Computers and Electronics in Agriculture 196 (2022) 106909

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Cultivating FAIR principles for agri-food data

Jan Top *, Sander Janssen, Hendrik Boogaard, Rob Knapen, Görkem Şimşek-Şenel

Wageningen University & Research, P.O. Box 9101, 6700 HB Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Agriculture
Food supply chain
Data sharing
FAIR principles
Ontology
Controlled vocabulary

ABSTRACT

Data generated by the global food system is crucial in the transformation towards sustainable, resilient, and high-quality food production. Although the amount of potentially useful data is growing rapidly, its (re)use is still limited. The FAIR-principles have been developed for making data findable, accessible, interoperable, and reusable both by humans and machines. This paper explores the further operationalization of the FAIR principles in agriculture and food. Experience shows that several conditions must be fulfilled before data can be effectively shared and reused. First, automated tools must be available for data providers and users. Secondly, we need a community-based approach in developing tools and vocabularies. Thirdly, data cannot be shared by an open-by-default policy only. Finally, scientific insight is needed in how data is actually (re)used in scientific communities. We conclude that bringing the FAIR-principles to full maturity requires a fair balance of efforts within the agri-food communities, supported by a proper infrastructure.

1. Introduction

The agriculture and food system aims to ensure a sustainable supply of healthy and nutritious food to nine billion people in 2050, while facing climate change and land degradation. It is widely assumed that data generated by the food system and digitalization of the agricultural supply chain have a role to play in facilitating the transformation towards a resilient, sustainable, and food-secure food system globally. Data and digitalization help by allowing for more fine-grained and holistic decision making by the farmer, consumer, business, or policy maker, leading to data driven solutions (Mey et al. (2019), Carolan et al. (2015)), and to increased trust, better strategic and operational decision making, and creation of new business. Networks like Global Open Data for Agriculture and Nutrition (https://www.godan.info/),[1] CGIAR's Big Data Program (https://bigdata.cgiar.org/) and UN Global Pulse (https://www.unglobalpulse.org/) give a vital position to data and digitalization in the transformation of food systems, moving towards a mature data ecosystem. Agricultural and food science have a role to play in supporting the emergence of these data and digitalization solutions, by (1) opening their data for other societal actors to validate, evaluate and apply the proposed solutions, (2) developing proof-of-principles or proof-of-concepts of potential solutions at the lower levels of technological readiness, for example by showing the value of machine learning on aggregated crop data (Paudel et al., 2021) and (3) exploring innovative digital and data solutions through blue skies research.

At the same time, already much attention is being paid to the role of data in all domains of science and applied research. There is a clear need to share data between researchers, but also between the research community and data users and providers in society. The underlying idea is that science and research should open, not only by producing papers but also by sharing data and models[2]. This is stimulated by developments like the European Open Science Cloud (https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud), (https://webgate.ec.europa.eu/funding-tenders-opportunities/display/OM/Online+Manual) and the fast spread of Open Access publishing methods as in Plan S (https://www.coalition-s.org/). The goal is to increase the effectiveness and transparency of scientific research and societal innovation (Australian_Academy_of_Science, 2021).

Currently, data sharing is considered as the main bottleneck in reaching this goal. Therefore, the FAIR-principles have been developed

* Corresponding author.
E-mail addresses: jan.top@wur.nl (J. Top), sander.janssen@wur.nl (S. Janssen), hendrik.boogaard@wur.nl (H. Boogaard), rob.knapen@wur.nl (R. Knapen), gorkem.simsek-senel@wur.nl (G. Şimşek-Şenel).

[1] All hyperlinks in this article have been accessed at August 28, 2021.

[2] We define *data* as a collection of values (numbers, text or images) for a set of variables or parameters, whereas *models* provide arithmetic or other relations between variables and parameters. When presenting the provenance of these data and models, this also extends to *methods*. The borders between data, models and methods are not always clear in practice, in particular if they are embedded in software.

https://doi.org/10.1016/j.compag.2022.106909
Received 5 November 2021; Received in revised form 13 February 2022; Accepted 23 March 2022
Available online 1 April 2022

(Wilkinson et al. (2016)). These principles express how data can best be prepared for sharing, i.e., by making it Findable, Accessible, Interoperable and Reusable. A key step is to link data from a wide range of data sources, as advocated in the five-star model proposed by Tim Berners Lee (https://5stardata.info/en/). Ideally these principles help others than the original creators of a dataset (or even the creators at a later point in time) to find, access, understand and automatically process the data. This would enable researchers in for example the following tasks:

- Understand each other's research claims,
- replicate research for verification and falsification,
- combine data from various sources to build new models or theories,
- save money by not repeating experiments unnecessarily,
- understand and verify published papers.

However, executing the FAIR principles in practice is still not trivial and needs further guidance (Jacobsen et al., 2020a). The first objective of this paper is to explore the operational requirements for the application of the FAIR data principles in agricultural and food, in particular with respect to using shared vocabularies for describing the data. The second objective is to provide recommendations for extending the adoption of the FAIR principles in these domains, based on the lessons we have learned in recent research projects.

In the following section we explore the particularities of food systems, and their implications for agriculture and food science and its data. Next, the paper provides some more background on the development of data sharing, also highlighting impediments in the practice of data sharing. Subsequently we present some case studies of data sharing in agriculture and food science in which we have been involved over the past years. We have selected these cases as they illustrate some issues that we ran into when applying the FAIR principles. Finally, in the discussion section the main lessons learned from these and other cases will be drawn, with implications for the FAIR principles, data sharing and data science in agriculture and food science.

2. Challenges in agriculture and food science data sharing

Data in agriculture and food science has some features that help to clarify the general context of this paper. First, agriculture and food science cover a wide range of scientific disciplines, from genetic research for breeding to nutritional science on the composition of food intake, with many different disciplines such as animal husbandry, food logistics, crop science, consumer behavior, etc. The data types in agriculture and food science are thus very divergent, depending on the disciplines considered. The overview given in (Janssen et al. (2017)) mentions the following issues.

- Governments collect data for monitoring purposes, management of information and administrative procedures. These data, which include ethnographic statistics, monitoring data for subsidies and taxes, and data on environmental performance and national health, are in general uniform in format. They are usually collected on a regular basis as a direct input for policy making.
- Research projects collect data (e.g., measurements in laboratories, field and household surveys, multi-dimensional panel data, soil sampling, population data) to meet specific project needs. These data are often incidental (i.e., collected on an irregular basis) and not well-structured (i.e., non-uniform in format and only partially annotated).
- Industries (including farmers and business-to-business service operators) collect data for their own operations and in their R&D activities. They often do not share data due to competitive concerns. Moreover, their data formats are often not harmonized.

Organizations collect data manually or automatically, using all kinds of devices. The upcoming Internet of Things adds new data sources, such as mobile technology, new sensors, crowd sourcing, drones, and remote sensing (Verdouw et al., 2019).

Over the past year, several reviews and position papers (Wolfert et al. (2017), Kamilaris et al. (2017); Lokers (2019); Marvin et al. (2017)) have described the state-of-the-art and challenges in data management in agriculture and food. A first challenge, often highlighted, is the lack of interoperability and standards across the agriculture and food sciences. Some parts of the agriculture and food sciences have strong implementations of standards, such as plant genetics through BRAPI (https://www.brapi.org/specification), geo-spatial through OGC (https://www.ogc.org/docs/is), thesaurus terms and language translations through GACS and AGROVOC (FAO, 2021), and farm management events through AgGateway (https://www.aggateway.org/). The links between these standards are weak, with many crucial parts lacking. For example, how do the thesaurus terms from AGROVOC relate to spatial concepts in OCC? There is no proper understanding of these white spots so far, leading to the conclusion that the advance of big data techniques in agri-food is a *variety* than a *volume* challenge (Lokers (2019), Marvin et al. (2017)). This impacts the sharing and use of research data; researchers lack cross-domain harmonized and accepted definitions to describe and annotate their datasets, and consequently combined analysis of existing datasets becomes a laborious and error prone task.

A second challenge is the lack of access to data and actionable solutions for trusted and privacy-safe governance of data. Many interesting data sources are still closed for commercial or other reasons, even if controlled access would not cause any disadvantage. This hinders the use of data in agri-food practice and research. On the positive side, farmer organizations across the globe have developed codes-of-conduct in sharing farmer data and laws like EU General Data Protection Regulation (https://gdpr-info.eu/) have provided more clarity on the management of individual data. Data sharing solutions pushed by industry have started to emerge (https://join-data.nl/, https://api-agro.eu/en/, https://my-agrirouter.com/en/, https://farmstack.digitalgreen.org) but are still in their infancy.

