Use of Semantic Artefacts in Agricultural Data-Driven Service Development

Silke Cuno[a] and Philipp Lämmel[b]

Fraunhofer Institute for Open Communications Systems FOKUS, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany

Keywords: Data Sharing, Agricultural Digital Integration Platforms, Semantic Artefacts, Agri-Food, Ontologies Decision Support, Interoperability, Digital Agriculture, Smart Farming.

Abstract: The paper is a survey of the resources and efforts in the field of semantic interoperability in the agricultural domain, and describes challenges and solutions for building federated digital agricultural integration platforms that provide service components based on shared and reused data. It also shows the state of the art of public semantic artefacts and their potential contribution to solving semantic interoperability within digital agricultural integration platforms by combining their main standards into common agricultural information models that can be used by developers and fit a wide range of agricultural use cases.

1 INTRODUCTION

January 2024 is marked by Europe-wide protests by farmers against bureaucratisation, subsidy cuts, environmental regulations and more. The reasons are many. Farmers say they are being asked to do more for less (Deutschlandfunk, 2024; FAO, 2021). At the first Agri-Digital Conference in Brussels in December 2023 (EC, 2023a), it was discussed that farmers' bureaucratic burdens, worries about yields, incomes and workloads could be reduced through the widespread use of digital decision support systems. For example, they could help farmers with specific crop, resource and production planning. Monitoring systems could detect early signs of disease or pests, allowing timely action to be taken and significant crop losses to be avoided. For rural areas in particular, digital services, ideally based on IT ecosystems, hold the promise of improving quality of life, ensuring geographical balance, food security, spatially balanced development, economic prosperity and the achievement of sustainability goals. An IT infrastructure shared by farmers and government agencies, providing embedded services based on an "agricultural data space", could be one solution, as both development agencies and farmers need an integrated view of the latest agricultural, economic and environmental data for analysis to facilitate their work and, for example, to maintain food security (EC, 2023e). There is a lack of decision support and reasoning systems based on integrated infrastructures of public data sources when it comes to planning to assess yield gains or what foods might become scarce. At the same time, the landscape of siloed agri-food IT services and systems is expanding. Farmers are increasingly using Farm Management Information Systems (FMIS), but the services are often isolated systems. The data produced is mostly stored locally or in closed vendor clouds, and is difficult to find and access. Solutions remain proprietary. For the EU, this hinders the development of the European Data Economy. To mitigate this, the EU Data Strategy promotes common data infrastructures with data spaces for data reuse and sharing to enable the rapid creation or integration of new innovative services. The EU's "Data Act 2023" legislation (EU, 2023) aims to: 1) ensure a fair distribution of the value of data among actors in the digital environment, 2) stimulate a competitive data market, 3) open opportunities for data-driven innovation, and 4) make data more accessible to all (EC, 2023b). These actions are supported by funding policies for the creation and sharing of findable, accessible, interoperable and reusable ("FAIR") data (Wilkinson et al., 2016; Jonquet et al., 2023). EU research initiatives and projects have produced successive results in recent years. Progress has been made towards solutions for "technical interoperability" through "reference architectures" and data-driven "data-space concepts" involving all sectors. However, data integration in agriculture still suffers from

[a] https://orcid.org/0000-0003-4796-7326
[b] https://orcid.org/0000-0002-4411-0557

Cuno, S. and Lämmel, P.
Use of Semantic Artefacts in Agricultural Data-Driven Service Development.
DOI: 10.5220/0012760400003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 346-357
ISBN: 978-989-758-707-8; ISSN: 2184-285X

numerous incompatible format specifications and encoding systems, as well as informal and undefined semantics at the data level (Subirats-Coll et al., 2022). In this domain, data integration is highly interconnected with other sciences such as meteorology, geography, space, biology, etc. Domain standards are typically not-interoperable in terms of semantics and syntax. And agriculture is producing increasing amounts of real-time raw data from multiple sources, such as soil sensors, drones, etc., which also need to be integrated. The question is how best to manage data the reuse and sharing of data in terms of "semantic interoperability" for all the different use cases. Can a "common information model" be the game changer for digital agricultural integration platforms? In this context, the numerous semantic artefacts and ontologies that have been developed over the last decades are once again the focus of interest in computer science research. To what extent can their knowledge, methods and standards support seamless data integration? Their study and reorganisation is experiencing a second renaissance in the age of data-driven service development and behavioural AI (Bochtis et al., 2022; Zeginis et al., 2023). Originally developed in the humanities, semantic artefacts were first used in IT after the turn of the century in connection with the Semantic Web and Linked Data, as reasoning tools for AI (Hoekstra, 2009). Now, in the context of agricultural digital integration platforms, it is being explored how they can be useful as components in federated service architectures for standardised data integration and, more recently, how they work in combination with AI technologies (Subramanian et al., 2024).

This paper discusses whether the content and standards of semantic artefacts, fed into standardised components of digital agricultural integration platforms as information models or similar, could be a way forward to facilitate service development and data reuse. A survey of semantic interoperability resources and efforts in the agricultural domain is presented in this paper. The paper is structured as follows: Section 2 presents related work in service architecture, Section 3 presents the initiatives on semantic artefacts and selected research projects integrating information models for service development, and finally Section 4 summarises the main findings.

2 THE DATA INTEGRATION CHALLENGE

2.1 Interoperability Aspects

The sharing and re-use of data on federated service platforms requires "interoperability". This is a critical enabler for interacting towards mutually beneficial goals, involving the sharing of information and knowledge between organisations, through the business processes they support, by exchanging data between their ICT systems (EC, DG Informatics, 2017). The "New European Interoperability Framework Model" (EIF) specifies six interoperability layers. This paper addresses semantic and technical interoperability. "Technical Interoperability" covers the applications and infrastructures that connect systems and services. Aspects include interface specifications, interconnection services, data integration services, etc. (EC, DG Informatics, 2017). "Semantic Interoperability" means that the precise format and meaning of exchanged data and information is preserved and understood throughout the exchange between the parties, in other words "what is sent is what is understood" (EC, DG Informatics, 2017). It involves the definition of data structures and elements to describe the exchange of data. One of the underlying keys to semantic interoperability is syntactic interoperability, which describes the ability to exchange data between systems (Bochtis et al., 2022). Interoperability issues are not unique to the agricultural sector.

