Food and Agriculture Organization
of the United Nations

Farm data management, sharing and services for agriculture development

©Adobe Stock/only_kim

Farm data management, sharing and services for agriculture development

Food and Agriculture Organization of the United Nations
Rome, 2021

Required citation:
FAO. 2021.
Farm data management, sharing and services for agriculture development. Rome. https://doi.org/10.4060/cb2840en

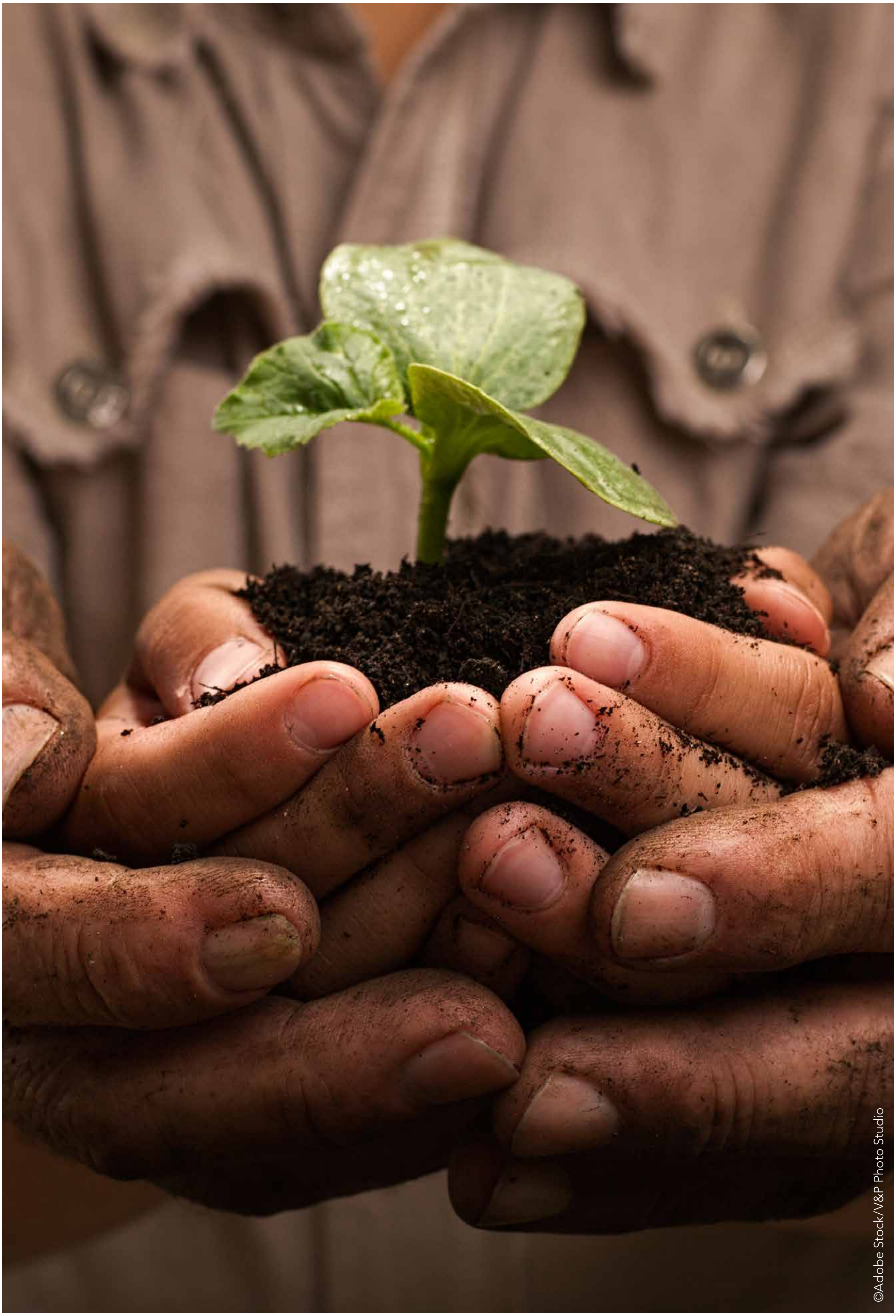
The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations (FAO) concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dashed lines on maps represent approximate border lines for which there may not yet be full agreement. The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by FAO in preference to others of a similar nature that are not mentioned.

The views expressed in this information product are those of the author(s) and do not necessarily reflect the views or policies of FAO.

ISBN 978-92-5-133837-7

Cover photograph: ©Adobe Stock/Kletr

Contents



©Adobe Stock/V&P Photo Studio

Preface

There is an exponential growth in data accompanying the digitalization of agriculture through the proliferation of mobile technology, remote sensing technologies and distributed computing capabilities. The effective management of data will open up new opportunities to better the lives and livelihoods of smallholder farmers by lowering cost and reducing information asymmetries. However, the lack of experience in data management or adoption of data driven services can limit the possibilities of digital transformation.

Better access to markets can result from added value to the crop. For example, a coffee grown above a specific altitude can command a premium. Hence the data on the farm and associated tracking of that consignment can mean new customers and higher prices for the farmers. Yields can be improved by knowing more about the farmer and farm and targeting extension advice and provision of fertiliser. Tea growers in Uganda who were profiled benefitted from field mapping which was used to more accurately provide fertiliser on credit. By registering farmers and aggregating purchase of inputs smallholder farmers purchased inputs at a discount.

Data can also be used to improve access to credit by registering the farmers' sales and mapping their fields by recording their history providing a dossier which improves confidence in the lender. Improvements in the way the value chain is organized come from knowing the association or agribusinesses members through data and arranging better collection of the crop through mapping and jointly planning crop calendars with targeted groups and improving trust within the group.

In this context, Food and Agriculture Organization of the United Nations (FAO) has a long experience in data management, curriculum development and training together to produce a valuable introduction and foundation to farm data management. Together with the Technical Centre for Agricultural and Rural Cooperation ACP-EU (CTA) and the Pan African Farmers Organisation (PAFO), FAO has demonstrated how farm data management is key to farmers' organizations supporting better access to markets, finance and inputs. These changes could improve productivity, farmer's livelihoods and resilience. Insights from the data were also crucial in managing the value chain and informing food security policy. This has been materialized with the publication of this book.

This book is a result of partnership between FAO, CTA and PAFO with the objective to develop a set of training programme for farmers to create awareness on information and communication technologies (ICTs) in agriculture, based on the experience from the Global Open Data in Agriculture and Nutrition Action Project (MTF / GLO/694/SDL). Whilst traditionally face to face workshops and training have been the key tools to disseminate knowledge on farm data management, FAO provided the opportunity to disseminate information and knowledge through a Massive Open Online Course (MOOC). This book features the rich content of this course.

The objective is to extend the audience to those interested in farm data management who were unable to attend the online course. In particular the book is aimed at those working in farmers' organizations as administrators or staff for example those collecting farmer data and managing the data; and development practitioners and technology providers, who assist farmers' organizations creating data services. They can benefit by applying some of the principles described in the book and hopefully benefit from some of the example lessons learnt.

Acknowledgements

The "Farm data management, sharing and services for agriculture development" was jointly prepared by the Food and Agriculture Organization of the United Nations (FAO) and the Technical Centre for Agricultural and Rural Cooperation (CTA), as a result of the work funded by CTA with the Pan African Farmers' Organisation (PAFO). *This book is dedicated to Hugo Besemer*

©Adobe Stock/Jacob Lund

About this book

This book consists of four chapters, with fifteen sections, and provides a guide to understand the value of data, the different types and sources of data and identify the type of services that data enables in agriculture.

The first chapter "Data, services and applications" focuses on the topics of value of data in agriculture to support farmers, how to increase their income and develop food production, digital farmer profiling and the strategies to design business models for profiling.

The chapter two "Data sharing principles" takes the readers to the principles and benefits of shared data, the potential of using and publishing data in agriculture, responsible data sharing practices for farm data, ethical and legal sensitivities of data-driven services and data protection. In detail, it addresses challenges in data sharing for smallholder farmers, issues regarding data ownership and data rights, outlines different roles of public and private data sources, challenges in reusing them in services for farmers.

The third chapter "Using data" guides readers on how and where to find open data, data quality elements, data analysis and visualisation with more technical background and in a broad sense, not only relevant for farm data. This chapter is based on the free online course on "Open data management in agriculture and nutrition" (GODAN Action, 2018).

The last chapter of the book "Exposing data" provides an overview on conceptual frameworks for sharing data and outlines comprehensively the ways to make data findable, accessible, interoperable, and reusable. The chapter goes deep into the interoperability and semantics as a key element in reuse of data by different systems at machine level.

Book readers can either go through the full content and embrace all aspects of the topic or may want to visit specific parts of the book based on their area of interest. The wider community at all levels in the agricultural domain can benefit from this book and build their own stories to tell from implementing some of the lessons learnt.

©Adobe Stock/kozlik_mozlik

Abbreviations and acronyms

ABCD	Access to Biological Collections Data
AEF	Agricultural Industry Electronics Foundation
AgMIP	Agricultural Model Intercomparison and Improvement Project
AgTrials	Global Agricultural Trial Repository and Database
ANACIM	Agence Nationale de l'Aviation Civile et de la Météorologie
APIs	Application programming interfaces
APSIM	Agricultural Production Systems sIMulator
ATPs	Agricultural technology providers
BARTOC	Basel Register of Thesauri, Ontologies and Classifications
BY	Attribution
CAP	Common Agricultural Policy
CAPAD	The Confederation of Agricultural Producers for Development
CC	Creative Commons
CC0	CC Zero
CEDEAO	Economic Community of West African States
CF	Climate and Forecast
CGIAR	Consultative Group on International Agricultural Research
IARD	Coherence in Information for Agricultural Research for Development
CKAN	Comprehensive Knowledge Archive Network
CRO	Crop Research Ontology
CSDGM	Content Standard for Digital Geospatial Metadata
CTA	Technical Centre for Agricultural and Rural Cooperation ACP-EU
DCAT	Data Catalog Vocabulary
DDI	Data Documentation Initiative
DISC	Data Interoperability Standards Consortium
DOIs	Digital Object Identifiers
DPLOS	Data Plot Sheet

DSSAT	Decision Support System for Agrotechnology Transfer
DWBP	Data on the Web Best Practices
EC	European Commission
EAC	East African Community
EDIFACT	Electronic Data Interchange for Administration, Commerce and Transport
EFSA	European Food Safety Authority
ESA	European Space Agency
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
FAO	Food and Agriculture Organization of the United Nations
FDP	Farmer digital profiling
FEPAB	Fédération des professionnels agricoles du Burkina
FGDC	Federal Geographic Data Committee
FMIS	Farm Management Information Systems
FOs	Farmers' organizations
GDPR	General Data Protection Regulation
GIS	Geographic information system
GiSC	Grower Information Services Cooperative
GODAN	Global Open Data in Agriculture and Nutrition
GPS	Global positioning system
HIMSS	Healthcare Information and Management Systems Society
ICASA	International Consortium for Agricultural Systems Applications
ICT	Information and Communications Technology
IFPRI	International Food Policy Research Institute
IGTF	Igara Growers Tea Factory
INRA	Institut National de la Recherche Agronomique
INSPIRE	Infrastructure for Spatial Data in Europe
IOT	Internet of things
IPR	Intellectual property rights
ISO	International Organization for Standardization
ITPGRA	International Treaty on Plant Genetic Resources for Food and Agriculture
IVR	Interactive voice response

KTBL	Association for Technology and Structures in Agriculture
LIVES	Local Inspector Value-Entry Specification
LOD	Linked Open Data
LOV	Linked Open Vocabularies
MoU	Memorandum of Understanding
NASA	National Aeronautics and Space Administration
NC	NonCommercial
ND	NoDerivatives
NetCDF	Network Common Data Form
NGOs	Non-governmental organizations
NOAA	National Oceanic and Atmospheric Administration
NUCAFE	National Union of Coffee Agribusinesses and Farm Enterprises
O&M	Observations and Measurements
OADA	Open Ag Data Alliance
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OATS	Open Ag Technology and Systems Group
ODB	Open Data Barometer
ODbL	Open Database Licence
ODC	Open Data Commons
ODC-By	Attribution Licence
ODC-ODbL	Open Database Licence
ODI	Open Data Institute
ODK	Open Data Kit
OGC	Open Geospatial Consortium
OGP	Open Government Partnership
OPeNDAP	Open-source Project for a Network Data Access Protocol
OURIndex	Index of Open-Useful-Reusable Government Data
PAFO	Pan-African Farmers' Organization
PAIL	Precision Ag Irrigation Language
PDDL	Public Domain Dedication and Licence
PDM	Public Domain Mark
PDP	Personal data protection
PPP	Public-private partnership

PSI	Public Sector Information
RDF	Resource Description Framework
REST	REpresentational State Transfer
ROSCAs	Rotating savings and credit associations
RTI	Global Right to Information Rating
SA	ShareAlike
SACCOs	Savings and credit cooperative societies
SensorML	Sensor Model Language
SMS	Short message service
SOSA	Sensor, Observation, Sample, and Actuator
SPADE	Standardised Precision Ag Data Exchange
SSN	Semantic Sensor Network
TBL	Tim Berners-Lee
TERRA-REF	Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform
TRIPS	Trade-Related Aspects of Intellectual Property Rights
TSML	Timeseries Profile of Observations and Measurements
UN	United Nations
UNECE	United Nations Economic Commission for Europe
URIs	Uniform resource identifiers
USAID	Agency for International Development
USDA	United States Department of Agriculture
USSD	Unstructured supplementary service data
W3C	World Wide Web Consortium

©Adobe Stock/Dusan Kostic

1

Data, services and applications

1.1 Data for agriculture

This section introduces the concept of the data revolution in agriculture and how data and information and communications technology (ICT) for agriculture services can support smallholder farmers to address their challenges, and in increasing their incomes and their yields. The number of such ICT for agriculture services has exploded in the last few years. As an example, the number of services identified in Africa was 41 in 2012, but 390 in 2018. However, they are still reaching a relatively low number of farmers. In Africa, around 33 million farmers have access to ICT services, which is less than 10 percent of the total number of farmers. However, it is expected that ICT services will reach 200 million farmers by 2030 (Tsan *et al.*, 2019).

This section presents the type of services that are potentially impactful, and the different elements required to build and deliver them. First, it introduces the state of agriculture, its role and the challenges involved in feeding the world in the next few decades. The potential of ICT services coupled with data exploitation to address these challenges will be also presented. Readers are then provided with more detail on what data consists of at farm level, and how it can be captured. The list of open datasets that are useful to provide actionable information[1] to farmers is also introduced and, finally, last section presents emerging applications and new trends on data analytics, artificial intelligence and machine learning.

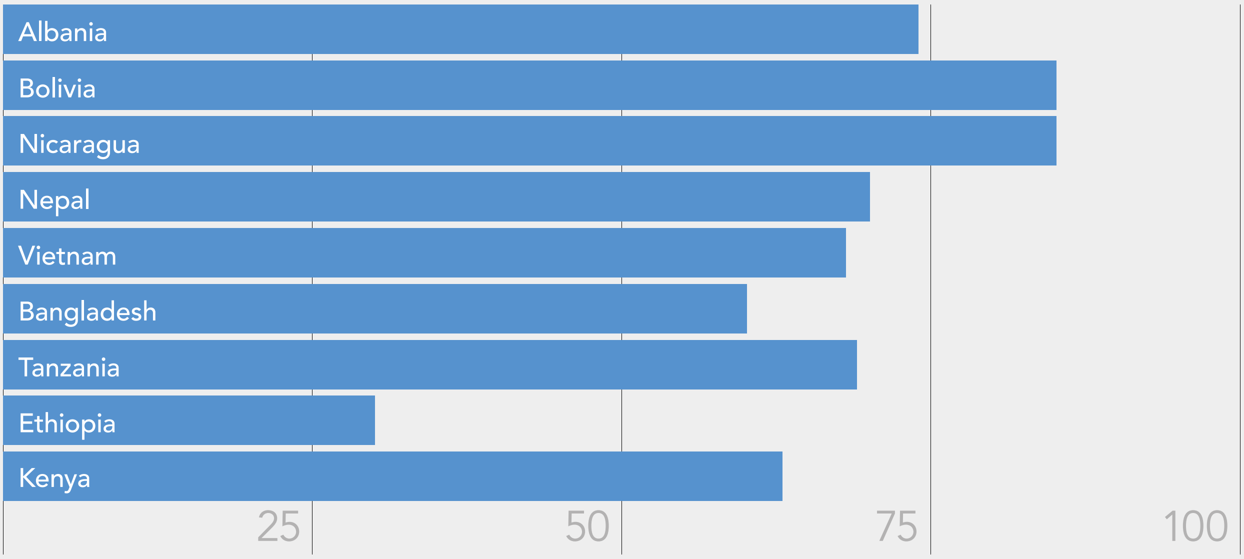
1.1.1 Opportunities for data in agriculture

Smallholder farmers represent the biggest employment sector in rural areas of the developing world, and they are also the most important contributors to global food production. More than 90 percent of farms in the world are family farms; they produce 80 percent of food and they operate 75 percent of the farmland (FAO, 2014). Figure 1 shows data from the FAO Smallholder Data Portrait (Rapsomanikis, 2015), which shows the importance of smallholder farmers and their contribution to food security. However, the evolution of global food production versus consumption and evolution of world demography shows that there is a strong need for increasing yield. The FAO 2019 report on the global state of food security and nutrition (FAO *et al.*, 2019) highlights some worrying facts:

- Since 2015, hunger and undernourishment in the world have stopped declining and almost 11 percent of the world population are still affected. This means that, in absolute terms, the number of people suffering from hunger is increasing.
- In 2019, more than 2 billion people "do not have regular access to safe, nutritious and sufficient food."

[1] Actionable information is data that is easily understandable and that can be directly used to make a decision or solve a problem. The concept of 'actionable information' is linked to the capacities, knowledge and environment of the person accessing the information, i.e. what is actionable to someone may be unusable by someone else with a different background.

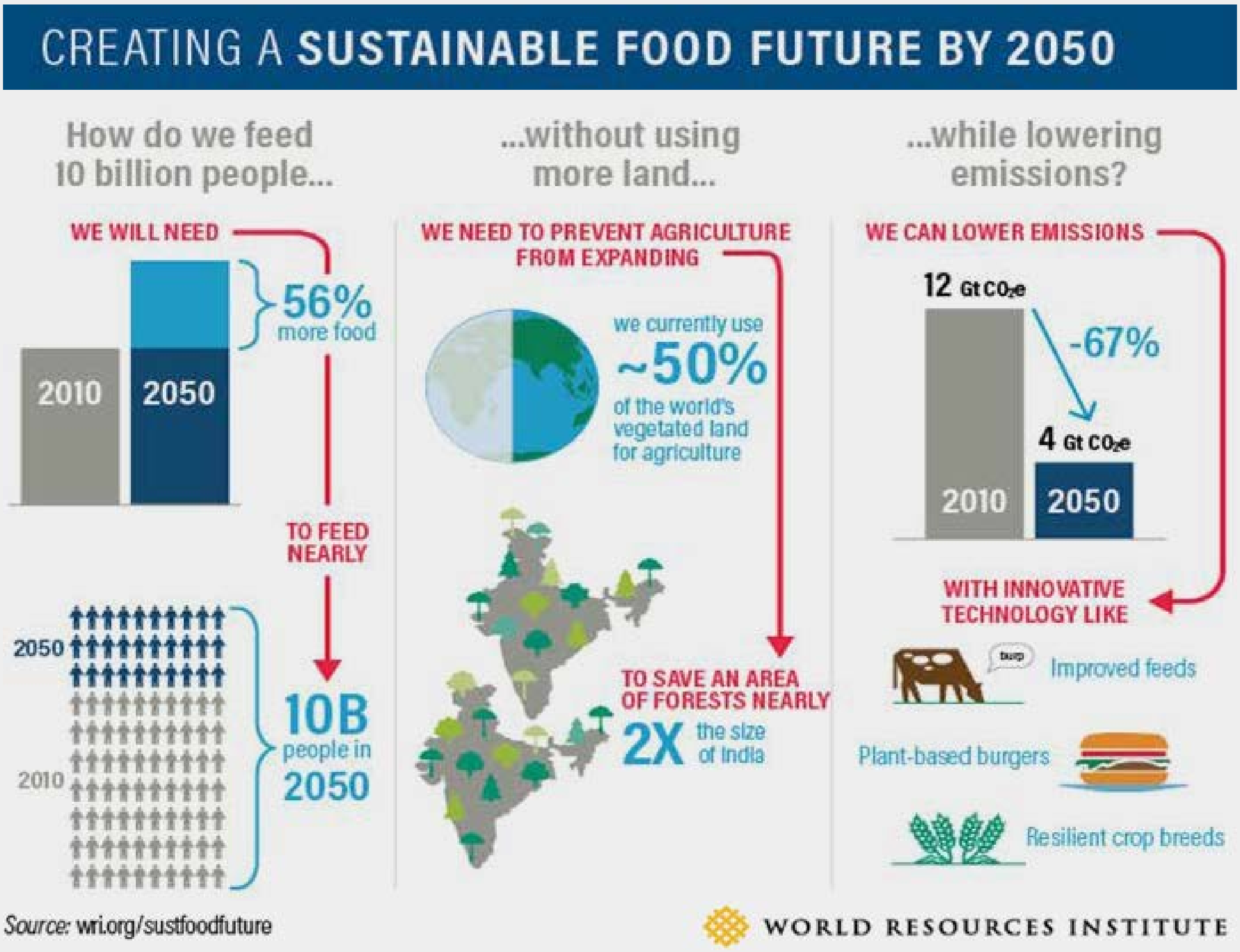
Figure 1. Proportion of national food production by smallholders



% of national food production

Source: Boyera, Addison and Msengezi, 2017.

Figure 2. Creating a sustainable food future for 2050


CREATING A SUSTAINABLE FOOD FUTURE BY 2050
How do we feed 10 billion people...
...without using more land...
...while lowering emissions?
WE WILL NEED
56% more food
2010
2050
TO FEED NEARLY
10B people in 2050
2050
2010
WE NEED TO PREVENT AGRICULTURE FROM EXPANDING
we currently use ~50% of the world's vegetated land for agriculture
TO SAVE AN AREA OF FORESTS NEARLY 2X the size of India
WE CAN LOWER EMISSIONS
12 Gt CO₂e
-67%
4 Gt CO₂e
2010
2050
WITH INNOVATIVE TECHNOLOGY LIKE
burp
Improved feeds
Plant-based burgers
Resilient crop breeds
Source: wri.org/sustfoodfuture
WORLD RESOURCES INSTITUTE

Source: Waite et al., 2018.

At the same time, climate change is also severely impacting yields. For example, the International Food Policy Research Institute (IFPRI) estimated in a report (Wiebe *et al.*, 2017) that rainfed maize yields in some regions of Africa could decrease by as much as 25 percent by 2050 compared with levels in 2000.

One of the most promising opportunities to address this multi-faceted challenge is to work towards increasing yields. Figure 3 illustrates the challenge of yield gaps and the opportunity:

These yield gaps could be addressed by providing more support to farmers and by enabling them to access more services such as extension services, trade services or financial services. Today, these services can be provided at scale through ICTs; Dalberg estimates 85 percent of farmers' households will have a mobile phone by 2025 (Tsan *et al.*, 2019). As an illustration of this opportunity, Dalberg shows that the bundling of three services (access to finance, advisory services and market linkages) can lead to a 57 percent increase in income for farmers, and up to a 168 percent increase in yield.

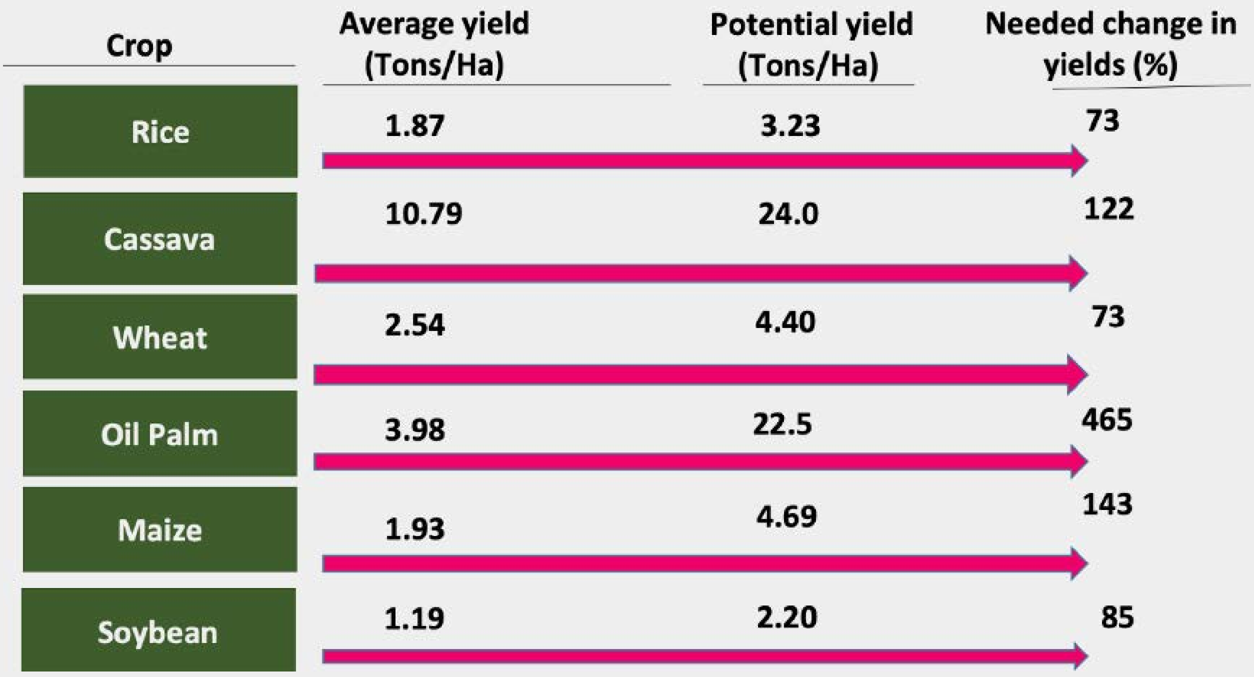
While the use of ICT technologies is the most promising way to deliver services at scale, the content of these services and their ability to deliver accurate actionable information or results depends on their ability to aggregate different data sources.