A third challenge is the lack of agreement on how to use the available technologies in IoT sensors, lab instruments and all kinds of equipment. They should collect (stream) data in a standardized and robust way. There are still many steps needed by the researcher to download data to a shared format, convert it, and bring it together from a diversity of devices. For example, there are several commercial solutions available for collecting data in field trials of crops for breeding, each with its own particularities, while an open, accepted setup is lacking. This has sparked off the development of AgroFIMS by CGIAR (https://bigdata.cgiar.org/divi_overlay/agrofims-your-new-companion-for-easy-standardization-of-data-collection-and-description/).

A fourth challenge is that little attention is being paid to the development of data management skills as part of educational degrees in agriculture and food sciences. Researchers are more trained in finding software-based solutions in the domain than in good practices and skills for working with data. A related challenge is that the food system requires an interdisciplinary and collaborative approach with researchers that act as knowledge brokers (Cash et al., 2003), also managing data across domains.

A consequence of the above challenges is that agriculture and food researchers are still facing quite some hurdles when it comes to data sharing. The abovementioned challenges have been covered in Section 3 by sharing the current practices and in Section 4 by presenting the cases which the authors have worked with.

3. Steps in FAIR data sharing in practice

How is data sharing in agriculture and food systems research currently arranged? As stated above, data can be generated by researchers, but also by companies, organizations, governments, and individuals. Ideally, this data would be properly annotated, using accepted

and harmonized standards. However, if this is not the case, users of data can adopt (part of) this shared data for their own specific objectives. This means that depending on their objectives, they can isolate and preserve datasets from external sources and 'make them more FAIR', just like the data generated by themselves.

Before discussing the practice of data sharing and the effort needed to realize some level of data reuse, we need to ask the question 'Which data is worth the effort?'. This question can be split into sub-questions related to benefit and cost: (1) is there potential interest in using this dataset and (2) how much work is needed to make it reusable? In principle, any dataset is potentially interesting, but specific or sloppy data (i.e., data obtained by inadequate methods or methods that are hard to explain) will be less interesting for reuse by others. In fact, in such cases persistent storage of data is not needed. In all other cases, in which data is deliberately stored with the idea that it may be useful in future (even if only for legal reasons), the question arises how much effort should be spent on making the data ready for reuse. For example, during an explorative research project, often raw data is (deliberately) collected in 'quick and dirty' trial experiments. At best, this data reaches the researcher's notebook as a few numbers, scribbles, and some quick drawings. Only the original researcher can explain such notes, which he or she even may find hard to do after some time. Most researchers would consider cleaning and documenting such a 'dataset' a waste of time. Even though at a later stage such data may prove to be groundbreaking, at the time of production this cannot be foreseen. So, spending time on FAIRification asks for a minimal level of quality and maturity of the data itself. We have seen many cases in which researchers did not explicitly describe the maturity and quality of their data, nor did they assess if and how their own data could be useful for others. This argues for defining a general data assessment task at the level of a research domain, as we will point out later.

Once a dataset or model is considered sufficiently interesting for sharing, work is needed to make it available for (re)use, either by the original researcher, by co-workers or by potential users of the data. Many of the datasets we have run into are stored and maintained by the individual researcher, at best in a digital format on a file system or in a structured data repository. In the agri-food domain, we see that datasets are scattered and not well organized. Documentation (metadata) is often at a minimum, and access is restricted. Learning that some dataset or model exists usually happens through personal contact, or when it is mentioned in a publication or report. Currently, the situation is rapidly improving due to the availability of public research repositories, at the national and international level (https://data.europa.eu/en, https://dans.knaw.nl/en, https://data.4tu.nl/info/, https://dataverse.harvard.edu/, https://gardian.bigdata.cgiar.org). Even then it can be hard to import the data into the user's tools for reuse.

FAIRification aims to take away the restrictions in data reuse, making datasets and data streams self-contained and interoperable. To understand what is needed, we need to answer the question: 'When do we consider the (digital) data sharing process to be successful?'. We list the following criteria:

- A potential user of a dataset can find the dataset based on subject, source, variables involved, format or other criteria, through a single access point on the web.
- She can decide from the available documentation (metadata) whether the dataset is sufficiently interesting for her task and whether she is allowed to use it for that purpose.
- She can import the dataset into her data processing environment, including mapping of variables from the imported set to those existing in her personal tools.
- She can add new metadata to the original data and link to new data, based on her own findings.

Provided that the required information is available in some form or another, it is in principle always possible to perform these actions *manually*, but this would require effort, expertise and time that is not available in practice. The point in following the FAIR-principles is that, if the data is stored in digital form and properly annotated, it can be accessed, understood, assessed, used, and edited *automatically*, with minimal effort and time for the user. This allows data users to focus on analyzing the data rather than spending time on acquiring and organizing it. Even more importantly, proper annotation will increase the quality of the data, i.e., once imported it will be more fit-for-purpose.

However, work is needed at the side of data *providers* to meet the above criteria. It should be noted that FAIR principles have been published for a broad audience and do not dictate specific ways of implementation. They are open to different interpretations (Jacobsen et al., 2020a). Based on our experience, we argue that data publishers in the agri-food domain at least need to take the following actions, for each which we indicate the associated FAIR principles (Wilkinson et al., 2016).

- Assign unique identifiers to datasets and data elements, and select publicly available repositories (F1, A1, A2).
- Review and formulate adequate metadata (F2, R1.2).
- Assign a license and express the metadata with machine actionable semantics (I1, I2).
- Track reuse of the data and extend metadata if needed (extension of I3).

In Section 4, we will show how these tasks can be result in achieving a specific 'data maturity level'.

Some steps for supporting these activities have already been made in terms of automated data management following the FAIR principles. For example, the GO FAIR initiative (https://www.go-fair.org/go-fair-initiative/) is promoting several tools for this purpose in the envisioned Internet of FAIR Data & Services (IFDS, https://www.go-fair.org/resources/internet-fair-data-services/). For example, a so-called *FAIRifier* assists in adding metadata, data license, data model, and linking the selected ontologies to a dataset. In addition, a metadata editor allows non-technical users to define and publish the metadata required by a FAIR data point. Finally, a FAIR Data search engine harvests the metadata available on FAIR data points or compatible data repositories (https://github.com/FAIRDataTeam/FAIRSearchEngine). In addition, editors for creating and editing ontologies are available, such as Protégé, TopBraid, ORKA, NeON Toolkit (https://www.w3.org/wiki/Ontology_editors). Also, numerous controlled vocabularies and ontologies are already provided on the web, as we will show in Section 5.1. We expect FAIRification features to become embedded in commonly used business and research applications and in general purpose software, such as spreadsheets (see for example (Wigham et al. (2015)) and databases.

Currently it is still unclear to many data providers and users how to perform these actions needed for sharing their data. Moreover, they still require quite some manual, labor intensive work. We cannot expect all data users to become data scientists. Hence, this calls for automated support that can ease the task of researchers, combined with support from local '(meta)data cultivators', as we will explain later.

4. Cases

In this section we discuss three cases of data sharing in agriculture and food research. These cases illustrate which steps are needed to reach the next level of data management in this domain as we have set out in Section 3. These cases have been selected as they address the challenges defined in Section 2 (i.e., 1. Lack of interoperability 2. Lack of access; 3. Link with (IoT) devices; 4. Data management skills) in the agriculture and food science domain in diverse ways, also attempting to produce FAIR datasets. Table 1 links the cases'Crop trial data', 'Agri-food data service' and 'Consumer behavior data' to the challenges.

Table 1
Link between three cases selected against the four challenges presented in 2, where +++ means that this challenge is strongly present in this case; ++ means there is some consideration of this challenge, and + means that only a weak link exists between the challenge and the case.

	Challenge 1: lack of interoperability	Challenge 2: lack of access	Challenge 3: Link with devices	Challenge 4: data management skills
Case 1: Crop trial data	+++	+++	+	+++
Case 2: Agri-food dataservice	++	+++	++	++
Case 3: Consumer behavior data	+++	+++	+++	+++

4.1. Crop trial data

Problem of the community: In-situ field data is used to benchmark productivity but also for building and validating new models and datasets. In the domain of near real-time monitoring of agricultural productivity, in-situ crop observations are indispensable. They are used to improve accuracy and reliability of global and regional studies and of monitoring systems. For example, data on when and where a crop was on the field can validate arable crop maps that are usually based on (un) supervised classification of remote sensing imagery. One can also think of deterministic crop models like WOFOST, DSSAT etc., which need local information on phenology (emergence, flowering, maturity) for calibration and as continuous input.