2.2 Developers' Perspective

IT developers tasked with building integrated solutions or reusing and sharing data are faced with an uncertain landscape of systems, formats, data sources and standards. Problems arise in interpreting the meaning of this data at a semantic and syntactic level. Documentation is often scarce or, in most cases, not aligned with common data models. Developers are challenged with conceptualising different syntaxes for identifiers, finding, interpreting and accessing digital data sources in databases or repositories. Much time is spent searching for the meaning in the context of the application being programmed and organising data from databases. Developers need in-depth knowledge of the structure of digital data sets and domain knowledge of the context. Having found the data for the relevant use case, the usual way to develop an IT service is to standardise the formats, data models and vocabularies of the data used between the concrete exchanging parties of the concrete service to be developed. This requires the exchanging parties to agree on the interfaces in advance, and this process usually leads to a lengthy standardisation process that is set up only for that specific service. Typically, the developer designs the underlying data or information model from scratch. A second way of developing services, data reuse and sharing, where semantic artefacts and knowledge repositories provide the concepts, emerged around 2010. This method promises to be more effective. The assumption is that if the semantic resources (such as ontologies) based on linked data standards are available online, service developers will be able to look up the form and meaning of the data they receive at runtime and react accordingly. This second option has been implemented in early research projects such as the "iGreen Project" (DFKI, 2014). The advantage is that developers don't have to develop a separate data format or data model for their service, but can take depending on their acitivity, for example, either an instance of the reference ontology (e.g. a knowledge graph) or an ontology as input. This would come in a JSON standard for linking data (JSON-LD) for the service to be developed. The benefits are clearly seen in the reduced development effort. The disadvantages of this approach are that a) developers are not trained in the use and combination of semantic artefacts and b) understanding the ontologies, semantic resources, data formats, their interfaces and dependencies may require time-consuming analysis using specific development tools and skills. Developers still need to spend a considerable amount of time studying scientific publications, standard specifications or learning from the ontologies and their documentation in order to reuse data. Although not directly related to the developers, another major disadvantage is that the adaptation workload - which can be quite extensive - is essentially shifted to the data providers meaning that they have to encode their own data in a way that is usable in the context of the Semantic Web (e.g. Web Ontology Language (OWL)).

2.3 Reference Architectures

Reference Architectures (RAs) are conceptual frameworks that play an essential role in building ICT infrastructures. Their design principles ensure secure, reliable and interoperable solutions. They describe a standardised architecture that provides a reference framework for vertical domains, such as agriculture, mobility or related industries. An RA provides a common understanding of processes, data structures and the underlying technologies in general. They provide a standardised blueprint for the design and implementation of ICT systems within specific domains. Most importantly, they provide guidelines for the integration of different technologies, data sources and stakeholders, facilitating the seamless flow of information and enabling efficient decision-making processes (LeanIX, 2023). The development of RAs began in the nineties in the telecommunications domain. The design principles were service orientation (SOA), object orientation, distribution, decoupling of software components and separation of concerns to ensure interoperability, portability and reusability of components. Technologies remain independent and the system is manageable by different stakeholders. Aspects such as security, flexibility and personalisation were built into RAs. They have subsequently been adopted in many sectors. First, the "IDABC" (EC, 2024) adopted the framework around 2005. The aim was to enable the public sector to deliver "interoperable services" across Europe. The use of RAs as a conceptual framework then became commonplace. The roll-out to other domains started around 2010. First versions of RAs for "Smart Agriculture" have been available since 2011 through the EU projects SmartAgriFood (SmartAgriFood, 2015) and FIspace (FIspace, 2015), later FIWARE (FIWARE, 2024). They developed a basic design for an agricultural RA (SmartAgriFood, 2015; FIspace, 2015) and brought together various solutions, information systems and networked devices in agriculture (FIWARE, 2018). FIWARE was then transferred to the FIWARE Foundation, which has continued developments around service integration, produced many standards and played the role of a major stakeholder. FIWARE RA has undergone continuous technical development. It has applied its RA developments to different sectors such as health and, since 2016, agriculture. In 2019, FIWARE defined the "smart agri-food domain" as a focus area, along with smart cities, smart energy, etc. (FIWARE, 2018). The "NGSI-LD" standard is an example of a result component of the FIWARE initiative that is considered to be of current relevance in the area of digital integration platforms for agriculture in the current Demeter Agricultural Information Model project (Rodriguez et al., 2018; SmartAgriFood, 2015; Bochtis et al., 2022).

2.4 Data Spaces

RAs laid the foundation for service-oriented technical interoperability. Around 2017, new challenges emerged with data-intensive AI and the need for interoperable data layer concepts - the proposed solution was Data Spaces. The term Data Space refers to a type of data relationship between trusted partners that agree on common governance rules for data storage and sharing within one or more vertical ecosystems (EC, 2020d). Data Spaces are about federated data sharing and reuse within federated interoperable services, ensuring so-called data sovereignty. Several Data Space initiatives were launched and IDSA (EC, 2020c), FIWARE, BDVA/DAIRO and GAIA-X formed an alliance. Their results became a strategic part of the "European Data Strategy" (EC, 2020a) 2020, which aimed to create a single "European Data Space". The first use case of a Data Space is to ensure data sharing across supply chains - in the case of agri-food, from "farm to fork". One of the most important aspects of the Data Space is that the data is not stored centrally, but at the source, with the data provider. Data is held exclusively by the members of the federation. Data Spaces are developed on a domain basis, but can be activated across domains. The Gaia-X framework enables the creation of a sovereign and federated agricultural data space that can be deployed securely and at scale. Through the data space concept, companies should have easy access to an almost infinite amount of high quality industrial data, creating value while minimising the environmental footprint. Data sovereignty and trust between members are essential for Data Spaces.

