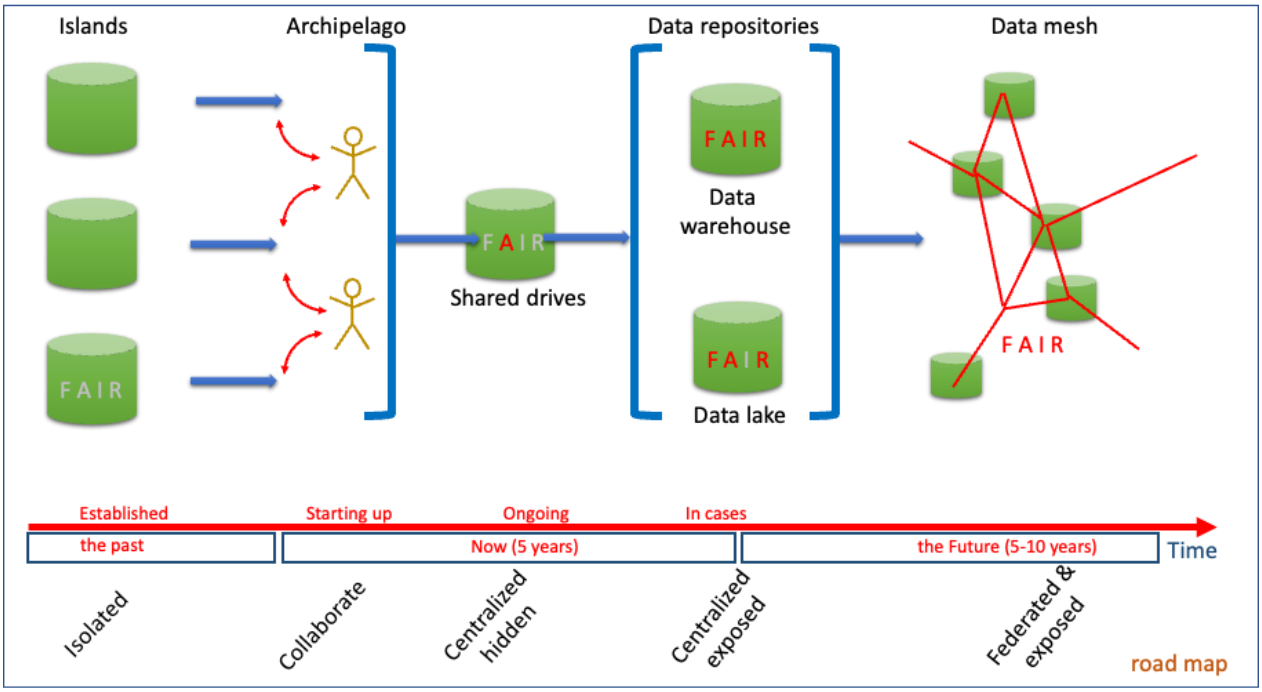

Figure 15 *A maturity model for data sharing environments*

At WUR, we are currently transitioning from using shared drives to data repositories, although time islands still exist within WUR. In terms of FAIR, we are gradually changing from situations where data is accessible towards data getting also more easily findable. At the same time, reuse is still often hindered by lack of data documentation and interoperability.

Today's trend is to move towards more federated infrastructures, where data stays at its source. These federated infrastructures, often referred to as a data mesh, offer many advantages. It is much easier to manage issues like data rights, data sovereignty, data privacy and handling sensitive data in general. From the WUR perspective, it would clarify the position of currently operational infrastructures as independent yet interconnected and interoperable components in a data mesh. Equally important, it would allow WUR to link with the larger external data landscape more easily, which is also gradually evolving towards a federated ecosystem. An important consequence of migrating towards a data mesh is that substantial efforts and investments are needed to improve interoperability. It requires improvements in the documentation of data and data interfaces but above all in semantic interoperability. The aspects required to achieve this are also interwoven into the recommendations provided in the previous chapters.

