Data in Brief 60 (2025) 111557

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Perspective Article

Reference framework for metadata description to commensurate data in grain production

Katariina Pussi[a,b,*], Petri Linna[c], Pasi Suomi[a], Kim Kaustell[a], Liisa Pesonen[a]

[a] Natural Resources Institute Finland (Luke), Production Systems, Latokartanonkaari 9, 00790 Helsinki, Finland
[b] LUT University, Department of Physics, School of Engineering Science, 53851, Lappeenranta, Finland
[c] Tampere University of Technology, Signal Processing Laboratory, Pori, Finland

ARTICLE INFO

Article history:
Received 25 November 2024
Revised 21 March 2025
Accepted 8 April 2025
Available online 14 April 2025

Dataset link:
Grain_production_data_samples (Original data)

Keywords:
Agriculture
Smart farming
Grain production
Data
Metadata
Data space

ABSTRACT

Data spaces will bring the need to harmonize the farm collected data for better interoperability. Attention needs also to be paid to data accessibility, since the value of data is strongly linked to its use. The evolving data space technologies will bring service providers that help farmers to translate farm born datasets to data products that can have measurable value. Data products can then be published to the data space catalog, allowing other people to discover and consume them. Data used for data products, decision-making, reporting, or analysis should be reliable and trustworthy. Common metadata standards, catalogs and ontologies will help to achieve this goal. The scope of this paper is to discuss requirements for metadata in grain production.

1. Introduction

Grain production is indispensable for meeting the nutritional needs of the global population and supporting the economy, especially in agricultural regions [1–3]. Efficient land use and

* Corresponding author at: Natural Resources Institute Finland (Luke), Production Systems, Latokartanonkaari 9, 00790 Helsinki, Finland.
E-mail address: katariina.pussi@luke.fi (K. Pussi).

https://doi.org/10.1016/j.dib.2025.111557

sustainable practices are key to balancing increasing production demands with environmental considerations. Data and metadata are fundamental for ensuring that grain production remains efficient, sustainable, and adaptable to future challenges. In the context of food security and climate change, data-driven approaches enable us to optimize agricultural practices, monitor environmental impact, and make informed decisions that benefit farmers, consumers, and global stakeholders. The future of grain production depends not just on the quantity of grain produced but on how we manage the vast amounts of data generated throughout the production process. By ensuring that data is collected, shared, and interpreted accurately, using proper metadata standards, we can create a more resilient, sustainable, and productive global agricultural system.

The 21st century agricultural system produces and depends on large amounts of data [4,5] and different user groups are involved in the process of data accumulation. Tools for data entry at the farm vary. Modern farms collect vast amounts of data automatically, using sensors and machinery. Manually input data are also still needed. Several platforms exist for collecting and managing farm data [6–8]. Data and data tools offer major benefits for farmers. Data can, for example, help farmers to better manage their operations and identify efficiencies that lead to higher productivity and profitability, lower input costs, and optimized chemical use.

Sustainable data management and analysis demands for “data about data” - or metadata. Metadata is data that fully describes the data and the areas they represent, allowing the user to make informed decisions [9]. Metadata gives answers to questions such as “who, what, when, where, why and how?”. Completeness of metadata refers to their sufficiency to fully describe a resource covering all its possible aspects [10]. Many metadata standards have been developed by scientific communities for distinct knowledge fields [11] along with controlled vocabularies, which provide a consistent way to describe data.

Agricultural data models, like Esri [12] and farmOS [13], are often related to specific data platforms created to help farmers to track their operations through data and combine data coming from different sources. Metadata and data standards are proposed and used at the processing [14] and higher levels of food chain [15]. The need for going deeper in standardizing data and metadata comes from the desire to share data created on privately owned farms and potentially turn it into a business. Data spaces are trust environments that allow farmers to share data securely and transparently, fostering collaboration and enabling new business opportunities.

The state of the art in data and metadata schemes related and applicable to agriculture is evolving rapidly, driven by technological advancements in data collection, sharing, and analysis. These schemes are becoming more sophisticated, incorporating a range of disciplines, including environmental science, economics, and policy. The trend is towards greater standardization and interoperability, allowing stakeholders across the agricultural value chain to collaborate more effectively. The lack of a single "universal" standard may be due to the diversity of stakeholders involved, the range of grain types and production practices, and regional variations in agriculture. This paper demonstrates how these standards and frameworks can work together to create a more integrated approach to data sharing and metadata management by introducing an exemplary metadata framework tailored for grain production, outlining key elements such as data reliability, completeness, and accuracy.

2. Metadata Initiatives Relevant to Agriculture

There are three notable metadata initiatives in the agriculture area [16]; The Agricultural Information Management Standards (AIMS) [17], which is a knowledge sharing platform that makes scientific information and digital data on food and agriculture available and accessible worldwide, The Agricultural Metadata Element Set (AgMES) [18], which describes metadata elements to be used for document-like information objects, for example publications, articles, books, web sites, papers, etc., and AGROVOC Multilingual Thesaurus (AGROVOC) [19,20], which offers a structured collection of agricultural concepts, terms, definitions and relationships which are used to unambiguously identify resources.

Several ontologies such as FoodWiki [21], AGROVOC, NAL Agricultural Thesaurus [22], Crop Ontology [23] and FoodOn [24] have been proposed to extract the semantics of food and agricultural data to share and reuse agriculture knowledge. In addition, sensor ontologies such as sensor node ontology [25], sensor-data ontology [26] and Sensor Model Language [27] are used for semantic interoperability of data collected from IoT devices. Agricultural data has often spatial aspects. The International Organization for Standardization (ISO) has multiple standards related to spatial metadata and metadata in general, ISO 19115, ISO 19119, ISO 19139, ISO 19110 and ISO 19157. ISO 19115 addresses the general definitions of geographic information metadata, ISO 19119 focuses on metadata for geographic information services, ISO 19139 provides the technical implementation (XML schema) for metadata, ISO 19110 describes feature catalogues and ISO 19157 addresses data quality. Dublin Core has a long history of going back to the 1990s, as a pioneering framework in defining metadata elements. The Dublin Core Metadata Initiative (DCMI) defines 15 core elements, known as Simple Dublin Core, which are also specified in standards such as ISO 15836. Nowadays there is a long list of elements [28] known as Qualified Dublin Core.