There is a wide scope for application of these ICT approaches. For example, understanding the cause of the underperformance of some crops will lead policy-makers to put in place appropriate legislation, subsidy schemes and interventions to address the issues. Access to detailed field information helps credit companies build a reliable credit profile and deliver loans to smallholder farmers more easily. At the same time, if farmers at each stage of the crop cycle can access timely actionable information, they will be able to take informed decisions on the best way to get the most out of their fields in a sustainable eco-friendly way.

The benefits of such approaches are not limited to farming and crops but also applies to other activities in the agriculture sector. An example is presented in the case of small-scale fisherfolk in Cape Town, South Africa, who are benefitting from the use of data to support their operation. Collection of data is also critical for them to access financial services.

The types of data-driven services that are potentially useful and impactful for farmers are represented in Figure 4.

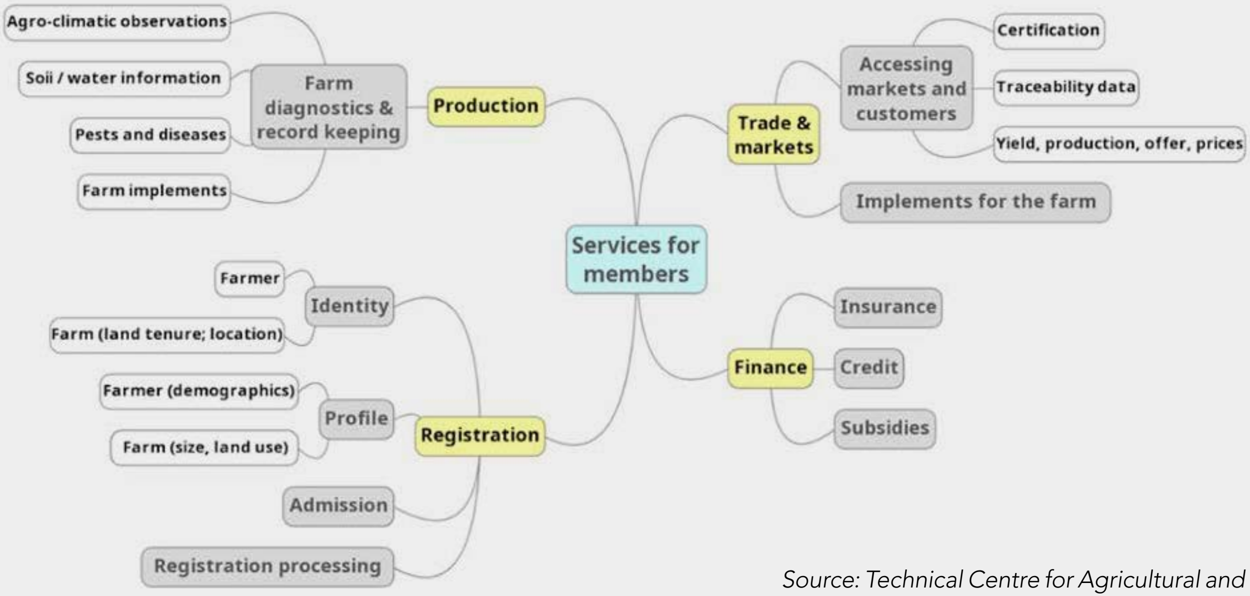
Figure 3. Average yields in Africa vs. best practices



Crop	Average yield (Tons/Ha)	Potential yield (Tons/Ha)	Needed change in yields (%)
Rice	1.87	3.23	73
Cassava	10.79	24.0	122
Wheat	2.54	4.40	73
Oil Palm	3.98	22.5	465
Maize	1.93	4.69	143
Soybean	1.19	2.20	85

Source: Technologies for African Agricultural transformation (TAAT), 2018.

Figure 4. A model for data-driven services and related farmer profiles



Source: Technical Centre for Agricultural and Rural Cooperation (CTA), 2019.

There are four main categories:

1. Production-related services: All the services from pre- to post-harvest to assist farmers to extract the greatest value from their assets and to combat any pest or disease that may endanger the harvest.
2. Financial services: A range of financial services are essential to support agricultural activities, including traditional banking services, microfinance and subsidy schemes.
3. Trade and market services: This area includes all services that enhance access to market and support farmers in getting the best prices for their commodities.
4. Registration services: These services encompass cooperatives and farmers' group services for their members, which includes membership management and communication.

Case Study 1: Alobi Fisher

Small-scale fisheries play an important role in providing sustainable food security for local, national and international markets. They are seen as stewards of the sea by some but largely remain marginalised and unrecognised by societies across the globe. In South Africa, the fisher community have collaborated with the University of Cape Town to co-design a suite of apps to support and improve the small-scale fisheries industry. Abalobi Fisher is an app that is free to download; it provides valuable information about the weather and climate from open sources, and also records data about fisher practice and catch information. This data has never fully been captured on a large scale before and it enables a new understanding of the small-scale fisheries sector, see Figure 5.

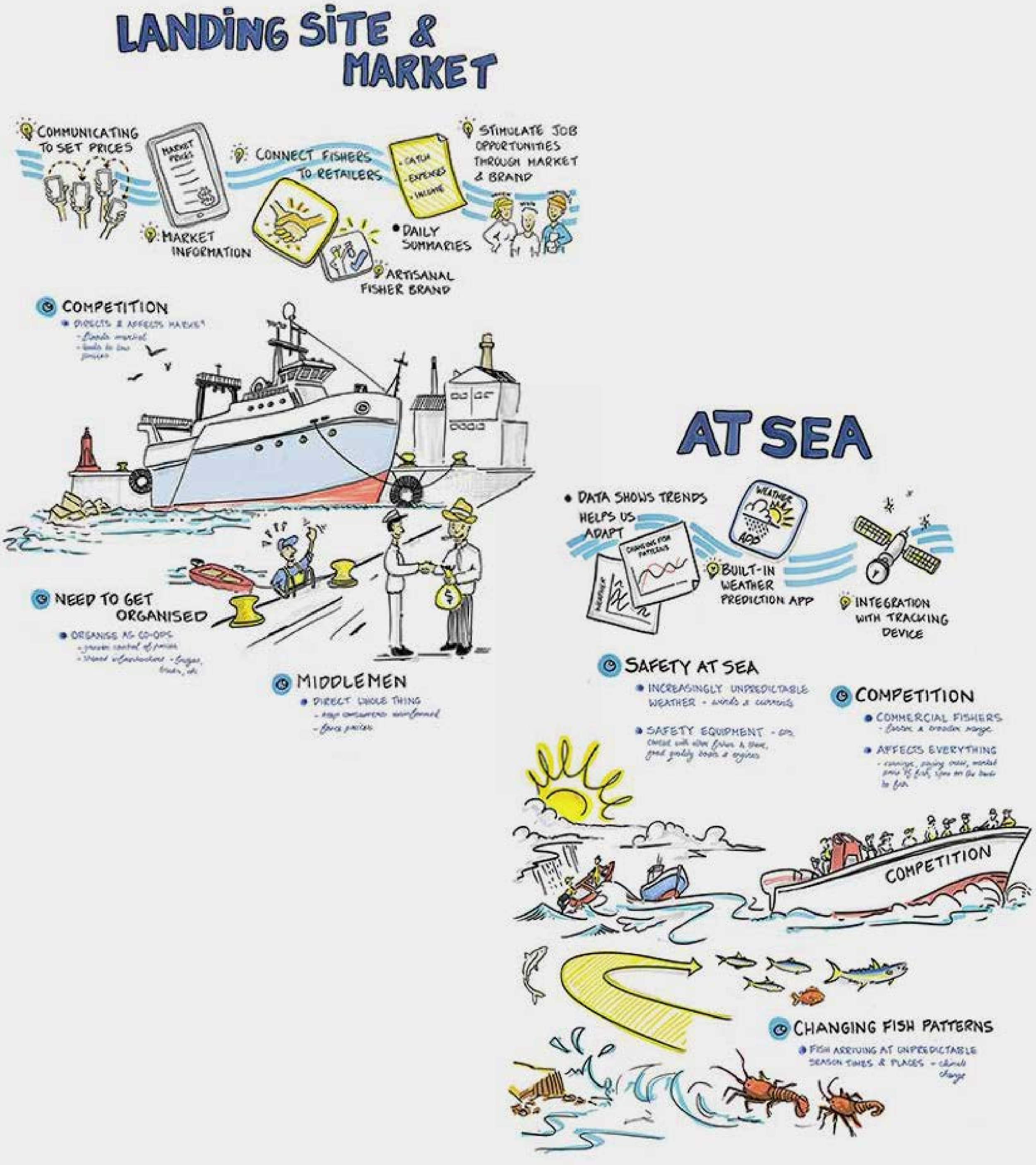
Five carefully co-designed apps collectively form the Abalobi 'From-Hook-to-Cook' system, which enables the processing and marketing of fish and other marine resources with an ecological and social 'story' in a manner that is fully traceable along the value chain. The goal is empowerment in the value chain and the development of fisher-community-based eco-labelling and social labelling. The app suite allows fishers to supply local and global markets interested in sourcing seafood that is fair, credible and has a low environmental impact.

The app also allows small-scale fishers to upload their daily catch onto a digital marketplace; information about location and quantity is included. Chefs have access to this marketplace and put in their catch requests. The Abalobi team facilitate the delivery and transportation of the seafood directly to some of South Africa's best restaurants. Each delivery is accompanied by a unique QR code that, when scanned by a smartphone, will direct the user to the Abalobi app. Here information about the fish, when and where it was caught in South Africa, and even the name of the fishing vessel can be accessed.

©Adobe Stock/Alex

With an integrated mobile payment platform and accounting functionality, the mobile app suite has the potential to maximise interoperability with a multitude of fintech services and tools to facilitate accessible, transparent business development for small-scale fishers. A key value proposition is that this component allows fisher groups to valorise non-quota, less 'mainstream' species towards community-supported fisheries and restaurant-supported fisheries.

Figure 5. The Abalobi innovation


LANDING SITE & MARKET
COMMUNICATING TO SET PRICES
MARKET PRICES
MARKET INFORMATION
CONNECT FISHERS TO RETAILERS
- CATCH
- EXPENSES
- INCOME
DAILY SUMMARIES
ARTISANAL FISHER BRAND
STIMULATE JOB OPPORTUNITIES THROUGH MARKET & BRAND
COMPETITION
DIRECTS & AFFECTS MARKET
NEED TO GET ORGANISED
ORGANISE AS CO-OPS
- greater control of prices
- shared infrastructure - fridges, trucks, etc.
MIDDLEMEN
DIRECT WHOLE THING
- keep consumers uninformed
- force prices
AT SEA
DATA SHOWS TRENDS HELPS US ADAPT
CHANGING FISH PATTERNS
WEATHER
WEATHER APP
BUILT-IN WEATHER PREDICTION APP
INTEGRATION WITH TRACKING DEVICE
SAFETY AT SEA
INCREASINGLY UNPREDICTABLE WEATHER - winds & currents
SAFETY EQUIPMENT - e.g. contact with other fishers & shore, good quality boats & engines
COMPETITION
COMMERCIAL FISHERS - faster & broader range
AFFECTS EVERYTHING
COMPETITION
CHANGING FISH PATTERNS
FISH ARRIVING AT UNPREDICTABLE SEASON TIMES & PLACES - climate change

Source: Abalobi - a Mobile App Suite for Small-Scale Fisheries Governance, 2020.

Some exciting initiatives by farmer-based businesses and organizations in Africa who realise benefits for smallholders through ICT services and digitalization in agriculture are explored below:

NUCAFE

A digital farmer profiling initiative, which enables coffee to be traced back to its roots, is paying off for smallholder farmers in Uganda. National Union of Coffee Agribusinesses and Farm Enterprises.

NUCAFE, CTA and the Pan African Farmers Organisation (PAFO) partnered in the Data4Ag project to improve data management systems and financial skills of NUCAFE and its hub members for development of the coffee value chain. The aim of this capacity building was to strengthen entrepreneurial and financial skills of the farmers union.

Six months after its launch, during the period of September 2018 to February 2019, the initiative produced promising results. NUCAFE generated farmer profiles and maps of coffee farms in order to develop a targeted and informed marketing strategy for the Union. Moreover, a memorandum of understanding with Centenary Bank was conceived to promote access to finance for NUCAFE farmers. Most importantly, by digitally profiling farmers, the traceability of coffee back to its roots has been enabled and is paying off for NUCAFE's 210 coffee farmers and farmers' organizations – totalling 205 120 farming families.

©Adobe Stock/Pasko Maksim

The design of a geospatial database and improved financial literacy as a result of the initiative, has helped Ugandan coffee farmers to access new markets as well as higher prices. Among others, international buyers from Italy and South Korea have offered higher prices for coffee produced by the profiled farmers, paying EUR 3.51/kg (USD 3.93) instead of EUR 2.16/kg (USD 2.42) or lower for untraceable coffee of similar quality. The premium increase of 24 percent on the basic price is directly related to product traceability, which ensures consumers that coffee farmers truly benefit from their purchase and that coffee beans carry specific geological and geospatial quality markers.

"We have benefited from the additional income we get from our coffee being traceable and certified by being able to take our children to school and working on the community health centre," says Mr Gibezi Yunus, a farmer from Bufumbo Organic Cooperative Association, which operates under NUCAFE. (See NUCAFE, 2020 for more information)

IGARA

The Igara Growers Tea Factory (IGTF) has also benefited from CTA's support in developing data-driven services for farmer members. The instalment of 40 digital weighing scales, for instance, is helping IGTF to pay for farmers' tea on receipt, and delivery records for over 4 000 farmers are stored by the new digital systems. Digital profiling of farmers has captured their location and farm size information, which means field extension services are better targeted and produce quality has improved. This is reflected in the price received at the tea auction market where IGTF is earning more than its competitors. Local youths are now also involved in the initiative, with over 40 young people using mobile apps to collect tea from the farmer members, and 70 youths are also using GPS-enabled tablets to validate farmer data for the profiling platform. (See Uganda Tea Development Agency, 2020. for more information)

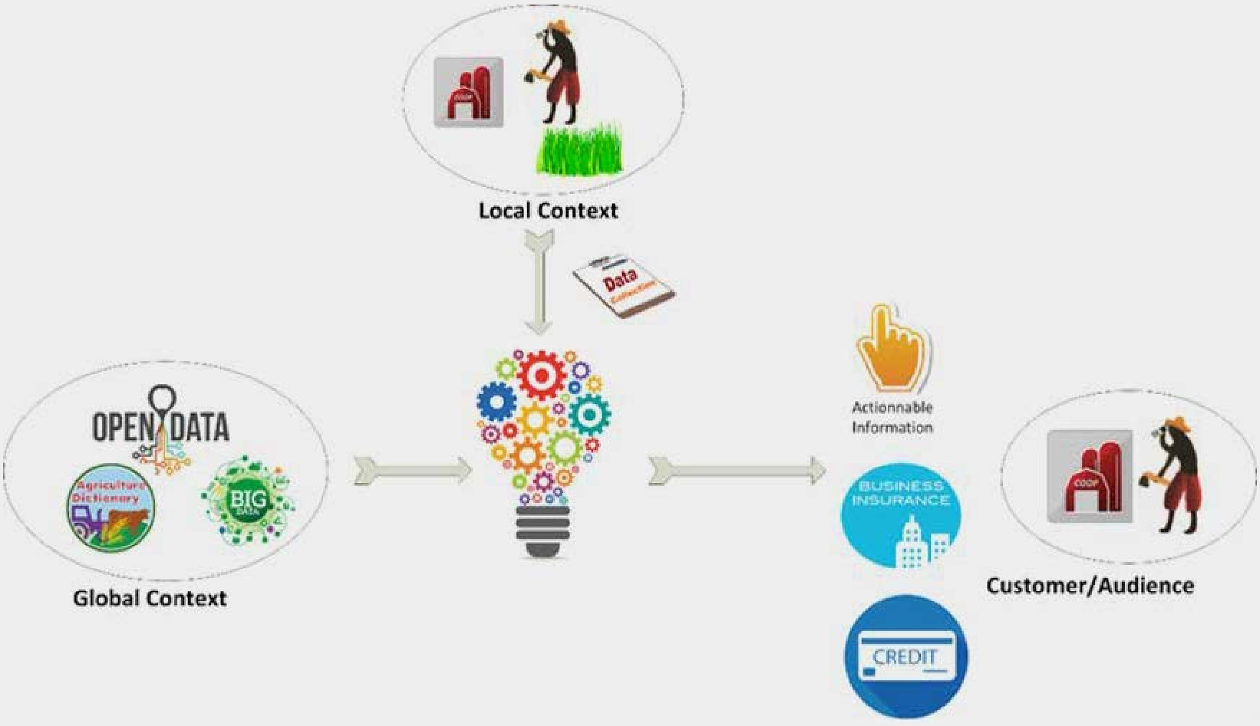
©Adobe Stock/Mauro Rodrigues

CAPAD

The Confederation of Agricultural Producers for Development in Burundi (CAPAD) has supported 39 smallholder cooperatives to register over 14 000 farmers (55 percent women), and has issued all members with an ID card. The data collected has enabled farmers' organizations (FOs) to better plan for the 2019 and 2020 seasons through bulk ordering of, for example, mineral fertilisers (1 059 tonnes), rice seed (27 tonnes) and corn seed (18 tonnes). Collection of the farmer information has also facilitated rapid processing of applications for agricultural credit and, as a result, 2 896 FO members have obtained BIF 214 275 049 (USD 111 500). Digital membership has also allowed cooperatives to better organize their post-harvest management and marketing of agricultural products; so far in 2019, 4 052 tonnes of rice, 132 tonnes of maize and 131 tonnes of beans have been sold collectively. (See Confédération des Associations des Producteurs Agricoles pour le développement, 2020 for more information)

For the majority of stakeholders, the design of digital services reside in the mash-up of global data (e.g. satellite images, research studies, databases of information about crops, seeds, pests and diseases, etc.) with farmer-level (credit records, field ownership documentation, etc.) and field-based information (e.g. soil information, geographic location, state of the fields, crops etc.) to determine the content (e.g. appropriate information to take decision). The results at farmer level are both the availability of new products to support their production (credit, insurance, etc.) and the availability of timely information to support decision-making (Boyera, Addison, and Msengezi, 2017), see Figure 6.

Figure 6. Design of services



Source: Boyera, Addison and Msengezi, 2017.

1.1.2 Data at farm level

Farm-level data is the essential element in delivering actionable tailored farmer-centric services and information to individual farmers. The information about the farm and the farmer can be categorised in different components that are useful for different types of services. The first part of this section, data and usage presents the categories of information and their usefulness for different types of services. The second part, data capture presents the different means to collect this data.

Data and usage

The main categories of data at farm-level are presented below. It is important to note that the exact set of information within each category may vary significantly depending on the specific service designed that may or may not require specific information. The content varies significantly from country to country. For example, in some countries, the name of a person is not meaningful without knowing the father's name. Some countries may have implemented a robust identity scheme that makes the ID number valuable information while, in other countries, this may not be reliable information.

- Personal information: This component contains the profile and lists the information about the farmer’s identity (name, ID number, birth date, gender, language spoken, income level, education level/literacy level, number of people in the household …). Note that some information, such as the literacy level or the languages known, is usually critical for the design of accessible ICT services.

- Communication information: Communication information covers all information to interact with the farmer, either directly or through broadcast media. This includes information such as phone number(s), phone type (smartphone, basic phone, etc.), phone literacy (ability to use different technologies on phone such as SMS or app), email, social networks used, or radio and TV listened (and at which time). This data is particularly useful in order to understand the most efficient way to deliver services and information to the farmer.

- Location: Location information is critical to locate the farmer. It usually includes information such as administrative address (split by administrative divisions such as region, district …) and GPS coordinates.

- Financial instruments: Information about financial instruments available at the farmer level is critical for financial services (e.g. credit, insurance or subsidy payments). It includes information about bank accounts, including mobile money accounts.

- Credit information: Credit information is critical to support access to credit. It includes information such as credit record, farm business plan (to identify cash needs and timing of repayment during a complete crop cycle), savings and credit cooperative societies (SACCOs), rotating savings and credit associations (ROSCAs) membership, active credit information.

- Insurance information: Insurance information is also an important set of information for different purposes, such as credit, but also to identify covered and uncovered risks. Information includes field(s) covered, risk(s) covered, cost, company, amount repaid in case the risk(s) materialise.

- Farm details: Information about the farm as an enterprise is critical to identify specific needs and interventions to support its activities. Key information includes registration number (if the farm is a formal registered business), labour force available on the farm, equipment (for planting, harvesting, post-harvesting) or the (list of) extension agent(s) associated with the farm/farmer. When the farm is a formal business, it is characterised by its financial data (turnover, benefit, etc.). In the case of smallholder farmers, the farm’s financial data is usually the same as the farmer financial data. In some cases, it may be appropriate to separate the two.

- Qualification and certification data: Qualification and certification apply to either the farm or the farmer and, sometimes, to specific fields. Most certification requires training first. However, some training does not lead to any certification. This information is critical for many purposes. First, most certifications have regulations on various activities from planting to applying treatment to harvesting. Extension services must adapt to these constraints. Then, certification provides added value to the end-product, and this is critical for marketing activities. Finally, knowing a farmer’s certification enables him to more easily access other certifications. This, for example, is the objective of a service like Standard Maps [standardsmap.org] that lets a farmer know, based on their current certifications and the ones they want to reach, the set of modules they need to follow. The information required for qualification and certification includes training/certification name/label, training/certification date or training/certification institution.

- Field information: In many cases, a farmer manages more than one field in different places, or even if he has one piece of land, the space is split into sections with different crops. Core field information includes location[2] , size (the size may be available on the land title, evaluated by the farmer or automatically computed if a field map is provided), elevation (important for some crops), soil, land title and crop history. Field information also includes crop information (crop, variety, type of seeds). The crop information is highly dependent on the type of commodity grown: trees (coffee, cocoa, coconut) or tea are very different compared to seasonal crops. The crop information must therefore be adapted.
- Production information: The production information is usually linked to a field. This information is usually useful for extension services and to prepare trade or post-harvest activities. It usually includes planting information (date, spacing, intercropping information, equipment, amount of seeds used), activities information (treatment applied, fertiliser, extension service interventions, pest and disease attacks and treatments, activities such as weeding, water usage, yield, loss, rainfall…). Here again the production information is directly related to the crop specificities.
- Business information: Business information is a critical element for marketing and selling of the yields or transformed products. This information describes the linkages between the farmer and other key stakeholders in the value chain for conducting his/her businesses. It includes information such as cooperatives/production cluster membership, markets the farmers are linked to, agribusinesses linkages, total amount of products sold (per trade channel such as cooperative, at market, at farm gate) and prices sold.

Data capture

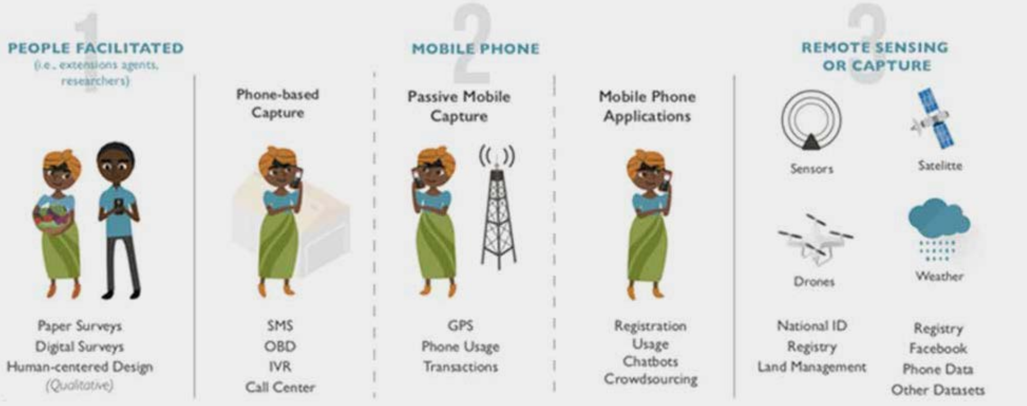
There are different ways to capture the data presented in the previous section (Data and usage). The main approaches are summarised in Figure 7.