10 Acknowledgements

We would like to acknowledge all the researchers that were involved in the different use cases explored in the Knowledge Base Data Driven High Tech (KB DDHT) project 'Smart infrastructures for farm generated data towards circular agriculture'. Without these researchers, there would have been no input for this report. We would also like to acknowledge Bert Lotz, Hans Marvin and Yamine Bouzembrak, who contributed significantly to the deliverable report of the aforementioned project that served as the basis for the current report. We would like to thank Willem Jan Knibbe too, who helped formulate the basis of this report by pointing out the nine building blocks of the Innopay framework. Bas van der Velden, Jene van der Heide, Willem Jan Knibbe, Shauna Ní Fhlaithearta and Wies Vullings are acknowledged for reviewing this report. Finally, we would like to acknowledge the Ministry of LNV for funding and, with that, enabling the publication of this report.

References

Abdulrahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C., & Guizani, M. (2021). A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. *IEEE Internet of Things Journal*, *8*(7), 5476–5497. https://doi.org/10.1109/JIOT.2020.3030072

Been, T. H., Kempenaar, C., Van Evert, F. K., Hoving, I. E., Kessel, G. J. T., Dantuma, W., Booij, J. A., Molendijk, L. P. G., Sijbrandij, F. D., & Van Boheemen, K. (2023). Akkerweb and farmmaps: Development of Open Service Platforms for Precision Agriculture. In D. Cammarano, F. K. Van Evert, & C. Kempenaar (Eds.), *Precision Agriculture: Modelling* (pp. 269–293). Springer International Publishing. https://doi.org/10.1007/978-3-031-15258-0_16

Dehghani, Z. (2022). *Data Mesh*. O'Reilly Media, Inc.

EU, A European strategy for data. (2020). *A European strategy for data* (COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS COM(2020) 66 final). EU Commission. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

Evans, E., & Fowler, M. (2019). *Domain-driven design: Tackling complexity in the heart of software*. Addison-Wesley.

Heimbigner, D., & McLeod, D. (1985). A federated architecture for information management. *ACM Transactions on Information Systems*, *3*(3), 253–278. https://doi.org/10.1145/4229.4233

Houweling, H., G.A.K. van Voorn, A. van der Giessen, & J. Wiertz. (2015). *Quality of models for policy support, WOt-paper 38* (WOt paper 38). http://edepot.wur.nl/362127

Janssen, H., Janssen, S. J. C., Knapen, M. J. R., Meijninger, W. M. L., Van Randen, Y., La Riviere, I. J., & Roerink, G. J. (2018). *AgroDataCube: A Big Open Data collection for Agri-Food Applications* [dataset]. Wageningen Environmental Research. https://doi.org/10.18174/455759

Skelton, M., Pais, M., & Malan, R. (2019). *Team topologies: Organizing business and technology teams for fast flow*. IT Revolution.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Appendix 1 Centralised versus distributed versus federated[59]

There is a lot of confusion about these terms as they can easily mean different things for different purposes. So, for databases, architectures, networks and even governments, these terms are used as adjectives and as such give different meanings to the definitions.

But there are also similarities. Federated means in all cases working together, to collaborate. Centralised and distributed have a spatial concept. That is, in one place or in different places, which can be spatial but also within one organisational unit.