Metadata catalogs and data-sharing frameworks are becoming increasingly important in other sectors, such as transportation, to facilitate data management, standardization, and interoperability. For example, in accordance with the EU ITS Directive 2010/40/EU, EU Member States are mandated to establish National Access Points (NAPs) to facilitate the management, standardization, and sharing of traffic and travel data across the European Union. The catalogue is designed to align with the INSPIRE and DCAT-AP specifications [29].

3. Paradigm Shift in Data Economy

There are many reasons to manage agricultural metadata, e.g. EU regulation, farm management and decision support, artificial intelligence, and data spaces. On farm level, managing data is of increasing importance since farm sizes are growing, decision making is becoming more complex and globally linked, and government and investors requirements for data sharing are increasing. Also, use of subcontractors and robots in farm works demands for efficient data sharing. The development of artificial intelligence-based tools or analyses for agriculture demands for large amounts of data, which emphasizes the role of metadata and standardization as it facilitates automatic processing of quality verified data. Perhaps the most significant thing, however, is that data spaces strongly attach metadata to its economic significance, i.e. monetization [30]. Without it, the data has no market value.

In the European Union, data regulation [31] and aspiration to enhance data enabled business by data spaces will significantly change data sharing, innovation and sovereignty. With distributed data networks data spaces increase the demand for describing data and its quality in a useful way, and to make it available to network participants. The role of metadata increases significantly, because new metadata marketplaces enable data to be found from catalogues and data demand and supply to meet. The collision of metadata and data users' needs opens up new possibilities for data reuse.

In the past decade, EU legislation has developed strongly regarding data. The development of standards is related to the EU's agendas, but more importantly to international cooperation. The portability of data between different countries is an important development angle. In the new operating environment organizations may be connected to several different data spaces which might mean that some level of consistency in data definitions will be required or at least recommended. An example of this is the soil health directive [32], where one of the goals is to get a unified view of the condition of the soil on EU level causing the need to improve comparability of soil health indicators between different countries.

Since data spaces are in an initial phase of definition, new research is needed to meet their requirements. Several efforts define guidance toward data space implementation, such as reference architectures and frameworks [33]. Noardo et al. [34] compare the available solutions, providing the mapping and integration of the proposed blueprints to available interoperable stan-

dards. A robust metadata framework is one of the foundational components of data space [30]. Metadata descriptions are also an essential part of building data ecosystems. A set of open-source software components for generating high-quality metadata has been presented e.g. by Conde et al. [35]. Wamhof et al. [35] developed a data service in the Agri-Gaia project using AGROVOC terminology for data retrieval. This approach does not require a predefined data structure but ensures accessibility if the terminology is correctly applied. Agri-Gaia project aims to create an AI ecosystem for the mid-sized agricultural and food industry based on GAIA-X, along with a standardized vocabulary for semantic data and algorithm description, enabling cross-manufacturer solutions. The importance of standards, vocabularies, catalogs and metadata has been well recognized and various other use-case examples are being implemented on a broad front [30,37–39].

The development of data spaces is generally fragmented, and this is particularly evident in the agricultural sector [40]. Metadata has been implemented in a variety of ways; on one hand, there are implementations that adhere to established standards, while on the other, there are entirely custom solutions. In between, some implementations use common terminology but lack full standardization. To support progress and foster a more coherent development of data spaces, an initiative called SIMPL has been launched [41]. SIMPL is the European Commission's initiative aimed at advancing the deployment and interoperability of various sectoral data space initiatives. SIMPL-Labs provides a comprehensive platform for developing and testing new data products. It is important to note that the SIMPL program began in 2024, so the work is still in its early stages. This development environment holds significant promise as a platform for testing components designed for data spaces, including the data product discussed in this paper. SIMPL-Open includes openly developed components designed not only to create more standardized implementations but also to accelerate the development of data spaces. This aligns with the objectives of this paper, supporting future testing of the data product's functionality, with the goal of eventually integrating it with data catalogs and marketplaces in the coming years.

4. Method Description

The reference frame of the study was grain production. The basic processes in Finnish grain production include preparation of soil, sowing (or often combined sowing and fertilization), adding fertilizers and manure, protection from weeds or pests, harvesting, drying and storage (see Fig. 1). The grain is then sold and transported to domestic consumption or international markets. Each process produces and uses data in modern agriculture. Grain producers are involved in a data ecosystem of various stakeholders from farm input and technology providers to grain buyers and the rest of the food chain (see Fig. 2).

Sample datasets that were used during development and testing of our metadata framework were collected from Natural Resources Institute Finland farm in Jokioinen, Finland. Our example process of grain production was performed on a 10 ha oats field for which the agricultural and

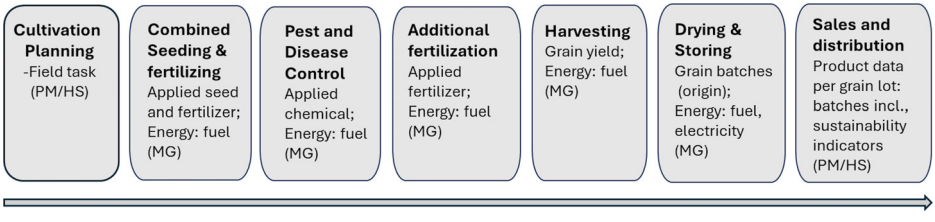

Fig. 1. Grain production process: Chain of processes in grain production during a growing season and created specific data types in addition to generic ones, i.e. identity codes, time, location and area information. Input and output materia data include quantity and specific quality information, i.e. grain species and variety, fertilizer or chemical type, grain moisture, etc. PM = process-mediated, HS = human-sourced, MG = machine-generated (according to Wolfert et al. [4]).