There are three main options, which are usually complementary:

1. Direct on-farm capture: The data is collected through human interaction by a data collector visiting the farmer. The data collection activity could be done on paper or by using more advanced smartphone tools.
2. Remote capture via mobile phone: The fact that many farmers now have a phone can be exploited to collect data remotely. There are predominantly two ways to capture data:
    a. Direct capture through farmer contributions: Data are collected from farmers (as in direct on-farm capture) but remotely. Such collection can take different forms: a person-to-person call (via call centre), a basic phone application (voice-based application, unstructured supplementary service data (USSD), short message service (SMS), or a smartphone application

[2] Location may have different formats: administrative location, the GPS coordinates of one point in the field or a map (geo-fencing) of the field. The latter offers more opportunities for specific services (forecast of production, evaluation of inputs required etc.) and is obviously more complex and more costly to acquire.

Figure 7. Ways of capturing data



Source: Gray et al., 2018.

b. Indirect capture through big data: Mobile operators are able to extract a large quantity of information from each of their clients, which includes e.g. usage of their phone, or usage of specific mobile services such as mobile money. This data is valuable and can complement other information in a farmer profile. Capturing, storing and exploiting this data requires partnership with mobile operators who are the only ones with access to this information. Of course, farmers have to be aware and have to consent to such data collection (see Section 2.4. Personal data protection).

3. Automatic capture using specific technologies: There are a series of new technologies that can be used to capture some information automatically. In particular, this includes drones for field mapping and analysis, and sensors (also known as Internet of things (IOT) technology. See for example, FarmBeats, an initiative around IoT supported by Microsoft (Microsoft, 2015). The project aims to make farmers more efficient by arming them with data to help them increase farm productivity, and also to reduce costs. FarmBeats project team is building several unique solutions to solve the problem of getting data from the farm by using low-cost sensors, drones, and vision and machine-learning algorithms.

Other data capture technologies include big data approaches used to analyse farmers behaviour online (e.g. on social media), but these approaches are still largely irrelevant for smallholder farmers in developing countries. Technologies such as satellites are more related to capturing global data.

1.1.3 Identifying key datasets in farming crop cycles

There are numerous datasets that are potentially useful to deliver information and services to farmers. Some datasets might be useful at different stages of the crop cycle[3], but with different requirements. One such example is market prices. Market prices are useful when selecting the crop to grow, if the market price data have a time series showing the recent evolution of prices in the past years. Market prices are also useful at the selling stage but, at this stage, this dataset must have real-time information (or as close to real-time) to be useful. Some datasets are obviously available at the country level only, but some others may be available in country or at the regional or international level (e.g. weather data or satellite images).

For further reading on this topic, the Global Open Data in Agriculture and Nutrition (GODAN) Agriculture Open Data Package (GODAN, 2016), and the presentations made at the big data session of the United States Department of Agriculture (USDA)/U.S. Agency for International Development (USAID) International Food Assistance and Food Security Conference in 2016 are recommended. Table 1 presents the different categories of datasets, the individual datasets[4], and the type of applications that can be built on top of them.

[3] In this context, the term 'crop cycle' is used as the set of stages that a farmer goes through from the selection of a commodity to plant in the field till the selling of the harvested product. The cycle is usually split into three main phases: (1) the pre-cultivation stage that includes access to finance, selection of crops, seeds, etc; (2) the cultivation and harvesting stage that goes from planting until harvest; and (3) the post-harvest stage that is mainly focused on trade and commerce, but also includes processing when appropriate.

[4] The list of datasets in each category is not exhaustive but mentions main elements.

Table 1. Categories of datasets, individual datasets and applications

Categories	Datasets	Examples of services
Government, agricultural law and regulations	(Phyto)sanitary regulations (list of quarantine organisms, etc.), environmental regulations, subsidy schemes, import/export regulations Example: **http://kenyalaw.org** public portal on all laws of Kenya	Advisory service on the crops and varieties to grow based on legal framework and subsidy schemes.
Official records	Land registration, licensed organizations (corporations, business, NGOs), import/export tariffs, permitted crop protection products. Example: Land registration in France (French only) **https://app.dvf.etalab.gouv.fr** California authorised pesticides **www.cdpr.ca.gov/docs/label/labelque.htm**	Determining ownership/rights to use the land, advocating for land reform and new land record management, etc.
Government finance data	Agricultural subsidy expenditure (direct payments, product support, tariffs etc.), agriculture-related tax income, penalties given to agricultural actors, investment in research and education (extension, research institutes, professional training and universities). Example: Albania Finance Portal: **http://isdatabank.info/albania** Uganda Budget data from 2003 till 2017: **https://bit.ly/2z1Eqmx**	Forecasting impact of change in subsidy schemes, raising awareness on penalties risk
Rural development project data	General project information, including financial data, location, beneficiaries, activities, project output, outcome and impact, project documents. Example: IATI data of all UKAID-sponsored projects. **https://devtracker.dfid.gov.uk**	Identifying existing initiatives to leverage
Land use and productivity data	Land use data, cultivated areas, current crop in the fields, harvested crop, crop types Example: India Land Use information: **https://bit.ly/2MdcNjn**	Ensure land use rights to cultivate the land, forecasting production, providing advisory services for traders on where crops are being grown, providing advisory services on pest and disease related to specific crops.

Categories	Datasets	Examples of services
Value chain data	Profiles of different value chain actors and organizations [1] Farm data, e.g., farming system, crops, land area, farm income, household composition, farm employment, farm holder's age, fertiliser use etc. [2] Cooperatives [3] Trade [4] Processors, e.g. type, size, turnover, capital, investments, environmental transparency indicators etc. [5] Retail, (Food) product data, e.g. food nutritional value, food composition, origin of produce, environmental factors, time and location of production, etc. (Safety) inspection results Certification Example: United States of America Livestock and Poultry industries analysis **www.gipsa.usda.gov/psp/publications.aspx** UL Mad Cow disease inspection data: **https://bit.ly/2MgmddD**	Leveraging linkages (farmers/ input dealers' link, market linkages, etc.)

©Adobe Stock/agnormark

Categories	Datasets	Examples of services
Infrastructure data	Road network and conditions, road maintenance schedule, public transport, waterways, internet connectivity map, mobile connectivity map. Example: Canada British Columbia publishes traffic data (that can e.g. be used to minimise transport for perishable goods) **www.th.gov.bc.ca/trafficData/index.html**	Transport services, applicability of different types of ICT services (GSM-based services vs internet-based services)
Market and price data	Global food prices, national stock exchange prices, regional market prices, local market prices, location of national, regional and local markets, import/export volume Example Ethiopia Commodity Exchange: **www.ecx.com.et**	Market price information, support for decision-making on market access, support on price bargaining, etc.[5]
Meteorological data	Short-term weather forecast, detailed agrometeorological data, seasonal weather forecasts (3-6 months ahead), real-time observations, historic archives of observations, historical simulated weather from re-analysis, climatological observations, climatological reference data, climate zones, climate change predictions, rainfall data Example: Australia real-time climate data **www.bom.gov.au/climate/data**	Identify the crops and the varieties to grow based on local conditions such as climate zone, agro-ecological zone, weather forecast, soil or global appropriateness of the field (e.g. flooding risks), supporting farmers in the schedule of their activities (seeding, harvesting, etc.). Alerting and preventing damage from severe meteorological events.[6]
Elevation data	Digital elevation model, elevation maps, height points, slope, aspect[7], catchments, drainage, erosion susceptibility Example: CGIAR STRM 90m digital elevation database: **http://srtm.csi.cgiar.org**	Identification of high-value production and setup of geographically certified products[8]

[5] The impact of market price information service is one of the most documented examples of the impact of information and ICT services on revenue increase and diminution of loss for perishable products. See e.g. example of farmers in Uganda (Svensson and Yanagizwa, 2010) or fishermen in Kerala (Jensen, 2007).

[6] See the example of ANACIM in Senegal sending weather alerts to fishermen (Anannya, 2018). Another example is Abalobi in South Africa, where a smartphone app allows fishermen to make the best decision to go fishing or not based on weather information (waves, tides, etc.)

[7] Aspect identifies the downslope direction of the maximum rate of change in value from each cell to its neighbours. It can be thought of as the slope direction. Aspect allows the followings: to find all north-facing slopes on a mountain as part of a search for the best slopes; to calculate the solar illumination for each location in a region as part of a study to determine the diversity of life at each site; to find all southerly slopes in a mountainous region to identify locations where the snow is likely to melt first as part of a study to identify those residential locations likely to be hit by runoff first; to identify areas of flat land. Source: ArcGIS Desktop, 2020.

[8] For an example of price increase offered to coffee maker due to their specific geographical position and their altitude see NUCAFE news.

Categories	Datasets	Examples of services
Hydrological data	Location of water sources flood zones, historical records on flooding, real-time water levels, water quality, water tables, water management Example: United Kingdom flood risks data: **www.shoothill.com/Floods&Rivers**	Alerting and decreasing the impact of flood, alerting farmers on water availability and quality for farming availability and quality for farming
Soil data	Soil maps, soil samples, soil classifications Example: ISRIC (International Soil Reference and Information Centre) soil database: **www.isric.org/explore**	Selecting the best crop and the best inputs based on soil information
Production advice data	Data on cultivars, landraces and farmer varieties including new releases; rop selection advice including new releases; crop calendars, intercropping, relay cropping, rotation; resource-related farm advice[9]; fertiliser recommendations Example: FAO TECA (Technologies and Practices for Small Agricultural Producers) database **www.fao.org/teca/new-search-result/en**	Extension services across the crop cycle from land selection, crop selection, up to harvest and post-harvest
Disease and pest management data	Occurrences and distribution of plant pests; treatment of pests and diseases; recognition of pests and diseases; biology of pests and diseases; toxicology or plant protection measures Example: Plantwise application from CABI **www.plantwise.org/KnowledgeBank**	Detection and curation of pest & disease, alert on disease outbreak

[9] Data related to crop selection, crop and land management as typically found in extension services information.

1.1.4 Data analytics, artificial intelligence and machine learning

The mash-up of farm-level data and global datasets allows the generation of a huge volume of information. To date, most available services have been relatively basic, consisting of human analysis of these data. New approaches, in particular, blockchain (Sylvester, 2019), data science, artificial intelligence and machine learning offer opportunities for the future. These opportunities include predictive analysis, such as yield forecasts, that will inform all value-chain actors, from public authorities with early warning on potential food security risks up to traders. These future approaches will be made possible through a greater availability of data. Farm-level data becomes more available as data collection becomes more automatised. For example, sensors start to spread at scale, and as governments, international organizations, and all actors including the private sector release more open datasets and increase access to big data streams, the volume of data will grow exponentially. This will offer more opportunities for more advanced predictive automatic services. These services provide greater added-value, and at lower costs, than the current generation of ICT services, making them more impactful and more sustainable. The trend is clear[10] and is likely to lead to a new wave of ICT services in the coming years with both the availability and the development of capacities on data science that is taking place in almost all countries worldwide.

1.2 Farmer profiling

In order to increase their production and income, smallholder farmers need various types of services including for extension, financial or trade services. The design, deployment and delivery of these services in physical or ICT format require the mash-up of data at the farm level. The data should have global context and be made available through open datasets released by national, regional or international organizations.

1.2.1 Digital farmer profiling

The delivery of tailored, actionable services is usually designated under the term 'precision agriculture' and they have been developed and deployed worldwide for more than a decade.

These services take different forms. In most developed countries, where farmers have the capacity, infrastructure and equipment, they can manage their own farm data collected directly or via sensors. Farmers are then able to select the services they need and share their own data with third parties who are providing those services.

In the majority of developing countries, the situation is very different. Farmers lack the capacity, infrastructure, and equipment and are not yet in a position to collect and manage their own data to be able to interact directly with service providers.

As a result, different service providers directly collect different types of data on the farm. However, this presents a series of challenges, such as:

1. Lack of sustainability for service-providers: At the moment, each and every service provider has to put in place data collection and the process to update data, which is extremely costly. As a result, the services provided are expensive for farmers and are rarely sustainable.
2. Farmers' rights are not respected: Currently, farmers are not really aware of the use of their data and, individually, cannot really protect their interests. In practice, they lose ownership of their own data and, at the same time, this data is in the hands of service providers, which is not necessarily used to maximise farmers' benefits.
3. Service provider lock-in: As service providers are usually putting in place an end-to-end service, it is difficult for farmers to switch to another provider. For example, if a financial service institution collects and builds a database of information about farmers to compute their credit scoring and their eligibility, this information (e.g. repayment rates or credit values) is stored at the institution and will not support the farmer who applies to other financial institutions for credit.

[10] Tsan *et al.* (2019) shows that 60 percent of the current ICT service providers plan to integrate these technologies in the next three years.

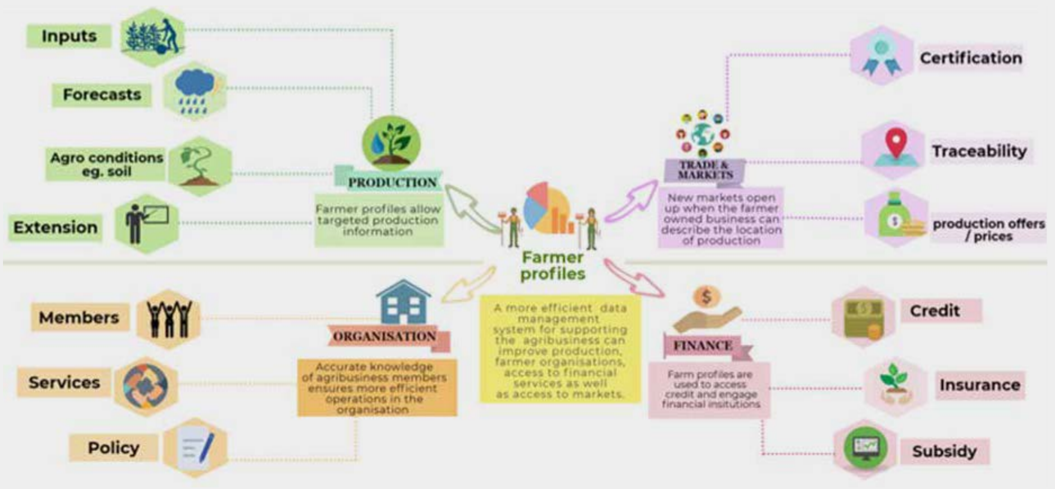
To address these issues, a new approach known as the farmer digital profiling (FDP) platform is now emerging. The concept is for an organization to aggregate all the farmers' profile information under its umbrella and then leverage this information to support service development as illustrated in Figure 8.

Organizations engaged in farmer profiling include farmers' organizations (e.g. Fédération des professionnels agricoles du Burkina (FEPAB) (Brussels Briefings, 2018a), cooperatives (Igara Tea Factory, Nucafé (Muwonge, 2018)), agribusinesses managing farmers under contract (e.g. Meridian in Malawi (TechnoServe, 2017), or even a government agency building a national repository (e.g. Fiji Crop Livestock Council (Daunivalu, 2018), Rwanda Ministry of Agriculture).

These different types of organizations can take advantage of a deep understanding and knowledge of their members and, in particular, who they are, what they do, where they live, and what they produce, etc. This information is essential for many reasons:

- Planning and strategy: Organizations can plan their services, their intervention and their areas of investment based on real data. They will be able to identify areas where they could expand, or where there are specific opportunities in terms of production or selling. A deep understanding of their members allows organizations to define their roadmap and identify new services. It also allows them to make financial forecasts and precisely evaluate potential markets for various services. In short, this information will allow organizations to plan their activities based on real facts and data.
- Easier membership management: For farmer's organizations, cooperatives or similar organizations, the use of an FDP platform helps in the management of membership for all internal activities, such as payments, elections and votes. The use of such platforms helps organizations to save time and money in those tasks.
- Easier communication: The capture of communication details, in particular phone numbers, allows the use of communication platforms that automate the sending of information in various formats (voice, SMS, etc.). The use of such platforms supports better and more regular communication between the central organization and its members. The use of new communication channels enables organizations to:
    - better understand the needs and demands of their members;
    - better understand constraints and pain points;
    - query their members on specific topics and obtain their feedback.
- Greater opportunities to identify and put in place new services: A better knowledge and understanding of its membership enables an organization to identify new targeted value-added services. These services not only provide new benefits to members, but also increase the value of membership and enable farmers' organizations (FOs) and cooperatives to recruit new members. ICT services developed by third-party suppliers have a specific place among potential new services.

Figure 8. Farmer digital profiling


Inputs
Forecasts
Agro conditions eg. soil
Extension
Members
Services
Policy
PRODUCTION
Farmer profiles allow targeted production information
ORGANISATION
Accurate knowledge of agribusiness members ensures more efficient operations in the organisation
Farmer profiles
A more efficient data management system for supporting the agribusiness can improve production, farmer organisations, access to financial services as well as access to markets.
TRADE & MARKETS
New markets open up when the farmer owned business can describe the location of production
FINANCE
Farm profiles are used to access credit and engage financial insitutions
Certification
Traceability
production offers / prices
Credit
Insurance
Subsidy

Source: Addison et al., 2020.

©Adobe Stock/Андрей Япанский

The availability of a maintained FDP platform is critical for ICT service providers as it saves huge costs, which makes services more affordable and sustainable. To reach this goal, these services must be designed jointly between FOs, cooperatives, service providers and farmers.

- Greater power in advocacy: A deep knowledge of its membership allows organizations to have a stronger voice in advocacy. At a basic level, an organization with an FDP platform can prove their membership and demonstrate the number of people it is representing, and who and where they are. This gives power to their voice based on their representativeness. At a more advanced level, an organization can exploit its membership to inform policy-makers in various ways:
  - *Simulating the impact of proposed measures:* based on farmers' information, it is possible for an organization to measure or simulate the impact of new measures (e.g. a new subsidy scheme) and define its position with regards to proposed measures based on real data.
  - *Quickly executing surveys to get members' positions on specific topics:* organizations can mobilise members and collect their opinions on specific topics. This process helps an organization to define and defend its position based on real data and on a clear mandate from members.
  - *New sources of revenue:* farmers' profiles are potential sources of revenue for third-party activities such as research; market surveys; and advertisement.

However, it is important to note that farmers must agree with the use of their data and their participation in such activities. These revenue streams are presented in detail in 1.2.5. Farmer profiling platform business model.

In short, an FDP platform is an opportunity for FOs, cooperatives and similar organizations managing a farmers' group, as well as for enablers of new farmer-centric services that are critical for increasing production, decreasing loss, and maximising income for smallholder farmers. An FDP platform is also a way to protect farmers' rights and ensure that they benefit from the disclosure of their farm-level data.

This video about a farmer profiling project in a tea factory in Uganda illustrates the value of an FDP platform to a cooperative (CTA, 2018a).

While an FDP platform is an important enabler for the design and delivery of services, it fits into a larger context, where other enablers could also provide an important impact.

It is important to note that an FDP platform is an enabler for both public services (e.g. planning, subsidiary scheme design, extension services) and private services (e.g. financial loans). Apart from information services, policy-makers are also potential consumers of data stored in profile information. The profiling platform can provide raw content to compute key policy indicators (e.g. land planted, size of land under irrigation, etc.). In the same way, the profiling platform can be used to forecast the impact of policy interventions.

However, while all these applications can potentially be enabled by a profiling platform, the success of such a platform and its ability to deliver expected results depends on several elements. These include the implementation context, the content of the profiles, the quality, timeliness and completeness of data stored, and the usability, reliability and effectiveness of the platform from a technical point of view.

1.2.2 Components of a farmer profiling project

Components of a farmer profiling project include five subsections, which introduce the different components or dimensions that must be considered to ensure the successful development of a farmer profiling project.

National context

In the case of a national-level profiling project, before engaging with the design and development of the FDP platform, it is important to explore and review other initiatives taking place in the country in order to identify existing public or private databases that could be linked to the profiling platform. This includes identity databases (e.g. Aadhaar system in India), land databases, subsidiaries databases, etc.

Legal and policy context

A farmer profiling platform collects and stores data about farmers and farms that are classified, by their nature, as personal data. In many countries, the collection, storage and management of personal data is regulated by specific legislation. It is therefore essential to review this legislation and to capture key requirements, such as official declarations, data sharing rights, obtaining farmer consent, which have to be considered for the building of the platform and to ensure that the project does not breach any regulations. See Section 2.4. Personal data protection for further details about this aspect.

In the same way, other legislation related to open data and official statistics may also have an impact on the ability to publish anonymised aggregated datasets, particularly if the platform is under the authority of a public agency. It is therefore essential to identify the legal constraints under which the technical platform is to be developed if it is to be compliant with national legislation. International and continental treaties, agreements and policies also need to be considered.

Operational dimensions

The success of a farmer profiling project largely depends on the operationalisation of the data collection tasks. There are two phases to consider:

- The setup phase: the profile information will be collected for the first time.
- The operational phase: the post-setup phase when the profiles are updated regularly.

Setup phase

The success of the setup phase relies on a series of factors:

- Engagement with targeted farmers: The collection of personal details of farmers and their farms is something that is not as easy as it first appears. The literature shows that farmers are reluctant, if not opposed, to providing their details if someone shows up at their farm. In order to support this task, and to ensure a faithful contribution by farmers, a series of activities have to be organized in advance.
    - *Awareness raising campaign*: Meetings with farmers, as well as radio spots, have to be organized to explain the concept, the process, the rationale, and the potential benefits for farmers. These campaigns should also explain, in detail, the information that will be collected and why.
    - *Data sharing agreement*: Farmers are usually not willing to share their data if they are not clear about who is going to use the data. From an ethical perspective, sometimes a legal perspective, and an operational perspective, it is important to present a data sharing agreement to farmers. Note that the data sharing agreement is the result of a careful analysis of the personal data protection regulations.
    - *Increase trust in data collectors*: Farmers need to be sure that those coming to their farms asking for data are authorised data collectors. It is recommended that enumerators are equipped with a professional card and an easily recognisable item of clothing such as a jacket or hat, which should be promoted during awareness raising campaigns. Other elements that increase trust and ease the work of the data collector include an announcement of the timing of the visit or the introduction of the collector to the farmers by a trusted person (e.g. extension agent, cooperative/farmer group leader).
- Training of enumerators: The training of enumerators, not only on the technical platform, but more importantly on the data collection process is essential for the success of the task. In particular, the training must include awareness raising on data security and the protection of their equipment, confidentiality of data collected and presentation of the data sharing agreement and capture of farmers' agreement. The training should also include best practices on the sending of profiles to the central platform for review. It is recommended to design a charter that data collectors must sign, which includes all elements of their tasks, as well as potential legal risks if they breach some of the legal requirements or confidentiality.

- Provision of robust equipment: The data collector's equipment is a critical element for the process so must be reliable. Different elements should therefore be considered:
    - Power: A data collector should be equipped with one or two power banks to ensure that they can conduct a full day of data collection without being affected by lack of energy issues. Depending on the location, organising how data collectors can recharge their power banks and equipment should also be implemented
    - *Tablet*: The robustness of the tablets and quality of their components, in particular the onboard GPS chip, are essential for ensuring the quality of the information collected.
    - *Memory*: Depending on the extensiveness of the profile, it may be advisable to plan for extra memory on the tablet to ensure that all data can be stored.
- Funding of data collectors: It is essential that data collectors have enough funding for travel and communication (e.g. to call for support, transmit profile data); inadequate support can have a major effect on timing.
- Monitoring: It is essential to monitor the quality of the profile collected in almost real time, particularly at the beginning of the process, to ensure that each and every data collector understands the tasks. When problems are detected on profiles, the information must be communicated to the collector so that the person completes the problematic profiles. Putting in place a payment per profile collected and validated is an efficient way to assure the profile quality.
- Support: Data collectors will always encounter problems that they are unable to fix or have questions that need to be answered for them to execute their tasks. It is critical to have a support mechanism in place (e.g. a hotline) to allow them to access the necessary support in the field when needed.

The gender dimension is an additional factor to consider in a farmer profiling project. The gender aspect is sometimes critical and may dramatically influence the quality of the data collection and the level of contribution by the farmer. For example, male data collectors interviewing female farmers or vice versa may impact the level of engagement and the quality of the data provided. In some countries, the issue of gender is not critical but, in others, it poses a major risk to data collection. Gender aspects have to be integrated at different levels. This includes, in particular:

- The gender of the data collector and the language enumerators speak
- The time and language of the awareness raising campaign broadcast by the radio station.
- The time (day and time) at which data collection is organized.