Under the umbrella of the European Single Data Space, the EC is funding the implementation of selected domain-specific data spaces, such as an "Agricultural Dataspace" (EC, 2020a), to improve the processing and analysis of related data, enabling precise and tailored application of production approaches at farm level. In describing the evolution of IT in agri-food (Wolfert et al., 2023), Wolfert et al. note that the nature of digitalisation has become more complex along two axes: Figure 1 shows that the level of IT integration on the x-axis has evolved from individual applications to a complex "system of systems." On the y-axis, the number of actors involved increases from individual process participants to complex business ecosystems with a large number of actors. This also expands the scope of RAs and Data Spaces: 1) from the former production process to the supply chain to the whole food system and data economy and 2) from individual applications to operational and chain information systems to Data Spaces as part of the "European Data Strategy" (EC, 2020a).

There are Data Space projects for agriculture ongoing: Germany is funding two called AgriGaia (AgriGaia, 2023) and NaLamKI (NaLamKI, 2023) and at EU level we have "AgriDataSpace" (AgriDataSpace, 2024). The aim is to map the current landscape of ongoing data sharing initiatives in agriculture, to design approaches based on Data Space concepts, to explore appropriate solutions for semantic interoperability, and to analyse and evaluate current governance models. AgriDataSpace is developing a

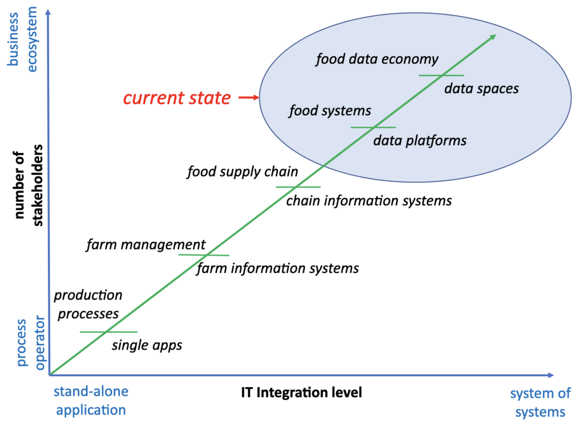

Figure 1: Evolution of IT in Agri-Food. Note: The ellipse indicates the current state, in which innovation ecosystems have become complex. Adapted from Wolfert et al. (Wolfert et al., 2023) based on (Wolfert et al., 2021).

conceptual RA with an "Information Model" (IM) to represent concepts and relationships, constraints, rules and operations to specify data semantics for the domain. In general, an IM should provide a shareable, stable and organised structure of information requirements or knowledge for the domain context. As a component, an IM should be open and technology and domain agnostic. It should enable the comprehensive description of data assets and the semantic interoperability required for this type of exchange. For an IM, the semantic interoperability of the domain needs to be based on standards so that data can be integrated for further decision support. A number of European projects using RAs for data exchange have indeed implemented such an IM for agriculture on a demonstrator basis using different approaches, as presented in section 3.3 of this paper. They haven't yet produced a standard IM for agriculture, but research is moving in that direction.

3 SEMANTICS FOR INTEROPERABILITY

An important motivation for the design of data spaces in agriculture is to support the efficient development of services for the entire food supply chain activities, ranging from agricultural production to distribution and consumption, while enabling analysis and reporting tools for health, food security and sustainability. Clearly, developers know the requirements and expected outcomes of the service they are developing. But beyond that, when it comes to providing openness and interoperability in the service to be implemented within a broader architecture, as required by the "European Data Strategy" (EC, 2020a), IT developers need to have a cross-domain knowledge, namely of core information models and APIs, so that their service can be used across application domains such as Smart Cities, Smart Industry, Smart Agriculture, and additionally they also need a deeper understanding of the terms in the specific domain and use cases in order to efficiently design or select the appropriate conceptual scheme for the required data model. As this requires a high level of expertise in domain concepts as well as cross-domain and core standards, and as such expertise appears to be rare, a standardised information model for the domain could make the process more efficient on the service development side. Without supporting standards, building the data model from scratch appears to be time-consuming, as the different concepts that may exist for a given term, depending on the semantic artefact in which it is found, need to be carefully investigated. For example, (Baker et al., 2019) shows a side-by-side visualisation of the data (properties and values) provided by the semantic artefacts GACS (GACS, 2024) and FoodOn (FoodOn, 2024), proposing comparable but not identical concepts for maize. The choice of an appropriate concept based on terms is crucial to ensure that the code is not only syntactically correct, but also logically meaningful and functional. Developers typically capture this understanding in "Information Models", from which a program can perform automated reasoning. Such models can also be called "knowledge representations" of the domain (Hoekstra, 2009) and are derived from semantic artefacts. Semantic artefacts in general use specific "semantic technologies" to describe the knowledge of a domain. They cover the perspectives, concepts and terminology of their provider's domain of interest, which may vary widely even within the same domain. They reflect "concepts" as building blocks of a specific domain theory. Semantic artefacts include "thesauri", "taxonomies", "controlled vocabularies" and "ontologies". Many have in common that they are based on research data. (Subirats-Coll et al., 2022) highlights two basic principles for using semantic resources to achieve semantic interoperability in an IT service: This is the use of a globally unique web identification - a "URI" (Uniform Resource Identifier) - to identify a logical/physical resource used by web technologies, and secondly, a data model that can accommodate all existing data sets and formats, essentially a graph-oriented data model. The "Resource Description Framework" (RDF) follows these principles of linked data and is seen as a catalyst for semantic interoperability (Subirats-Coll et al., 2022). The "Linked Data Concept" (Berners-Lee, 2006) is based on this, dealing with redundancy and handling it with links. Semantic resources can appear in different knowledge representation languages such as OWL, RDFS, SKOS, OBO and UMLS-RRf etc.