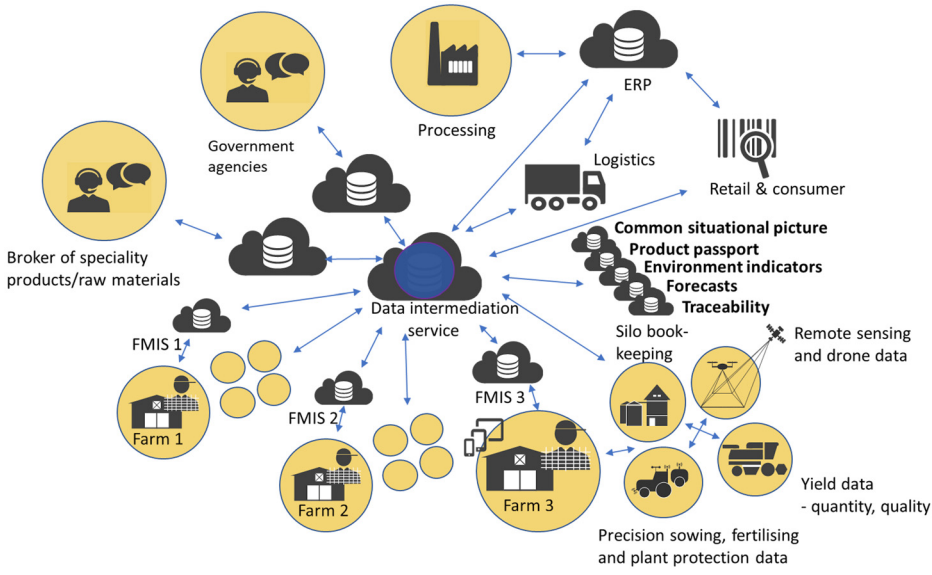

Fig. 2. Grain chain data ecosystem: Grain chain data ecosystem includes various stakeholders from farm input and technology providers to grain buyers and the rest of the food chain.

Table 1
Example data sheet for a field in Jokioinen Finland. Variables column lists example variables that can be extracted from data sources. Data samples from different data sources used in this work are included in the dataset [50]. The data samples are not necessarily from the same processes as listed in the table but are produced with the same machinery.

LUKE JOKIOINEN FIELD DATA SHEET			
FIELD ID		169-02384-37	
NAME		PV	
YEAR		2023	
AREA		10,69 ha	
CULTIVAR NAME		Oats Taika	
	DATE	VARIABLES	DATA SOURCE
SOIL SAMPLE	2.9.2022	Soil type	Soil Fertility analysis
TILLAGE	29.5.2023	Fuel consumption	ISOBUS task data, Valtra Connect
SEEDING + FERTILIZING	30.5.2023	Seed, Fertilizer, Fuel consumption	ISOBUS task data, Valtra Connect
SPRAYING	27.6.2023	Chemicals, Fuel consumption	ISOBUS task data, Valtra Connect
HARVEST	27.10.2023	Yield, Fuel consumption	CERES yield mapping
DRYING	27.10.2023	Energy consumption	Dryer process data

data details are shown in Table 1. ISOBUS-task files were used to plan and implement the field operations, Valtra Connect [42] was used to monitor fuel consumption and Ceres yield monitor [43] was used for yield mapping. The field operations included tillage prior to the seeding. Fertilization was done simultaneously with seeding which is a common practice in Finland, and the field was sprayed once against weeds. Valtra T-163 tractor [44] equipped with Junkkari Maestro 4000 Plus seed-driller [45] and Amazone UF-1501 sprayer [46] was used for seeding, fertilizing and spraying. The different ISOBUS equipment were controlled with Topcon X-35 terminal [47]. Grain was harvested with Sampo Comia C6 [48]. After the harvest the grain was dried in Mepu grain dryer [49] using oil as an energy source. Samples of data collected during each farming process are included in the dataset [50]. The role of the dataset is to give examples of the different types of data that are produced during grain growing season.

Our approach to creating an exemplary metadata framework was technology-based requirements engineering in which the design context was created by defining regulating factors such as user, purpose of use or operating principle. Majority of the farm born data is produced by machines or sensors. The data formats and accuracies vary depending on the equipment used, and interpreting the raw data often requires intensive studying of operation manuals etc. The need for metadata thus became evident when the grain production case was monitored at the level of data and its further processing and reuse. At the same time, it also became very clear that the existing vocabularies were only sufficient up to a certain point - there were shortcomings and gaps. However, the existing vocabularies provided a structure for further development and ideas.

The work was started by searching standard description for machine, or sensor-based measurement data. At the beginning of the metadata engineering process, we looked for ontology framework from Sensor Model Language (SensorML) [27] and vocabulary and classifications from LusTre [51], AGROVOC [19] and AFO [52]. SensorML was chosen because the models and schema within the core SensorML specification provide a "skeletal" framework for describing processes, aggregate processes, and sensor systems, which fit well for processes involved in grain production. These desired features were not included in other frameworks explored. Various datasets, like tractor ISOBUS data and grain dryer process data, were tabulated and ways to describe properties and reliability of different measurements were examined. After time the data tables evolved towards our goal; a minimum viable set of metadata fields to be used to describe accuracy and reliability of components associated with different datasets produced in grain production.

The final metadata framework uses ontology framework from SensorML and FAIR Principles [53] as baseline and amended it with INSPIRE metadata elements [54], since geospatial reference is important in farm data. At the end we added some case specific metadata fields, like reliability and spatial resolution, to better describe our datasets. The datasets produced in grain production process vary between years and farms, depending on the cultivation tasks performed and equipment used. We tested the feasibility of our framework against data produced in Luke Jokioinen research farm. The equipment and farming practices there represent quite well a typical Finnish farm that has done some investments to modern farming technology and data production capability. Metadata for some of our test datasets are shown in Table 5 and Table 6, and in the Appendix.