Operational phase

Organising one-off collection of data (for example, by means of a census) is relatively easy, but putting a system in place, which will ensure that profile data are updated regularly, is more challenging. There are different models that could be explored, and different potential organizational and institutional arrangements that should be considered. The options are presented below.

- Centralised versus decentralised model: There are different ways of organising an FDP platform. One way to do it is by using a centralised model, where a given organization decides to build the platform and map all farmers. This model fits well for any local organization in direct contact with their farmers (a local FO, an agribusiness, a cooperative, etc.). Another way to do it is through a decentralised model, where the profiles are provided by a series of organizations. This model fits well for nationwide FDP platforms or for decentralised organizations (e.g. a group of cooperatives). In such a setup, it is then more efficient to rely on an individual cooperative or farmer group to map their own members. The decentralised model is usually more efficient, for two main reasons: (1) the mapping organization is in direct contact with farmers, and a trust relationship already exists; (2) the mapping organization can use profile information for their activity, creating an important incentive to do a good job. However, such a model is only possible when farmers are well organized, and each entity has enough capacity to conduct the tasks and exploit the profile.

©Adobe Stock/Jacob Lund

- Specific task versus part of other tasks: Usually the setup phase, as presented above, is run like a census. It is possible to organize profile updates in the same way with annual or seasonal profile update tasks. However, this model presents two challenges: (1) such a task requires a similar mobilisation of funds as the launch phase, and it is therefore costly; (2) the data are updated once a year at best, usually every 3 to 5 years and this may not be appropriate for specific tasks or value chains. For example, for tea and coffee growers or coconut producers, the profile information will not vary much over the years; for rain-fed crops, information must be updated at least once a year, and usually twice or three times in a year.

  The second option is to give the task of updating profiles to people who are regularly visiting farmers. Such a model allows a far more flexible data update and does not require major funding. The approach is easily implementable in countries where there is a strong extension agents' network and where the tasks can be added to their job description, or in the case of a decentralised model presented above where mapping organizations are closely linked to farmers. The best approach is usually a mixed approach where a full census is run every 3 to 5 years, and a seasonal update is in place but focuses only on a limited number of types of information and uses more basic technologies.

- Farmer-led process versus organization-led process: Updating profile information can be conducted as a top-down task with a mapping organization collecting information as presented above. Alternatively, farmers can update their own profiles.[11] This model requires two main elements to be successful. The first is that different channels must be put in place to enable all farmers to update their profiles. This may include a call centre as many farmers in developing countries will not be able to use smartphone applications, SMS and USSD services or even voice-based (also known as interactive voice response (IVR)) services. A farmer-led model also requires regular communication campaigns to remind farmers to update their profiles. The second and more important element is the incentive that rewards farmers for updating their profiles; farmers will spend time updating their profiles if there is a direct benefit. Incentives can assume various forms, including free extension services, financial rewards, access to a subsidy scheme, etc. The incentive plan is the cornerstone of the farmer-led model.

It is important to note that the selection of specific options for data collection has a major impact on the technical platform, its functionalities and the required infrastructure such as setting up a call centre or IVR services. It has also an impact on the overall project organization like who to train or how to train them, and on the overall budget. It is therefore essential that these elements are selected before technical choices are made.

Technical dimension

A farmer profiling platform is, by definition, an ICT platform and the technical and technological dimension is therefore an important component to consider. A profiling platform consists of three main elements:

1. The data collection module: This is usually in the form of a tablet application. In most countries, internet connectivity in rural areas is not reliable enough to opt for a connected application, and an offline application is therefore recommended. The application has to have a series of functionalities that include (but are not limited to):

   - *Ability to synchronise the profile information when connectivity is available.*
   - *Security:* The tablet and/or the application have to be secure to protect data:
     - the tablet should have antivirus protection;
     - the tablet should be erasable remotely when it connects to the internet;
     - the tablet should be protected and dedicated to data collection to avoid collectors sharing with others for e.g. pictures, games, etc. that risk allowing others to access confidential data;
     - the application has to have an authentication feature to ensure that data is only accessible to authenticated data collectors.
   - *Flexibility* to adapt to profile changes: The profile content is likely to evolve over time and the application should offer such flexibility.
   - *Geolocation:* As an application is very likely required to capture GPS coordinates (farms, fields, etc.), it should be able to handle an on-board GPS chip.
   - *Validation:* The application should support the data collector and detect at the time of entry any typos or errors by conducting a data check.
   - *Update:* The application has to offer a way to review and update existing profiles (operational phase). The application and the central profile repository have to synchronise their profiles, and specific data collectors have to have access to specific profiles assigned to them.

[11] It means that farmers would proactively get in touch with the organization running the FDP platform to update their profile.

2. The central profile repository: This is the heart of the platform. Key functionalities or elements to consider include:
    - *Agility:* The platform should enable an easy update of the profile template as the profile content is likely to evolve over time as new applications emerge that require specific data to be collected.
    - *Robustness:* The platform and underlying infrastructure should be designed based on the targeted number of profiles, the targeted number of data collectors and the likely number of simultaneous connections from third-party applications.
    - *Security:* The security of the central platform relies on different elements:
        - *Software security*: The platform must be up-to-date with regards to the operating system, and have updated antivirus software.
        - *Protection*: The platform must be hosted in a data centre that provides protection related to power, physical intrusion, temperature, water, etc.
        - *Backups*: These must be conducted at least on a daily basis and content must be stored securely at different physical sites.
        - *Access*: The platform should offer the latest technology for secured access in particular the use of encrypted password from end-to-end and should have a fine-grain authorisation model that offers a flexible framework for different user categories. In particular, the authorisation model should at least work at the individual profile level providing access to specific profiles and at the profile content level providing access to specific parts of profiles.
    - *Validation:* The platform should have a series of built-in checks that will highlight potential inconsistencies on stored profiles.
    - *Analytics:* The platform should offer a series of analytics to visualise different elements, such as the data collectors' activities, the database content (e.g. to drive the extension or prioritise tasks) or the age of the information. These analytics are important to support and monitor the data collection tasks, and to evaluate the quality and completeness of the stored information.
    - *Technologies and standards:* The evaluation of the technical platform should also cover the technologies and standards selected for the platform to ensure that the code can be easily maintained and updated over time, and to avoid any vendor lock-in.
3. The data access and publication module: As the role of the profiling platform is to enable other ICT services, the platform should offer a series of access channels. At the very least, this should include a web interface for human access, an API for software access, and potentially an export of open data to provide anonymised statistics and content for policy-makers and other stakeholders. Depending on the existence of a national open data policy, the platform may include a functionality to publish datasets automatically on a national open data portal.

The exact set of functionalities largely depends on the operational setup and, in particular, the model selected for the running phase. The most common approach includes the following elements:

- A smartphone-based data collection tool for the first data collection. The reference free and open source tool for this purpose is Open Data Kit (ODK) [**opendatakit.org**].
- A central repository developed as a web platform: For this part, there is not a reference free and open source package although there are different modules that could be used which would need to be integrated. If there is no geospatial data, packages such as ODK collect (part of ODK solution), ONA [**ona.io**] or KoBo Toolbox [**www.kobotoolbox.org**] are potentially interesting options. A comparison of the different tools used in August 2017 is available (CartONG, 2017). If geospatial data are included, the reference free and open source geographic information system is QGIS [**qgis.org**].
- A service for information update: As previously presented, such a module can be implemented in different ways, such as a USSD service, a voice-based service or a call centre. There is no reference package for such services, and they require ad hoc development. If there is a large agent network on the ground, and if they are equipped with smartphones, the same tool as the one for the first collection should be used.

An example of a complete architecture is presented in an article entitled *"Farmer Registration and Profiling: How Did it Go?"* published by CTA. This architecture has been used in several projects in Africa (CTA, 2019).

1.2.3 Profile content

The earlier subsection 1.2.2. Components of a farmer profiling project presented the core elements to successfully conduct farmer profiling tasks. However, as presented in the introduction, a farmer profiling platform is not a goal in itself but is an enabler that supports and eases the delivery of targeted information and other services to farmers. An FDP platform can support those services depending on the information stored in each profile. In that regard, it is important to note two points:

1. A profile cannot be exhaustive and capture everything about a farmer and his or her farm. Some crops require specific data; livestock and crop farming are, for example, very different both in terms of the available data and the interests of service providers. The profile content is therefore usually guided by how the profile data will be used.
2. There is a trade-off between, on the one hand, collecting a vast amount of data that makes the collection process longer, more expensive, more difficult and will likely create more resistance from the farmers, and on the other hand, the collection of basic data that are easy to capture may not vary much over time, and may not allow the development of advanced information services.

Based on these elements, it is critical to define the content of profiles based on stakeholders' perspectives and plans. It is not possible to be exhaustive, nor it is possible to identify all possible applications[12], but the project's success will rely on a bootstrapping phase that will demonstrate how such a platform can enable useful services for farmers, who, in turn, become more open to providing additional data to access more services. The best practice to define the profile content can be summarised in the following steps:

1. The organization in charge of the FDP platform should identify the set of services they want to implement which also represent the underlying rationale for setting it up.
2. For each service (e.g. access to credit, trade, etc.), the organization should organize a workshop with other parties involved in the service (e.g. microfinance institutions) to identify the set of information that would need to be collected.

This process will ease the design of the profile content.

Apart from the profile content, another critical element is the identification of the farmer. An FDP platform stores a large set of profiles and it is therefore important to know how to retrieve the specific record attached to a farmer. Some countries implement a national scheme for personal identification (e.g. ID cards in European countries and the Aadhaar system in India). Such schemes can easily be used as the index in the profiles database.

However, in most developing countries, there is no reliable ID number or other unique identifier to identify a specific farmer. In such a case, it is critical to understand the element of information that could uniquely identify a person. In some countries, first and last names are unreliable identifiers, so too are farm addresses. The use of biometrics, such as fingerprints, is relatively difficult to put in place and presents several challenges outside the cost dimension. The identifier usually requires a series of elements such as name, address related to a specific point of interest (school, health centre, etc.), phone number, which are useful to identify the person. Note that a picture is a potentially useful element to verify the profile information, but it is useless as a search criterion.

The design of the identification process is an extremely important element that should be explored during the early days of setup to avoid duplicate records and a split of information between multiple records. In the absence of a robust identification scheme, different use-cases have to be considered:

1. Identification of the person to update a profile in a face-to-face interaction.
2. Identification of the person without face-to-face interaction (via a call, or via a USSD service). In this case, the technology may also have a role to play (e.g. whether the technology used allows the capture of GPS coordinates or not).

Finally, the profile is very valuable information for farmers, and each profile records individual farmer data. Outside the services provided by the organization setting up the FDP platform, a farmer may be interested in using his/her profile information for other purposes. At the same time, the ability to see the profile information depends on an element of trust between the organization and the farmer. It is therefore highly recommended to include the design and delivery of a paper-based profile to each farmer.

An example of such a dossier provided to farmers is implemented by IGTF (Uganda) in their profiling initiative (Brussels Briefings, 2018b).

[12] See Section 1.1. Data for agriculture for more details on categories of applications.

1.2.4 Farmer profiling platform business model

The sustainability of an FDP platform is a critical element to consider during the early days of the project. There are three elements to consider as part of the business model:

1. The platform development and operational business model: whether the organization who wants to setup the FDP platform will cover the costs of development, operation of the platform and collect all revenue, or whether a public-private partnership model should be put in place?
2. The revenue streams: who are potential customers of the FDP platform?
3. The operational costs: what are the recurring costs?

Concerning the operational costs, they depend on the data collection and update model that is selected. At the very least, the different cost elements include:

- the hosting of the FDP platform;
- the hosting of the update services;
- the incentives for the different stakeholders (farmers, intermediaries in charge of updating profile information, etc.);
- the depreciation and renewal of equipment;
- the communication costs between various stakeholders;
- the staff in charge of monitoring the platform.

Public-private partnership (PPP) vs. organization-owned platform

The term PPP is used in a broad sense where a (public) entity needing an investment does not pay for the investment, but then either pays a fee annually to the other (private) entity that makes the investment or allows it to make revenue from the investment. The choice of a PPP model versus an organization-owned model should be driven by a careful evaluation of each option. The factors that should drive the selection of an appropriate model are:

- Funding: One of the main factors for the choice of model is usually driven by the available funding. If specific funding from a development partner is available, such investment does not fit well with a PPP model. In contrast, in the absence of specific funding for the platform, a PPP model is often the only option given the size of the required investment.

- Operational costs vs. investment costs vs. potential revenue streams: The choice of the model should be influenced by the evaluation of the operational costs, the investment costs and the potential revenue streams. All these elements must be balanced to evaluate the best option. This factor should be evaluated only at the end of the requirement phase, depending on the identification of the real costs versus non-financial incentives.

©Adobe Stock/sculodoc

- Feature evolutions: In a PPP model, the software is the property of the private sector partner and therefore the evolution of functionalities is not under the authority of the organization in charge of the FDP platform. Each extension will require a specific discussion. In both cases, for any extensions, the organization will likely have to pay for it but, in the case of the PPP option, there will be only one possible provider. In that regard, the PPP model offers slightly less freedom and is likely to be more expensive.

- Data governance: The governance issue includes data ownership and access. In the case of an FDP platform, the collected data can be monetised. At the same time, the protection of personal data should also be considered.

- Technical capacity: One of the key advantages of the PPP model in technology platforms is that the platform is operated and maintained by the private sector partner and does not require technical expertise or transfer of ownership to an IT team that may or may not have the required capacities.

The choice is not really a binary choice of a PPP versus an organization-owned model; mixed models should also be considered. A typical mixed model for technical platforms is a model where the organization pays for the development, owns the platform, data and the revenue, but outsources the operations. If the organization in charge does not have the staff or the capacity to manage the platform, this is a suitable option.

Revenue stream

Different products can be designed from an FDP platform which can generate revenue. Table 2 presents an overview of the needs, requirements, and potential uses of main actors in the agriculture sector.

Table 2. Overview of actors and requirements in the agricultural sector

Potential clients	Examples of data needs	Data available in the profiling platform	Client likely to access the profiling platform	Potential customer's requirements for the profiling platform	Potential customer's activities based on profiles data
NGOs (and donor partners)	Performance data to support programme or intervention designs, contact data to support advocacy and outreach campaigns	Yes.	Possible	Data quality, timeliness and completeness;[*] neutrality in data access;[#] personal data in compliance with open data and privacy principles; affordability;[$] real-time web and API access.	Define programme interventions (community, geography, targets, etc.); support programme activities (communication, support of beneficiaries; monitoring and evaluation)
Financial institutions	Data on unbanked farmers; data on farmers unable to purchase inputs due to financial constraints; financial (assets) and performance (yield) data to design agri-specific financial products, and to issue loans and credit facilities	Mostly	Possible	Data quality, timeliness and completeness; data provided by neutral party; personal data; affordability; real-time web and API access.	Feed credit scoring engine with data design new financial product Conduct market analysis targeted advertisements
Insurance companies	Performance (yields) and environmental data to design gri-specific insurance products	Yes	Possible	Data quality, timeliness and completeness; geolocalized data;[%] neutrality in data access; affordability; real-time web and API access.	Implementing new insurance products at lower cost; market research; product design; targeted advertising

[*] The data collection and validation chain should be transparent with various elements of trust added to the profile, such as the time of update, the evolution of the profile over time, the name of the operator updating the profile, and the GPS coordinates of where the profile update was made.

[#] The data provider should be a neutral party without any bias and not providing preferential access to specific parties.

[$] The concept of affordability is hard to define because affordability is not only a function of a potential customer's ability to pay but of the value they attach to the product or service being offered to them. During the interviews conducted, interviewees were not in a position to provide indications of what they are prepared to pay for data because they were uncertain which data would be available, and whether this is accurate and reliable. Most organizations have access to some data via their own or other existing systems (e.g. access to credit information from the Credit Reference Bureau in the case of financial institutions) and they cannot yet evaluate how a new approach could ease their work and increase their opportunities.

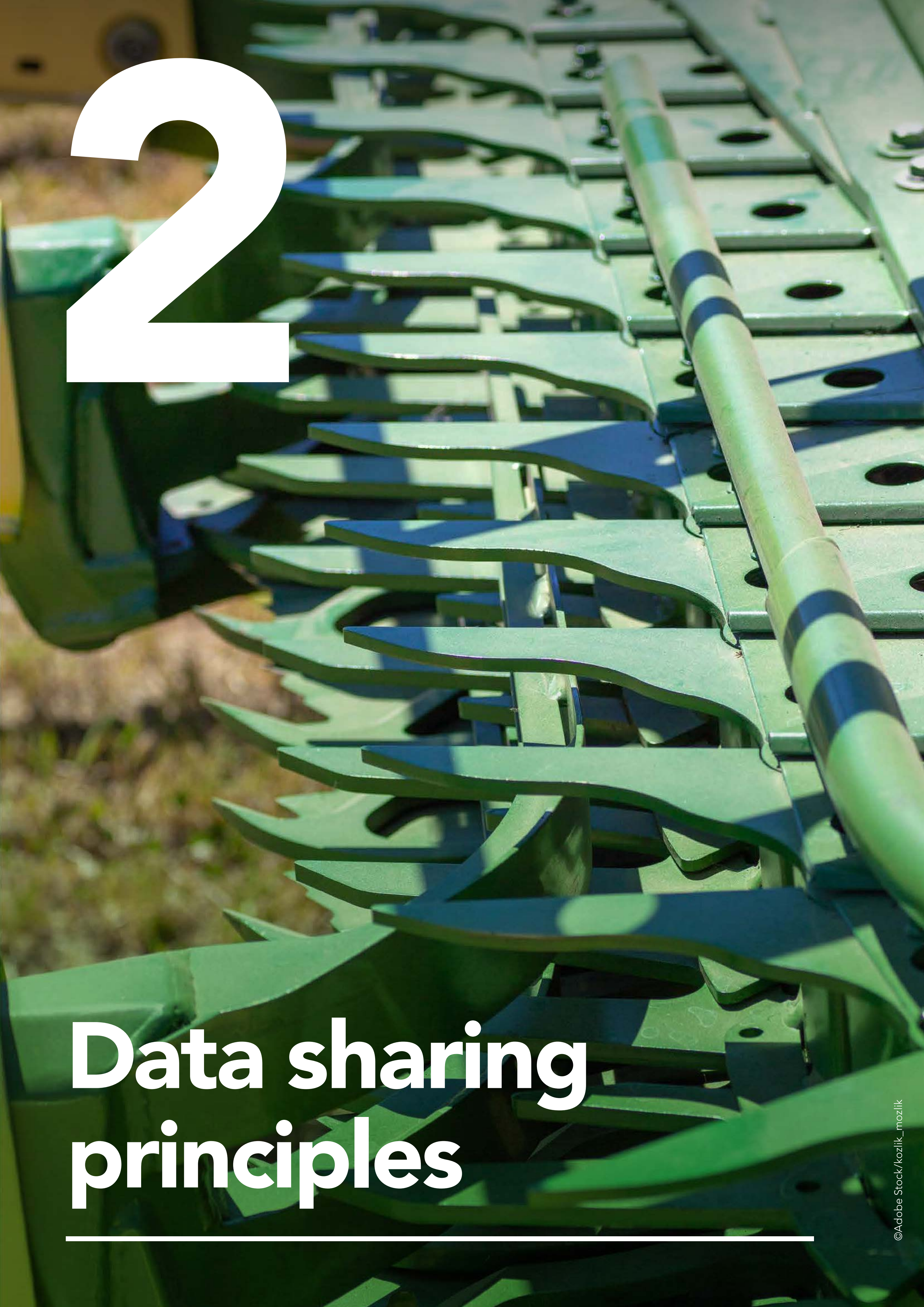
[%] The trend in agriculture at the moment is on weather-based insurance, which requires information about the exact location of fields owned by the farmer or cooperative applying for insurance.

Potential clients	Examples of data needs	Data available in the profiling platform	Client likely to access the profiling platform	Potential customer's requirements for the profiling platform	Potential customer's activities based on profiles data
Food processing and exporters	Data on certification, yields, prices for planning and to ensure quality	In part	Unlikely	Data quality, timeliness and completeness; neutrality in data access; affordability; real-time web and API access; specific data needs&	Identifying new production opportunities; developing agri-food supply chains
Agricultural products dealers	Data on crops, land, pests, inputs, etc. to provide farmers with targeted products and services	Mostly	Likely	Data quality, timeliness and completeness; neutrality in data access; affordability; real-time web and API access	Market-research design new product targeted advertisement
Researchers	Empirical data to support scientific studies; data to design innovative agri-products to support transformation to the knowledge economy	Mostly	Likely	Data quality, timeliness and completeness; open data@	Research planning and analysis; replication; quantitative surveys; modelling; design of research and development; prototypes

Agritech startup	Profile information to complement specific activities (e.g. drone mapping, IoT/ sensor, tools, lending, etc.)	Mostly	Likely	Data quality, timeliness and completeness; neutrality in data access; affordability;
				real-time web and API access
					market research;
					product design & deployment;
					targeted advertising

& Each value chain has specific data needs depending on the commodity and a farmer profile may therefore accommodate customised value-chain specific data. For example, the altitude of a wash station is an important data point in the coffee value chain but does make sense in other value chains.

@ Agricultural researchers are not usually interested in personal data, but more in anonymised open data for analysis and publication.



©Adobe Stock/kozlik_mozik

2

Data sharing principles

2.1 What is shared and open data

Section 2.1. What is shared and open data aims to introduce readers to the principles of sharing data, what makes data open and the benefits of opening data. While the focus throughout the book is on shareable, structured data, Section 2.1. What is shared and open data makes a particular link between shared and open data on a spectrum.

In Section 1.1. Data for agriculture, it was explained that the availability of more data at global and farmer level helps to enhance extension, trade and financial services, which can increase income and yield. The use of ICT makes it possible to forecast the future much better than before or to answer seemingly complicated questions much more quickly based on data. Such questions might be: Where does our food come from? Can we manage risks in our farm and take control measures against droughts or pests? Are we able to predict problems such as floods or low yields? Can we make informed decisions on what to grow, what treatment to apply, when to plant, treat or harvest? Technologies today allow us to build services to answer these questions, but data only offers these opportunities when it is usable.

The notion of open data has been around for some years. Considerable amounts of data today are generated by the public sector, e.g. soil surveys, cultivar registrations, pesticide residues, healthcare, defence industries, infrastructure, public education, and telecommunications. See the categories of datasets presented in Section 1.1. Data for agriculture, including individual datasets accessible on public portals. In 2009, various governments, including Canada, United Kingdom and the United States of America, launched open government initiatives to open up their public information.

In addition to public data, for which there is a general demand for openness, private sector data is also becoming more important for decision-making. While it is not always feasible to make this data completely open, many of the principles of open data (access, reuse, interoperability) apply also to the sharing of private sector data even if under different access conditions.

Open access to research and sharing of data are vital resources for food security and nutrition, driven by farmers, researchers, extension experts, policy-makers, governments, international agencies and other private sector and civil society stakeholders participating in 'innovation systems' and along value chains. Lack of institutional, national and international policies and openness of data limit the effectiveness of agricultural and nutritional data from research and innovation. Making open data and data exchange in the value chain work for agriculture requires a shared agenda to increase the supply, quality, and interoperability of data, alongside action to build capacity for the use of data by all stakeholders.

From mobile technology used by health workers to open data released by government ministries, data is becoming ever more valuable, as agricultural business development and global food policy decisions are being made based upon it. But the agriculture sector is also home to severe resource inequality. The largest agricultural companies make billions of dollars per year, in contrast to subsistence farmers growing just enough to feed themselves, or smallholder farmers who grow enough to sell on a year-by-year basis (Ferris and Rahman, 2016a).

The scarcity of available data prevents us from identifying and learning from real progress at the global and national levels. It also hides inequalities within countries, making it more difficult for governments to know about them or for others to hold governments fully accountable (IFPRI, 2016). National averages are not enough to see who is being left behind, as nutritional levels can vary even within households. Data should be used actively to make better choices and inform and advocate decision-making from the household level all the way up to policy level.

2.1.1 Notion of shared and open data

Data exists on a spectrum and it can be closed, shared or open. Datasets may include sensitive information for security, personal or commercial reasons. For instance, health records may cover sensitive data, which raises privacy issues. For these reasons, data can be closed or can be shared with limited persons or groups but not licensed to permit anyone to access, use and share it. The Data Spectrum in Figure 9, developed by The Open Data Institute (ODI), illustrates the degree of openness of data and helps to understand the language of data (The Open Data Institute (ODI), 2019a). Data can be shared within a closed or partially closed group or even publicly on the web without being identified as 'open data'. What makes it shareable is the structure of data and machine readability.