(Baker et al., 2019) arranges semantic artefacts or resources on a "semantic spectrum" ranging from "lightweight" to heavyweight". Leightweight resources include e.g. "controlled lists" as standardised and organised arrangements of words or phrases or as thesauri and taxonomies with a hierarchical structure. Languages such as RDF/RDFS, SKOS, UML are suitable for representing these, as no reasoning is required (Hoekstra, 2009). In the "semantic spectrum" given by (Baker et al., 2019), the semantic resources GACS and ARGROVOC (FAO, 1981) are labeled as "concept schemes" and classified as "middleweight", whereas the well-known FoodOn, Crop (Ontology, 2024a) and Gene Ontology (Ontology, 2024b) are labelled as "Ontologies" and "heavyweight". They represent knowledge and are reusable terminological knowledge representations (Hoekstra, 2009). Ontologies are semantically described with rich metadata (Drury et al., 2019). Languages suitable for representing ontologies are developed to support ontology-based reasoning services within knowledge-based applications. The main standards used are the OWL-DL family (Web Ontology Language Description Logic) (Hoekstra, 2009; Baker et al., 2019). Over the last few decades, international organisations have built up large semantic artefacts for agronomy and have carried out a multifaceted collaboration on editorial service tools, concept mapping, etc. Examples include the prominent AGROVOC of the Food and Agriculture Organisation (FAO), the NAL Thesaurus (NALT) (NALT, 2024) of the National Agricultural Library (NAL) and the CAB Thesaurus of the Centre for Agricultural Bioscience International (CABI, 2024). Developers can find them in repositories such as AgroPortal (AgroPortal, 2024) or Fairsharing.org etc. In addition to these general agronomic semantic resources, there is a wide range of specific semantic resources, all of which provide specific exchange formats and associated services. A good overview can be found in (Arnaud et al., 2020). Nevertheless, the use of ontologies in IT service development remains underexploited (Drury et al., 2019). As argued above, one reason may be that the proliferation of semantic resources that redefine the same concepts, even when based on linked data, is not straightforward for the average developer. Key questions are on how to improve the linking of various resources (Subirats-Coll et al., 2022) and how to encourage developers to use this complex and rich knowledge of semantic artefacts for service development. The next two sections describe some domain-specific semantic artefacts and repositories that provide implementations of Semantic Web technology, rich concepts and datasets, and SPARQL endpoints that may be relevant to developers.

3.1 Selected Semantic Artefacts

AGROVOC. (FAO, 1981; Caracciolo et al., 2013) is known as the world's most widely used RDF-based semantic resource for agriculture. It covers the topics food, nutrition, agriculture, forestry, fisheries, environment, etc. It is hierarchically organised under 25 top concepts with definitions and relationships. It is a multilingual "controlled vocabulary". It was created in 1981 and has since then been continuously developed and maintained by a network of institutes. Technically, it has moved from databases to the Semantic Web with Linked Open Data (Caracciolo et al., 2013). It collaborates with other knowledge resources and is constantly expanding its coverage (Brewster et al., 2023; Baker et al., 2019). It is edited using the web-based platform VocBench, a software for managing controlled vocabularies such as ontologies, thesauri and generic RDF datasets (Stellato et al., 2020). AGROVOC can be accessed, searched for concepts or browsed by hierarchy. It can be downloaded as an RDF dataset. Available web services can be used, searches can be performed via SPARQL queries, using public SPARQL endpoints and as an interface for submitting RDF queries. Recently, AGROVOC has been brought into the context of a "Data Space for Agriculture" in (Dörr et al., 2023). Data Spaces can be seen as data sharing spaces, but AGROVOC, even if it is based on linked data, would be a "vocabulary" component within a Data Space. And a data space is an organised community of data providers and data users who contractually agree to share their data according to rules, and whose members have commercial objectives. In contrast, AGROVOC is open to researchers and knowledge managers for indexing, retrieving and organising data in agricultural information systems, and seems to be less used by IT developers, as discussed by (Baker et al., 2019).

CGIAR - Big Data in Agriculture. is a global initiative of Semantic Web technology for agriculture, with datasets connecting experts around the world. The CGIAR (Consultative Group on International Agricultural Research) platform companies apply big data approaches to agricultural use cases. CGIAR supports selected ontologies across domains, including the Crop Ontology, the Agronomy Ontology (AGRO), the Environment Ontology (ENCO), the Plant Ontology (PO) and the Socio-economic Ontology (Arnaud et al., 2020). It stimulates the exchange of knowledge between stakeholders. It has made its 40 separate open data and publication repositories discoverable in one place through GARDIAN, a data harvesting and annotation service. The GARDIAN ecosystem has expanded to include data from the US Department of Agriculture, the Indian Council of Agricultural Research, the World Bank's Microdata Library and USAID Development Data. Linked stewardship should enable broader data discovery. 15 CGIAR research centres and 12 research programmes are partners in the platform, along with 70 external partners from research to industry, such as Bayer and Amazon. It covers work from public to industrial, from analysis to ICT deployment.

FoodOn - for Food Tracebility. Created in 2019, FoodOn is aimed at "food traceability". It integrates existing ontologies and reuses terms from OBO Foundry, environmental terms from ENVO, agricultural terms from AGRO, etc. Conversely, FoodOn terms are reused in a growing list of ontologies. FoodOn extends traceability standards such as the GS1 standards that enable traceability in the food supply chain with agronomic and distribution sensor data (Dooley et al., 2018). FoodOn is part of the open source OBO Foundry (Open Biological and Biomedical Ontology Foundry) of interoperable life science-oriented ontologies. It supports the goals of FAIR to annotate and share data across research, government and commercial sectors. It provides a vocabulary for nutritional analysis, including chemical food constituents, which are factors in nutrition, health and plant and animal agriculture research. It aims to build a comprehensive and easily accessible global farm-to-fork food ontology that accurately and consistently describes foods commonly known in cultures around the world. Although the FoodOn consortium includes industry partners, it appears to be focused on research data. Much of FoodOn's vocabulary comes from the transformation of LanguaL, a mature and popular food indexing thesaurus, into an OWL vocabulary that provides system interoperability, quality control, and software-driven intelligence.

SAREF4Agri. created in 2019, is the "Smart Agriculture and Food Chain" domain extension of SAREF (Smart Applications and Reference Ontology). This ontology is maintained by the European Telecommunications Standards Institute (ETSI). It specifies recurring core concepts, main relationships between concepts and axioms to constrain the use of these concepts and relationships. It is based on a) reuse and alignment of concepts and relationships defined in existing assets b) modularity to allow separation and recombination of parts of the ontology depending on specific needs and c) extensibility for growth d) maintainability to identify and correct defects, accommodate new requirements and cope with changes. It has a SPARQL endpoint and can be downloaded in various formats.