The main benefit of using our own metadata model instead of SensorML or INSPIRE or any other existing metadata models is that the accuracy of the cultivation data depends on many different variables. Each farm has their own processes and ways of working. There are no standard equipment or sensor for measuring data etc. Adding e.g. reliability parameter to metadata helps data user to decide if the data produced in some cultivation process is accurate enough for e.g. carbon footprint calculation or some other desired purpose. At the same time, our proposed method ensures compliance with existing standards, enabling interoperability while allowing the flexibility needed for farm-specific variations. The reliability parameter is described in more detail in Chapter 5 and in Table 4.

5. Metadata Framework for Grain Production

To enhance data use and interoperability in crop production ecosystem, ecosystem participants need to understand what metadata needs to be produced and what should be available to understand a resource. Since metadata must be understandable by anyone using it, standardization of metadata is important. A metadata standard is a requirement intending to establish a common understanding of the meaning of the data and ensuring correct and proper use and interpretation of the data by its owners and users.

In Tables 2 and 3 we propose a preliminary list of metadata fields toward building metadata standards specific for the data relevant to grain production. Metadata schema starts with parameters describing the overall dataset properties, like keywords, language and location, and

Table 2
Overall description of the dataset. Ontology framework from SensorML and INSPIRE metadata elements have been used as guidelines.

DESCRIPTION	A textual description of the dataset
UNIQUE ID	A globally unique identifier for the dataset
NAME	A common name for the dataset
KEYWORDS	A list of keywords that enhance the findability of this particular dataset
DATE	Date of dataset creation
LEGAL CONSTRAINS	A variety of licensure and other regulatory requirements
CONTACTS	Contact information for dataset owner. This can include name, e-mail, phone number, physical address etc.
LOCATION	Geographical area or location where data has been collected.
TEMPORAL COVERAGE	Temporal coverage for data included in dataset.
COORDINATE SYSTEM	Reference coordinate system for **LOCATION** and spatial data
LANGUAGE	The language in which the document is written. The value of this attribute will be a two-letter code which conforms to ISO 639-1
IDENTIFIERS	Identifiers are primarily used for search and discovery. Identifiers can include e.g. **CATEGORY** and **FORMAT** fields that help to find field specific data in desired format.
RELIABILITY	Describes the overall reliability level for the dataset. Individual data elements can have different levels of reliability. Exemplary reliability classification is shown in Table 4
DOCUMENTATION	An external document related to dataset creation, equipment or other relevant information.
NUMBER OF VARIABLES	Number of different data variables included in dataset

Table 3
Data variable specific metadata fields. Ontology framework from SensorML and INSPIRE metadata elements have been used as guidelines. * If the spatial resolution of data is not constant, which is the case e.g. for raw ISOBUS-task data, average value is reported.

CAPABILITIES	Property information that further clarify or qualify the data variable. This can include e.g. **MEASUREMENT PROPERTIES (TYPE, SAMPLING FREQUENCY, START DATE, END DATE), SPATIAL RESOLUTION*, COVERAGE** / **REPRESENTATIVENESS** and **RELIABILITY**.
CLASSIFIERS	Describes various aspects of the data variable. These might include e.g. **EQUIPMENT** used, **ACCURACY** of the equipment, or its **INTENTED APPLICATION**.
CHARACTERISTICS	Important for understanding the nature of the data variable e.g. **UNITS** and **TARGET**.

Table 4
Reliability of the data source classified in five different categories.

RELIABILITY
1= Data verified on the basis of measurements
2= Data based partly on assessment, supplemented by measurements
3= Non verified data based on expert assessment/manufacturer's guideline value
4= Expert assessment
5= Default value

reliability (see Table 2). The data variables, such as fuel consumption in field operations, included in the dataset are then described more detailed (see Table 3). Data reliability refers to how well the data reflect reality for a given use-case. In our grain cultivation framework we have classified data reliability to five categories (see Table 4). The machine measured data is considered to be most reliable. This would include for example amount of seed used which is recorded by ISOBUS task controller during seeding. Second category of reliability is measurement data that is complemented with assesment. This could be e.g. fuel consumption measurement from farm tank divided between tractor work done during one day. Third category is machine data that is not measured for the use-case but e.g. taken from manufacturers manual. This could be e.g.

Table 5
Example metadata for field operator fuel consumption including 4 variables: tillage, seeding & fertilizing, spraying and harvest. Variable specific metadata for Tillage is shown in Table 6. (**) See Table 4 for reliability classification.

DESCRIPTION		Dataset for total fuel consumption from field operations including 4 variables: tillage, seeding & fertilizing, spraying and harvest. Data was collected with Valtra Connect.
UNIQUE ID		e687cb1e-d6ce-4af5-9eab-2dbd14161c9a
NAME		Fuel Consumption
KEYWORDS		Fuel consumption, Carbon footprint, Agriculture, Smart Farming
DATE		1.11.2023
LEGAL CONSTRAINS		Use of this information is without limitation
CONTACTS		N.N@luke.fi
LOCATION		[60.805162808°N, 23.490085206°E]
TEMPORAL COVERAGE		29.5.2023 – 28.10.2023
COORDINATE SYSTEM		WGS 84
LANGUAGE		En
IDENTIFIERS		
	CATEGORY	Agriculture
	FORMAT	JSON
RELIABILITY (**)		1
DOCUMENTATION		Manufacturer manual
NUMBER OF VARIABLES		4

Table 6
Fuel consumption measurement specific metadata for tillage. The other three variables in Fuel consumption dataset (seeding & fertilizing, spraying and harvest) would each have similar metadata. (**) See Table 4 for reliability classification.