To make data open, the important thing is how it is licensed. For data to be considered open, it must be:

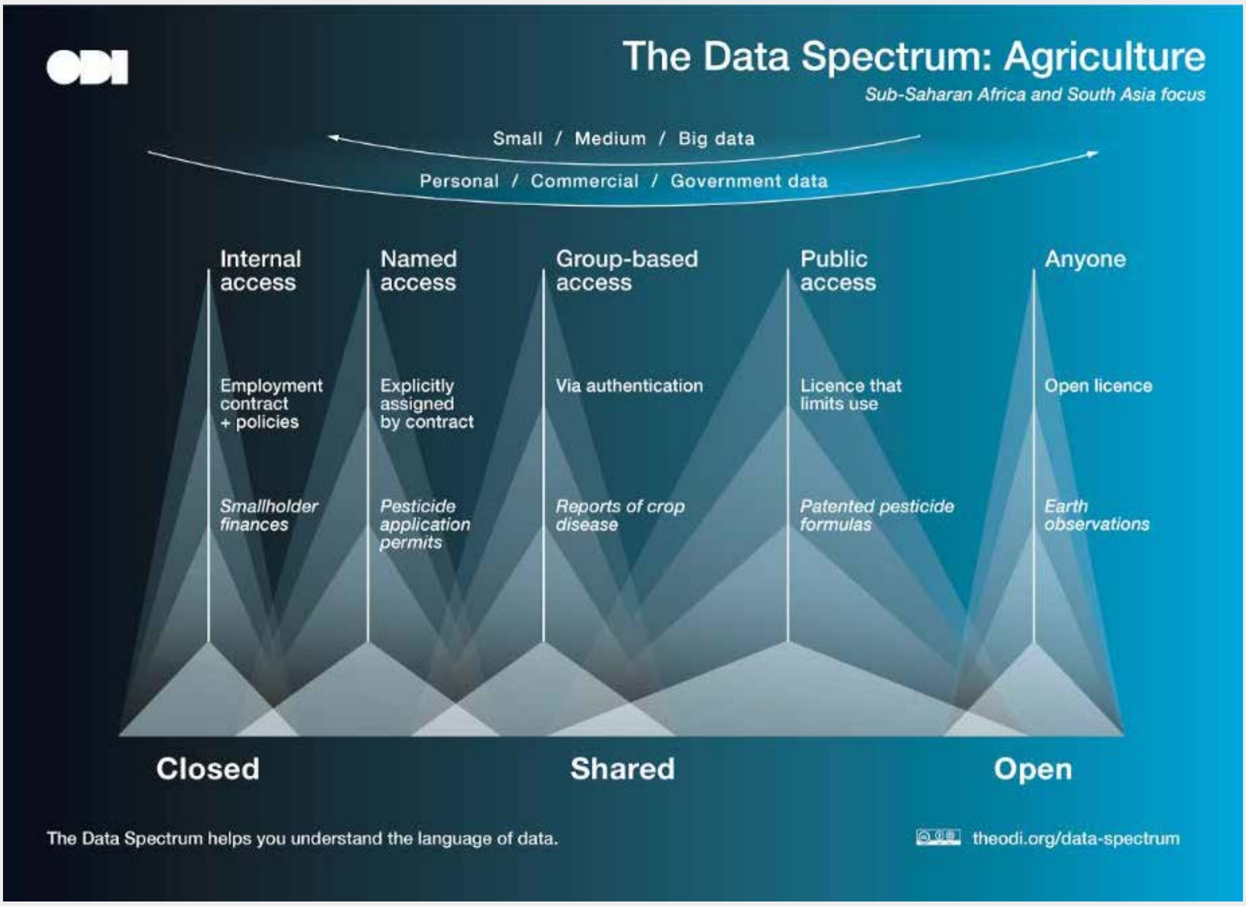
- accessible, which usually means published on the web;
- available in a machine-readable format;
- with a licence that permits anyone to access, use and share it – commercially and non-commercially.

Many individuals and organizations collect a broad range of different types of data in order to perform their tasks. Government is particularly significant in this respect, both because of the quantity and centrality of the data it collects, but also because most of that government data is public data by law, and therefore could be made open and available for others to use (Open Knowledge Foundation, 2020a).

The open data movement has been advocated strongly by governments to allow others to benefit from their data and their desire to be transparent, but research institutions and the private sector also generate data which they are willing to share as a common good (Gray, 2014).

Open data is *"data that can be freely used, reused (modified) and redistributed (shared) by anyone"* as defined in the Open Data Handbook ('What Is Open Data?', 2020). The Open Data Handbook emphasises the importance of the definition of open data and highlights key features as follows:

Figure 9. The ODI Data Spectrum: Agriculture


ODI
The Data Spectrum: Agriculture
Sub-Saharan Africa and South Asia focus
Small / Medium / Big data
Personal / Commercial / Government data
Internal access
Employment contract + policies
Smallholder finances
Named access
Explicitly assigned by contract
Pesticide application permits
Group-based access
Via authentication
Reports of crop disease
Public access
Licence that limits use
Patented pesticide formulas
Anyone
Open licence
Earth observations
Closed
Shared
Open
The Data Spectrum helps you understand the language of data.
theodi.org/data-spectrum

Source: The Open Data Institute (ODI), 2019a.

Availability and access: The data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form. Managing data can be costly in terms of time and resources needed. An example of costing for data management can be seen at the UK Data Service (UK Data Service, 2020).

Reuse and redistribution: The data must be provided under terms that permit reuse and redistribution including intermixing with other datasets.

Universal participation: Everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

2.1.2 Open data principles

The Open Definition makes precise the meaning of 'open' with respect to knowledge, promoting a robust common in which anyone may participate, and interoperability is maximised. Knowledge is open if anyone is free to access, use, modify and share it – subject at most to measures that preserve provenance and openness (Open Knowledge Foundation, 2020b).

Open data must comply with an open licence or a status; it must be in a public domain or under an open licence. Without a licence, the data cannot be reused. Open data must also be accessible and downloadable via the internet. Any additional information necessary for licence compliance must accompany the work, such as an attribution to say that people who use the data must credit whoever is publishing it, or a share-alike requirement to say that people who mix the data with other data have to also release the results as open data.

Open data must be in a machine-readable form which can be processed by a computer and where the individual elements of the work can be easily accessed and modified. It must also be in an open format which places no restrictions, monetary or otherwise, upon its use and can be fully processed with at least one free/libre/open-source software tool.

The licence used for the open data should be compatible with other open licences and should permit free use, redistribution, creation of derivatives, and compilation of the licensed work. The licence must allow any part of the work to be freely used, distributed, or modified separately from any other part of the work or from any collection of works in which it was originally distributed. The licence must not discriminate against any person or group.

The Open Data Charter, which is a collaboration between over 70 governments, agrees on six principles for how governments should be publishing information. Each of them is briefly explained below. On their website, the Charter also provides detailed action items to achieve each of these principles (Open Data Charter, 2015).

Open by default: Free access to and use of government data (data held by national, regional, local, and city governments, international governmental bodies, and other types of institutions in the wider public sector) brings a significant value to society and the economy, and the government data should, therefore, be open by default. Resources, standards, and policies for the creation, use, exchange, and harmonisation of open data should be globally developed, adopted and promoted so long as citizens are confident that open data will not compromise their right to privacy.

Timely and comprehensive: Data may require time, human and technical resources to be released and published. It is important to identify which data to prioritise for release by consulting with data users. The data must be comprehensive, accurate, and of high quality.

Accessible and usable: Opening up data enables stakeholders to make informed decisions. The data should be easily discoverable and accessible, and made available without any barriers.

Comparable and interoperable: The data should be published in structured and standardised formats to support interoperability, traceability and reuse. It should also be easy to compare within and between sectors, across geographic locations, and over time in order to be effective and useful.

For improved governance and citizen engagement: Open data strengthens governance and provides a transparent and accountable foundation to improve decision-making and how land markets operate. Open data enables civic participation and better-informed engagement between governments and citizens.

For inclusive development and innovation: Openness stimulates creativity and innovation. Open data, by its nature, offers an equitable resource for all people regardless of where they come from or who they are and provides a less digitally divided environment to access and use the data.

2.1.3 Benefits for shared and open data

The benefits of open data are diverse and range from improved efficiency of public administrations, economic growth in the private sector to wider social welfare and citizen empowerment.

Performance can be enhanced by open data and contribute to improving the efficiency of public services in health and nutrition. Greater efficiency in processes and delivery of public services can be achieved thanks to cross-sector sharing of data which can, for example, provide an overview of unnecessary spending. Resources can be better targeted thanks to local-level, disaggregated data, showing which areas and populations have the greatest needs.

The economy can benefit from easier access to information, content and knowledge, in turn contributing to the development of innovative services and the creation of new business models.

Social welfare can be improved as society benefits from information that is more transparent and accessible. Open Data enhances collaboration, participation and social innovation (What Is Open Data, 2020).

The availability of detailed open data is essential to improve delivery of services at the local level. Examples include mySociety [mysociety.org], the Hungarian 'right to know' portal, and Fix my Street fixmystreet.com] in Norway. To support the emergence of new data-driven businesses and the growth of existing ones, governments need to publish key datasets. By growing economies and improving services, open data allows governments to make savings in key areas, such as provision of healthcare, education and utilities. In the UK, open data helped reveal GBP 200 million (USD 248 million) of savings in the health service. In France, energy data is being used to drive more efficient energy generation practices (The Open Data Institute (ODI), 2019b).

GODAN's report *How can we improve agriculture, food and nutrition with open data?* specifies three ways that open data can help solve practical problems in the agriculture and nutrition sectors (Carolan *et al.*, 2015).

- Enabling more efficient and effective decision-making. Open data enables computers to pull data from various sources and to process it for us. It does not rely on humans to interpret and integrate information contained in web pages. Open data underpins new products and services by presenting information from a wide range of sources that helps everyone from policy-makers to smallholders find gaps in markets or fine-tune their products or services. There are good examples of it in fisheries like the Abalobi Initiative described in Chapter 1.
- Fostering innovation that everyone can benefit from. As a raw material for creating tools, services, insights and applications, open data makes it inexpensive and easy to create new innovations. When data is open for all to experiment with, there is no need to invest large amounts in repeating trials that have already been completed. When data is openly licensed, it also allows for novel combinations with other data to gain new insights. A good example of using data to apply precision farming tactics to their land is the story of Andrew from Allington, United Kingdom, who works alongside his family on their arable and dairy farm. The family has taken advantage of new tools and technology that allow them to easily view satellite data of their land, which has been opened up by the European Space Agency. His story is available in the Open Skies video, which is part of the GODAN Documentary Web Series (GODAN Secretariat, 2017a).
- Driving organizational and sector change through transparency. Transparency around targets, subsidy distribution and pricing, for example, creates incentives which affect the behaviour of producers, regulators and consumers. By requiring companies, government departments and other organizations to publish key datasets – performance data, spend data or supply-chain data, for example – governments, regulators and companies can monitor, analyse and respond to trends in that sector. More importantly, publishing this data across a sector can ultimately transform how products and services are delivered. We can refer here to the same example as given above, Abolobi Fishers from Open Water, of using data about fisher practice for the small-scale fisheries industry to make informed decisions.

Providing farmers with more accurate, accessible, timely information – from large agriculture groups to individual smallholders – will help to ensure food commodity markets function well in future. Progress will be driven largely by providing better access to accurate, timely information for individual smallholder farmers, businesses and policy-makers alike. Open data can and should be part of the solution. Open data promotes transparency across the sector to accelerate progress, identify areas for improvement and help create new insights *(Carolan et al., 2015).*

Agmarknet [agmarknet.gov.in] in India is a good example of providing market information with more than 2 700 data sources to farmers, traders, policy-makers and other stakeholders for better production and marketing decisions. The rice producer's federation of Colombia keeps data sets historically and helps small and medium-scale farmers by measuring climate, yields and farming practices related to rice-growing in the country. Readers can watch the story of Blanca, who runs her farm outside the town of Ibagué with the assistance of early-warning systems and weather data in the Open Climate video, which is also part of the GODAN Documentary Web Series (GODAN Secretariat, 2017b).

2.1.4 Open data acts as change agent

Open data acts as a change agent as implementing an open data initiative often involves cultural and institutional change. Opening data goes far beyond putting data on a website under an open licence. Applying the technology is relatively easy when compared with bringing about a cultural change, which can be much harder (European Commission, 2020) and requires consulting with potential data users internally within an institution, as well as external stakeholders.

The same is true for private sector data sharing projects that may not adopt fully open data approaches but still need to change their attitude towards their data and engage with other actors for its reuse.

However, this difficulty of adopting a change does not stop the amount of data which is increasingly becoming openly available. There are still challenges related to data management, licensing, interoperability and exploitation. There is a need to evolve policies, practices and ethics around closed, shared, and open data.

The challenges involved in opening agricultural data are best addressed at the level of a particular problem in a specific field, where standards can be identified or developed, and data released as part of solving a problem. This is especially true when advocates can point to a clear theory of change. GODAN addresses this issue with care and sets out five strategic steps (Carolan *et al.*, 2015) for pursuing solution-focused open data initiatives for agriculture and nutrition:

- Engage with the growing open data community, including key problem owners and GODAN experts to identify the challenges that open data can help solve.
- Build open data strategies and projects with a focus on finding solutions to land tenure, agriculture and nutrition problems.
- Develop the infrastructure, assets and capacities for open data in relevant organizations and networks.
- Use open data and support users of relevant data.
- Learn through ongoing evaluation, reflection and sharing to ensure we can all continue to improve our practice.

2.2 Challenges for smallholders in data value

Challenges for smallholders in data value aims to make both service providers and farmers' organizations aware of the challenges and risks that smallholders face with the data flows in different value chains for example, through precision agriculture or with any actor that needs to profile them, like farmers' associations or governments.

Understanding these challenges is essential to be able to create services and negotiate business models that meet farmers' needs and address their concerns.

For smallholders, the two main challenges are: (a) to gain access to relevant and usable data and services; and (b) to make sure that any data they share does not actually weaken their position in the value chain (and ideally that sharing data actually benefits them). In both data sharing directions, smallholder farmers face big data asymmetries in relation to other actors in the value chain. Section 2.2. Challenges for smallholders in data value illustrates these data asymmetries and the related ethical, legal and policy issues.

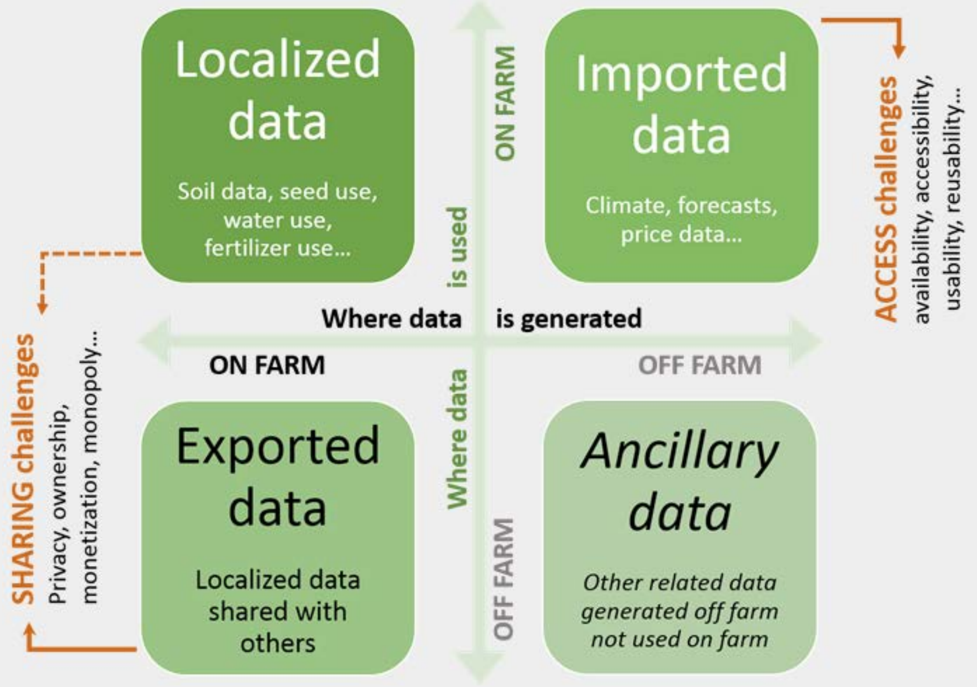
Figure 10 illustrates the different streams of data from and to the farm (the fourth stream, completely off farm, is not relevant for this topic) and the related types of challenge that farmers are facing (Maru *et al.*, 2018).

2.2.1 Data asymmetries and power imbalances

The data challenges that we describe in this book have probably always existed: farmers always needed to find good sources of information for decision-making and they have always shared information about their farm in order to get advice while trying not to lose any competitive edge. However, such challenges have been amplified in recent years by new technologies that collect and process data at drastically higher volumes and velocity, and by the difficulty of tracking data and data rights across the many flows and transformations generated by ICTs. In addition to these data flow challenges, there are also socio-economic factors that add to the complexity: the ownership and administration of such technologies and related amounts of data are, of course, linked to the power dynamics already present in value chains.

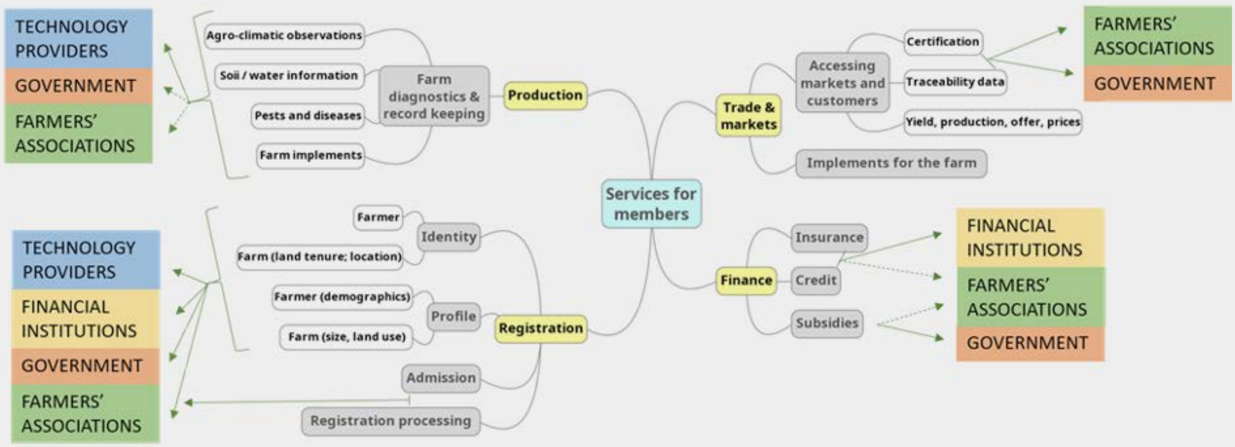
Such power dynamics are not the same across all regions of the world as they depend on economic and political conditions that determine the structure of the value chains. However, with the market economy and similar liberalist policies adopted widely, there tend to be similar trends worldwide. Power imbalances are especially strong in regions where the vast majority of farms are small family farms and where other actors in the value chain are stronger, for instance, where actors upstream of the farm (e.g. input providers, technology providers) or downstream of the farm (processors, distributors, retailers) are more concentrated in large companies, often multinationals. Concentration means less competition and more bargaining power. Besides, there is the growing phenomenon of 'vertical integration', by which big companies integrate other steps in the supply and value chain, thus becoming, in some cases, both a supplier and a buyer for the farmer, if the production itself is not integrated as well, for example through contract farming. Vertical integration brings less dependence on other actors, even more bargaining power and better knowledge of the whole value chain.

Figure 10. Streams of data from/to farm


Localized data
Soil data, seed use, water use, fertilizer use...
Imported data
Climate, forecasts, price data...
Exported data
Localized data shared with others
Ancillary data
Other related data generated off farm not used on farm
ON FARM
OFF FARM
Where data is used
Where data is generated
ON FARM
OFF FARM
SHARING challenges
Privacy, ownership, monetization, monopoly...
ACCESS challenges
availability, accessibility, usability, reusability...

Source: Maru et al., 2018.

Figure 11. Farm data shared with other actors



Source: Technical Centre for Agricultural and Rural Cooperation (CTA), 2019.

The reason why digital technologies and related data services have come to play an important role in these dynamics is that they are now mainly administered by large companies (concentrated or vertically integrated) that sell these technologies to farmers with much less bargaining power, which has added data asymmetries to the already existing value-chain disparities. Besides the fact that data, software and equipment are managed by the technology provider, smallholder farmers would not have the skills to operate the equipment and infrastructure, and to collect, manage and give informed access to their data.

Data asymmetry means a disparity in access to data and knowledge, as well as a disparity in how much data each actor shares and how dependent they are on other actors' data. Farmers share a lot of valuable data with different actors in different value chains (supply chain, farmers' organizations, finance, government data collection for subsidy schemes etc) and often have little access to other valuable data and knowledge, even to the knowledge generated by their data once aggregated. At the same time, their data, aggregated and combined with other data from the value chain, gives data collectors precious knowledge and foresight to make decisions, a privileged position to tailor other actors' needs and to influence important aspects of the value chain (e.g. the seeds market, prices in general - including price discrimination, supply chain disruptions, etc.).

2.2.2 Challenges with farm data sharing

Farmers share a lot of valuable data with several other actors in different data value chains, e.g. with technology providers for precision agriculture decision support systems; with suppliers and distributors for data exchange in the supply chain; with farmers' associations for the purpose of registration and service provision; with banks for financial assessment; and with governments for subsidy eligibility and compliance, etc. Figure 11 provides an overview of data flows. Dashed lines mean that the specific data flow is still not very common but is being experimented with.

This data sharing is more frequently managed through digital technologies. New digital technologies – such as sensors, drones and the Internet of things (IoT) in general or the blockchain – create an automatic flow of data from the farmer to the data collectors, over which the farmer has very little control. Besides, such flows are designed in ways that are not clearly explained in contracts, and contracts are often non-negotiable.

For this reason, farmers are often wary about the use of technologies like drones or sensors whose data is under the control of the technology provider.

Table 3 provides a summary of challenges presented by sharing farm data, followed by a more in-depth analysis.

Table 3. Challenges of sharing farm data through the value chain

Challenge		Specific relevance to the farmer
Risk of unfair data practices	• Uncertainty on the ownership of data collected through digital technologies and related rights on these data • Lack of legal protection for sensitive non-personal data • Issues of data privacy, security, manipulation, veracity, validation, liability • Lack of awareness of and consent from the farmer • Monetisation (e.g. actors down the line in the value chain reusing acquired data for commercial purposes) • Lack of clear legal framework for new ICTs (especially blockchain and IoT)	• The farmer is, in most cases, the primary originator and the subject of the data and therefore the most exposed to data rights uncertainty, data manipulation, veracity and liability • Farmer in a weak contractual position, often not aware of data reuse down the line • Monetisation: the farmer is the actor that generates most of the data and profits the least from it, while farm data is easily monetised by other actors
Risk of data power imbalances	• Unfair competition (data giving some actors more knowledge and a privileged position to sell tailored services, risk of lock-in)	• Actors upstream and downstream of the farmer have more knowledge about the market and about farmers' needs; they can sell farmers tailored technologies and products and potentially lock them in
Risk of widening digital and socio-economic gaps	• Risk of 'excessive transparency' of weak actors' data • Unbalanced data value chains and different degrees of dependence on external data sharing • Risk of concentration of power: ○ Cost of infrastructure (telecoms, secure protocols, 'ledgers', clouds etc): risk of natural monopoly for big actors and first movers ○ Possible unfair trading practices (lock-in, price discrimination, opaque algorithms hiding biased decision-making and lock-in mechanisms)	• Actors upstream and downstream of the farmer have more knowledge about the market and about farmers' needs; they can sell farmers tailored technologies and products and potentially lock them in • Excessive transparency is a phrase (DLG, 2018) regarding the excessive and not always justified amount of data shared by farmers with government (but it applies to transparency towards the rest of the value chain as well) • Big multinational consortia and vertical segments of the agricultural value chain are data self-sufficient and do not need to share • Farmers in a weak position to negotiate • Technologies too expensive for small farmers • Risk of opaque (biased) algorithms removing all decisional power from farmers (devaluation and loss of farmer's knowledge) • Risk of infringing on farmers' and/or indigenous rights (traditional knowledge, indigenous seeds etc.)

Source: Pesce, 2019, modified by the author.

Legal uncertainty and unfair data practices

Sensitive data

Besides personal data, which is normally protected by legislation, there are other types of data shared by farmers that are of a sensitive nature, such as confidential data concerning specific farming techniques. Besides the fact that precision agriculture equipment can reveal details about farming conditions and techniques and other potentially sensitive business data, there is also data collected by other actors, especially the government, such as agricultural censuses, satellite data and geospatial data in general. A lot of information about a farm and its activities can be inferred by combining all this data.

Currently, there is no legal protection for this type of sensitive non-personal data, unless they are classified as trade secrets, in which case they should not be shared (see Excessive transparency of farmers' data in 2.2.2. Challenges with farm data sharing).