It was recognized before that currently the agricultural community is fragmented in its data management and lacks commonly available reference data on agricultural production (Janssen et al. (2011)). Although there are several data portals (e.g., https://gardian.bigdata.cgiar.org/#/) and repositories (e.g., https://dataverse.harvard.edu/), where datasets are described and published, these datasets are scattered over various sources, lack standardization, and have incomplete metadata. This hampers the re-use of this data by others, causing an inefficient use of resources, while also limiting the calibration and validation work which in turn affects product quality.

FAIRness of the deployed technical solution: In different EU-funded projects and global initiatives (GEO, AGMIP), we developed an on-line global crop trial repository, called AgroSTAC (https://agrostac.org/en), to open key crop observations in a FAIR way. AgroSTAC aims for a minimum data set (MDS) for calibration and validation of methods, models and tools in the crop modelling and monitoring domain. Published, open data sets, for example from ODJAR.org, an open data journal for agricultural research, and DataVerse (https://dataverse.harvard.edu/) are screened for crop type, phenology, leaf area, biomass, and yield. Data are curated (see next paragraph) and stored in a database, designed such that it can store all aspects of an agricultural field trial. It requires that the trial has a location, and the observations and management events are timestamped. On top of the database, we developed tools and procedures to process, load, maintain and publish the data. to process, load, maintain and publish the data.

Process for data collection: Selected data goes through a dedicated data curation procedure. This is a crucial step to enable use of the data beyond its original purpose of collection. In this procedure metadata is checked and completed using all information available in the data files, supporting documents, publications in data and scientific journals, etc. If needed and feasible, we convert data to the required units and correct date format, we map phenology events to the BBCH scale and we link the data to variables based on the ICASA v2 master variable list (White et al. (2013)). The curation is documented in a wiki.

To test the quality of (open) data sets on open repositories eleven data sets were downloaded from the repository Harvard DataVerse and cleaned, out of a selection of 42 relevant data sets. Three data sets were further processed and imported into AGROSTAC. Some observations from this test are:

- From the initial list of 42 data sets identified (created 29 December 2017 by searching on 'maize trials') many data sets have restricted access (around two-third).
- None of the data sets is based on a predefined code book for field trials, only some had a dictionary.
- Almost all datasets had incomplete descriptions of variables and units. To a certain extent, data sets could be completed based on the data itself, additional data in other sheets or other documents, the related publication (if available) and domain knowledge of the data curator. Often obvious characteristics (e.g., year) were omitted.
- Some technical issues arose, in particular with floating point values due to different regional computer settings.

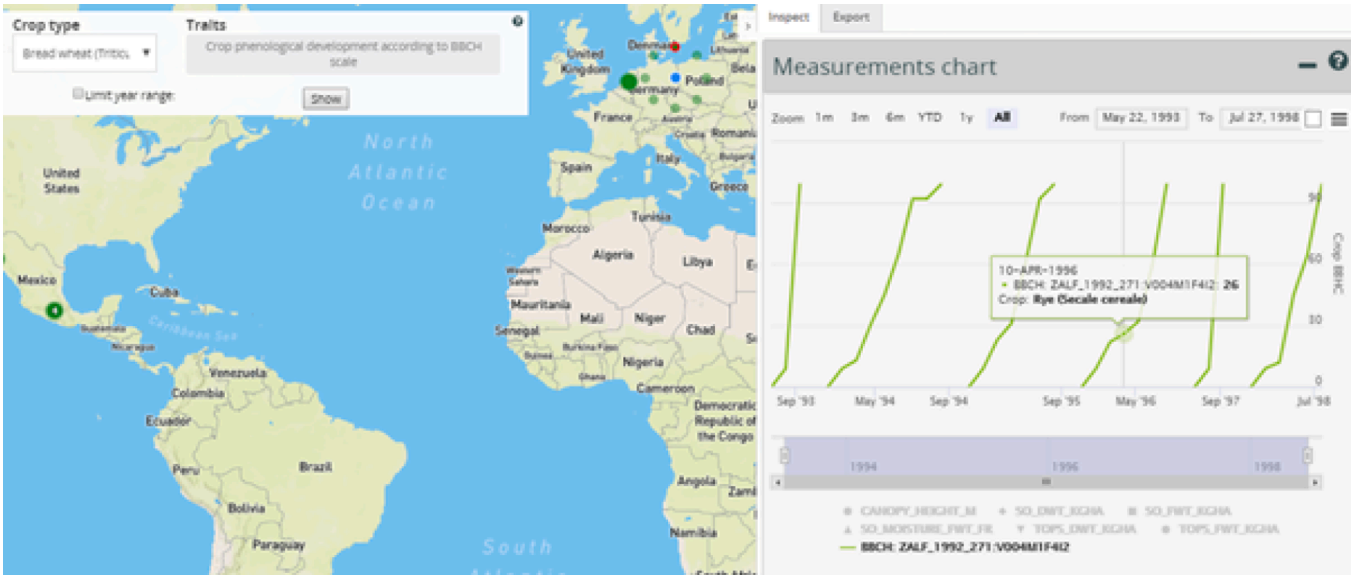

Fig. 1. Visual presentation of AgroSTAC data.

- At that moment, no ontology for relevant agronomic terms was yet available.
- For loading data into the AgroSTAC database, a text file format (SIF) was developed as well as a loader application that can read this format and load the data, while carrying out the necessary checks on the input data.

In addition to this data from Harvard Dataverse we also curated datasets from ODJAR.org. This curation work was easier as datasets published in ODJAR.org already underwent a review. However, these datasets still lack interoperability as they are stored in closed formats (no web endpoint) and are not yet in a standardized format.

The AgroSTAC data management system enables fast and easy data retrieval and facilitates sound data management and has a flexible data model. Coding of queries (e.g., in support of a REST interface) is straightforward and the queries run fast. The database can be exported as a triple store. For exploring the data, a viewer application is available (https://agrostac.vito.be/), see Fig. 1.

Lessons learned:

From this case we have learned that.

- It is useful to define a Minimum Data Set (MDS) for all trials that covers the aspects that are most needed by the community.
- It is crucial for the community to set standards for data formats and ontologies.
- Existing (legacy) in-situ data requires substantial effort to FAIRify.
- A data journal is needed as well as quality control of *meta*-data.

We recommend focusing on a limited set of variables and datasets that have most to offer for the invested time and resources, given typical applications in the domain.

To extend the FAIRification of existing in-situ data the following protocol is proposed. It includes downloading data from data repositories, saving original data, checking descriptions and units, mapping to the ICASA v2 master variable list, conducting additional visual and statistical quality checks, and adding metadata (reference to data repository, crop type). It also covers document curation, i.e., encoding, referencing external resources used, and describing any assumptions and ambiguities encountered. We suggest installing a curation team with an agronomy expert for supervising and reviewing the (meta) data. Currently the FAIRification of data for AgroSTAC relies mostly on project funding. To sustain AgroSTAC we are exploring sustainable business models and partnerships. Besides, we stimulate data providers to share curated data by emphasizing the advantages (e.g., attribution, recognition, re-use of data, new collaboration opportunities etc.) and offering a user-friendly guided data curation and publication procedure.

We know AgroSTAC is not the only and final solution. Publication of superior quality data for wider use and discovery requires integration of several actions in a workflow. Several functions need to be ensured at the same time, including archiving, quality checking, data curation, search and finally visualization and storage. These functions will be specific per domain and problem. To cover all of them, different platforms and capabilities need to play complementary roles. Within this landscape, we envision AgroSTAC as a vital component to clean and share a minimum dataset to the outside world following the FAIR principles.

4.2. Agri-food data service

Problem of the community: In the agri-food domain there is an increasing interest in applications that process data at the crop field, and even sub crop field level, instead of the previously more common gridded level of e.g., 1–10 km^2 scales. Naturally, this goes together with technological requirements to collect data at finer scales and increased storage and computing power available to researchers (see for example (de Wit et al., 2019). Many datasets are available as open data, such as yearly crop field data by RVO (https://english.rvo.nl), daily weather data by KNMI (https://www.knmi.nl), soil data by WUR and ISRIC (https://www.isric.org), terrain height AHN (https://www.ahn.nl, in Dutch) and satellite imagery (NASA, ESA), which can for example be used to calculate a vegetation index time series. Typically, a combination of these datasets is needed, requiring for each research project to carry out the time-consuming steps of collecting relevant data based on spatial area and time range, figure out what the data means, its lineage, and what e.g., the units are, find ways to harmonize it into a coherent dataset, store it and make it available for actual use.