Variables			
Tillage			
DESCRIPTION	Fuel consumption during tillage measured with Valtra Connect.		
CAPABILITIES			
	MEASUREMENT PROPERTIES		
		TYPE	Volumetric fuel flow meter
		SAMPLING FREQUENCY	2 Hz
		START DATE	29.5.2023
		END DATE	29.5.2023
	SPATIAL RESOLUTION	0.005 ha	
	COVERAGE, REPRESENTATIVENESS %	100	
	RELIABILITY (**)	1	
CLASSIFIERS			
	INTENTED APPLICATION	Fuel consumption measurement	
	EQUIPMENT	Valtra Connect	
	CALIBRATION DATE	1.1.2020	
	ACCURACY	0.1 l / h	
	SOURCE FOR ACCURACY	Manufacturer manual	
CHARACTERISTICS			
	UNITS	l/h	
	TARGET	Field	

grain dryer energy consumption per hour in certain circumtances. Fourth category is expert assesment, e.g. soil type from manual assesment. The most unreliable category, default value, refers e.g. to use of national or regional averages for yield. Each variable in the dataset can have different reliability classification due to the fact that some values needed e.g. for the carbon footprint calculation are not or can not be directly measured. One example of this kind of variable is grain dryer energy consumption, which can be complicated parameter, depending on the drying energy source used. The reliability and accuracy of each data variable in a dataset heavily influence the reliability and accuracy of the indicator or other output that the dataset is used for. Tables 5 and 6 show metadata example for diesel consumption in field operations. Other three metadata examples for datasets formed from Table 1 parameters (Soil type, N fertilization and Grain

drying energy consumption) are shown in Appendix (Table A1, A2 and A3). These exemplary metadata tables demonstrate how our scheme is suited for different farm born datasets.

6. Discussion

Economic activity centered around data collection, analysis, and sales is expanding in all sectors including agriculture. There are many different stakeholders in grain chain e.g. agricultural input companies (like seed and fertilizer companies), agricultural retailers, farmers, agricultural credit institutions, government agencies, crop consultants and advisors, aggregators, processors, distributors, transportation companies and ingredient manufacturers. The need for better data and metadata management applies to them all. The technical framework for data and metadata sharing in agriculture has been discussed e.g. in Agri-Gaia project [36], but the actual content of agricultural metadata has received less attention.

Rich metadata enables better data management and sharing, and it should be considered as an important component of data and created during data development. Metadata also makes data more transparent and defines the usability and value of the datasets. For data trading, metadata is one of the most important factors [55], since it e.g. can provide a complete profile and detailed insight into each data set. Metadata can also facilitate data protection and data governance. It gives to data its markers or characteristics that help to understand what kind of restrictions, controls, and handling requirements apply.

Any metadata framework should be such that computational systems can easily find, access, interoperate, and reuse data. Special attention should be paid to defining the reliability of the dataset, since it further determines the value and usability. In our example we have classified data reliability to five categories (Table 4). In this paper our exemplary metadata framework is shown in table format for better human readability, but it could be easily expressed e.g. in XML language which makes it machine readable.

The framework was built using data generated by crop production as a test case, but its use is not limited to agricultural data. Since the proposed metadata framework is built on SensorML ontology and INSPIRE metadata elements, it is best suited for reporting metadata for static spatial datasets generated by machines, sensors or other devices. It can also be used for static non-spatial datasets or datasets collected even by hand. The proposed framework is not usable for dynamic datasets, since our framework does not include metadata version management fields.

7. Conclusions

The evolution of metadata schemes for grain production is driven by technological progress, cross-disciplinary integration, and the need for better collaboration among stakeholders. While there is no universal standard, the integration of various existing standards is gradually leading to a more cohesive and interoperable data ecosystem. As these standards continue to evolve, the future of grain production will likely see enhanced data-driven decision-making, greater transparency, and more sustainable practices. These improvements will help address global challenges such as food security, climate change, and the optimization of agricultural resources.

The paper aims to foster discussion and drive standardization in metadata capture and sharing. It presents an exemplary metadata framework for grain production, demonstrating the feasibility of a common schema at the farm level. The framework effectively communicates dataset reliability and accuracy through proper metadata description and has been tested with various grain production datasets. Dataset [49], containing diverse data samples, is linked to the paper. By enabling flexible metadata modeling while ensuring compliance with existing standards, our approach enhances data usability, accuracy, and interoperability for applications such as carbon footprint calculations and sustainability assessments.

The concept of data spaces strongly attaches metadata to data monetization [30]. Besides the financial incentives, the efficiency gained from sharing data, by reducing redundancies and

speeding up processes, could be a major benefit to farmers in making their operations more streamlined and effective. Farmers, like other entrepreneurs, need to have user friendly and compatible data tools to mastermind a successful business. Common metadata standards, catalogs and ontologies will help to achieve this goal. Machines and equipment represent an important part of a farmer's data production capacity and therefore, farmers will in the future expect that the new machine/device purchased will automatically add meta-descriptions to the data it produces. In this context, it is also important to further develop metadata for dynamic datasets in agriculture, as well. Next step would be to test the feasibility of described metadata framework in the metadata marketplaces in data spaces. Further metadata research should also include the viewpoint of automatic/dynamic AI assisted creation of data products to enable real-time metadata updates.

CRediT Author Statement

Katariina Pussi: Writing - Original Draft. **Petri Linna:** Conceptualization, Methodology, Writing - Original Draft. **Pasi Suomi:** Conceptualization, Methodology, Investigation, Data Curation, Writing - Original Draft. **Kim Kaustell:** Supervision, Writing - Review & Editing. **Liisa Pesonen:** Conceptualization, Methodology, Writing - Review & Editing.

Funding Sources

This work was supported by Business Finland [grant number 6199/31/2023].

Data Availability

Grain_production_data_samples (Original data) (Mendeley Data).