In our case, it is best to have the meaning it is used for federated learning, usually used for machine learning (ML) algorithms, but the concept can be used in a generic manner. It seems to fit to the 'Farm Data Train' metaphor. The figure below depicts this:

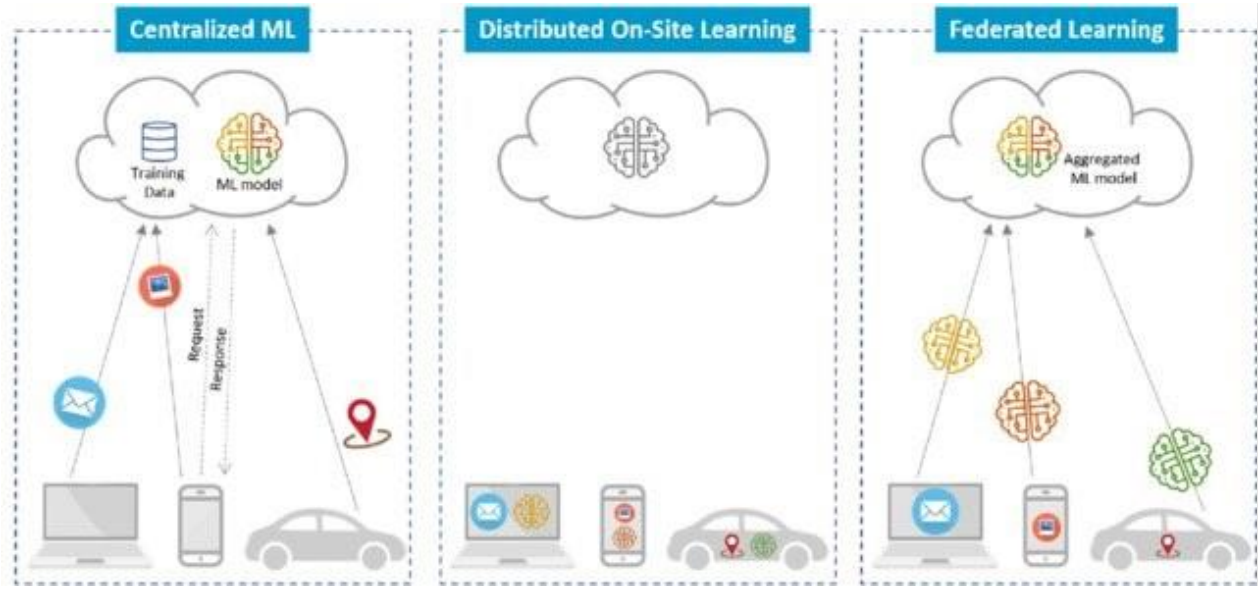

Figure 16 *Centralised vs Distributed On-Site vs Federated Learning Architectures*

In Centralised ML (left), data is sent to the cloud, where the machine learning model is built. The model is used by a user through an API by sending a request to access one of the available services. For Distributed On-Site Learning (middle), each device builds its own model using its local data set. After the first interaction with the cloud to distribute a model to the devices, no more communication with the cloud is needed. In Federated Learning (right), each device trains a model and sends its parameters to the server for aggregation. Data is stored on the devices and knowledge is shared through an aggregated model with peers.

[59] https://www.researchgate.net/publication/344871928_A_Survey_on_Federated_Learning_The_Journey_From_Centralized_to_Distributed_On-Site_Learning_and_Beyond

Appendix 2 Legal framework on sharing data and information

Aarhus Convention

Regulation (EC) No 1367/2006[60] of the European Parliament and of the Council of 6 September 2006 on the application of the provisions of the **Aarhus Convention** on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters to Community institutions and bodies Key points:

- It requires the EU's institutions and various bodies to implement the obligations contained in the Aarhus Convention. These obligations guarantee the public access to information, participation in decision making and access to justice on environmental issues.
- EU institutions and bodies must:
  - guarantee the public access to environmental information they receive or produce
  - ensure that environmental information is progressively made available and disseminated to the public
  - provide for early and effective public participation on environmental plans and programmes
  - grant the public access to justice on EU environmental matters
  - avoid any discrimination based on citizenship, nationality or domicile when treating a request for environmental information
  - organise the information in databases which the public can easily access
  - update the information as appropriate and ensure it is accurate and comparable
  - inform an applicant within 15 working days if they do not have the information being requested.
- Environmental databases or registers must contain:
  - texts of international treaties, conventions or agreements, policies, plans and programmes
  - progress reports on implementation of the above items
  - steps taken in proceedings for infringements of EU law
  - monitoring data of activities that could affect the environment
  - authorisations given which could affect the environment
  - environmental impact studies and risk assessments.
- Non-governmental organisations which meet certain criteria may request an EU institution carry out an internal review of an environmental matter.
- Requests for information are considered to be open to the public. They may only be refused in specific circumstances, such as ongoing legal proceedings or if they might harm the environment by, for instance, revealing breeding sites of rare species.