Ownership, access and control rights

Ownership

It is common to read that farm data belongs to the farmer. It is sometimes even stated in contracts. However, 'ownership' is a legal assertion and data ownership is not addressed by legislation except for copyright for datasets as intellectual products. This is partly due to the peculiar nature of data compared to other goods that can be owned. In legal terms, it is non-rivalrous: the same data can be in different places and be owned by different people because, when data is copied or migrated to other platforms, it remains the same. In addition, there is a difference between data collected in a structured dataset, which can be considered as an intellectual product by law, and raw data as individual, unstructured bits before they are collected and made sense of. These raw data are similar to facts, for which no copyright and no ownership is legally applicable.

Thus, the concept of ownership is not strictly applicable to farm data. Besides the lack of legal applicability of the concept to raw data in general, machine-generated farm data presents additional complexities: (a) it is generated on the farm and is about the farm, but is generated by machines without the intervention of the farmer so the farmer is not considered the generator or collector; (b) it is raw data, so not an intellectual product, but is then transmitted and processed and combined with other data in aggregated datasets, which are intellectual products and can therefore be 'owned'.

It may be useful to also point out additional difficulties in using the concept of ownership:

- Depending on the type of data, some of it may fall under other established ownership-like rights with precedence on any contractual statement of 'ownership': copyright or database rights in the case of 'datasets', personal data, property rights, trade secrets and patents (de Beer, 2017). Copyright and database rights are often applied to aggregated data, which would technically belong to the company responsible for its collection and processing (Ferris and Rahman, 2016b; Rasmussen, 2016).
- Focusing on the traditional ownership concept in farm-data value chains that require data sharing and data transformation and aggregation may be counterproductive and restrict the flow of data.
- It is not obvious whether the ownership of farm data should be attributed to the farmer or to the landowner, when they are not the same person.

This does not mean that, in the future, there may not be new legislation addressing all the difficult aspects of data ownership and in particular raw data and machine-generated data. But at the moment, *"legal and regulatory frameworks around agricultural data ownership remain piecemeal and ad hoc"* (Townsend *et al.*, 2019).

The next section in this chapter – Section 2.3. Responsible data sharing in agricultural value chains – explains that the current predominant attitude towards data ownership in agriculture follows the approach of recognising 'attribution' (the word ownership is used reluctantly, for the reasons mentioned above) of farm data to the farmer or, as in the European Union (EU) code of conduct illustrated later, the 'data originator'.

Access

In the absence of a legal framework for farm data ownership, and considering the difficulties above, many experts agree that it is much more important to clarify rights of access and control on the subsequent versions of processed data (Sanderson, Wiseman and Poncini, 2018). In a way, if ownership cannot be asserted legally, it can be attributed 'by proxy' by defining other key aspects of ownership, like control and access.

Right to access attributes the right to 'see' and retrieve the data at any stage, even when collected and processed by other actors. For instance, with IoT equipment collecting data, manufacturers may or may not grant data access and reuse rights to the user of the object. A farmer using a drone to get images and data from his field may only have the right to use the drone software and not the right to actually 'see' or store the underlying data. Quite often, once aggregated and in further stages of reuse, data is no longer retrievable by the farmer (Kritikos, 2017).

Control rights

Right to control: This attributes the right to decide on the sharing and further reuse of the data. Regarding further reuses of data, under personal data protection laws, it is very common to apply the principle of purpose limitation (no reuse for purposes other than those to which the right-holder has originally consented) and this principle is sometimes recognised also for non-personal data.

However, its implementation seems difficult in the management of digital agriculture data, where data needs to be transformed and combined with other data in order to be useful for decision-making. For instance, if a farmer gives consent to a company to use farm data on soil, crop growth and pests in aggregation with data coming from other farmers in order for the company to provide back production advice, if at a certain point the company wants to share this data with an input supplier to get recommendations on fertilisers and pesticides, the company should again ask the farmer for consent for the new data sharing purpose.

Since there is no legal framework for these rights either, except when they are about personal data, in practice, the definition of rights is currently left to contractual agreements.

Contracts quite often recognise data 'ownership' to the farmer, but without specifying specific access and control rights: since legislation does not define data ownership rights, the recognition of ownership in contracts without access and control rights does not guarantee the enforceability of any rights.

Regarding contracts as an instrument, it has been observed that they are not the ideal solution: on the one hand, contract law is not harmonised internationally; on the other hand, contractual negotiations may not be a suitable solution for stakeholders that have very little bargaining power (de Beer, 2017). However, contractual agreements may work if farmers negotiate collectively and/or if contracts are based on a strongly endorsed code of conduct.

Data portability and interoperability

An important aspect of the right to access and control the data is the right to exchange the same data again with other actors. For example, a farmer using precision agriculture equipment that collects data on soil properties, irrigation, weather and crop health, may want to share this data with an insurance company for negotiating better premiums or with a bank for demonstrating the viability of his business. Precision agriculture systems do not always allow the information to be repackaged for further sharing.

A more specific and technical aspect of the right to reuse and re-exchange the data is the technical implementation of data portability which is the ability to port the data from one system/provider to another. Without this feature, there is little freedom to switch providers, which again weakens the farmer's bargaining power and is a limitation of fair competition in general. While this right is normally recognised for personal data, it is not always recognised for non-personal data.

The issue of data portability is linked also to the issue of interoperability between farm instruments and tractors and the data they generate which is often only compatible with other machinery of the same brand. Beyond the right to portability itself, the lack of interoperability between pieces of agricultural equipment and software systems across the value chain also contributes to locking farmers into one technological solution. While interoperability standards exist, they are not legally enforced.

Liability, veracity

With large amounts of data collected and transmitted by machines and used to make decisions throughout the value chain, one single error, transmitting incorrect or intentionally manipulated data, can have a potentially disastrous domino effect (Van der Wees *et al.*, 2017). It is true that such risks can actually be mitigated by data-driven technologies themselves: for instance, data collected by sensors and drones is more accurate and reliable than data collected manually, and blockchain technology can ensure that data are not manipulated in subsequent transactions. However, the legal value of data collected by IoT equipment is not universally accepted and legislation is still not clear about liability in the case of damage caused by incorrect IoT behaviour.

One aspect that farmers may want to consider is the legal framework which is still not stable and more importantly the contractual clauses on their potential liability for data generated by digital technologies on their farm or for their data in aggregated datasets.

Monetisation

Data monetisation is the profits that farm data can generate for data processors once aggregated, if they are sold or reused in paid services. An example would be a precision agriculture provider selling aggregated data on farm soil, crop growth and crop pests to agricultural input suppliers, who would be able to sell tailored fertilisers and pesticides.

Monetisation is not forbidden. However, it is useful to consider a couple of aspects that may be covered in contracts:

- If the contract includes a purpose limitation clause (e.g. data cannot be used for purposes different from the ones initially agreed upon) and monetisation was not agreed upon, in theory data cannot be monetised without asking again for consent from the farmer.
- It has been often claimed that any financial benefit generated thanks to data contributed by the farmer should be shared with the farmer, or that farmers should, in some way, benefit from it as well. However, this has not been successfully tried so far although there are a few initial examples in the United States of America and in Canada of platforms for farmers to share and sell their data. The main difficulties are along the lines of the ODI's reasoning around personal data (Tennison and Scott, 2018): (a) while the total value of all farm data from all farmers is high, the value of data from the individual farmer would probably be very small; (b) mainly poor farmers would resort to selling data, while richer farmers would maintain the privilege of data control.

Given these two challenges, the idea of collective platforms for farm data sharing and selling could work, if they reach a critical mass (so that aggregated data can become interesting for buyers) and are governed transparently, but there does not seem to be a market yet (Block, 2018).

Data power imbalances

Excessive transparency of farmers' data

Whether it is for receiving advice and services, or for the sake of certification, or for ensuring traceability, or for demonstrating compliance for subsidies, farmers have to share a lot of potentially business-sensitive data.

A position paper of the German Agricultural Society mentions the impression of farmers' *"excessive transparency vis-à-vis the public authorities"* but the same could be said of transparency towards other actors. The same paper notes that it is as though farmers were not considered *"on an equal footing with other economic operators whose operational and business data are recognized as worthy of protection"* (DLG, 2018).

Some of these business data might even be considered trade secrets. As noted in another study (Carbonell, 2016), *"details on soil fertility and crop yield have historically been considered akin to a trade secret for farmers, and suddenly this information is being gathered under the guise of technology and miracle yield improvements."* This statement may be considered too strong, as farmers have always shared data in some way, especially if this could benefit them, for instance in receiving advice.

With precision agriculture, data is shared for the same reason: to gain something back. Perhaps what has changed is that rather than being shared selectively and on a case by case basis, data is automatically 'taken' by technology providers directly from machinery at a rate and level of precision that practically reveals to the technology provider everything about the farmer's practices, even more than the farmer himself knows.

In any case, the fact itself that certain data are necessary for the functioning of precision agriculture or for obtaining subsidies or for certification makes it impossible to recognise them as trade secrets: data underlying a trade secret should remain secret and should never become readily accessible to third parties (Van der Wees *et al.*, 2017).

Risk of natural monopoly

The fact that agriculture technology providers are becoming bigger and fewer in many developed countries increases the risk of monopolies, more precisely, oligopolies. This is particularly the case as some digital technologies need costly infrastructure, although it should be noted that not all of them have high costs and some digital technologies can actually facilitate the entry of new actors in the market. Another risk is for agriculture technology providers accumulating huge amounts of data before other competitors, which becomes a big advantage. Network effects and switching costs can also create barriers to scale for new entrants (Townsend *et al.*, 2019).

Imbalance of contractual power and risk of unfair trade/data practices

The risk of oligopolies also affects farmer's freedom of choice in selecting technology providers and also his/her contractual power. The farmer may become dependent on the provider or be subject to unfair practices because of their lack of choice. Companies with more data and more insights than others can enact *"anti-competitive practices including price discrimination and speculations in commodity markets that may affect food security"* (Kritikos, 2017).

Risk of lock-in and biased algorithms

A risk that many farmers perceive is the opaqueness and potential bias of the algorithms used to process their data and provide advice. Since the algorithms by which decision support tools produce advice are almost always closed, the farmer can easily feel that he cannot exercise control over the decision-making process (European Commission, 2020). Potential consequences of this that have to be considered are: (a) the devaluation of farmers' knowledge and the weakening of his/her decisional power; and (b) the potential bias of algorithms, which can lock farmers into solutions chosen by the service provider. It has also been noted (European Commission, 2020) that impartial advisory services provided by governments or farmers' organizations could counterbalance the domination of private sector agronomic advice, but thus far they cannot offer the same level of tailored advice and are not well organized.

2.2.3 Challenges with accessing and reusing necessary data

Farmers need a significant amount of data or alternatively, since farmers do not use data directly, they need services that, in turn, need to access a large amount of data that is generated or aggregated by other actors. Examples include weather forecasts, climatic data, market prices, crop growth data, pest alerts etc (see 1.1.4. Identifying key datasets in farming crop cycles). This data is usually owned, managed and controlled by a third party and made available, directly or through intermediaries, to farmers and their representatives. It can be managed by governments or, more often, by private companies. For farmers and subsequently those who provide services to the farmer, this type of data presents the common challenges of availability (Is it available? From whom?), accessibility (Is it free? Is it open?), reusability (Can it be reused in other services? Is it interoperable? Are there licenses?) and quality (Is it reliable? Does it fit the purpose?), many of which are introduced in Chapter 3.

Chapter 2 Data sharing principles is about data in value chains, and Section 2.2. Challenges for smallholders in data value focuses on challenges related to the relative weight of different actors in the value chain, their willingness to share, their dominating position and the role of the public sector in providing data to level the playing field.

Table 4 summarises the challenges related to the availability of data for farmers and its accessibility and use by the farmer.

Table 4. Challenges for farmers to access relevant data and services

Challenge		Specific relevance to the farmer
Relevance of and lack of access to private sector's data	• Lack of access to private sector's public-interest data; the private sector often holds the highest-impact datasets • Lack of incentives for private sector to share data publicly • Cost of private sector's data • Lack of public alternatives to private sector's data and services • Lack of public data as a level playing field for smaller service providers	• Lack of access to private sector's agricultural data of public interest (e.g. product tracking, sensor data, prices…), which would help foresee market crises, epidemics etc. • Farmers dependent on private sector's data and services, with a few big providers and no competitive market
Little use and usability of publicly available data	• Difficulties in reuse: lack of comprehensive coverage, quality, veracity, standardisation; lack of applicability and fitness for use • Lack of publicly available high-impact datasets • Public data not responding to needs • Lack of real-time data	• Public agriculture-related datasets not much reused by private sector • Public data not responding to the needs of the farmer • Lack of agricultural real-time data (real-time data very relevant for agriculture; e.g. on product safety status through the supply chain, or pest alerts)

Source: Pesce, 2019, modified by the author.

Accessing private-sector data

The private sector holds a huge amount of data, most of which has very high public interest. In particular, data that can be of high value to farmers is collected or aggregated by private companies (e.g. reliable weather data, timely market data, precision agriculture aggregated data – soil, water, use of fertilisers and pesticides, plant and animal health, sales data, product tracking data etc.). In some cases, this data or more often related services can be purchased, but often at prices that are too high for smallholder farmers. In most cases, only services are sold, not the raw data or data at the level of aggregation needed by other actors to gain insights or build services. If governments had access to such data, they could offer better services to farmers.

A related issue is the little amount of business-to-business data sharing, especially between big data holders and small companies: such exchange would encourage innovation and lower the barrier for new market entrants, therefore increasing competition and offering more choice to farmers (avoiding monopoly and lock-in). At the moment, incentives are lacking for the private sector to share data. Some ideas that have been put forward to encourage data sharing are pre-competitive spaces, innovative business models, and leveraging social responsibility.

Accessing public data

Open data for farmers

In many countries, there are policies (41 as of September 2019 by Open Data Barometer (World Wide Web Foundation, 2020)) prescribing that public sector data – or, better, data that can be described as a public good – should be open and reusable. In most cases, the objective of these policies is to provide free useful data for the development of innovative services. However, due to difficulties in collecting/digitising the data and unclear criteria for the prioritisation of types of data, not all sectors are equally covered by the provision of open data: even in countries where agriculture represents a big percentage of GDP, agricultural data is only now starting to be explicitly mentioned in open data policies (FAO, 2018).

Types of data that are useful in agriculture and are traditionally prioritised in open data policies are: geospatial data, soil data and soil maps, cadaster data, sometimes weather data (although the private sector has a much bigger role here) and, more recently, price data. However, even this data is not always published or not always in a useful way. Besides, other data that would be especially useful for farmers is normally not published: agronomic data (e.g. crop growth data, pest and disease management data), value chain data, land productivity data etc).

Open data have an important role in mitigating data asymmetries. It enables access to data for the less resourced actors, like small farmers who can only get data from expensive providers. Open data also contributes to levelling the playing field for small companies that could reuse the data to offer competitive services to farmers and make farmers less dependent on big concentrated providers.

In order to achieve this important role in mitigating data asymmetries, public data should include data that is really needed by the industry, i.e. data that can make a big impact on the agricultural value chain. However, mitigating data asymmetries would entail an impact assessment exercise to determine which types of data are most needed by the agricultural sector, and by farmers in particular. However, so far, the public sector has not engaged much in impact assessments or consultations with society and industry to determine what data is needed and not enough 'high-value' datasets (datasets that would have a high impact) are published. This is why private sector data is still essential for the development of high-quality agricultural data services.

The major challenges why public open data does not make the desired impact are:

- Little usefulness and use of public data: Publicly available data may not be the most useful for farmers: e.g. governments may have price time series but not real-time price data which clearly has more value.
- Limited data usability: Public data does not always meet the needs of the various audiences that need it: e.g. data needed by farmers or their service providers may be available but not at a suitable level of granularity or standardisation.

An additional source of data and services that is becoming more and more relevant is farmers' associations: they can have the same role as public data in supporting farmers, especially in combination with farmer profiling, which allows them to tailor services to the needs of their members (see Section 1.2. Farmer profiling).

Data for subsidy schemes

At the intersection between farmers' data sharing and public open data is the bi-directional data flow around subsidised input schemes. This is the data in the public sector where farmers contribute directly. For example, farmers share data with the government about their expenses, their farming practices and management of natural resources etc. as evidence for subsidy payments. Some of the compliance data is aggregated in statistics for monitoring the results of the subsidy scheme and becomes public data.

Traditionally, collection of this data has been done via paper forms, or more recently with electronic forms manually filled in by farmers. However, some countries are starting to allow automatic subsidy payments based on digitally collected and submitted data for administrative simplification and lower transaction costs.

Regarding data sharing, the main challenge for farmers is to share the data for specific purposes without losing their competitiveness and without weakening their position in the value chain.

Regarding access to required valid data, farmers, their representatives and the technology/service providers have to understand the challenges of accessing public and private-sector data. Currently, public data and services as well as those provided by farmers' associations are often not sufficiently useful or reusable and cannot compete with private-sector data and services, which are often expensive and potentially biased. However, the private sector is not willing to share data either.

There is clearly a need for measures such as policies, platforms, and guidelines that build trust between value chain actors, especially small farmers, in order to mitigate data asymmetries.

2.3 Responsible data sharing in agricultural value chains

Responsible data sharing in agricultural value chains outlines the policy spaces and instruments to be considered when dealing with farmers' data sharing. The policy spaces that are relevant here are different from those relevant for the open data lifecycle. Data shared along the value chain is normally not open and not designed for public use but for mutual transactions for the provision of specific services, which raises a different range of issues, mostly contractual.

Some relevant legislation areas will be outlined briefly, but it must be noted that many important aspects of agricultural data sharing such as attribution, access, portability, interoperability, benefits, risk of lock-in are not covered by legislation. Examples of relevant legislation areas (tangentially) are digital policies in general, and personal data protection (PDP) laws in particular, open data policies, agricultural policies – especially with respect to the data sharing entailed by subsidy schemes, competition law, and unfair trade practices.

Beyond legislation, there are self-regulatory instruments that can help negotiate fair conditions for this type of data sharing: Section 2.3. Responsible data sharing in agricultural value chains illustrates existing examples of such instruments, such as codes of conduct and guidelines agreed upon by different value chain actors, and their potential role in making data sharing fairer. Some experts also envisage collaboration at the global level to create international guidelines for agricultural data sharing.

It is important to highlight governance components that form the 'data ecosystem' which supports smallholders. The data platforms in the ecosystem can be managed by different actors and with different purposes: in most cases, farm data is still managed on technology providers' platforms, but there are examples of farm data cooperatives and the potential role of Trust Centres with trusted governance, which are recognised as trusted organizations.

2.3.1 Legal and policy spaces for addressing the challenges of farm data sharing

Agricultural data sharing does not have a dedicated policy space but there are broader policy instruments to be used to ensure fairness of farm data sharing.

1. Digital strategies address important issues like data protection, data portability, data standards, access to data, public data and open data, movement of data across countries. These are now a common component of public policies, and related legislation.
2. Agricultural policies address issues related to extension services, interventions on agricultural value chains, subsidy compliance data, etc.
3. Competition law in general, and legislation on unfair trade practices addresses aspects like unfair data practices and power imbalances, although the specific data dimension is not yet addressed.
4. Codes of conduct or governance models like PPPs are self-regulatory instruments that address the more specific aspects of contractual fairness and fair data governance.

Although public policies do not address agricultural data sharing explicitly and do not offer solutions for most of the issues highlighted in the previous sections, it is useful to be aware of the existing policy spaces to understand where, in the future, these issues might be addressed and to be able to influence these policies and push, for instance, for a better coverage of a) the data dimension in agricultural policies and b) the value chain and data asymmetry dimension in digital strategies.

After a brief overview on policy spaces in general, readers will be introduced to codes of conduct and innovative data governance options as potential stakeholder-led and bottom-up solutions to more equitable data sharing in agriculture.

Table 5. Examples of policy spaces relevant for farm data sharing, linked to the issues highlighted in the previous section

Data asymmetry/conflict	Data flows/technologies	Risks for actors	Policy space
Within private sector			
Within value chain • Farmers vs. agricultural technology providers (ATPs) • Farmers vs. processors and distributors • Farmers vs. financial institutions	Farm data > Agricultural Technology Providers (ATP) *Precision agriculture* Farm data > supply chain *IoT in some cases* Farm data > banks *So far, submission, now initial use of satellite/ sensor data* Very little raw data towards the farmer, only in the form of services/information, often through the ATP	Farmers: • excessive data transparency, sensitive data more exposed than other actors' • lock-in • unfair contractual practices • no capacity to process the data • sharing more than receiving	**Data protection** (Data ownership/rights, business data) (IPR and copyright) **Agricultural policies** (Interventions on agricultural value chain, extension services, subsidies for digital technologies, possible use of blockchain data) **Public data policies** (More data for farmers to empower them in the value chain) **Contract law & Codes of conduct** (Data rights, data sharing guidelines, contractual power)
Within same segment of value chain • ATPs • Data processors • Any actor that aggregates data	Little flow, concentration *Cloud, big data*	New companies, small companies: unfair competition Farmers: lock-in, unfair contractual practices	**Competition law** (Avoid monopoly, facilitate market entry, unfair trade practices) **Public data policies** (More provision for levelling the playing field)
Private vs. public interest			
Farmers vs. government e.g. subsidies management Private sector data holders vs. government e.g. public-interest data, SDG monitoring, now also raw data for AI	Farm data > farm registries Farm data > ag policy monitoring *So far, submission, now initial use of satellite/ sensor data*	Farmers: excessive data transparency, sensitive data more exposed than other actors Private sector: risk of losing the competitive advantage of exclusive access to the data they have	**Data privacy policies** (Data ownership/rights, sensitive business data; use of public data for administrative simplification)

Data asymmetry/conflict	Data flows/technologies	Risks for actors	Policy space
Private vs. public interest			
	Private sector data à Open data systems *Very little so far, manual submission, APIs* *Rare transfer of raw data so far, now initial idea of sharing raw IoT data*		**Agricultural policies** (Consider sensitive data) Public data policies (Incentives for private sector to share, identification of pre-competitive spaces) **Public-private partnerships** (PP data platforms, data sharing agreements) **Codes of conduct** (Private sector data sharing agreements)
Public data for private sector			
Public data to level the playing field, foster fair competition and stimulate innovation	Public data à private sector applications Open data, big data	Government: high investment, lack of high-impact/real-time data	**Public data policies** (prioritise high-impact data, big data platforms, real-time data) PPPs (share investment, prioritise)
For all data asymmetry issues, given the impact on fundamental rights and the cross-border nature of data flows			**International Treaties** (idea of an IT on agricultural data flows)

Source: Pesce, 2019.

Digital strategies

Data protection laws and non-personal data rights

Data protection used to be addressed as part of privacy law, or under trade law in the case of consumer data protection. Nowadays, the core issues of data protection concern online, or digitally transmitted data and data protection is often addressed under a dedicated policy area in countries and regions where there is a digital strategy.

Data protection laws protect personal data. For an in-depth analysis of personal data in farm data sharing, see Section 2.4. Personal data protection.

There is no legal framework to allow non-personal data to be protected unless it is protected by copyright or trade secret; so, the rights illustrated in Section 2.2. Challenges for smallholders in data value including access, control, portability are not enforced. Non-personal sensitive data can be protected in contractual practice or in agreed codes of conduct, but it is not covered by legislation.

One important aspect of PDP that is rarely extended to other sensitive data but is very relevant for farm data is the right to portability. The right to retrieve and reuse the data, sometimes granted in contracts, is only a part of the full implementation of portability. In order to be reused, data should be standardised and interoperable. Conversely, machine-generated data is almost always in a format that is compatible only with the machinery and software sold by the technology provider. This is related to the lack of interoperability between machinery of different brands.

For both agricultural machine and data interoperability, more solutions come from international technical organizations (like the International Organization for Standardization (ISO), Agricultural Industry Electronics Foundation (AEF) or AgGateway) than from public policy. The so-called ISOBUS standard (ISO standard 11783) has become the de facto interoperability standard between tractors and equipment from different manufacturers, while AgGateway and the Open Ag Data Alliance (OADA) provide good practices and specifications for farm data standardisation.

Another aspect that digital policies often cover and is relevant for farm data are data localisation requirements. Considering the different data protection regimes in different countries, such policies prescribe that initial collection, processing, and storage of data (primarily about citizens) occurs first within national boundaries (Wikipedia, 2019). In some cases, data about citizens stays in the country or may be transferred only to countries that have the same level of data protection. This usually applies to personal data and other sensitive data, which is relevant for taxation or justice, but some countries have applied it more broadly.

Data protection is particularly important to consider because it affects cloud services and it is also considered as a protective measure. In some trade agreements, data localisation is considered in contrast with fair competition legislation: in the European Union, the recent "Regulation on a framework for the free flow of non-personal data in the EU" explicitly prohibits national governments from creating unjustified data localisation requirements (European Commission, 2018a).

Open data policies

In many countries there are policies that prescribe that public sector data should be open and reusable. While many of these policies are similar in approach and objectives, priorities and actual implementation vary in each country.

In order to assess the availability of free open data and therefore to be able to determine the feasibility of services which may need additional paid data, service providers, farmers and FOs need to be aware of the open data policies (if any) and data publication status in their country. It is also necessary to understand the licensing clauses and the interoperability standards used by an open data platform to determine the legal and technical feasibility of services.