Acknowledgments

We acknowledge Kirsi Usva (Natural Resources Institute Finland) for valuable discussions related to sustainability indicators. This paper reports results from research conducted as part of the OurData project, an initiative aimed at developing a functional data ecosystem for the grain industry. Coordinated by the Natural Resources Institute Finland (Luke) and involving multiple partners, including businesses and research institutions, the project focuses on creating a transparent, fair, and efficient framework for data exchange in data space environment across the grain value chain.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Table A1
Metadata for N fertilization applied during combined seeding and fertilization. (**) See Table 4 for reliability classification.

DESCRIPTION		Dataset for N fertilization. Data source ISOBUS task file from combined seeding and fertilization. Fertilizer agent: YARA MILA Y 26.		
UNIQUE ID		e687cb1e-d6ce-4af5-9eab-2dbd14161c9a		
NAME		N fertilization		
KEYWORDS		Nitrogen, Fertilization, Carbon footprint, Agriculture, Smart Farming		
DATE		30.5.2023		
LEGAL CONSTRAINS		Use of this information is without limitation		
CONTACTS		N.N@luke.fi		
LOCATION		[60.805162808°N, 23.490085206°E]		
TEMPORAL COVERAGE		30.5.2023 - 30.5.2023		
COORDINATE SYSTEM		WGS 84		
LANGUAGE		En		
IDENTIFIERS				
	CATEGORY	Agriculture		
	FORMAT	ISOXML		
RELIABILITY (**)		1		
DOCUMENTATION		Manufacturer manual		
NUMBER OF VARIABLES		1		
VARIABLES				
	Fertilization			
		DESCRIPTION	Fertilizer rate from ISOBUS task data. N rate can be calculated from used fertilizer agent.	
		CAPABILITIES		
			MEASUREMENT PROPERTIES	
			TYPE	Precision seed-driller
			SAMPLING FREQUENCY	2 Hz
			START DATE	30.5.2023
			END DATE	30.5.2023
			SPATIAL RESOLUTION	0.005 ha
			COVERAGE, REPRESENTATIVENESS %	100
			RELIABILITY (**)	1
		CLASSIFIERS		
			INTENTED APPLICATION	Precision fertilization
			EQUIPMENT	Junkkari Maestro seed-driller
			CALIBRATION DATE	1.1.2020
			ACCURACY	1 kg/ha
			SOURCE FOR ACCURACY	Manufacturer manual
		CHARACTERISTICS		
			UNITS	kg/ha
			TARGET	Field

Table A2
Metadata for grain dryer process data with direct energy consumption measurement. (**) See Table 4 for reliability classification.

DESCRIPTION	Energy consumption measurement from grain drying process. Energy source is diesel. The diesel generator uses 0.4 liters of fuel for every kWh produced.			
UNIQUE ID	e687cb1e-d6ce-4af5-9eab-2dbd14161c9a			
NAME	Grain dryer process data with direct energy consumption measurement			
KEYWORDS	Grain drying, Energy, Diesel, Carbon footprint, Agriculture, Smart Farming			
DATE	27.10.2023			
LEGAL CONSTRAINS	Use of this information is without limitation			
CONTACTS	N.N@luke.fi			
LOCATION	[60.805162808°N, 23.490085206°E]			
TEMPORAL COVERAGE	27.10.2023			
COORDINATE SYSTEM	WGS 84			
LANGUAGE	En			
IDENTIFIERS				
	CATEGORY	Agriculture		
	FORMAT	JSON		
RELIABILITY (**)	1			
DOCUMENTATION	Manufacturer manual			
NUMBER OF VARIABLES	1			
VARIABLES				
	Energy consumption			
		DESCRIPTION	Dryer fuel consumption for a drying batch converted to energy consumption (kWh).	
		CAPABILITIES		
			MEASUREMENT PROPERTIES	
				TYPE Fuel flow meter
				SAMPLING FREQUENCY 1 Hz
				START DATE 27.10.2023
				END DATE 27.10.2023
			SPATIAL RESOLUTION	-
			COVERAGE, REPRESENTATIVENESS %	100
			RELIABILITY (**)	1
		CLASSIFIERS		
			INTENTED APPLICATION	Fuel consumption real-time monitoring
			EQUIPMENT	Autonomous flow meter DFM
			CALIBRATION DATE	1.1.2020
			ACCURACY	1 kWh
			SOURCE FOR ACCURACY	Manufacturer manual
		CHARACTERISTICS		
			UNITS	kWh
			TARGET	Drying batch

Table A3
Metadata for soil type determined by sieving and sedimentation. (**) See Table 4 for reliability classification.

DESCRIPTION		Soil type classification from soil fertility samples. 15 samples have been collected to represent the field soil variability.		
UNIQUE ID		e687cb1e-d6ce-4af5-9eab-2dbd14161c9a		
NAME		Soil type		
KEYWORDS		Soil, Carbon footprint, Agriculture, Smart Farming		
DATE		2.9.2022		
LEGAL CONSTRAINS		Use of this information is without limitation		
CONTACTS		N.N@luke.fi		
LOCATION		[60.805162808°N, 23.490085206°E]		
TEMPORAL COVERAGE		2.9.2022		
COORDINATE SYSTEM		WGS 84		
LANGUAGE		En		
IDENTIFIERS				
	CATEGORY	Agriculture		
	FORMAT	JSON		
RELIABILITY (**)		1		
DOCUMENTATION		Soil fertilization service producing company		
NUMBER OF VARIABLES		1		
VARIABLES				
	Soil type			
		DESCRIPTION	Organoleptic determination of soil type and soil fertility.	
		CAPABILITIES		
			MEASUREMENT PROPERTIES	
				TYPE Organoleptic determination
				SAMPLING FREQUENCY -
				START DATE 2.9.2022
				END DATE 2.9.2022
			SPATIAL RESOLUTION	0.05 ha
			COVERAGE, REPRESENTATIVENESS %	50
			RELIABILITY (**)	4
		CLASSIFIERS		
			INTENTED APPLICATION	Soil type determination.
			EQUIPMENT	-
			CALIBRATION DATE	-
			ACCURACY	-
			SOURCE FOR ACCURACY	-
		CHARACTERISTICS		
			UNITS	-
			TARGET	Field