FAIRness of the deployed technical solution: AgroDataCube (Janssen et al., 2018), (https://agrodatacube.wur.nl) is a repository that discloses the kind of data described above, using crop fields as the main assets. From the mandatory yearly crop registration by farmers, a limited subset of properties of all registered fields is made available as open data and imported into the repository. The geometry of the fields is used to relate them to other available assets in the repository. For example, the soil types for a crop field in a certain year are found by the spatial intersection of the two datasets. The database used handles such spatial queries when needed. A DOI (needed for 'F' and 'A' of FAIR) is available, which refers to a digital catalogue entry at the Wageningen University & Research library and directs the reader to the associated web page of AgroDataCube. This web page contains the information on how data can be retrieved from the repository, using standard web technologies and formats such as HTTP requests, REST and (Geo)JSON ('I', 'R'). A more complete implementation of the OpenAPI specification (https://swagger.io/specification/) is needed to increase the FAIRness for machines; currently they need to be AgroDataCube-aware. Humans will have less problems finding their way to the repository. The web page of AgroDataCube also provides information about the used license and the access token needed for data retrieval ('R').

Crop fields as digital assets have some inherent challenges, as they have a limited lifespan. This lifespan is based on the registration process that produces a draft version at the beginning of the growing season and a definitive version of the dataset near the end of the year. To our knowledge there is no publicly available global or national schema for providing crop fields with IDs that can serve as Digital Object Identifiers. Within the scope of AgroDataCube, each crop field is uniquely numbered ('F'), but typically a first matching needs to be done on spatial attributes to combine an external dataset with assets in the AgroDataCube repository. Here again FAIRness can be improved, as machines still need some AgroDataCube-awareness.

Metadata is available both at the dataset and data item level ('F', 'I'). For the latter it is included as a header block in all retrieved data. However, currently this metadata is mostly provided in the form of the original source and depends on that for its FAIRness. This is for example clearly visible in the crop field information that uses a vocabulary ('I') defined and maintained by RVO (https://english.rvo.nl) for the purpose of registration and e.g., provisioning of subsidies to farmers. Consequently, the crop descriptions are only available in Dutch, and the codes and descriptions can carry combined information (e.g., a mix of type of crop, type of soil it is grown on, and the purpose it is grown for), which can moreover fluctuate over time due to changing wordings and spelling.

In the EU Cybele project (https://www.cybele-project.eu), an experiment was carried out to add an additional API for accessing data from the AgroDataCube repository. It provides on-the-fly transformations between the semantic representations (SPARQL and RDF) and the web technologies (REST and JSON) used by the repository. It also specifies a mapping between the AGROVOC (FAO, 2021) vocabulary and the crop codes defined by RVO, instantly providing crop name translations into all languages supported by AGROVOC. Furthermore, it provides a connection to other datasets that are mapped to this vocabulary as well. This increases the 'F', 'A', and 'I' for machines considerably and reduces the required AgroDataCube awareness. The proof of concept has been realized using Metaphactory (https://www.

metaphacts.com/product) and can be extended with further mappings for other data in the repository, such as soil types and weather observations (see Fig. 2).

Process for data collection: Data is ingested into AgroDataCube in several ways, depending on the source and characteristics of the data:

- The RVO crop field data is made available yearly in two versions, a draft one in spring and a final one near the end of the year. Due to crop fields not having unique identifiers in the (open) datasets, geometric processing must be performed to add the data to the repository in a validated and consistent way.
- The KNMI daily weather data is collected periodically by a script, validated, and inserted into the repository. The data however is added as-is, which means that failing meteorological stations and sensors result in gaps in the time series.
- The soil data is taken from Wageningen University & Research datasets and only updated when new versions are made available, which is not very frequent (soils do not change rapidly).
- Remote sensing (satellite) based data such as the vegetation index is added when new images become available and have been internally pre-processed.
- Besides the above there is ongoing processing and improvement of the data (and the API) based on new insights and feedback from users.

Lessons learned: Distinct types of users will try to access and use published data in diverse ways, and it is difficult and expensive to serve them all. For example, the choice for GeoJSON as a response format prohibits GIS specialists from accessing the AgroDataCube as a regular OGC (https://www.ogc.org) compliant data source. The semantic API helps semantic specialists in using the AgroDataCube, but it is less clear how 'FAIR' the repository is, since it can depend on the perspective and knowledge of the user.

Source data might be available as open data through government resources. However, practical use reveals inherent problems in the data, e.g., geometric inconsistencies in crop fields or typos in crop descriptions or even a complete lack of identifiers over the years. These issues are repaired in subsequent versions but may mess up text search. Having harmonized metadata description helps, but it also can mask shortcomings of the underlying data, that are ideally solved at the source.

For the AgroDataCube itself, besides providing an API with documentation it is also needed to provide coding examples (https://github.com/AgroDataCube) on how to use the API properly. This gives confidence to data scientists and programmers when working with the data sources.

4.3. Consumer behavior data

Problem of the community: Consumer dietary behavior has a major impact on health and environment. Research in the field of consumer food choice can make considerable progress if more data on consumer behavior becomes available. This data can help generate insight into the influence that specific policies, but also food products and national diets have on human health and on our environment. It can also provide understanding of how psychological and social factors, lifestyle, and culture influence food choice. Do consumers take sustainability and fair pricing into account? Next to traditional studies based on self-reporting and surveys, such data is increasingly generated by electronic devices in real life settings (point-of-sale data, apps, food scanners, wearable technology, computer vision, etc.). In addition, data on food products is becoming available, not only in terms of ingredients, nutritional values, but also about sustainability (e.g., CO_2 footprint) and consumer acceptability factors. By linking and analyzing these different data sources, researchers can propose innovative solutions for important societal challenges, such as preventive health, environmental impact reduction and minimizing food waste. For consumers, digital dietary coaches can be developed to support decision making by individual

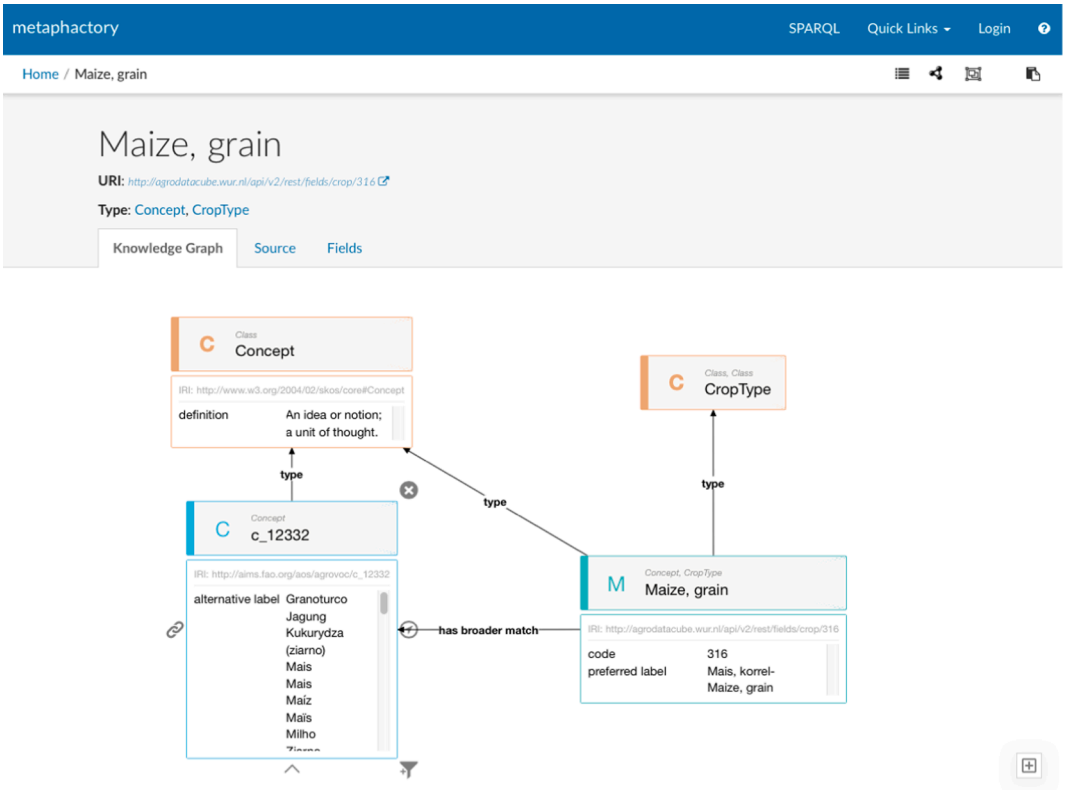

Fig. 2. Screenshot of Metaphactory, a tool for knowledge graph management.