PAEI[61]

Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC (**PAEI**). Key points:

- It fully adapts European Union (EU) countries' national laws to the 1998 Aarhus Convention on access to information, public participation and access to justice in environmental matters.
- It guarantees the public access to environmental information* held by, or for, public authorities*, both upon request and through active dissemination.
- It sets out the basic terms, conditions and practical arrangements that a member of the public must respect when granted access to the requested environmental information.

Access upon request:

- Public authorities must make available any environmental information they possess to an applicant without the person having to state a reason.

[60] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=celex:32006R1367
[61] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32003L0004&qid=1634570845589

- The information should be provided at the latest 1 month after the request is received. This may be extended to 2 months for voluminous and complex requests.
- Public authorities must make every reasonable effort to ensure the information they have can be readily reproduced and accessed electronically.
- The information should be supplied in the form or format the applicant specifies unless it is already publicly available in another format.
- EU countries must ensure civil servants help the public seeking access to information and maintain a list of accessible public authorities.
- Practical arrangements for dealing with requests include:
  - appointment of information officers;
  - facilities for examining the information; and
  - registers or lists of the information held and details of information points.
- Requests may be refused if they are:
  - manifestly unreasonable;
  - too general;
  - relate to unfinished material; or
  - concern internal communications.
- They may also be refused, in full or in part, if the disclosure could be damaging to one of the exhaustive grounds envisaged, for instance:
  - international relations;
  - the course of justice;
  - intellectual property rights; or
  - commercial or industrial confidentiality.
- Access to public registers or lists should be free of charge. Public authorities may charge for the environmental information they make available upon request, but the amount should be reasonable.
- Applicants who consider their request has been ignored or wrongfully refused may have access to remedies, including a court of law or another independent body.

Active dissemination

- Electronically accessible environmental information must contain at least:
  - texts of international treaties, conventions or agreements, policies, plans and programmes relating to the environment;
  - progress reports on implementation of the above items;
  - reports on the state of the environment;
  - monitoring data of activities that could affect the environment;
  - authorisations which could have a significant impact on the environment;
  - impact studies and risk assessments.
- For items other than those above, active dissemination may be done progressively taking account of the human, financial and technical resources required.
- EU countries must ensure any information compiled by them or on their behalf is up-to-date, accurate and comparable.

Transposition into Dutch Law:

- Wet openbaarheid van bestuur. Official publication: Administrative measures
- Wet milieubeheer. Official publication: Administrative measures
- Wet van 23 juni 2005, houdende wijziging van de Wet milieubeheer, de Wet openbaarheid van bestuur en de Archiefwet 1995 ten behoeve van de implementatie van richtlijn nr. 2003/4/EG van het Europees Parlement en de Raad van 28 januari 2003 inzake de toegang van het publiek tot milieu-informatie en tot intrekking van Richtlijn 90/313/EEG van de Raad (PbEU L 41) en van richtlijn nr. 2003/35/EG van het Europees Parlement en de Raad van 26 mei 2003 tot voorziening in inspraak van het publiek in de opstelling van bepaalde plannen en programma's betreffende het milieu en, met betrekking tot inspraak van het publiek en toegang tot de rechter, tot wijziging van de Richtlijnen 85/337/EEG en 96/61/EG van de Raad (PbEU L 156) (Implementatiewet EG-richtlijnen eerste en tweede pijler Verdrag van Aarhus). Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 341; Publication date: 2005-07-07

- Besluit van 27/6/2005, houdende vaststelling van het tijdstip van inwerkingtreding van de Implementatiewet EG-richtlijnen eerste en tweede pijler Verdrag van Aarhus. Official publication: Staatscourant (Journal Officiel néerlandais); Number: 2005/342; Publication date: 2005-07-07; Page: 00001-00001

Open Data directive[62]

Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on **open data and the re-use of public sector information**.