References

[1] World Bank Group, "Cereal production (metric tons)," 2022. [Online]. Available: https://data.worldbank.org/indicator/AG.PRD.CREL.MT. [Accessed 2025].
[2] The Food and Agriculture Organization (FAO), "The food and agriculture organization (FAO)," [Online]. Available: https://www.fao.org/home/en. [Accessed 2024].
[3] Our World in Data, "Global agricultural land use by major crop type," [Online]. Available: https://ourworldindata.org/grapher/global-agricultural-land-use-by-major-crop-type. [Accessed 2025].
[4] S. Wolfert, L. Ge, C. Verdouw, M.-J. Bogaardt, Big data in smart farming – a review, Agric. Syst. 153 (2017) 69–80, doi:10.1016/j.agsy.2017.01.023.
[5] L. Ahmad, F. Nabi, Agriculture 5.0: Artificial Intelligence, IoT, and Machine Learning, CRC Press, 2021.
[6] L.A. Pesonen, F.K.-W. Teye, A.K. Ronkainen, M.O. Koistinen, J.J. Kaivosoja, P.F. Suomi, R.O. Linkolehto, Cropinfra – an internet-based service infrastructure to support crop production in future farms, Biosyst. Eng. 120 (2014) 92–101 4, doi:10.1016/j.biosystemseng.2013.09.005.
[7] J. Backman, R. Linkolehto, M. Koistinen, J. Nikander, A. Ronkainen, J. Kaivosoja, P. Suomi, L. Pesonen, Cropinfra research data collection platform for ISO 11783 compatible and retrofit farm equipment, Comput. Electr. Agric. 166 (2019) 105008 11, doi:10.1016/j.compag.2019.105008.
[8] M. Reddy, R. Reshma, S. Kumar, S. Krithika, S. Manokaran, Improving the efficiency of farm management using advanced software technology, SGSES 1 (1) (2021).
[9] C. Santos, A. Riyuiti, An overview of the use of metadata in agriculture, IEEE Latin Am. Trans. 10 (1) (2012) 1265–1267 1, doi:10.1109/TLA.2012.6142471.
[10] M. Margaritopoulos, T. Margaritopoulos, I. Mavridis, A. Manitsaris, Quantifying and measuring metadata completeness, J. Am. Soc. Inform. Sci. Technol. 63 (4) (2012) 724–737 4, doi:10.1002/asi.21706.
[11] F.M. Soares, B.C.M.D.S. Maculan, D.P. Drucker, A.M. Saraiva, Methodological principles to create a metadata extension to the Darwin core standard for agrobiodiversity data, Braz. J. Inform. Sci.: Res. Trends 14 (4) (2020) 12.
[12] Esri, "Agriculture data model," 2003. [Online]. Available: https://downloads.esri.com/support/datamodels/Agriculture/AGDM_Poster.gif. [Accessed 2024].
[13] farmOS.org, "farmOS data model," [Online]. Available: https://farmos.org/model/.
[14] E. Everz, M.S.M.G. Vaz, Application of metadata standards for grain classification, Ibero Am. J. Appl. Comput. 10 (2) (2020).
[15] GS1, "GS1 global data model," [Online]. Available: https://www.gs1.org/standards/gs1-global-data-model. [Accessed 2024].
[16] C. Bahlo, P. Dahlhaus, H. Thompson, M. Trotter, The role of interoperable data standards in precision livestock farming in extensive livestock systems: a review, Comput. Electron. Agric. (156) (2019) 459–466, doi:10.1016/j.compag.2018.12.007.
[17] J. Silva, D. Leite, M. Fernandes, C. Mena, P. Gibbs, P. Teixeira, Campylobacter spp. as a foodborne pathogen: a review, Front. Microbiol. (2) (2011) 200, doi:10.3389/fmicb.2011.00200.
[18] Food Agriculture Organization, United Nations, "Agricultural metadata element set (AgMES)," 2010. [Online]. Available: http://aims.fao.org/standards/agmes. [Accessed 2024].
[19] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, J. Keizer, The Agrovoc linked dataset, Semant. Web (4) (2013) 341–348.
[20] I. Subirats-Coll, K. Kolshus, A. Turbati, A. Stellato, E. Mietzsch, D. Martini, M. Zeng, AGROVOC: the linked data concept hub for food and agriculture, Comput. Electr. Agric. 196 (2022) 105965, doi:10.1016/j.compag.2020.105965.
[21] D. Çelik, FoodWiki: ontology-driven mobile safe food consumption system, ScientificWorldJournal (2015) 475410, doi:10.1155/2015/475410.
[22] U.S. Department of Agriculture, "NALT concept space," [Online]. Available: https://lod.nal.usda.gov/nalt/en/. [Accessed 2025].
[23] CGIAR, "Crop ontology for agriculture," 2012. [Online]. Available: https://cropontology.org/. [Accessed 2025].
[24] D.M. Dooley, E.J. Griffiths, G.S. Gosal, P.L. Buttigieg, R. Hoehndorf, M.C. Lange, L.M. Schriml, F.S.L. Brinkman, W.W.L. Hsiao, FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, NPJ Sci. Food 2 (2018) 23, doi:10.1038/s41538-018-0032-6.
[25] S. Avancha, C. Patel, A. Joshi, Ontology-driven adaptive sensor networks, UMBC Stud. Collect. (2004) 194–202, doi:10.1109/MOBIQ.2004.1331726.
[26] M. Eid, R. Liscano, A. Saddik, A novel ontology for sensor networks data, in: Proceedings of the 2006 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, La Coruna, Spain, 12–14 July 2006, 2006, pp. 75–79, doi:10.1109/CIMSA.2006.250753.
[27] "Sensor model language (SensorML)," [Online]. Available: https://www.ogc.org/standard/sensorml/. [Accessed 2024].
[28] DublinCore, "DCMI metadata terms," [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/. [Accessed 2024].
[29] P. Lubrich, J. Ansorge, M. Böhm, L. Hendriks, T. Hoffmann, J. Jaderberg, A. Kochs, C. Lüpges, L. Rittershaus, S. Schwillinsky, B. Witsch and T. Vlemmings, "EU EIP SA46 coordinated metadata catalogue, monitoring and harmonisation of national access points in Europe," EUEIP, 2019.
[30] B. Otto, M. T. Hompel and S. Wrobel, Designing data spaces: the ecosystem approach to competitive advantage., 2022.
[31] T. Bräutigam, M. Aholainen, F. Cunningham, M. Geus, F. Kukorelli, M. Toivanen, R. Aarnio, L. Halenius, T. Rastas and J. Kippo, "EU regulation builds a fairer data economy. The opportunities of the Big Five proposals for businesses, individuals and the public sector," 2022. [Online]. Available: https://www.sitra.fi/app/uploads/2022/06/sitra-eu-regulation-builds-a-fairer-data-economy.pdf. [Accessed 2024].
[32] European Commission, "Proposal for a directive on soil monitoring and resilience," 2023. [Online]. Available: https://environment.ec.europa.eu/publications/proposal-directive-soil-monitoring-and-resilience_en. [Accessed 2024].