consumers. The models and data in food and nutrition are for example used in studies to promote a fiber-rich diet (Rijnaarts et al. (2021)). The same underlying data is used by food processing companies to innovate their products. The current interest in 'digital twins' for monitoring and control of industrial processes, but also for supporting human decision making, clearly increases the demand for high quality data in this field. However, in this domain the range of data sources, formats and properties is even more diverse than in the first phase of the supply chain (primary production). Consumer data are collected by many organizations, from retail to NGO, food data originates from all different suppliers, environmental data requires complex life cycle analysis, etc. Reusing and combining data from such diverse sources requires that the data and its metadata be well defined *and* used. The FAIR principles can help to realize this if they are accompanied by guidelines that indicate which data, metadata and ontologies are suitable for reuse. Moreover, in this field we are often dealing with confidential personal or proprietary data. It is then crucial that data protection is properly respected.

Deployed technical solution: FNH-RI-platform, GS1 Data Source and food ontology. The European project RICHFIELDS (https://www.richfields.eu/richfields-final-conference-2018-in-brussels-belgium/) aimed to design a research data platform for scientists, businesses, policy makers and people to connect and share information about consumers' food behaviors. This initiated the development of a research infrastructure on Food, Nutrition and Health (FNH-RI) in Europe. Part of the platform has been realized, with a focus on setting up, storing, and analyzing consumer surveys. In addition, a repository for metadata provides access to models and data developed in this field and the broader area of agri-food.

In addition to requiring *consumer* data, decision support related to food also calls for *food* data. For example, to understand the impact of introducing new plant-based food products, we need to know which nutrients they contain, but also how they affect the texture and taste of the final product. So, properties that relate to consumer acceptance are crucial. This type of data is typically scattered and incomplete. Actors in the supply chain provide the minimal information that is legally required, using local, text-based standards. Additional data on, for example, health and environmental is still only available to a certain extent and in standard formats.

Process for data collection: Data is provided by consumers partly manually and partly using smart devices. NGOs collect population data on food consumption and food choice. Food producers record product data in their running processes, varying from manual notes to automated sensor input.

Collecting data on human behavior is notoriously difficult. For example, food intake is measured by so-called food frequency questionnaires, 24-hour recalls or intake diaries. However, these are well-known to be inaccurate and incomplete and considered by consumers cumbersome to complete. Measuring the consumer's health status is even more problematic, since it requires invasive techniques, for example taking blood samples. These complications have sparked off numerous initiatives to collect such data in a non-intrusive way, for example using computer vision technology or lab-on-a-chip solutions. However, only when data from such devices is combined with circumstantial and personal data, meaningful conclusions can be drawn. Applying the FAIR principle helps make these combinations. It will support automated inference on the merged data. This is not widespread practice yet.

In terms of collecting food product data, other issues arise. Food companies are legally obliged to declare for their products which ingredients they contain, and what their nutritional values and energy content are. Centralized organizations such as GS1 (https://www.gs1.org/) provide text-based standards to express this information. They act as channels to share data with retailers and other organizations. However, the quality of the data is not always optimal and paper labels are still considered as the 'ground truth'. A change towards distributed sourcing of this data directly from the original sources (the providers) and the use of a machine readable linked-data approach would significantly increase the quality and usability of this data in many applications. This would be beneficial for all food supply chain actors, including the consumer.

Lessons learned: In the domain of food and nutrition many opportunities exist to valorize data on consumer behavior and food products and processing. Challenges are (1) lack of sufficiently detailed controlled vocabularies and ontologies, (2) limited support for respecting privacy and confidentiality and (3) availability of non-intrusive measurements. Fig. 3 shows a web service that provides access to environmental footprint data. Several services are currently being developed, covering a wide range of food products and food product properties. The data disclosed by these services are expressed in terms of a food ontology. The FoodOn initiative (https://foodon.org/) is one of the most promising

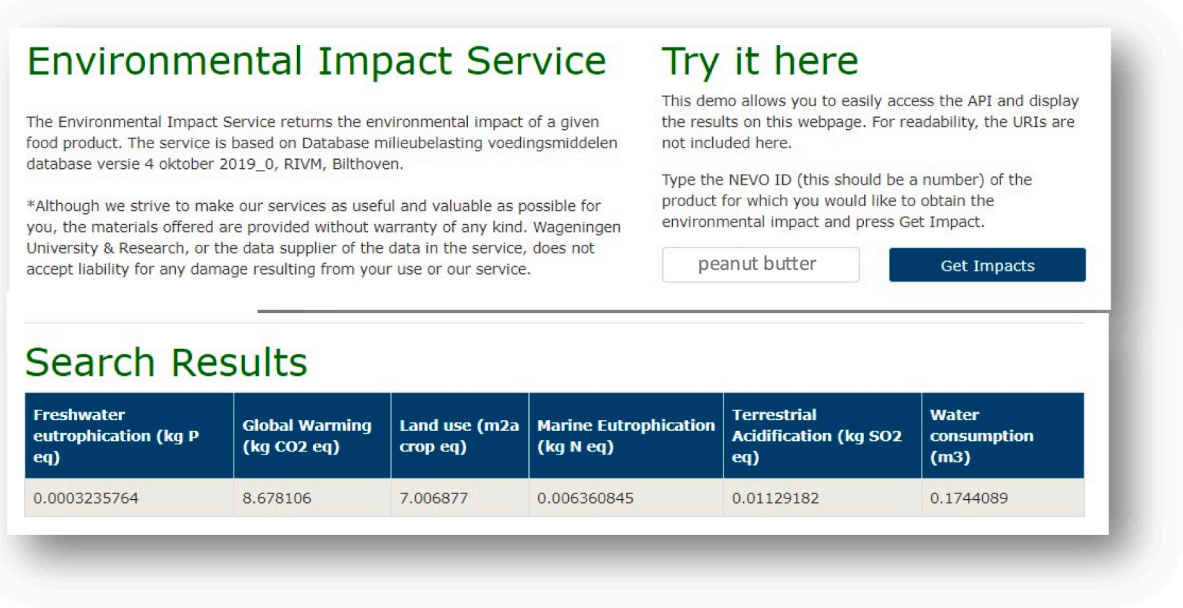

Fig. 3. Example of a web service providing environmental footprint data for a food product.

initiatives for ontology development in this field.

5. Solution paths

The above and other case studies have revealed some issues in making agricultural and nutrition data FAIR. They are practical examples of how this currently works in addressing the given global challenges. From such use cases important lessons can be drawn on how to extend the application of the FAIR principles for data sharing in the food and agriculture domain.

5.1. Controlled vocabularies and ontologies in food and agriculture

As identified in earlier sections, there is a need for further standardization and harmonization of concepts in the agriculture and food domain through common controlled vocabularies and ontologies. As shown in Table 1, there are several models available, each with its own application area, expressiveness, and scope. Every application has its specific needs and will require (parts of) the selected models, but also dedicated additions and modifications. Moreover, it is essential to make a distinction between terminology and general classification on the one hand, and complex data structures, with properties and constraints, on the other. Whereas a *controlled vocabulary* provides a light-weight taxonomy of terms, an *ontology* provides a logical model with formalized concepts and relations in a domain. Vocabularies are typically used to support search in unstructured data sources, whereas ontologies support querying and reasoning over complex but structured datasets. Vocabularies and ontologies can directly be mapped to software structures. For example, a controlled vocabulary such as AGROVOC (see Table 2) has been developed to structure terminology in the agriculture and food domain with the purpose of finding resources in a library of books and articles. It is not intended for connecting data in precision farming and automatic control of machines. On the other hand, the ontology OM for representing units and quantities expresses which quantities can have which units and how units can be converted in associated software services (Rijgersberg et al. (2013), Keil et al. (2018)).