The recast directive is based on the general principle that public and publicly funded data should be reusable for commercial or non-commercial purposes.

Open Data
- The directive promotes the use of open data (data presented in open formats* that one can use freely and share for any purpose). Public-sector bodies and public undertakings must make their documents available in any pre-existing format or language and, where appropriate, by electronic means in formats that are open, machine readable, accessible, findable and reusable, complete with their metadata.

Practical arrangements for reuse
- Public-sector bodies must, through electronic means where appropriate, process requests for document reuse, making them available within a reasonable time.
- At the same time, they should make necessary arrangements to facilitate the online search and discovery of the documents they keep.
- EU countries must also facilitate effective reuse of documents, in particular by supplying information on the rights outlined in the directive and by offering assistance and guidance.

Dynamic and real-time data
Dynamic ***data*** must be made available for reuse immediately on collection via an application programming interface (API) and, where relevant, as a bulk download.

Research data:
- EU countries must adopt policies and take action to make publicly funded research data openly available, following the principle of 'open by default' and support the dissemination of research data that are findable, accessible, interoperable and reusable (the 'FAIR' principle).
- Concerns relating to intellectual property rights, personal data protection and confidentiality, security and legitimate commercial interests must be considered in accordance with the principle of 'as open as possible, as closed as necessary'.
- Directive not transposed legal acts in NL available yet

Fair trading and non-discrimination
- The reuse of documents is open to everyone in the market and any applicable reuse conditions should be non-discriminatory.
- As a general rule, arrangements between public-sector bodies or public undertakings holding the documents and third parties cannot grant exclusive rights.
- In narrowly defined cases where the directive allows for the conclusion of such arrangements, their validity is subject to regular review and special transparency requirements apply.

High value data sets
- The European Commission is given the possibility to adopt a list of high-value data sets which should be made available in machine-readable formats and free of charge through APIs. The data sets will be selected from within 6 thematic categories set out in Annex I:
  - geospatial;
  - earth observation and environment;

[62] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32019L1024&qid=1634571003714

- meteorological;
- statistics;
- companies and company ownership; and
- mobility.

- New thematic categories may be added by the Commission, by way of a delegated act.

INSPIRE[63]

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (**INSPIRE**). The aim of the directive:

- It lays down general rules setting up an infrastructure for spatial information* in Europe for the purposes of European Union (EU) environmental policies and for policies or activities which may have an impact on the environment.
- The European infrastructure builds on that of spatial information that is established and operated by EU countries.

Key points

The legislation applies to spatial data that:

- cover areas where EU countries have jurisdictional rights;
- exist in electronic format;
- are held by, or on behalf of, a public authority or another body using the network;
- relate to environmental information.
- EU countries are responsible for ensuring that metadata* are created for the various environmental spatial data* sets and services listed in the legislation. Depending on the subject, these must be established within 2 or 5 years of the rules coming into force.
- The European Commission, assisted by a committee, adopts the technical arrangements to ensure spatial data sets and services can work together and harmonises them where this can be put into practice.
- These implementing rules were to be adopted no later than 15 May 2009 or 15 May 2012, depending on the subject.
- EU countries must establish and operate a network with the following services:
  - Discovery: to search for spatial data sets and services.
  - View: to display, navigate, zoom in and out, pan or overlay viewable spatial data sets.
  - Download: where this can be put into practice, to access directly and download copies of spatial data sets.
  - Transformation: to transform spatial data sets to achieve interoperability.
- Public authorities must be able to link their spatial data sets and services to the national network.
- EU countries may limit public access to spatial data sets and services for various reasons, such as legal confidentiality, public security, intellectual property rights or protection of the environment.
- The Commission operates an EU Inspire geoportal. This gives access to the national networks.
- The legislation does not require the collection of new spatial data.