GACS - Global Agricultural Concept Scheme. was created in 2014 as a terminology hub, incorporating terms from the three main sources: AGROVOC from the FAO, CABT (CAB Thesaurus) from the Centre for Agriculture and Biosciences International (CABI) and NAL Thesaurus (NALT) from the National Agricultural Library of the USA as a lightweight "shared concept scheme" (GACS, 2024). GACS includes in its interoperable concepts identities related to agriculture from AGROVOC (32,000 concepts), the CAB (140,000 concepts) and the NAL (53,000 concepts). According to (Baker et al., 2019), GACS was motivated by the fact that the major thesauri were too detailed in terms of custom relation properties, so that they couldn't be fully exploited. They lacked reasoning tools, so it was unclear to developers what purpose their relationships served. Although they provided standards for data exchange, developers were unable to extract the knowledge (Baker et al., 2019). GACS consisted of the most commonly used (i.e., important) concepts in agriculture (Short et al., 2023). GACS included an interoperable layer that transformed data silos into a more reusable web of data, making resources more discoverable, and its relative simplicity was intended to make it cheaper to create and maintain (Baker et al., 2019). However, major ontologies such as FoodOn, AGRO and CROP were no longer mapped to GACS, and maintenance of GACS was discontinued in 2019. Mapping required too many resources, process concepts were mapped algorithmically, but results were manually checked and inconsistencies were resolved by discussion. GACS lives on in VocBench for manual editing and maintenance. In 2022 it was integrated into NALT as a resource for the "machine age" and is now a multi-issue concept space with added structural features for machine readability.

3.2 Ontology Repositories

The number of semantic artefacts is increasing, they are scattered in different formats with different structures and from overlapping domains. This is where AgroPortal comes in as a repository for ontologies and semantic artefacts (Jonquet et al., 2018). In January 2024, there are 160 ontologies and semantic artefacts in AgroPortal (AgroPortal, 2024).

AgroPortal. is conceived as a one-stop shop and as a "home of ontologies and semantic artefacts in agri-food and related domains". It was initiated by French institutes such as INRAE (French National Research Institute for Agriculture, Food and Environment) etc. and involves major international organisations such as FAO, CGIAR etc. IT collaborates with Stanford's BMIR (Center for Biomedical Informatics Research) in the USA and the EU research infrastructure consortium LifeWatch within the OntoPortal Alliance. (Jonquet et al., 2018; Drury et al., 2019; Jonquet et al., 2023). Ontologies and semantic artefacts can be uploaded. In January 2024, its statistics count 160 ontologies, 59 projects and 409 users, collected over the last 11 years. AgroPortal provides SPARQL endpoints as an interface for submitting RDF queries. AgroPortal provides browsing of features of related ontologies, it summarises information of ontologies e.g. calculates a FAIR score. It contains mappings of classes and concepts. AgroPortal and has continued the work of GACS. It provides "recommendations" for relevant ontologies from a text excerpt or a list of keywords, one can "annotate" ontologies and sort ontologies by the formats. However, in January 2024, there are only 409 registered users. The most popular ontologies in AgroPortal in February 2024 are AGROVOC and DEMETER AIM. (DEMETER, 2023) The developers of AgroPortal are actively involved in AGROVOC as users and editors. The international ontology community is small and interconnected, and also provides OntoPortal as a generic technology for building ontology repositories. The AgroPortal researchers at INRAE, namely Clement Jonquet and others, have provided software for creating ontology repositories that are domain-agnostic, customisable and open. The "Ontology Portal Alliance" presents (Jonquet et al., 2023) software for creating ontology repositories and semantic artefact catalogues that can support resources ranging from SKOS thesauri to RDF-S and OWL ontologies. The AgroPortal software is based on OntoPortal, which emphasises the importance of ontologies and semantic artefacts being FAIR, and seeks to counter a trend that has been observed, namely the proliferation of other semantic artefact catalogues, often developed with their own ad hoc technology and code.

D2KAB-Project (2019-2023). The research team of LIRMM (Laboratory of Informatics, Robotics and Microelectronics), INRAE and others responsible for AgroPortal and OntoPortal are running the D2KAB project transforming "Data to Knowledge in Agronomy and Biodiversity" (D2KAB, 2023) to develop methods and technologies for ontology lifecycle and alignment, to build a "Linked Open Data Cloud" for agronomy and biodiversity, enabling semantically driven agronomy and biodiversity science. The focus is on ontology-based services, ontology alignment as in GACS, and on building a knowledge graph for agriculture and biodiversity. Within D2KAB functionalities have been developed (e.g. SKOS support, FAIRness assessment, interoperability mapping, etc.).

3.3 EU Research on Agricultural Digital Integration Platforms

To date, ontologies have mainly focused on the clear definition of domain concepts and terms in the form of a vocabulary for annotating research publications or research datasets. Only a few projects have reused existing semantic standards within a technological IT architecture to enable limited data exchange between different agri-food stakeholders. The use of semantic standards from industry organisations for information exchange, such as the GS1 EPICS standard, ISO (International Organisation for Standardisation), "AgGateway", "agrirouter", does not appear to be widespread or even beginning to be adopted in research (Brewster et al., 2023). However, the ISO website states that a new ISO standard for agriculture and food is in preparation. In the development community, there appears to be little use of ontologies in the context of sharing and exchanging data across the agri-food chain, for example for traceability and data analysis. In the challenge to build "Agricultural Digital Integration Platforms", some developers are now exploring data models derived from ontologies as an essential element for semantically interoperable agri-food platforms. Key questions for research projects in this context are how to improve the linking of different resources and, secondly, from a developer perspective, to what extent do the advantages outweigh the disadvantages of using data models to share data for service development. In 2019, the EU invested around €80 million in projects for "Agricultural Digital Integration" (EC, 2020b). The projects experiment with different approaches to technical and semantic interoperability, including the use of data models, service and data management approaches involving many partners and pilots. They applied RAs and incorporated data space concepts. They developed and integrated service components of the FMIS, involving many services and stakeholders in different roles, depending on the service, data or process chain orientation. They reused several FIWARE solutions, e.g. for authentication or access control. The central infrastructure components are currently located (Wolfert et al., 2023) on the IT "transformation ladder" between "data platforms" and "data spaces". Projects that exemplify this challenge are mentioned below, namely CYBELE (CYBELE, 2022), PLOUTOS (PLOUTOS, 2023), DEMETER (DEMETER, 2023) and ATLAS (ATLAS, 2023), as well as the SmartAgriHubs network (SmartAgriHubs, 2022) to promote the technological approaches described above (EC, 2023c). For an overview of the state of the art considered in the proposed solution of each EU agritech project, see (Roussaki et al., 2023).