However, even if vocabularies and ontologies facilitate finding and connecting different data sources, it still requires some effort to apply them in practice, for example for merging distinct datasets. Moreover, vocabularies and ontologies need to be updated continuously, given feedback from practical applications. One issue is that the human readable labels they provide for concepts are often ambiguous. Terms that have similar labels can refer to different concepts and different labels are used for similar terms. This happens for example in a list of crops, animals, fertilizers, pesticides, foods, etc., as explained in more detail in Janssen et al. (2011). The term 'maize' can refer to maize as a crop, but also to maize as a single food product. In such cases the neighboring concepts in the graph explain what is meant by the single node. For example, either 'crop' or 'food product' being the broader term clarifies the difference. Reusing vocabularies and ontologies is therefore not straightforward if certain subdomains are not well covered in quantity and quality. The following table lists a number of vocabularies and ontologies in our domain.

More effort is needed to further standardize and harmonize vocabularies and ontologies in the agri-food domain and to share experience on applications of these models. Until now this has depended on

Table 2
Overview of available vocabularies and ontologies in the agriculture and food domain. Several relevant sources can be found at AgroPortal (http://agroportal.lirmm.fr/).

Vocabulary	Reference	Characteristics
ADAPT	https://www.aggateway.org/GetConnected/ADAPT(inter-operability).aspx	Toolkit to link to farm data, defining a common object model. Support from machinery and sensor producers, focus on technical protocols
AgroFIMS	https://agrofims.org/about	Crop management, complete coverage of field trial design and activities based on a field book. Tied to CG Core Metadata Schema and the Agronomy Ontology (AgrO)
AGROVOC	https://aims.fao.org/agrovoc	Food and agriculture-related terms, controlled vocabulary, comprehensive (more than 38k concepts), multilingual (40 languages)
ask-Valerie	https://www.foodvoc.org/page/Valerie-9	Innovations in agriculture and forestry. Multilingual controlled vocabulary, covers agriculture and forestry for several themes
ATO	https://www.animalgenome.org/bioinfo/projects/ato/	Animal Trait Ontology. Animal genetic traits for distinct types of animals. Limited scope.
BRAPI	https://www.brapi.org/specification	Plant genetics ontology, limited scope. Compatible with MCPD, MIAPPE, GA4GH Variants Schema, GeoJSON and Crop Ontology.
Crop Ontology	https://www.cropontology.org/	Small ontology per crop with different variables, limited alignment, no master list of crops.
eCrop	https://unece.org/fileadmin/DAM/cefact/brs/BRS_eCROP_v1.pdf	Agricultural management. Lacks master lists of fertilizers, pesticides, biological control agents, etc.
FAOSTAT	https://www.fao.org/faostat/en/#definitions	Text-based standards for worldwide statistical data on food production, security, trade, sustainability, etc.
FoodEx2	https://www.efsa.europa.eu/en/data/data-standardisation	Food safety, standardized system for classifying and describing food
FoodOn	https://foodon.org/	Food and nutrition ontology, including organisms, production, processing, packaging, product hierarchy, etc.
GACS	https://aims.fao.org/fr/global_agricultural_concept_scheme_gacs	Global Agricultural Concept Scheme, multilingual, connects AGROVOC, CAB and NAL
GO	https://geneontology.org/	Gene Ontology. Controlled vocabulary, extensive (more 43k terms) and many applications
GPC	https://gpc-browser.gs1.org/	Global Product Classification. Logistics, traceability. More than 130k food products
GS1 Web vocabulary	https://www.gs1.org/voc/	Controlled vocabulary for GS1 standards, extension to Schema.org.
ICASA variable list	https://dssat.net/data/standards_v2/	Focus on crop trials and models. Comprehensive list of variables used in crop simulation models
Langual	https://www.langual.org/default.asp	Food product data. Predecessor of FoodOn
NAL	https://agclass.nal.usda.gov/	Controlled vocabulary, partly multi-lingual, more than 265k terms
OBO - Open Biological and Biomedical Ontologies	https://www.obofoundry.org/	Open Biological and Biomedical Ontology Foundry. Extensive collection of ontologies for the biological sciences
OGC	https://www.ogc.org/docs/is	Geospatial (location) information, implemented by many organizations, readily accepted
OM	https://www.wur.nl/en/product/Ontology-of-units-of-Measure-OM.htm	Comprehensive ontology on units of measure, quantities, dimensions.
QUDT	https://www.qudt.org	Units of measure, quantity kinds, dimensions, and data types

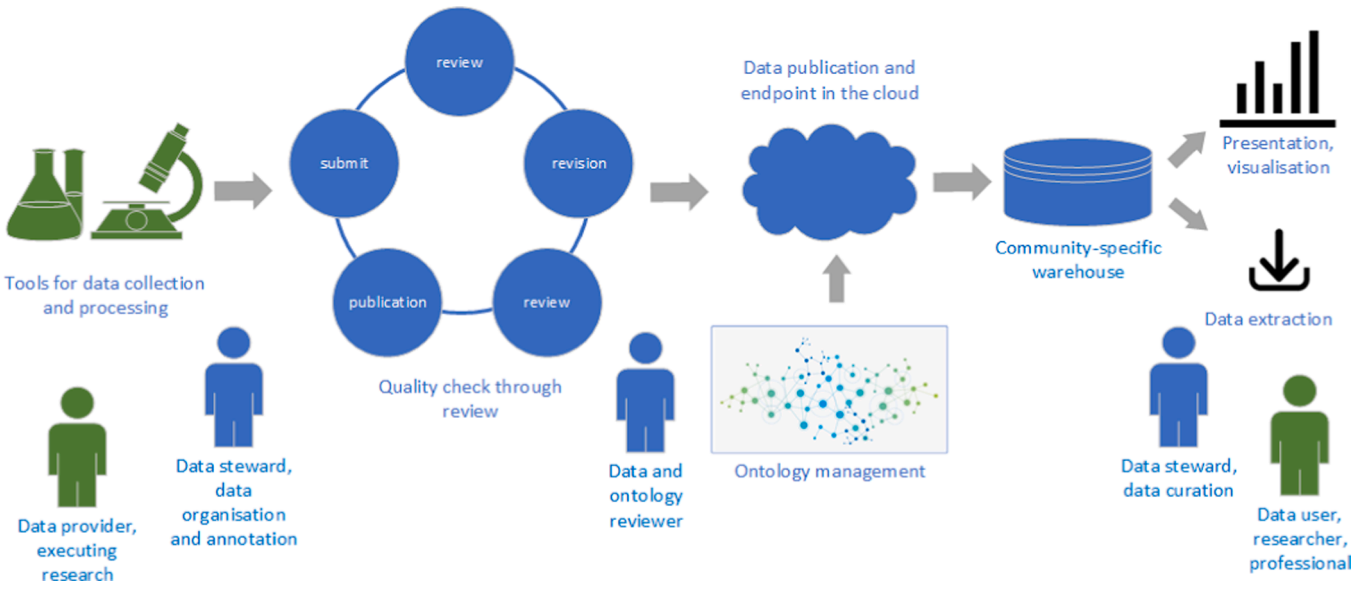

Fig. 4. Stepwise overview of the process for data collection, sharing and use.

scattered and haphazard actions of different players in the field. We can also hypothesize that the agriculture and food researchers and their respective communities do not take sufficient ownership of the common standardization and harmonization efforts required to achieve easy data use. This role of the scientific communities was clear in the cases presented above, where efforts are often dependent on a single university, research institute or project team taking steps to define a list of commonly agreed terms. However, harmonization efforts by the broader community remain undefined, and the original creative process undocumented. Many of the vocabularies in Table 2 are the result of initiatives to achieve some sort of alignment in a community. A successful case is BRAPI, in which a consortium of universities manages to jointly maintain and align a set of vocabularies. Also, in the field of geo-spatial data, the OGC plays a leading role in defining relevant standards that are commonly followed in the community. Here the leading role of these standards in many applications plays a key role. In the food domain, FoodOn is gradually taking up a leading position and is supported by several research organizations world-wide. For adoption by the food industry and trade organizations a transition from existing text-based to machine readable standards is needed. GS1 can be a key player in this process.

Undoubtedly the role of the communities involved needs to be strengthened to ensure that relevant and robust vocabularies emerge. This requires the recognition by the formal and informal leaders of these scientific communities of the importance of such vocabularies for the general advancement of their science. Funders need to allocate resources for the long-term development, and universities and research institutes should recognize that contributing to vocabulary development is an addition to a scientific career and not a distraction.

5.2. Fairification

The FAIR principles have been instrumental in getting good practices on data management on the agenda of the scientific and science funding community. The increased awareness of the good practices has at least led to more publication of data sets gradually, however, it is unclear whether it has also led to more data use. The application of the FAIR principles to different cases leads to new challenges once one descends into the particularities of a specific scientific domain and use case. Achieving interoperability (the I in FAIR) by expressing metadata in terms of controlled vocabularies and ontologies is still a demanding task. From the use cases lessons can be drawn towards the implementation of the FAIR principles.