Transposition into Dutch law:

- Wet van 2 juli 2009 tot implementatie van richtlijn nr. 2007/2/EG van het Europees Parlement en de Raad van de Europese Unie van 14 maart 2007 tot oprichting van een infrastructuur voor ruimtelijke informatie in de Gemeenschap (Inspire); Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 310; Publication date: 2009-07-27; Page: 00001-00006
- Besluit van 12 augustus 2009, houdende vaststelling van het tijdstip van inwerkingtreding van de wet van 2 juli 2009 tot implementatie van richtlijn nr. 2007/2/EG van het Europees Parlement en de Raad van de Europese Unie van 14 maart 2007 tot oprichting van een infrastructuur voor ruimtelijke informatie in de Gemeenschap (Inspire) (Implementatiewet EG-richtlijn infrastructuur ruimtelijke informatie) (Stb. 2009, 310); Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 343; Publication date: 2009-08-25; Page: 00001-00002

[63] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32007L0002&qid=1634571767360

- Besluit van 30 oktober 2009 tot uitvoering van de Implementatiewet EG-richtlijn infrastructuur ruimtelijke informatie; Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 454; Publication date: 2009-11-10; Page: 00001-00008
- Besluit van 23 november 2009, houdende vaststelling van het tijdstip van inwerkingtreding van het Besluit Inspire; Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 502; Publication date: 2009-12-02; Page: 00001-00001

GDPR[64]

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation (**GDPR**) The aim of the directive:

- the general data protection regulation (GDPR) protects individuals when their data is being processed by the private sector and most of the public sector. The processing of data by the relevant authorities for law-enforcement purposes is subject to the data protection law enforcement directive (LED) instead (see summary).
- It allows individuals to better control their personal data. It also modernises and unifies rules, allowing businesses to reduce red tape and to benefit from greater consumer trust.
- It establishes a system of completely independent supervisory authorities in charge of monitoring and enforcing compliance.
- It is part of the European Union (EU) data protection reform, along with the data protection law enforcement directive and Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by the EU institutions, bodies, offices and agencies

The GDPR strengthens existing rights, provides for new rights and gives citizens more control over their personal data. These include:

- easier access to their data — including providing more information on how that data is processed and ensuring that that information is available in a clear and understandable way;
- a new right to data portability — making it easier to transmit personal data between service providers;
- a clearer right to erasure ('right to be forgotten') — when an individual no longer wants their data processed and there is no legitimate reason to keep it, the data will be deleted;
- right to know when their personal data has been hacked — companies and organisations will have to inform individuals promptly of serious data breaches. They will also have to notify the relevant data protection supervisory authority.

The GDPR is designed to create business opportunities and stimulate innovation through a number of steps including:

- a single set of EU-wide rules — a single EU-wide law for data protection is estimated to make savings of €2.3 billion per year;
- a data protection officer, responsible for data protection, will be designated by public authorities and by businesses which process data on a large scale;
- one-stop-shop — businesses only have to deal with one single supervisory authority (in the EU country in which they are mainly based);
- EU rules for non-EU companies — companies based outside the EU must apply the same rules when offering services or goods, or monitoring behaviour of individuals within the EU;
- innovation-friendly rules — a guarantee that data protection safeguards are built into products and services from the earliest stage of development (data protection by design and by default);
- privacy-friendly techniques such as pseudonymisation (when identifying fields within a data record are replaced by one or more artificial identifiers) and encryption (when data is coded in such a way that only authorised parties can read it);
- removal of notifications — the new data protection rules will scrap most notification obligations and the costs associated with these. One of the aims of the data protection regulation is to remove obstacles to free flow of personal data within the EU. This will make it easier for businesses to expand;

[64] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32016R0679&qid=1634572034814

- impact assessments — businesses will have to carry out impact assessments when data processing may result in a high risk for the rights and freedoms of individuals;
- record-keeping — SMEs are not required to keep records of processing activities unless the processing is regular or likely to result in a risk to the rights and freedoms of the person whose data is being processed.