Besides the obvious recommendation of consulting national laws, there are some international initiatives that try to keep track of the implementation of open data policies worldwide:

- The Open Government Partnership (OGP) is a formal partnership with specific eligibility criteria, which includes "*an access to information law that guarantees the public's right to information and access to government data is essential to the spirit and practice of open government.*"(Eligibility Criteria & OGP Values Check Assessment, 2020). If a country is an OGP member (in 2019, 79 countries were members), this gives good leverage to agricultural stakeholders, and FOs in particular, to advocate for the publication of data useful for agriculture.
- The Open Data Charter is an initiative that invites national and local governments to adopt a common charter of six principles: public data has to be open by default; timely and comprehensive; accessible and usable; comparable and interoperable; for improved governance and citizen engagement; and for inclusive development and innovation (Open Data Charter, 2015). The Charter has been adopted by 22 countries. If a country has adopted the charter, agricultural stakeholders can leverage this to advocate for the publication of data useful for agriculture, especially pushing for the principle for inclusive development and innovation.
- Country ranking can be looked up in one of the indexes created by different international initiatives that assess the level of openness of public data, like the Open Data Watch [opendatawatch.com], the Open Data Barometer (not actively maintained), the OECD Index of Open-Useful-Reusable Government Data (OURIndex) for OECD countries [**oecd.org/gov/digital-government/open-government-data.htm**]; or the European Data Portal "*European Open Data Maturity report*" (Cecconi and Radu, 2018) for European Union countries; or for the more general dimension of right to information, the Global Right to Information Rating (RTI) [**rti-rating.org**].

In some cases, open data sources do not necessarily have to be from the farmer's country. There may be sources from other countries that either are not geospatially sensitive (pest treatment, some general aspects of crop growth) or cover different countries and regions or have global coverage (e.g. many services rely on weather data from National Aeronautics and Space Administration (NASA)).

Open data policies are very important to enable equitable data sharing, however:

- Policies rarely address usefulness and usability of the data as reported in Section 2.2. Challenges for smallholders in data value, especially the prioritisation of high-impact datasets and the challenge of providing real-time dynamic data.
- Normally, open data policies do not address issues of data asymmetries and how public data can counterbalance data concentrations and contribute to levelling the playing field for new actors.

These issues have been very recently addressed by European Union policy-makers in policy briefs (see the 2018 EC Communication "*Towards a common European data space*" (European Commission, 2018)) and, also partly, in the new Public Sector Information Directive which focuses on reusability and impact of data and encourages the identification and prioritisation of high-value datasets and the publication of real-time data (European Commission, 2019). In general, it is not easy to find advanced open data policies that foresee public real-time dynamic data and prioritisation processes for high-impact datasets tailored to the needs of farmers. However, it appears that things are moving:

- In some countries, prioritisation 'models' and processes are becoming part of open data strategies and include demand from and consultation with stakeholders (two examples: public engagement and prioritisation methodologies in the Open Data Project of United States of America and the prioritisation model in the Open Data Strategy of Macedonia (Government of the Republic of Macedonia, 2018)).
- Regarding real-time data, some developed countries have started recommending the publication of real-time data, although this is often limited to transport data. In some less developed countries, for instance, Ghana, Ethiopia and Tunisia (Boyera, van Schalkwyk, and Grewal, 2017), there are new draft open data policies including plans to design and implement a data inventory that mandates periodic data update.

A significant proportion of the data that governments have already opened or are expected to open for the benefit of farmers, is quite static or changes over longer periods like soil maps, cadastre data. However, there are additional types of data that governments may be asked to collect and open that are very sensitive to timeliness and require periodic if not real-time updates such as granular weather forecasts, market data including price information at all stages and early pest warnings. These types of data are often covered by private sector services. However, governments could either start collecting this data as a public service or could explore ways to induce private companies to share it.

Private sector data sharing

A high amount of public interest data is held by the private sector. Data that have high value to farmers is collected or aggregated by private companies e.g. reliable weather data, market data, precision agriculture aggregated data on soil, water, use of fertilisers and pesticides.

Many governments are trying to negotiate the publication of private-sector data of public interest and to explore grounds in which the private sector might be willing to share data, both with other businesses and with the government for boosting innovation and public interest. The difficulty is to strike a balance between the privacy/business value of this data and its public interest or social responsibility value.

The paths that have been proposed so far are:

- Claiming public interest based on specific criteria: some examples include: (a) the European Statistical System which suggests providing a clear legal framework recognising "*a general principle of access to privately held data of public interest*" (European Statistical System, 2017); (b) claiming public interest based on the level of public or collective contribution to the value of certain private data assets (Mazzucato, 2018); (c) enforcing in public contracts the open access publication of all data generated with public money (following the now broadly adopted approach of open science, enforcing the publication of all publicly-funded research data as open data); (d) the enforcement of open data publication of data generated by companies that provide public services – as in the French "données d'intérêt général" (data of public interest) policy.
- Leveraging companies' sense of social responsibility by social certification schemes or leveraging 'data philanthropy'. Initiatives such as DataKind [**datakind.org**] and the Global Partnership for Sustainable Development Data [**data4sdgs.org**] champion the use of private sector data for social and humanitarian purposes, while Data Collaboratives proposes "*a new form of collaboration, beyond the public-private partnership model, in which participants from different sectors exchange their data to create public value*" (Data Collaboratives, 2020).
- Identifying pre-competitive spaces for sharing private sector data; for instance, companies sharing early-stage data, without much added value, to be combined with other datasets for new insights; or sharing data for improved value chain efficiencies.

- Public-private partnerships where benefits for private partners are identified and compensation for data may be considered.
- General 'crowd-sourcing' of public data from citizens and companies, like in France, where the government opens its national open data portal to allow anyone to publish open data sets.

The relevance of this trend towards private sector data sharing for agriculture, and for farmers in particular, is still very low because, so far, there are examples in other sectors (transport, telecommunications etc.), but not many in agriculture. Some private sector data shared publicly is already useful for agriculture (e.g. weather data, transport data) and a few agricultural input and technology companies have started publishing some datasets (e.g. Syngenta with the Good Growth Plan, which publishes datasets of productivity and soil data from the field). The potential would be very high if more companies started sharing data on precision agriculture-aggregated data on soil, water, use of fertilisers and pesticides, plant and animal health, etc.

Agricultural policies and other policy spaces

Agricultural policies

Agricultural policies do not cover the same issues in all countries. Some aspects that are often covered and are relevant for farm data include:

- *Subsidies* and the related data conferment for subsidy eligibility and compliance. In some countries, governments have started accepting data from new data-driven technologies for subsidy compliance. Two examples include: the European Union which is now allowing the use of Sentinel data for compliance evidence for payments under the Common Agricultural Policy (CAP) scheme (The Good Growth Plan Progress Data, 2020); and India, where the government and an input supplier are implementing a proof-of-concept application using blockchain technology for fertiliser subsidy management. An example of an interactive platform is the Smart Nkunganire System in Rwanda, where farmers provide data to the government to get subsidised inputs. Issues have been raised regarding privacy for this type of data conferment (Smart Nkunganire System, 2020). For instance, three German farmers that won a case against the European Union regarding the publication of data of farms that were not firms (BBC, 2010), and the European Union Common Agricultural Policy underwent an assessment of the European Data Protection Supervisor with a positive outcome.
- *Agricultural advisory services*, which can counterbalance the dependence of farmers on private services and can include the provision of public data services.
- *Agricultural information systems*, like market observatories, farm registries, plant variety databases, animal monitoring/tracking systems.
- *Rebalancing agri-food value chains*, strengthening the position of farmers and strengthening cooperation among farmers: this could be a space for unfair data practices in the value chain and for supporting data cooperatives.

This overview shows that some data aspects of agricultural policies overlap with aspects covered by digital policies. It is not straightforward to say which data issues are or should be addressed in agricultural policies or in digital strategies. It is interesting that the FAO e-agriculture project invites policy developers to promote national e-agriculture strategies "*as part of national ICT and/or agriculture strategies*" and "*to map the relevant existing policy environment that can be sometimes fragmented, within the agriculture and information sectors*" (FAO, 2018).

Agricultural public data services, market observatories, farm registries and crop, animal monitoring, tracking systems may be part of general digital strategies, but issues that are very specific to agriculture and to value chain dynamics may require a dedicated policy space.

In theory, agricultural policy could be a space where issues of farm data and data asymmetries in agricultural value chains are addressed, extending and specifying more general digital policy provisions on data rights or open data.

Intellectual property rights and copyright law

Intellectual property rights (IPR) is an overarching term for a wide variety of different legal instruments. IPRs protect the results of intellectual efforts or, express differently, products of the human mind. It is a broad concept as indicated in Table 6.

IPRs may vary in different national legislations, but there are international treaties with which signatory countries' legislation has to comply and which they must enforce. Examples include the Berne Convention and the World Trade Organization's Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement for patents.

Copyright and database rights are the most relevant property rights in relation to data: they apply mainly when there is either a clear creative effort in the creation of an artefact (copyright) or a clear compilation effort (database copyright).

So, in the context of agricultural data, these rights can apply to compiled datasets for which an intellectual and unique effort in the design or collection of data can be demonstrated. But they do not apply to raw data.

Table 6. Types of intellectual property rights

Intellectual Property	
Industrial property	**Copyright**
Patents	Literary work
Trademarks	Film
Industrial design	Music
Geographical indication	Artistic work
	Architectural design

Table 7. Summary of a number of areas that are relevant to data legislations

Type of law	What does it protect?	Differences between legislations	Applicable to data?
Patent law	Inventions	Most legislations protect inventions	No, but data may underlie patent applications
Copyright law	Creative, intellectual, artistic works	Generally, legislations protect copyrights	Yes
Database law	Effort to compile data collections	European Union legislations and Mexico; in some countries (e.g. India, South Africa) seen as part of copyright	Yes
Trademarks and 'trade dress'	Signs, names and expressions that identify marketable products or services	Generally, legislations protect trademarks	No, but there are concerns that such rights may be infringed, when reusing data from the private sector
Breeders' rights	Plant cultivars and animal breeds	In most legislations, breeders' rights are protected, but the way cultivars or breeds are registered varies	No, but data may underlie registrations

Competition law

Issues of monopoly and concentration of power in the same sector fall under legislation on fair competition and trade laws. Data concentration is not an infringement of competition rules, only its abuse is; for instance, using a dominant position for price discrimination, lock-in, denial of service etc. It is important to be aware of the competition law and, in particular, legislation on unfair trade practices that are applicable in the farmer's jurisdiction.

2.3.2 Focus on self-regulatory instruments: codes of conduct

Legislation does not address or solve many of the challenges described in Section 2.2. Challenges for smallholders in data value, in particular rights on non-personal data in data value chains.

While laws and regulations that govern personal data are becoming more common, legislation still does not cover data flows in many industries where different actors in the value chain need to share data and, at the same time, protect all involved from the risks of data sharing. Data in these value chains is currently governed through private data contracts or licensing agreements, which are normally very complex and that provide data producers with very little negotiating power.

Examples of the current common contractual practices on farm data include:

- Data ownership: there may be no clauses on data ownership, or clauses state that IoT-generated data belongs to the IoT producer. In other cases, clauses state that raw IoT data generated on the farm belongs to the farmer, while processed and aggregated data belongs to the technology provider.
- Data reuse: in most cases, either uses of farm data are not clarified and data is subject to unlimited reuse, or uses of farm data are clarified, but not negotiable; in some cases, need for consent from the farmer is required for reuse.

Self-regulatory instruments have started to emerge to set common standards for data sharing contracts. These instruments have taken slightly different shapes and names (codes of conduct, voluntary guidelines, principles): they will hereafter be just called 'codes' for ease of reference. Codes provide principles that signatories, subscribers and members agree to apply in their contracts.

In the agricultural sector, there are three codes that have been published recently and are known in the community of experts worldwide; in chronological order, these are:

- American Farm Bureau Federation's Privacy and Security Principles for Farm Data (2014). A set of seven principles around consent and disclosure in farm data sharing, providing companies that collect and analyse farm data (Agricultural Technology Providers, ATPs) with a few generic principles that should be applied in contracts.
- New Zealand Farm Data Code of Practice (2014). A set of six guidelines for data sharing in the New Zealand agriculture industry.
- European Union Code of Conduct on Agricultural Data Sharing by Contractual Agreement (2018). The European Union Code focuses on contractual agreements and provides guidance on the use of agricultural data, particularly data rights, access rights and re-use rights. Its aim is to create trust between partners.

The three codes have some common aspects: they all have a self-regulatory and voluntary nature; they are principle-based, they focus on the outcome of agriculture data practices rather than the exact process or actions by which this is to be achieved; they have been prepared by a combination of stakeholders (different combinations of farmers' associations, ATPs, machinery suppliers and other input suppliers); and they revolve around three core common points: consent, disclosure and transparency.

Table 8. Summary of the key points of the three codes (in red, the points that are specific to one code)

United States of America	New Zealand	European Union
Farmers are the **owners** of farm data and continue to be the owners of non-aggregated farm data down the line Responsibility of service providers to **inform farmers** that their data are being collected, and how they are used Collection and reuse require **consent** from farmer; do nothing without the consent of the farmer Right to **retrieve** own data for storage or use in other systems	Make **disclosures** to primary producers and other end users about the rights that the parties have **Disclose practices** and policies around: data rights, data processing and sharing, data storage and security Implement practices to ensure data is managed according to **agreed terms** and for **agreed purposes**, and accessible under appropriate terms and conditions	The data originator continues to be the **owner** of the data down the line and can determine who can access data and use it Originator's right to **know the purpose** of data collection and sharing Collection and reuse require **consent** and reuse is subject to **purpose limitation** Right to retrieve their data down the line Originators' right to benefit from their data (even financially) Aggregated data belongs to the aggregator

It is interesting to note that most of the rights attributed to the farmer in these codes are an extension of the rights attributed to the data subject by PDP laws. The most important points of these codes, which address some of the issues identified in the previous section, are:

- Data ownership assertions. Codes in the European Union and the United States of America consider the farmers as the 'owners' of information generated on their farms and, as such, the farmers are entitled to decide on data use and sharing with other stakeholders. The codes also recognise the 'data generating' role of the precision agriculture system, but still consider the farmer as the owner. A particularly interesting concept in the European Union code is that of the 'data originator': "*The data originator of all the data generated during the operation is the one who has created/collected this data either by technical means (e.g. agricultural machinery, electronic data processing programs), by themselves or who has commissioned data providers for this purpose*" (Copa Cogeca *et al.*, 2018). This definition avoids the complications of the ownership concept and also bypasses the issues related to the perception of the farmer as either the cultivator or the land owner: the originator is the person who collected the data or commissioned the data collection. The New Zealand code does not assert any ownership rights (if anything, given the fact that it is agribusinesses that have to disclose which rights are asserted on the data, they might assert their own ownership rights).
- Rights to access and control. For the European Union and USA codes, collection, access and use of farm data should be allowed only with the explicit consent of the farmer and the farmer maintains control of the data down the line, while the New Zealand code leaves it to the agribusiness to decide and disclose to primary producers what rights the organization asserts in relation to the data and what rights the primary producer has in relation to the data. In all three codes, control down the line also means that no reuse of the data is allowed for purposes other than those that had been originally agreed (purpose limitation).
- Transparency and choice. All three codes require that farmers be informed that their data is being collected, for what purposes and how it will be used, and that they be allowed to opt out of the agreement and halt the collection.
- Disclosure. All three codes prevent agribusinesses from disclosing non-aggregated farm data to third parties without the farmer's consent and without the same bounding legal conditions as the agribusiness has with the farmer.
- Retrieval and portability. All three codes require that farmers be able to retrieve their data for storage or use in other systems. As for standards and interoperability, the European Union and the New Zealand codes mention that data should be made available in a structured, frequently used and machine-readable format.

Certification

The codes in the United States of America and New Zealand foresee some form of certification:

- The USA code is associated with the Ag-Data Transparency Evaluator, a process to certify those ATPs whose contracts comply with the code and to award them with the Ag Data Transparent Seal of Approval.
- The New Zealand code provides a compliance checklist, which is then evaluated by a review panel: compliance is awarded by an annual licence and certificate, as well as the New Zealand Farm Farm Data Code trademark.

A data certification scheme can enhance trust because producers are assured that an independent and objective party has evaluated the provider's practices and deemed them worthy of certification.

The role of farmers' organizations

The organizations producing a code should carefully consider the balance of perspectives represented. In particular, farmers' associations should negotiate for the most vulnerable actors, those who risk the most from data sharing and might therefore be most reluctant to share. Endorsement and co-creation of codes by farmer-led associations can ensure that the farmers' perspective becomes central.

The existing codes, although co-written by farmers' associations, have the declared objective of gaining producers' trust for agribusinesses, so they seem to reflect the perspective of agribusinesses and the impression is that codes include what agribusinesses are ready to accept.

Regarding the target audience of these codes, it is important to note that the existing farm data codes do not have farmers or farmers' organizations as primary target audience – not to mention smallholder farmers – but rather the agribusinesses and agtech companies that work with farmers and use their data. So, while being prepared by bodies that also represent farmers (so far, big farmers' associations of developed countries) and indirectly raising farmers' awareness of their data rights, they are not written primarily for farmers. This is an important point for farmers' organizations: they have an important role in making farmers aware of the codes and, for instance, assessing contracts against the codes for their farmers.

Advantages of codes of conduct

Codes are not mature enough and their adoption is not broad enough to evaluate their success so far. One study on agricultural data codes of practice identifies some key positive aspects of codes (Sanderson, Wiseman and Poncini, 2018):

- They build trust.
- They fill normative gaps.
- They simplify the assessment of behaviours like other forms of accreditation when companies want to demonstrate compliance with social responsibility requirements. This is true especially if they are accompanied by some form of certification.
- They build awareness among technology providers as well as farmers.
- They foster participation and inclusiveness. Codes of conduct are normally co-developed by different organizations representing the concerned stakeholders; this in turn fosters trust and increases credibility.

International guidelines

Existing codes of conduct have regional or national coverage, which makes sense considering that they concern contractual practices and are quite sensitive to local contract laws. However, it has already been stated that they share many common points, which indicates that there may be a need for some general guidelines worldwide.

Considering the cross-border nature of agri-food systems, there have been repeated suggestions from policy studies (Schrijver, 2016; Maru *et al.*, 2018; BMEL, 2019) to coordinate guidelines at the international level, perhaps under the umbrella of the United Nations (UN) and more precisely, as suggested by Global Forum for Food and Agriculture, the Food and Agriculture Organization of the United Nations (FAO).

Such coordination could lead to international voluntary guidelines, or a set of standards, or an international agreement or treaty based on the model of the International Treaty on Plant Genetic Resources for Food and Agriculture.

2.3.3 Focus on governance options for a 'data ecosystem' for farm data

Agri-food data ecosystems are a combination of governance (from policies to laws, codes of conduct, community norms etc.), institutions and infrastructures dedicated to the management and flows of agri-food data, as well as the actors providing and using the data.

The data platforms in the ecosystem can be managed by different actors and with different purposes: in most cases, farm data is still managed on ATP platforms, but some new platforms have recently been launched for farm data to be shared independently. Many experts agree that the use of independent platforms should be encouraged. "*Farmers, consultants, advisers, and related companies need a data infrastructure that can collect, store, visualise, exchange, analyse and use large amounts of data, and they require a legal framework to deal with the ownership and the use of data outside of the farm premises*" (Kritikos, 2017).

The governance of such platforms is key to making them 'trusted' platforms. There are two types of governance suggested in different policy recommendations: public governance and stakeholder governance.

Public sector-led data platforms

With regards to the role that the public sector could have in the provision of trust-enabling platforms, such as blockchain-based platforms and e-infrastructures for data collaboration among farmers and other actors, there does not seem to be any explicit policy. There have been suggestions of independent, farmer-centric data repositories under public governance, which could be either general or organized by scope (commodity-specific, value-chain segment-specific etc.). There is no known example of public sector-led collaborative or interactive farm data platforms, although some of the stakeholder-led platforms listed in Chapter 3 Using data are supported, endorsed (like JoinData in the Netherlands) or partly funded (like AgBox in Canada or the Fiji Crop and Livestock Council) by governments.

Similar public sector approaches are: (a) government-led platforms for interactive conferment of subsidy-compliance data: an example is Smart Nkunganire System in Rwanda; (b) databases of farmers' profiles maintained by governments, which often include a lot of farm data. An example is Rwanda where they plan to put in place a national farmer digital profile platform; (c) government-led market/price observatories, where data are provided by producers' associations.

Stakeholder-led data platforms

Trust in non-public data platforms can be built if platforms are governed by a trusted organization of network members. Examples include data platforms governed by farmers' aggregations or consortia, including other value-chain actors, or any form of ' data cooperative' owned by their members (see a few examples below). The bodies governing these platforms should be recognised as 'trust organizations' that are entitled to verify, validate and authenticate data flows as well as assuring fair, inclusive and equitable data and information flows in agri-food systems (Maru *et al.*, 2018).

Governance models for these trusted platforms are based on negotiation, transparency, and innovative business models and would facilitate equitable flows of agri-food data. PPPs could also be considered for both the governance and the funding of such data platforms. There are already a few examples of stakeholder-led platforms:

- *In the Netherlands*, the Dutch JoinData platform allows agricultural actors to share data on the basis of clear agreements about access to and use of the data. JoinData is not a public initiative, it is a cooperative, but the government sees it as an example of a good type of agreement that can work for sharing private data.
- *In Jamaica*, the Slash-Roots Foundation is currently working on a project to take the data from the Farmer's Registry in Jamaica and turn it into a platform for transactions (Ferris and Rahman, 2016b).
- *In the United States of America*, a few years ago, the Iowa Farm Bureau had already proposed a farmer-controlled data warehouse. Recently, the Grower Information Services Cooperative (GiSC) and the Ag Data Coalition have created the AgXchange platform, a "*grower-owned and governed data cooperative*" whose vision is to provide cooperative members with an independent data platform, state-of-the-art tools for decision-making and a market for farm data (GiSC, 2020). GiSC also partnered with Farmobile, an "*independent farm data company*" that provides a collect–share–monetise strategy for farm data and a technology to read and harmonise all data (Farmobile, 2020).
- *In Canada*, AgBox, managed by a consortium of actors and partially funded by the government, is envisioned as a farmer-owned data cooperative, which gives farmers a confident and secure Canadian blockchain platform for the storage of on-farm data, featuring data connections to several precision farming data platforms (Agbox, 2020).

- In the European Union, the Declaration of cooperation on 'A smart and sustainable digital future for European agriculture and rural areas' encourages the creation of "*a European data space for smart agri-food applications*" and mentions the revision of the Public Sector Information (PSI) Directive. The 2017 study from the European Parliamentary Research Service recommends an European Union-wide independent, farmer-centric data repository (Kritikos, 2017).
- Some other stakeholder-led platforms are more focused on managing databases of farmers' profiles and not (perhaps yet) on letting them share farm data, but they are still good examples of how producers' associations can manage data platforms and can substitute or act as intermediaries with governments in managing farmers' registries. For instance, the Fiji Crop and Livestock Council, which is made up of commodity associations and supported by the government and the European Union, manages the farmers' registry for all commodity associations. Or in Colombia, the Colombian Coffee Growers Federation channels government subsidies to farmers and maintains a geospatial database with profiles of more than 520 000 coffee growers and their farms.

Some of the existing platforms are owned by farmers. Farmers' associations or co-operatives as trust organizations can have an essential role in shepherding farmers' data, negotiating access to other actors' data, and ensuring equitable data flows.

Public policy spaces, primarily digital strategies, are relevant under two aspects: (a) data sharing safeguards – nowadays most data protection laws, as well as laws on data localisation, are formulated under digital strategies; (b) access to data – open data policies play a very important role in providing data to less resourced actors, and some recent trends in the most advanced open data policies go in the direction of providing data with more impact potential: 'high-value' data based on industry demand and even data from the private sector.

In the area of digital policies, for issues of portability and interoperability of data across systems, standardisation organizations and industry collaboration have a stronger role than public policies.

There are some aspects of agricultural policies that can be relevant for data sharing, like data conferment for subsidy compliance, agricultural advisory services and agricultural information systems like market observatories, farm registries, plant variety databases, animal monitoring/tracking systems.

However, it was noted that certain key challenges regarding data ownership, data control and bargaining power – in general, trust issues among actors in value chains – have been better addressed by stakeholder-led initiatives, like codes of conduct or data platforms with a trusted governance.

Codes of conduct prepared in a participatory way by different stakeholders, including farmers' representatives, build trust through the provision of agreed guidelines on how digital agriculture contracts should address farmers' rights on farm data. Building trust is also the objective of platforms managed either directly by farmers or farmer-led associations or by third parties with transparent governance.