The first lesson is that the FAIR principles address only a part of the data publication process. They define what is needed to publish a suitable dataset and associated vocabularies. They do not prescribe all steps needed to proceed from making the initial observations to providing high quality data and metadata, ready for use in software. An additional description of the FAIRification implementation process is given in (Jacobsen et al. (2020b)). This, however, still leaves the different roles of individuals involved implicit. To clarify these roles, we propose to distinguish between ‘data provider or user’, ‘data steward’ and ‘data reviewer’. We link these roles to the process of data collection, annotation, review, publication, curation, and use, as visualized in Fig. 4.

First, we have the *data provider* or *and the data user*. This role is typically played by researchers and professionals performing their daily tasks, generating data, and using data from other sources. This can range from the scientist performing lab experiments to the farmer ploughing his land, and from the journalist authoring an in-depth article on the environmental impact of modern farming to the consumer following a healthy diet. They typically approach the data through the tools for data collection and processing they are accustomed to in their everyday work, assuming that the data involved are properly understood and valid for the cases considered. The *data provider* must decide which data to publish; what is needed to verify or replicate his work? Is it of interest for peers (for example, should one share ‘failed experiments’)? What are potential risks of misunderstanding and misuse? On the other hand, the *data user* decides which part of the available data to use, how to process and present them. They estimate the quality of the data, i.e., what is its fitness-for-use in their own case? These are tough questions to answer in general and may take considerable effort to answer. The advantages of sharing data are clear, but they come at a price. In the research community for example this price cannot only be paid by the researcher, also considering that there are already high demands on the researchers in terms of writing grant proposals, scientific papers, and teaching students.

The role of the *data steward* to take up part of this burden has already been amply described in the FAIR related literature (Mons (2021)). Data stewards assist researchers in implementing the FAIR principles from an organizational point of view. Each research group or unit employs a data

Table 3
Five-star model of data quality and use.

Number of stars	Status	Quality level
*	Raw	Published in a data repository that is externally accessible and findable, contains raw data with some textual meta-description (institution, publisher), with a DOI for the entire dataset.
**	Reviewed	One star + published in a data journal with reviews and feedback; metadata as textual description with more information on variables and provenance.
***	Linked and licensed	Two stars + metadata and variables linked to public ontologies or controlled vocabularies in linked data format; license attached for use of the data.
****	Operationalized	Three star + a data expert from a specific community has created a version of the data according to the agreed common data model for that community and has added community-specific metadata.
*****	Used	Four star + use of the dataset by others in another study, tracked through citations; added metadata on scope of re-use: for what does the data work and for what not?

steward as a gatekeeper for creating and implementing data management plans. It is stated that 5% of a research budget should be spent on data stewardship (Mons, 2020). On the data production side, the steward helps data providers to publish high-quality and FAIRified data effectively with endpoints on the web (cloud). She assists in finding ways to describe and store data (if possible in community-specific data warehouses), to identify potential data publication outlets, and to find and use relevant vocabularies. Whereas the data steward assists in properly annotating the data for publication, the data itself should not be touched. The original creator of a dataset is responsible for performing experiments in a proper way and making accurate observations. On the data consumption side, the data steward assists users in preparing data from external sources for research goals. The data user is responsible for selecting data elements and making proper inferences over the data. We can also consider *data curation* for a particular application context as part of the data steward's task (covering both metadata and data), but this requires a close connection of the steward with the cases considered.

Hence, the role of the data steward alleviates the task of the researcher. However, the institutional and scientific effort involved still causes much pressure on the capacity and budget available in research projects. This pressure can be reduced by (1) partly automating the process (for example checking parameter compatibility or for unit conversion), and (2) by introducing the role of the *data reviewer*.

We propose the role of the *data and ontology reviewer* as the intermediate actor between data providers and data users. When data publication becomes more of a widespread practice, like manuscript publication, its value for scientific careers will become apparent. Data reviewers are scientific peers that can evaluate the general fitness-for-use and understandability of a dataset for use by other researchers in the same domain, based on the metadata provided by the researcher who originally collected the data. These reviewers operate in the same way as reviewers of scientific manuscripts, being independent from the original author and not coming from the same lab or research group. Data reviewers can define a minimal information level that prescribes which variables and meta-data is required for any dataset to be published in a specific domain. An example of such a standard in proteomics is MIAPE (https://www.psidev.info/miape).

The second task of data reviewers, as domain experts but not tied to specific projects, is to ensure that ontologies in this domain are available, properly published, updated, and documented. If needed they initiate new ontology development projects. Finally, they check whether the considered datasets properly link to these accepted, community-driven vocabularies and ontologies.

Researchers and non-academic professionals can act as data reviewers, provided they distinguish their data needs from the needs of the community in their domain – not unlike manuscript reviewers. In that way, the organizations they are embedded in can carry their part of the burden either financially or through time resources, to ensure that shared data becomes usable. Here again the importance of scientific and professional communities emerges as a mechanism to support the dialogue between data providers and users.

5.3. Data maturity levels

Data sharing is all about the quality of data and metadata, and the best proof of success is in the automated use of data in other studies than for which they were originally created. However, the ideal situation of fully automatic reuse will not always be achieved in practice, given limited resources, capabilities, and tools available. To our best knowledge no maturity model for automated data reuse has been defined yet. Based on the existing FAIR principles we propose a five-star model for data quality, inspired by the five-star model of Tim Berners Lee for Linked Data (https://5stardata.info/en/). Whereas the latter refers to the technical level of linking data elements, our five levels for data maturity express to which extent a dataset is reused or at least ready to be reused. The levels are shown in Table 3.

Although we addressed data in particular, similar conditions hold for sharing computational models, algorithms, and tools. The latter relate directly to computation, in contrast to more descriptive research outputs, such as descriptive theories, explanations and debates that are shared through traditional publications (Goldacre et al. (2019)).

6. Conditions for easier data use

Based on the above observations, we identify four conditions that will have to be met for data sharing across research domains to succeed.

A first condition is the availability of automated tools for data providers and users, but also for stewards and reviewers. Data annotation tools are needed to facilitate easy alignment with commonly agreed vocabularies during data production. For example, an analyst capturing data in a lab journal or a crop trial manager who logs activities in a field book should be supported in connecting their data definitions to concepts defined in a shared vocabulary. Ideally this functionality is embedded in applications that are already commonly used, such as spreadsheets. Some initiatives have been taken in this direction (Wigham et al., 2015; Wolstencroft et al., 2011), but most commercial (generic) solutions lack such functionality. An additional benefit of this would be that data providers avoid mistakes already at the point of creating data and become aware of potential reuse of the data at an early stage.

A second condition is that funding, tools, and ontology development are *community-driven*. It is an institutional effort to finance and build the required infrastructure. Domain-specific communities must manage pools of data sets, following the needs and particularities of each domain, including a clear governance model for sensitive data. Problems such as missing data, out-of-range data, missing metadata, concept matching and ontology alignment are specific for a domain or even the actual problem to be solved. Tools for annotation, review and curation of data can best be developed at a community level. Such communities must be small, for example within agri-food farm-household modelers require quite different data with different challenges than crop-modelers (Hammond et al. (2017)), see Section 4. Considering agri-food as just one community will not lead to substantial progress, as it is too diverse internally.

A third condition is that we acknowledge that sensitive data (e.g., because of confidentiality, privacy, implicit assumptions that may lead to misuse) cannot be opened through an open-by-default policy. In food

systems, private sector parties play a key role in creating and applying data from the supply chain, which are valuable for others, for example in research. These private parties often insist on non-disclosure of data that researchers have used to generate their articles. However, there should be a way to give restricted access to reviewers and secondary users to at least the metadata to verify the quality of the data for a particular purpose. Here, transparent governance mechanisms need to be adopted and shared across communities. As an example of a rule for data access, usage could be restricted to accredited research organizations or consultants only, or to a particular research project. These rules and criteria need to be objectified and universally applicable, also hold beyond the research project's lifetime, implemented in clear authorization mechanisms that also avoid stagnation in data use.

One last condition for improving data sharing practices is to create insight into how data is actually (re)used in scientific communities. Such insight could for example come from a serious study of the role of data in scientific meta-reviews, an example being the Cochrane reviews managed by the Cochrane Collaboration in the health domain (https://www.cochranelibrary.com/). Most other scientific domains do not yet have such review mechanisms, thereby lacking mechanisms to connect widely available data (see, for example, (Suškevičs et al. (2017)). These scientific communities need to explicitly discuss and think about the next-generation research questions that can be answered and how they are going to make shared data pools available as a scientific community.