ATLAS-Project (2020-2023). is a "service-oriented solution for interoperability" that solves the problem without a semantic data model. The "Agricultural Interoperability and Analysis System" integrates FMIS with sensor systems and data analysis on agricultural equipment. The technology connects ATLAS-enabled systems and establishes a data flow between them. ATLAS is a distributed, decentralised network of systems. Each participating system remains independent, based on its own technical infrastructure. Semantic interoperability is achieved at the "service level" through standardised services described in the "ATLAS Service Templates". They define specifications of abstract elementary agricultural processes with clear semantics and define vendor and technology-agnostic formal specifications of the APIs and data formats to which associated ATLAS services must conform. Templates allow a service consumer to communicate with any compliant ATLAS service without requiring vendor-specific code, allowing farmers to choose the compliant ATLAS service that best meets their specific needs in terms of quality/accuracy, geographic/crop specificity, budget, etc. Formal API specifications claim that any client code designed to integrate the functionality of an ATLAS service template will interoperate with any ATLAS service implementing that template. However, this approach places the burden of the required conversion work on the side of the customers dealing with the service templates.

CYBELE-Project (2019-2022). stands for "Fostering Precision Agriculture and Livestock Farming" is a semantic metamodel that is conceptualised from use cases and it is based on general "W3C metadata standards" to facilitate discovery, exploration, integration and access to data, with "cross-domain models" such as the Semantic Sensor Network (SSN) ontology and "domain-specific models" such as AGROVOC. CYBELE assumes that existing vocabularies can individually express most of the model concepts, but so far there is no single model that satisfies all the requirements derived from the identified use cases. In addition, the merged metamodel defines concepts that facilitate access to and querying of data stored in databases. The model has been tested and claims to be open to support further structured or semi-structured data (e.g. JSON) stored in different types of databases, e.g. noSQL or graph databases. The use of different types of databases is solved by following linked data approaches to transform or query different sources on the fly in an integrated manner. The proposed model has been shown to achieve interoperability and homogeneous access to data sources. However, it could also be used in other domains with similar data and requirements (Zeginis et al., 2023). The validation of this approach still needs to be done (Roussaki et al., 2023).

PLOUTOS-Project (2020-2023). stands for "Data-driven sustainable agri-food value chains" and uses a data sharing architecture based on the reuse of existing semantic standards from different ontologies and an architecture that computes data along a supply chain and guarantees digital sovereignty. The project defines a "PLOUTOS Core Semantic Model" based on use cases to achieve data sharing along the supply chain. The system will allow queries to be made across a federated network of agri-food stakeholders (Brewster et al., 2023) that are identified as common to the supply chain as a whole. The supply chain data sharing architecture formalises the GS1 EPICS standards as ontologies and enables SPARQL queries across distributed triple stores. The classes and relations were selected from the requirements of the pilot projects. The data model is based on RDF and OWL to provide flexibility for modular reuse or extension. The architecture is based on so-called "Interoperability Enablers" (PIEs), which use "graph query patterns" to traverse the network and collect the required data to be shared. A core part of the PIE is the "Knowledge Mapper", which enables the interoperability of the data in the value chain at the semantic level by providing a data translation service that translates data streams provided by the hosting system into selected standardised data formats and vice versa. The translation functionality of the Knowledge Mapper is customisable. The data sharing approach promises to be extensible, but it's unclear how many resources this will require. The full version of the PCSM is available from the GitLab repository.

DEMETER-Project (2020-2023). is building an "Interoperable, Data-Driven, Innovative, Sustainable European Agri-Food Sector". It's based on a federated RA referencing IDSA, capable of integrating the communication, sensing and computing technologies used in agritech scenarios. It proposes an "Agricultural Information Model" (AIM) as a core service to enable semantic interoperability between different FMIS. AIM uses a layered, modular approach that extends the CYBELE approach. AIM provides predefined mappings between legacy models such as FIWARE, Safe4Agri, ADAPT (AgGateway's ADAPT Framework), AGROVOC, INSPIRE and Earth Observation standards. AIM is built on layers (Bochtis et al., 2022; Roussaki et al., 2023):

- AIM core metamodel following the NGSI-LD standard (ETSI) as IM approach and API.
- The "cross-domain ontology" layer using a set of relevant cross-domain standards such as Time Ontology , SOSA/SSN, GeoSparql (Geographic Query Language for RDF Data), etc.
- The "domain-specific ontology" layer modelling agricultural concepts such as crops, animals, products, farms etc.
- The "pilot-specific ontologies layer" for use cases.
- The "metadata schema layer" refers to meta-information based on cross-domain and domain-specific ontologies.

AIM is being tested in 20 pilot projects in 18 countries. It is said to be easy for developers to use. It is uploaded to AgroPortal where it is the second most visited after AGROVOC. It is being standardised by the OGC (Open Geospatial Consortium). IT developers are invited to evaluate it. Its successor is DIVINE (DEVINE, 2023), which demonstrates the costs and added value of sharing agricultural data.