NIB[65]

Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union (**NIB**).

It proposes a wide-ranging set of measures to boost the level of security of network and information systems (cybersecurity*) to secure services vital to the EU economy and society. It aims to ensure that EU countries are well-prepared and are ready to handle and respond to cyberattacks through:
- the designation of competent authorities,
- the set-up of computer-security incident response teams (CSIRTs), and
- the adoption of national cybersecurity strategies.

It also establishes EU-level cooperation both at strategic and technical level.
Lastly, it introduces the obligation on essential-services providers and digital service providers to take the appropriate security measures and to notify the relevant national authorities about serious incidents.

Improving national cybersecurity capabilities;
- EU countries must:
  - designate one or more national competent authorities and CSIRTs and identify a single point of contact (in case there is more than one competent authority);
  - identify providers of essential services in critical sectors such as energy, transport, finance, banking, health, water and digital infrastructure where a cyberattack could disrupt an essential service.
- EU countries must also put in place a national cybersecurity strategy for network and information systems, covering the following issues:
  - being prepared and ready to handle and respond to cyberattacks;
  - roles, responsibilities and cooperation of government and other parties;
  - education, awareness-raising and training programmes;
  - research and development planning;
  - planning to identify risks.
- The national competent authorities monitor the application of the directive by:
  - assessing the cybersecurity and security policies of providers of essential services;
  - supervising digital service providers;
  - participating in the work of the cooperation group (comprising network and information security (NIS) competent authorities from each of the EU countries, the European Commission and the European Union Agency for Network and Information Security (ENISA));
  - informing the public where necessary to prevent an incident or to deal with an ongoing incident, while respecting confidentiality;
  - issuing binding instructions to remedy cybersecurity deficiencies.

The CSIRTs are responsible for:
- monitoring and responding to cybersecurity incidents;
- providing risk analysis and incident analysis and situational awareness;
- participating in the CSIRTs network;
- cooperating with the private sector;
- promoting the use of standardised practices for incident and risk-handling and information classification.

[65] https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32016L1148&qid=1634572298830

Transposition into Dutch law:

- Besluit van 30 oktober 2018 tot aanwijzing van het CSIRT voor digitale diensten en tot vaststelling van het tijdstip van inwerkingtreding van de Wet en het Besluit beveiliging netwerk- en informatiesystemen Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 389; Publication date: 2018-11-08; Page: 00001-00002
- Besluit beveiliging netwerk- en informatiesystemen. Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 388; Publication date: 2018-11-08; Page: 00001-00019
- Wet van 17 oktober 2018, houdende regels ter implementatie van richtlijn (EU) 2016/1148 (Wet beveiliging netwerk- en informatiesystemen). Official publication: Staatsblad (Bulletin des Lois et des Décrets royaux); Number: 387; Publication date: 2018-11-08; Page: 00001-00012

Wageningen Environmental Research
P.O. Box 47
6700 AA Wageningen
The Netherlands
T 0317 48 07 00
wur.eu/environmental-research

Report 3357
ISSN 1566-7197

To explore the potential of nature to improve the quality of life

Wageningen Environmental Research
P.O. Box 47
6700 AB Wageningen
The Netherlands
T +31 (0) 317 48 07 00
wur.eu/environmental-research

Report 3357
ISSN 1566-7197

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 7,600 employees (6,700 fte) and 13,100 students and over 150,000 participants to WUR's Life Long Learning, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.