[33] European Data Spaces Support Center, "Data spaces blueprint v1.5," 2024. [Online]. Available: https://dssc.eu/space/bv15e/766061169/Data+Spaces+Blueprint+v1.5+-+Home.
[34] F. Noardo, R. Atkinson, L. Bastin, J. Maso, I. Simonis, A. Villar, M.-F. Voidrot, P. Zaborowski, Standards for data space building blocks, Remote Sens. (16) (2024) 3824, doi:10.3390/rs16203824.
[35] J. Conde, A. Pozo, A. Munoz-Arcentales, J. Choque and A. Alonso, "Fostering the integration of European open data into data spaces through high-quality metadata.," 2024.
[36] T. Wamhof, A. Bernardi, D. Martini, M. Leinberger, A. Sinha, H. Tapken, A. Schliebitz, H. Graf, Metadata management and asset exchange in the agricultural data ecosystem of the project Agri-Gaia, Datenbank Spektrum (23) (2023) 107–115, doi:10.1007/s13222-023-00444-3.
[37] Sitra, "State of Finnish data spaces," 2024. [Online]. Available: https://www.sitra.fi/en/publications/state-of-finnish-data-spaces/. [Accessed 2024].
[38] P. Subramaniam, Y. Ma, I. Mohanty and R. C. Fernandez, "Comprehensive and comprehensible data catalogs: the what, who, where, when, why, and how of metadata management," 2023.
[39] L. Ehrlinger, J. Schrott, M. Melichar, N. Kirchmayr and W. Wöß, "Data catalogs: a systematic literature review and guidelines to implementation," pp. 148–158, doi:10.1007/978-3-030-87101-7_15, 2021.
[40] AgriDataSpace Consortium, "D3.3: report on data, models and interoperability solutions as technology enablers for the implementation of agricultural data space," 2024. [Online]. Available: https://agridataspace-csa.eu/deliverables/. [Accessed 2024].
[41] European Commission, "Simpl: cloud-to-edge federations empowering EU data spaces," [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/simpl. [Accessed 2024].
[42] AGCO Ltd, "Valtra connect," [Online]. Available: https://www.valtra.com/technology/connect.html. [Accessed 2024].
[43] RDS Technology Ltd, "Ceres 8000i yield monitor," 2015. [Online]. Available: https://www.rdstec.com/custom-content/uploads/2015/01/Ceres-8000i-Ag-EN.pdf. [Accessed 2024].
[44] AGCO Ltd, "Valtra T series," [Online]. Available: https://www.valtralita.lt/files/Valtra_T_Series_EN.pdf. [Accessed 2024].
[45] Junkkari Oy, [Online]. Available: https://junkkari.fi/en/combi-seed-drills/. [Accessed 2024].
[46] Amazonen-Werke, "Operators manual Amazone UF 1501," 2019. [Online]. Available: https://downloadcenter.amazone.de/file/view/252935. [Accessed 2024].
[47] Topcon, "X35 all-in-one, premium console for leading control," [Online]. Available: https://mytopcon.topconpositioning.com/ie/agriculture-gnss-and-guidance/cab-consoles/x35. [Accessed 2024].
[48] Sampo-Rosenlew Oy, "Comia C6," [Online]. Available: https://www.sampo-rosenlew.fi/combine-harvesters/comia-c6-2/. [Accessed 2024].
[49] Mepu Oy, "Grain handling solutions for agricultural industries and farms," [Online]. Available: https://www.arskagroup.com/mepu/en/. [Accessed 2024].
[50] K. Pussi, "Grain_production_data_samples," 2025. [Online]. Available: doi: 10.17632/nrhhcsc6bt.1.
[51] R. Albertoni, M. De Martino, P. Podestà, A. Abecker, R. Wössner, K. Schnitter, LusTRE: a Framework of Linked Environmental Thesauri for Metadata Management, Earth Sci. Inform. (11) (2018) 525–544, doi:10.1007/s12145-018-0344-8.
[52] Finto.fi, "AFO - natural resource and environment ontology," 2018. [Online]. Available: https://finto.fi/afo/en/index. [Accessed 2024].
[53] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. sa Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Sci. Data (2016) 160018, doi:10.1038/sdata.2016.18.
[54] European Commission, "Commission regulation (EU) No 1311/2014 of 10 December 2014 amending regulation (EC) No 976/2009 as regards the definition of an INSPIRE metadata element," 2014. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2014.354.01.0006.01.ENG. [Accessed 2024].
[55] S. Lawrenz, P. Sharma and A. Rausch, "The significant role of metadata for data market places," 2019. [Online]. Available: https://dcpapers.dublincore.org/files/articles/952141811/dcmi-952141811.pdf. [Accessed 2024].