Research Data Infrastructures (RDIs). provide resources and services to research communities to conduct research and foster innovation in their fields. They include major equipment or instruments-sets and knowledge-related facilities such as collections, archives or scientific data infrastructures such as ontologies or semantic artefacts. As part of the EU Strategic Plan for Research Infrastructures, RDI will contribute to achieving the key strategic orientations of the Horizon Europe Strategic Plan by consolidating and developing a European Research Infrastructures landscape and by opening up, integrating and interconnecting research infrastructures. As part of the research infrastructure landscape, the EU maintains a complete list of Horizon 2020 research infrastructures by domain, including some food and agriculture research facilities. In this context, RDI initiatives at national level are a concrete area of intervention. In agriculture in Germany, the NFDI4Agri initiative can be seen as a first step in the field of agrosystems research, which operates the FAIRagronet (EC, 2023d).

4 CONCLUSION AND FUTURE WORK

The paper discusses the current state of IT-based data sharing solutions in public agricultural research in Europe, highlighting key sections that play a crucial role in advancing this field. First, the paper presented the current state of service development for IT-based data sharing solutions in public agricultural research in Europe hereby highlighting the need for IT-based solutions. Based on this need, the following sections presented solutions for technical interoperability in the development of federated service and data architectures, showing that the agricultural sector provides a large number of rich and constantly evolving semantic resources that are organised in a free and open way. Their development has been mainly driven by academia and is accompanied by standards and service solutions for data representation. As a result, agritech developers can now make general use of them, for example to reason and deduce on a logical basis. Based on these semantic resources and common interoperability standards, current European research projects are developing and evaluating extensible and layered agricultural information models as standardised components for Agricultural Digital Integration Platforms for broad service development. They reuse and combine standards and best practices from existing ontologies that are suitable for a wide range of common use cases and cross-domain requirements, promising ease of use and a shortcut to service development. Nevertheless, the shift of workload to data providers needs to be carefully considered when evaluating the benefits of Semantic Web technologies for service development.

Future research could focus on two key areas: First, there is a need to further investigate how artificial intelligence can improve the use of semantic web technologies in the agricultural sector. By exploring innovative solutions that automate semantic annotation processes using AI algorithms, stakeholders can streamline data integration, improve knowledge representation and enhance information retrieval capabilities within agricultural systems. In that sense, the use of AI for semantic annotation seems to be promising, but, further research is needed to be able to make well-founded decisions. Secondly, it is crucial to bridge the current research gap between academia and industry. Investigating ways to facilitate knowledge transfer and collaboration between the agricultural and research sectors can ensure that research outcomes are effectively translated into practical solutions for the industry and vice versa.

ACKNOWLEDGEMENTS

This work was created as part of the research and development project "Stadt-Land-Fluss (SLF)". The project with the funding code 2821KI002 is funded by the German Federal Ministry of Food and Agriculture on the basis of a resolution of the German Bundestag.

REFERENCES

AgriDataSpace (2022-2024). Building a European framework for the secure and trusted data space for agriculture. https://agridataspace-csa.eu/. [Online; last access May 9, 2024].

AgriGaia (2023). KI-Ökosystem für die mittelstandsgeprägte Agrar- und Ernährungsindustrie in Anknüpfung an GAIA-X. https://www.agri-gaia.de/agri-gaia/das-oekosystem/. [Online; last access May 9, 2024].

AgroPortal (2024). AgroPortal. https://agroportal.lirmm.fr. [Online; last access May 9, 2024].

Arnaud, E. et al. (2020). The Ontologies Community of Practice: A CGIAR Initiative for Big Data in Agrifood Systems. *Patterns*, 1(7):100105.

ATLAS (2019-2023). Agricultural Interoperability and Analysis System. https://cordis.europa.eu/project/id/857125. [Online; v May 9, 2024].

Baker, T. et al. (2019). Global agricultural concept space: lightweight semantics for pragmatic interoperability. *npj Science of Food*, 3(1):16. Number: 1 Publisher: Nature Publishing Group.

Berners-Lee, T. (2006). Design Issues Linked Data. https://www.w3.org/DesignIssues/LinkedData. [Online; last access May 9, 2024].

Bochtis, D. D. et al., editors (2022). *Information and Communication Technologies for Agriculture—Theme III: Decision*, volume 184 of *Springer Optimization and Its Applications*. Springer International Publishing, Cham.

Brewster, C. et al. (2023). Data sharing in agricultural supply chains: Using semantics to enable sustainable food systems. *Semantic Web*, pages 1–31.

CABI (2024). CABI Improving lives by solving problems in agriculture and the environment. https://www.cabi.org/. [Online; last access May 9, 2024].

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., and Keizer, J. (2013). The AGROVOC Linked Dataset. *Semantic Web*, 4(3):341–348.

CYBELE (2019-2022). Fostering Precision Agriculture and Livestock Farming. https://cordis.europa.eu/project/id/825355. [Online; last access May 9, 2024].

D2KAB (2019-2023). Data to Knowledge in Agronomy and Biodiversity. https://d2kab.mystrikingly.com/. [Online; last access May 9, 2024].

DEMETER (2023). Building an Interoperable, Data-Driven, Innovative, Sustainable European AgriFood Sector. https://cordis.europa.eu/project/id/857202. [Online; last access May 9, 2024].

Deutschlandfunk (2024). Großdemo in Berlin. Warum die Bauern auf der Straße sind. https://www.deutschlandfunk.de/bauernproteste-102.html. [Online; last access May 9, 2024].

DEVINE (2023). Demonstrating value of agridatasharing for boosting data economy in Agriculture,. https://cordis.europa.eu/project/id/101060884. [Online; last access May 9, 2024].

DFKI (2014). iGreen. http://www.igreen-projekt.de/iGreen/. [Online; last access May 9, 2024].

Dooley, D. M. et al. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2(1):23. Number: 1 Publisher: Nature Publishing Group.

Drury, B., Fernandes, R., Moura, M.-F., and De Andrade Lopes, A. (2019). A survey of semantic web technology for agriculture. *Information Processing in Agriculture*, 6(4):487–501.

Dörr, J., Favier, M.-A., and Möhrle, F. (2023). Interoperabilität im Digital Farming – Status quo und zukünftige Herausforderungen. *Wirtschaftsinformatik & Management*.

EC (2020a). A European strategy for data. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066. [Online; last access May 9, 2024].