7. Conclusion

Data sharing in the agriculture and food domain is gradually becoming accepted practice. With some cases we have demonstrated concrete efforts. We have emphasized the need for distributing the work involved in annotating data over different players in the information chain. We have distinguished three complementary roles to execute this work: *data provider/user*, *data steward* and *data reviewer*. The latter needs some level of public commitment in specific domains. Interaction between researchers producing the data and those using the data is crucial to create pools of datasets that have the potential to be 'fit-for-purpose' and 'within-scope' for related applications that use that data. Finally, we have defined five levels of data sharing maturity to create awareness and listed four general conditions that need attention.

To conclude, while the FAIR principles have been instrumental in raising awareness on data sharing with researchers and other professionals, they have not yet shown to be particularly helpful in implementation of operational data sharing within agri-food communities. Many agri-food communities still lack the crucial building blocks required, such as shared vocabularies, sufficient quality data sets and shared data handling practices. Our findings suggest that a casuist-approach (i.e., case-by-case investigation) to data-sharing in concrete scientific communities is preferred to depositing generic principles that are hard to operationalize.

Funding

This work was supported by the Dutch Ministry of Agriculture, Nature and Food Quality in the Wageningen UR research program *Data Driven & High Tech*. They were not involved in this research and in the preparation of this article.

CRediT authorship contribution statement

Jan Top: Conceptualization, Supervision, Writing – review & editing, Writing – original draft. **Sander Janssen:** Methodology, Writing – review & editing, Writing – original draft. **Hendrik Boogaard:** Investigation, Writing – review & editing. **Rob Knapen:** Investigation, Writing – review & editing. **Görkem Şimşek-Şenel:** Investigation, Writing – review & editing.

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.

References

Australian_Academy_of_Science, 2021. Advancing data-intensive research in Australia.
Carolan, L., Smith, F., Protonotarios, V., Schaap, B., Broad, E., Hardinges, J., Gerry, W., 2015. How can we improve agriculture, food and nutrition with open data? Open Data Institute 2015.
Cash, D.W., Clark, W.C., Alcock, F., Dickson, N.M., Eckley, N., Guston, D.H., Jager, J., Mitchell, R.B., 2003. Knowledge systems for sustainable development. Proc. Natl. Acad. Sci. USA 100 (14), 8086–8091. https://doi.org/10.1073/pnas.1231332100.
de Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agric. Syst. 168, 154–167. https://doi.org/10.1016/j.agsy.2018.06.018.
Fao, 2021. AGROVOC – Semantic data interoperability on food and agriculture. FAO, Rome, Italy.
Goldacre, B., Morton, C.E., DeVito, N.J., 2019. Why researchers should share their analytic code. BMJ 367 (6365).
Hammond, J., Fraval, S., van Etten, J., Suchini, J.G., Mercado, L., Pagella, T., Frelat, R., Lannerstad, M., Douxchamps, S., Teufel, N., Valbuena, D., van Wijk, M.T., 2017. The Rural Household Multi-Indicator Survey (RHoMIS) for rapid characterisation of households to inform climate smart agriculture interventions: Description and applications in East Africa and Central America. Agricultural 151, 225–233. https://doi.org/10.1016/j.agsy.2016.05.003.
Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C.T., Goble, C., Guizzardi, G., Hansen, K.K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R.W.W., Imming, M., Jeffery, K.G., Kaliyaperumal, R., Kersloot, M.G., Kirkpatrick, C.R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L.O.B., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M.D., Willighagen, E.L., Wittenburg, P., Roos, M., Mons, B., Schultes, E., 2020a. FAIR principles: interpretations and implementation considerations. Data Intelligence 2 (1–2), 10–29. https://doi.org/10.1162/dint_r_00024.
Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L.O.B., Mons, B., Schultes, E., Roos, M., Thompson, M., 2020b. A generic workflow for the data FAIRification process. Data Intelligence 2 (1-2), 56–65.
Janssen, H., Janssen, S.J.C., Knapen, M.J.R., Meijninger, W.M.L., van Randen, Y., la Riviere, I.J., Roerink, G.J., 2018. AgroDataCube: a big open data collection for agri-food applications. Wageningen Environ. Res. https://doi.org/10.18174/455759.
Janssen, S.J., Athanasiadis, I.N., Bezlepkina, I., Knapen, R., Li, H., Domínguez, I.P., Rizzoli, A.E., Ittersum, M.K.v., 2011. Linking models for assessing agricultural land use change. Computers and Electronics in Agriculture 76(2) 148-160. DOI 10.1016/j.compag.2010.10.011.
Janssen, S.J., Porter, C.H., Moore, A.D., Athanasiadis, I.N., Foster, I., Jones, J.W., Antle, J.M., 2017. Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology. Agric. Syst. 155, 200–212. https://doi.org/10.1016/j.agsy.2016.09.017.
Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. Comput. Electron. Agric. 143, 23–37. https://doi.org/10.1016/j.compag.2017.09.037.
Keil, J.M., Schindler, S., Brodaric, B., 2018. Comparison and evaluation of ontologies for units of measurement. Semantic Web 10 (1), 33–51.
Lokers, R., 2019. Guidelines for analysing pathways to impact: Evaluation of open data for development F1000Research.
Marvin, H.J., Janssen, E.M., Bouzembrak, Y., Hendriksen, P.J., Staats, M., 2017. Big data in food safety: an overview. Crit. Rev. Food Sci. Nutr. 57 (11), 2286–2295. https://doi.org/10.1080/10408398.2016.1257481.
Mey, L., Berdou, E., Ayala, L.M., Lokers, R., 2019. Open Data Impact Narratives – Stories of Impact of Open Data in Agriculture and Nutrition, GODAN F1000 Gateway. GODAN. DOI 10.7490/f1000research.1117566.1.
Mons, B., 2020. Invest 5% of research funds in ensuring data are reusable. Nature 578 (7796), 491. https://doi.org/10.1038/d41586-020-00505-7.
Mons, B., 2021. Data stewardship for open science: implementing FAIR principles. Chapman and Hall/CRC.
Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016. https://doi.org/10.1016/j.agsy.2020.103016.
Rijgersberg, H., van Assem, M., Top, J., 2013. Ontology of units of measure and related concepts. Semantic Web 4 (1), 3–13. https://doi.org/10.3233/SW-2012-0069.
Rijnaarts, I., de Roos, N.M., Wang, T., Zoetendal, E.G., Top, J., Timmer, M., Bouwman, E. P., Hogenelst, K., Witteman, B., de Wit, N., 2021. Increasing dietary fibre intake in healthy adults using personalised dietary advice compared with general advice: a single-blind randomised controlled trial. Public Health Nutr 24 (5), 1117–1128. https://doi.org/10.1017/S1368980020002980.
Suškevičs, M., Hahn, T., Rodela, R., Macura, B., Pahl-Wostl, C., 2017. Learning for social-ecological change: a qualitative review of outcomes across empirical literature in natural resource management. J. Environ. Plann. Manage. 61 (7), 1085–1112. https://doi.org/10.1080/09640568.2017.1339594.

Verdouw, C., Sundmaeker, H., Tekinerdogan, B., Conzon, D., Montanaro, T., 2019. Architecture framework of IoT-based food and farm systems: a multiple case study. Comput. Electron. Agric. 165, 104939. https://doi.org/10.1016/j.compag.2019.104939.

White, J.W., Hunt, L.A., Boote, K.J., Jones, J.W., Koo, J., Kim, S., Porter, C.H., Wilkens, P.W., Hoogenboom, G., 2013. Integrated description of agricultural field experiments and production: the ICASA Version 2.0 data standards. Comput. Electron. Agric. 96, 1–12. https://doi.org/10.1016/j.compag.2013.04.003.

Wigham, M., Rijgersberg, H., Vos, M.d., Top, J., 2015. Semantic Support for Tables using RDF Record Table. International Journal on Advances in Intelligent Systems 8(1&2) 16.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3 (1). https://doi.org/10.1038/sdata.2016.18.

Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J., 2017. Big data in smart farming – a review. Agric. Syst. 153, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023.

Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J.L., du Preez, F., Goble, C., 2011. RightField: embedding ontology annotation in spreadsheets. Bioinformatics 27 (14), 2021–2022. https://doi.org/10.1093/bioinformatics/btr312.