EC (2020b). Agricultural digital integration platforms. https://cordis.europa.eu/programme/id/H2020_DT-ICT-08-2019/en. [Online; last access May 9, 2024].

EC (2020c). International Data Spaces Association. https://internationaldataspaces.org/. [Online; last access May 9, 2024].

EC (2020d). What is Gaia-X. https://gaia-x.eu/what-is-gaia-x/. [Online; last access May 9, 2024].

EC (2023a). Agri-Digital Conference. https://agriculture.ec.europa.eu/eu-agri-food-days/eu-agri-food-days-conference-material_en\#ref-8-december-agri-digital-conference. [Online; last access May 9, 2024].

EC (2023b). Data Act Council adopts new law on fair access to and use of data. https://www.consilium.europa.eu/en/press/press-releases/2023/11/27/data-act-council-adopts-new-law-on-fair-access-to-and-use-of-data/. [Online; last access May 9, 2024].

EC (2023c). Digitalisation of the European Agricultural Sector: Activities in Horizon 2020. https://digital-strategy.ec.europa.eu/en/policies/digitalisation-agriculture-horizon-2020. [Online; last access May 9, 2024].

EC (2023d). FAIRagro. https://fairagro.net/en. [Online; last access May 9, 2024].

EC (2024). Interoperable Delivery of European eGov Services to Public Administrations Businesses and Citizens. https://ec.europa.eu/idabc. [Online; last access May 9, 2024].

EC, J. R. C. (2023e). *Digital transition: long term implications for EU farmers and rural communities*. Publications Office of the European Union, LU.

EC, DG Informatics (2017). *New European interoperability framework promoting seamless services and data flows for EU public administrations.* Publications Office, LU.

EU (2023). Regulation (EC) 2023/2854 of the European Parliament and of the council of 13.12.2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). Technical report.

FAO (1981). Multilingual Thesaurus. https://agrovoc.fao.org/browse/agrovoc/en/. [Online; last access May 9, 2024].

FAO (2021). *The State of Food and Agriculture 2021*. FAO.

FIspace (2013-2015). Future Internet Business Collaboration Networks in Agri-Food, Transport and Logistics. https://cordis.europa.eu/project/id/604123. [Online; last access May 9, 2024].

FIWARE (2018). Smart Agrifood - The new revolution in farming. https://www.fiware.org/community/smart-agrifood. [Online; last access May 9, 2024].

FIWARE (2024). FIWARE About us. https://www.fiware.org/about-us/. [Online; last access May 9, 2024].

FoodOn (2018-2024). FoodOn: A farm to fork ontology. https://foodon.org/. [Online; last access May 9, 2024].

GACS (2014-2024). Global Agricultural Concept Scheme. https://agroportal.lirmm.fr/ontologies/GACS. [Online; last access May 9, 2024].

Hoekstra, R. (2009). Ontology representation - design patterns and ontologies that make sense.

Jonquet, C. et al. (2018). AgroPortal: A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144:126–143.

Jonquet, C. et al. (2023). Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology. In *The Semantic Web – ISWC 2023*, volume 14266, pages 38–58. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

LeanIX (2023). Guide Referenzarchitekturen. https://www.leanix.net/de/wiki/ea/reference-architecture. [Online; last access May 9, 2024].

NaLamKI (2023). Nachhaltige Landwirtschaft mit Künstlicher Intelligenz. https://nalamki.de/. [Online; last access May 9, 2024].

NALT (2022-2024). NAL Agricultural Thesaurus. https://agclass.nal.usda.gov/. [Online; last access May 9, 2024].

Ontology, C. (2024a). Crop Ontology for agricultural data. https://cropontology.org/. [Online; last access May 9, 2024].

Ontology, G. (1978-2024b). Gene Ontology Unifying Biology. https://www.geneontology.org/. [Online; last access May 9, 2024].

PLOUTOS (2019-2023). Data-driven sustainable agri-food value chains. https://cordis.europa.eu/project/id/101000594. [Online; last access May 9, 2024].

Rodriguez, M. A., Cuenca, L., and Ortiz, A. (2018). FIWARE OS Standard Platform in Smart Farming - A Review. In Camarinha-Matos et al., editors, *Collaborative Networks of Cognitive Systems*, volume 534, pages 581–589. Springer International Publishing, Cham. Series Title: IFIP Advances in Information and Communication Technology.

Roussaki, I. et al. (2023). Building an interoperable space for smart agriculture. *Digital Communications and Networks*, 9(1):183–193.

Short, N. M., Woodward-Greene, M. J., Buser, M. D., and Roberts, D. P. (2023). Scalable Knowledge Management to Meet Global 21st Century Challenges in Agriculture. *Land*, 12(3):588. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

SmartAgriFood (2013-2015). Smart Food and Agribusiness: Future Internet for Safe and Healthy Food from Farm to Fork. https://cordis.europa.eu/project/id/285326. [Online; last access May 9, 2024].

SmartAgriHubs (2018-2022). Unleashing the innovation potential for the digital transformation of the European Agrifood Sector. https://www.smartagrihubs.eu/. [Online; last access May 9, 2024].

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., and Keizer, J. (2020). VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11(5):855–881. Publisher: IOS Press.

Subirats-Coll, I. et al. (2022). AGROVOC: The linked data concept hub for food and agriculture. *Computers and Electronics in Agriculture*, 196:105965.

Subramanian, C., Liu, M., Khan, N., Lenchner, J., Amarnath, A., Swaminathan, S., Riegel, R., and Gray, A. (2024). A Neuro-Symbolic Approach to Multi-Agent RL for Interpretability and Probabilistic Decision Making. arXiv:2402.13440 [cs].

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Wolfert, S. et al. (2021). Navigating the Twilight Zone: Pathways towards digital transformation of food systems. Publisher: Wageningen University & Research.

Wolfert, S. et al. (2023). Digital innovation ecosystems in agri-food: design principles and organizational framework. *Agricultural Systems*, 204:103558.

Zeginis, D. et al. (2023). A semantic meta-model for data integration and exploitation in precision agriculture and livestock farming. *Semantic Web*, pages 1–29.