2.4 Personal data protection

Profiling farmers and capturing farm-level data is an essential step towards building services that are critical for smallholder farmers to increase their production and their income. As presented in Chapter 1 Data, services and applications, profiling activity can be conducted by different types of actors such as agribusinesses, farmers' groups, cooperatives, or ICT service providers. This activity is about collecting and storing data about farmers and farms that are, by their nature, classified as personal data. In many countries, the collection, storage and management of personal data is regulated by specific legislation at the national, regional and/or continental level. Even in countries where there is no regulation on this matter, the international trend shows that more countries are moving towards adopting such legislation, particularly with regard to electronic communications. Some examples have demonstrated that the implementation of PDP measures, beside its ethical dimension, is also a powerful way to develop trust between farmers and organizations collecting and managing farm-level data. It is therefore strongly recommended that anyone implementing a farmer profiling platform or farmers' registry implement best practices and common approaches to PDP, even when the country-specific law does not make those approaches mandatory.

Section 2.4. Personal data protection presents the core principles of PDP legislation, the obligations that organizations collecting personal data have to follow, and the best practices that should be implemented.

2.4.1 Definitions

A common description of personal data is all information that can be attributed to a living individual person. Data is also considered personal if it can be combined with other data to make attribution to living individuals possible. This information can take various formats such as an identification number (e.g. social security number) or one or more factors specific to his/her physical, physiological, mental, economic, cultural or social identity (e.g. name and first name, date of birth, biometric data, fingerprints, DNA etc.). Based on this definition, the process of collecting farmers' data, as soon as it includes elements such as name or phone number or address or GPS coordinates, falls into the category of processing of personal data. Figure 12 shows the types of data classified as personal.

PDP is commonly defined as laws designed to protect citizens' personal information. As of 2018, 120 countries around the world had data protection/privacy laws and 40 other countries had pending bills or initiatives (Banisar, 2019).

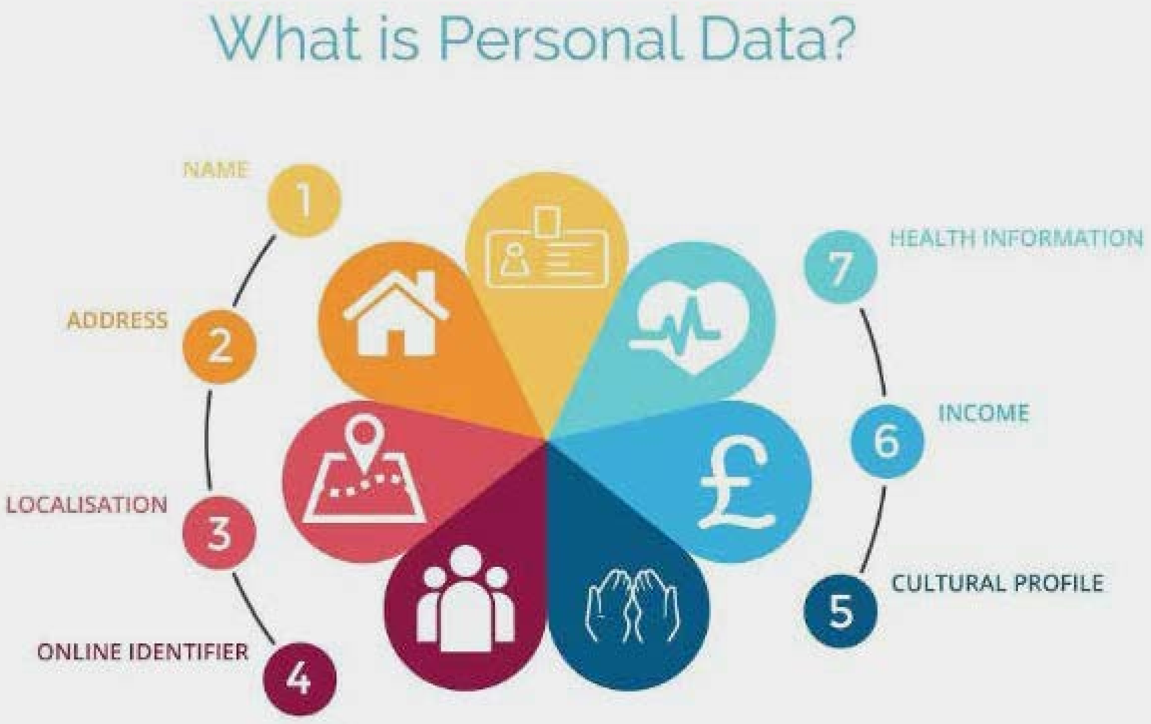
It is important to note two important elements:

1. Even in countries without specific national regulation, some regional, continental or international treaties ratified by the country may provide a legal framework for PDP. Some of the best-known treaties include:
   (a) Convention 108 "Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data" from the Council of Europe, ratified by 55 countries;

   (b) African Union "Convention on Cybersecurity and Personal Data Protection" (African Union, 2014), ratified by 14 countries;

   (c) CEDEAO additional act A/SA.1/01/10 on Personal Data protection within CEDEAO (Economic Community of West African States) (CEDEAO and ECOWAS, 2010).

Figure 12. Types of data classified as personal



Source: Rivenes, 2019.

2. Most regulations apply to both electronic and paper-based collection and management of personal information. For example, a farmers' organization may have kept track of its members' details on paper for decades and may not have realised that such a repository falls under PDP legislation that was enacted in the meantime; the farmers' organization is now infringing the law and risks penalties provisioned in that law.

2.4.2 Commonalities and differences between legislation

Each country that has adopted a form of PDP legislation has used its own template and wordings. However, across all legislation, there are numerous commonalities that appear in all laws, and some elements that appear in several laws when the legislation is more protective. The paragraphs below summarise these commonalities and additional elements.

It is important to note that the PDP landscape is currently evolving very rapidly all over the world. This evolution is triggered by two main factors:

- The European Union General Data Protection Regulation (GDPR): The European Union enacted a new regulation and the GDPR came into force on 25 May 2018. This new regulation is now serving as a reference for many countries and they are updating their older legislation to meet this new standard now perceived as the most protective regulation for citizens (European Commission, 2018b).
- The numerous cases of illegal exploitation of personal data collected by major companies (e.g. Facebook, Google) and used by international firms and governments for non-ethical activities. The best-known recent case is the Cambridge Analytica story (Wikipedia, 2020a) where personal data of millions of Facebook users were used for political advertising purposes. Similarly, but more farmer-related, examples include the use of farm-level data (e.g. availability of specific commodities) by intermediaries that are able to maximise their profit in business matching due to information on location and availability of goods.

These factors are now creating a momentum for all countries to adopt or update their PDP regulations.

Commonalities between legislations

The following elements are at the core of any PDP legislation.

- Scope of the law: PDP legislation always covers data collection and exploitation. The term 'processing personal data' usually used in legislation covers any activity related to collection, storage and use of personal data. For example, if a farmers' organization collects and stores farmer profiles, and shares this information with an ICT service provider, it must obviously comply with PDP legislation. But an ICT provider who has access to and use of personal data, even if it does not directly collect it, must also comply with the PDP legislation and, for example, must get individual consent for further sharing or use not covered by this consent.
- The need to make explicit the data to be collected: An organization collecting personal data must inform the person involved of the list of information that has been collected, explicitly or not, and stored. Some information may indeed be explicitly requested during an interview (e.g. name, commodity grown, etc.), but some information is not explicitly captured (e.g. address or GPS coordinates that could be filled automatically or by the collector without asking the person directly). The person has therefore to be informed on all the information collected.[13]
- The need to make explicit the purpose of the collection: The person must be informed about the purpose of the collection and the use that the organization collecting the information will make of it. From country to country there are differences in the level of details for usage. Some legislation requires a detailed description of usage, which does not allow any other application without a renewed consent. Under other legislation, the description of use can be a wide-open statement (e.g. the information will be used to design and build services for the surveyed farmers). It is obvious that the latter does not bring much confidence to the person; the current trend (e.g. with GDPR) and the recommended best practices is the former.

[13] It is important to note that the level of detail depends on each legislation. In some legislation, information is defined in broad terms such as household information, or production information. In some others, each piece of information must be explained.

- The need to make explicit the data sharing policy: The organization collecting the data must make explicit its sharing policy. From country to country, there are differences in the level of detail for the sharing policy. Sometimes legislation requires a detailed description of the list of third parties that will have access to the data, the part of the data they will have access to and their data usage. A change of usage, of data access level, or of the list of third parties will require renewed consent. Under other legislation, the description of the sharing policy can be a wide-open statement (e.g. the information may be shared with any third party that has the objective of designing and building services for the surveyed farmers). It is obvious that the latter policy does not bring much confidence to the person. There is a trade-off to ensure that individuals can provide their informed consent to ensure the maximum impact of the data collection, and to support the emergence of new innovative services, without requiring massive investment in renewing consent. As a best practice, it is recommended that different categories of actors are identified (public agencies, ICT service providers, financial institutions, etc.) and a specific sharing policy with a rationale is explained to the person. The GDPR is aligned with this approach and requires that each data use-case is individually presented and agreed by the person.
- The need to collect explicit informed consent: People providing their personal data must explicitly give their consent for usage and sharing. Some legislation, like GDPR, forbids global consent and requires individual validation in each case. Most legislation requires that the presentation of the data usage and sharing is done in understandable and intelligible terms. The organization doing the collection must ensure that the person gives informed consent.

"Profiling farmers and capturing farm-level data is an essential step towards building services that are critical for smallholder farmers to increase their production and their income."

©Adobe Stock/Jacob Lund

- The need to protect collected data: All legislation requires that an organization collecting and storing personal data put in place all measures to protect the personal data along the whole chain from the data collection point to the central repository and on to the third parties where the data is accessed.
- The need to offer a means of verification and update: All legislation requires that an organization collecting and storing personal data puts in place a means for people whose details are collected to query the organization; anyone whose details are stored has the right to know what those details are and has the right to update them.

Additional components in more protective legislation

Apart from the provisions listed in Section 2.3. Responsible data sharing in agricultural value chains, which appear in most PDP legislation, there are other elements that appear in a significant amount of legislation. In particular, it is worth mentioning the following:

- Official declaration: much legislation establishes an independent authority to control data processing by public and non-public organizations and to give fines. When such an authority is established by law, that law also usually makes mandatory the declaration of any personal data collection or processing. The authority provides official forms and processes to be followed for this declaration.
- Opt-out mechanism: much legislation requires an entity collecting data to offer a mechanism to people who were recorded to opt out of the repository, a posteriori after the data collection.
- Security breach report: when legislation establishes an independent authority, it is usually mandatory to report to this authority any security breach that may have led to unauthorised access to personal data. Some legislation even requires that the entity informs individuals whose details have been accessed.

2.4.3 Best practice for capturing and managing farmers' data

PDP legislation creates obligations and duties for organizations collecting and managing farmers' data. In countries where such legislation applies, they always include penalties and sanctions for entities infringing the law. Those penalties provide a strong incentive for organizations to comply with the law. In countries where there is no legislation, organizations profiling farmers are strongly encouraged to implement the measures presented below for two main reasons:

- The quick evolution of the PDP landscape across the world is likely to lead to all countries adopting such legislation in the next decade. While the implementation of PDP measures in the design of a farmer data collection project does not bring extra costs, the implementation of such measures at a later stage is far more costly. Indeed, not only does it require all platform components (data collection forms, central repository, etc.) to be updated, but it also means that a new complete data collection for all members in the repository is required.
- The implementation of PDP measures has significant benefits for the data collection task. Indeed, these days, farmers are reluctant to provide accurate farm data without understanding why these data are collected and with whom they will be shared (e.g. for tax risks, etc.). The trust relationship induced by the implementation of PDP measures is a critical success factor for such tasks. This video about a farmer profiling project in a tea factory in Uganda illustrates this point (CTA, 2018b).

The measures to comply with PDP regulations and to implement best practice spread over the different stages of the creation of any data collection and exploitation project:

1. Stage 1: design of the data collection process;
2. Stage 2: data collection;
3. Stage 3: exploitation of data collected.

They also involve/apply to different actors, as a data collection and exploitation project usually involves different actors:

- the organization responsible for the repository of information (e.g. a farmers' organization or a cooperative);
- the technical partner in charge of implementing the ICT elements[14] (e.g mobile data collection tools, central repository application);
- the data collectors who are in direct contact with people from whom the personal data are collected;
- third parties accessing the repository of information for reuse.

Sometimes one actor has two roles, e.g. the organization in charge of data collection has ICT capacities and does not need a technical partner. However, this set of actors is the most common in real-life projects.

Stage 1: Design of a data collection process

The preparation stage is usually the most important phase and is also the weightiest in terms of measures to implement. After the list of data to be collected is finalised, the following steps have to be implemented:

1. Official declaration: If the law requires it, the first action is to fill in the official declaration and submit it to the authority appointed by the law.

2. Memorandum of understanding (MoU) with technical partner(s): Before any activities are started, there is usually an MoU signed between the organization in charge of the data collection process and the technical partner. This MoU must have a few specific sections:

    - A section on data ownership: the MoU should explicitly give full ownership of the data to the organization in charge of the process. The technical partner should explicitly agree not to use the data it will have access to for its own commercial interest or to share it with third parties without the organization's consent. Data reuse by the technical partner(s) should follow the same rules and processes as all other data sharing agreement between the organization in charge of the farmers' data and third parties interested in accessing and using the data.
    - The technical partner should explicitly commit to raising awareness and training those of its staff assigned to the project on the sensitivity of personal information, and the need for complete confidentiality.

    Sometimes the technical partner is aware of the administrative procedures that are required in the country. The MoU may therefore assign the execution of the official declaration to the technical partner, on behalf of the organization in charge of the data.

3. Data collection and sharing agreement: One of the key steps in PDP is to inform individuals about the rationale for the data collection, the information to be collected, and the data sharing agreements with third parties. In order to address these three points, the most efficient approach is to design a data collection and sharing agreement that will integrate these elements which will then be presented to each person who has had personal data collected. The agreement must be written in simple, clear terms and not in legal jargon so that it is easily understandable for the person involved. For data sharing aspects, there is a trade-off to make. It is neither practical nor efficient to list all organizations that will have access to the information, because any new third-party agreement will require an update of the agreement and a new capture of the consent on the new version. However, it is critical to develop trust to explicitly identify the different categories of actors such as public agencies, extension service providers, financial institutions, etc. that are considered, and the authorised uses of information for each of these actors.

4. Data access and verification: The organization in charge has to implement a process for anyone to be able to access and update all the data about themselves and their business. The process should be clearly stated in the data collection and sharing agreement and should be easily accessible to people. One easy way to implement this process is the provision of a phone number to call to access data and provide updates.

[14] Note that most of the recommended measures in the sub-sections below are related to the case of a digital ICT-based repository where all information is centralised in a software platform. Some specific measures apply to specific cases when the data is collected using a mobile application or when the central repository is online.

©Adobe Stock/Sezer66

5. Opt-out procedure: As a best practice or required by the law, the organization in charge should implement a process for anyone to opt out from the repository. As above, an easy way to implement this process is the provision of a phone number to call to opt out.

6. Data collection form: The data collection form, paper or digital, should include specific questions:

   - an explicit capture of the fact that the person was given a detailed presentation of the data collecting and sharing agreement and understood it;

   - an explicit capture of the consent of the person to participate in the data collection;

   - it is also recommended to add a question capturing the fact that the person understood how to opt out from the process.

7. Training of organization staff: To ensure that they understand the importance of protecting people's privacy, the organization staff in charge of the repository should be trained about the sensitivity of personal data and the need for confidentiality. If the local PDP regulation defines penalties and sanctions, they should be included in the training.

8. Training of data collectors:

   - Data collectors should be trained on the sensitivity of the personal data and for confidentiality, similar to organization staff. In the case of paper-based collection, specific paper protection measures should be included. In the case of ICT-based collection, the training should include measures to protect the equipment, the need to notify the organization as soon as a breach is detected, and the importance of not sharing access logins.

   - Data collectors should be specifically trained on the presentation of the data collection and sharing agreement, which is the cornerstone of the PDP.

9. Protection of mobile equipment: For ICT-based data collection, the software and mobile equipment used must implement basic protection features. This includes:

   - *The use of login credentials to access data on the equipment:* Many data collection tools do not require the authentication of the data collector and the central repository to authenticate the mobile equipment. Such an approach should be banned. In such situations, if a data collector loses their mobile equipment, anyone finding it can access the personal data already collected that is available on the tablet and could even pollute the central repository by sending bogus data. Even if the login process is cumbersome, it is an essential security element. The login should also include an automatic logout after a time of inactivity.

   - *The ability to remotely erase the equipment:* These days, all modern equipment offers a mechanism for remotely erasing all the data they have after the equipment is lost or stolen. Operating systems like Android or iOS offer users an option to declare their equipment lost, and, at the next online connection, the equipment can be erased and blocked. However, such functionalities have to be installed and activated before it becomes available. The equipment has therefore to be prepared accordingly before being provided to data collectors.

10. Central repository requirements: The central repository where all data is stored should also implement a series of specific measures. These measures are only related to PDP, there are other requirements. They include:

    - *The implementation of different access levels:* it is critical to ensure that specific people access only the information they are supposed to. For example, a data collector should be able to access only the data about the people they survey. The central repository should therefore implement a multi-dimensional access mechanism to ensure that various categories of people can use the repository of information. But they can access only the information they are authorised to access and described as such in the data collection and sharing agreement. The access level should at least integrate the following dimensions:

  - Per category of information: the repository should allow access to only a subset of an individual's record. For example, access may be granted to access details about production, but not about individual details (name, phone, gender, etc.).

  - Per criteria on the individual record: the repository should grant access to a subset of all records based on specific criteria (geography, gender, commodity, etc.).

- *The monitoring of access and detection of security breaches:* The central repository should put in place monitoring processes to detect unusual activities and proactively detect any security breach.

Stage 2: Data collection

At data collection time, when all elements are in place, the most important tasks to conduct before starting data collection are:

- the presentation of the data collection and sharing agreement;
- the capture of explicit consent during the data collection process.

The data collection should then start only after those steps have been completed.

Stage 3: Exploitation of data collected

Finally, when the repository of information is populated, the organization in charge of the repository may grant access to third parties for them to exploit and reuse the data. At that stage, the most important element is the MoU with each third party that must include some specific paragraphs:

- The MoU should include the authorised usage of the information. The piece of information that will be used, the rationale and objectives. These elements have also to appear in the data collection and sharing agreement.

- The MoU should explicitly forbid the third party to share the information with any other parties, or to publish the personal information.

- The MoU should explicitly require the third party to raise awareness and train its staff accessing the information on the sensitivity of personal information, the need for complete confidentiality and the associated legal risks and penalties.

- The MoU should explicitly require the third party to report any security breach.

3

Using data

©Adobe Stock/Seventyfour

3.1 Using open data

As the web has evolved, so has its continued use for sharing ever more complex resources and data, but it challenges existing paradigms. The World Wide Web has been central to the development of the Information Age and is seen as an information space where documents and other web resources are accessed (Wikipedia, 2020b).

Seeing the web in this way has led to the development of a web of documents, or webpages as they are more commonly known, designed for humans to access and read; by June 2020 there were about 5.49 billion (World Wide Web Size, 2020). Many of the documents are linked together, and those connections (hyperlinks) add value. In a blog, newspaper article or academic paper, links can be used to build on previous discussions or point to factual sources, which helps users to explore the web of documents.

Section 3.1. Using open data illustrates the shift from a web of documents to a web of data and how data can be discovered on the web, from simple downloads from data portals, to searches on data aggregators, to web scraping and application programme interfaces (APIs) (see Application programming interfaces (APIs) in 3.1.7. Obtaining data from 'in the web' for more details).

This section, like the following sections in this chapter, is mainly a general technical introduction to the use of data on the web, with only a few explicit references to agricultural data; it provides the general background that is necessary to use any type of data from the web. The chapter is particularly relevant for those professionals who develop services for farmers and need to find and reuse data made available from various sources, for instance weather data or market data.

3.1.1 The generations of the web

In the early days of the web, there were far fewer documents, but people still needed to be able to discover things. The first efforts at manually maintaining an index were performed by Sir Tim Berners-Lee, the ODI's President and Co-Founder. People could go to the list and then jump to pages that looked interesting.

Portals were then created, such as DMOZ and the early Yahoo! (Wikipedia, 2020c). These were curated lists of websites and pages organized by particular topics. As the web scaled up, portals were no longer viable, and people moved to metadata search engines, such as Altavista (Wikipedia, 2020d) and Lycos (Wikipedia, 2020e), which used metadata that had been manually set in the webpage and provided information about the document.

Search was more scalable because pages were discovered automatically, but results were unreliable and easily manipulated.

The next generation of web discovery (ODI, 2016) came with PageRank-style (Wikipedia, 2020f) search, such as Google, which used many more cues for search, including an understanding of content, usage and linking. This third generation learnt how to look within the web of documents to discover how relevant each document would be for users.

It was vital that we learnt how to build these different types of search. The web of documents could not scale until search became the primary means of discovery (Jansen and Pooch, 2000). All these methods still exist and meet different needs, one of them being to fulfil the requirements of putting open data on the web.

3.1.2 The generations of open data on the web

The early open-data publishing techniques mirror the same generations as those of the web of documents. Portals were first created, such as **data.gov.uk** and **data.sncf.com**, which are curated catalogues of datasets organized by particular topics or organizational structures.

As the amount of open data began to grow, data aggregators began to appear that provided services to ease the discovery of data related to particular topics or regions. Examples include enigma.io, European data portal and **transportapi.com**. Such services rely on the availability of metadata from other portals, websites and services to provide information about and access to the data.

The third generation of search engines for data is still very focused on the web being an information space, thus we have to rely on the same search engines to find data as well as information. Development of search engines for data is still in its infancy and this is greatly related to the methods being used to publish data on (or in) the web.

3.1.3 Data is just another resource on the web

As the web evolved, it started to become a place to share multimedia resources. The inclusion of images, audio and video unlocked new potential to deliver services such as streaming services. Audio was the pioneer here, where early sites like last.fm used metadata about music tracks to build customised recommendations for people. This technology precedes and forms the basis of many of the recommendation systems in use today. Last.fm provided recommendations but did not allow people to listen to the music itself. This functionality did not emerge until three years later with the launch in 2005 of Pandora, a personalised radio station application. Another three years on, the launch of Spotify, which was the first streaming service that used web technologies to deliver a dedicated audio platform outside the web browser.

Wind forward and the web and internet are now the delivery platform for huge amounts of different resources. Sticking inside the web browser, search engines like Google have added multimedia-specific search capacity to find images and videos, while specific portals like YouTube now provide web-based access through their website, as well as via applications and connected TVs. This is not quite true of data. Search engines still do not have specific searches for data, perhaps they never will. But finding data on the web can be a challenge, one that starts with the definition of data itself.

What is data?

An image is a visual representation of something (a picture). An audio file makes a sound when played. A video combines multiple images with audio to make a moving picture. Data is difficult to define and thus search for.

Data is the lowest level of abstraction from which information and knowledge are derived. These are abstract terms and thus data could be an image, or spreadsheet, or audio file. In addition, if the web is an information space then is data something which is of a lower level?

Traditionally, data is thought of as a spreadsheet or set of numbers that can be analysed in some way. On the web, such data is often shared via data portals simply as a file that can be downloaded. Some portals provide YouTube-like functionality where the data can be explored without downloading; however, the data itself is still a static resource, uploaded ready to be downloaded by someone else.

While data remains a static resource, second- and third-generation web services are perfectly suited to harvest metadata about these static resources and provide entries in search results. Static resources can be linked to directly, and thus algorithms like PageRank remain relevant. This approach suits the web of documents approach, where the metadata is still the key way in which the data can be found, which means that existing search engines can be used to find data with the correct query.

However not all data is static. When visiting a shopping website, travel website or weather website, the content is likely to be different each time, depending on the raw data. On a shopping website, certain products may vary in price and stock level; on a travel website, options will vary in availability and price depending on the search criteria; while the weather changes constantly.

The data that powers these sites is vast and hidden in the web. Machine-readable data is creating a new web of data which experts claim has the potential to unlock a data age. Perhaps the web will then transform from an information space into a data and information space.

Applications that use such data are already prevalent from travel planners, weather and shopping apps, to games. Such applications exchange data to help them function; however, this data is often hidden, making it difficult for others to access and use.

The rest of Section 3.1. Using open data explores the two approaches of data on the web of documents and how to search it, before exploring the web of data and how to begin to unlock its potential.

3.1.4 Government portals

The history of open data is closely connected with laws that govern access to public information. Such laws ensure that the public has a right to access information from those providing public services. The open data movement attempts a complete reversal of this logic. Rather than having to request data, that data should already be 'open by default'; to close a dataset should require a good reason rather than the opposite. Governments and public service providers should proactively work in the open.

Going a step further, governments in many countries signed up to the Open Government Partnership (OGP). OGP was launched in 2011 to provide an international platform for domestic reformers committed to making their governments more open, accountable, and responsive to citizens. A key part of OGP is the commitment to being 'open by default' and to open data. This led to the launch of many government open data portals that were built to hold government data and make it easily accessible for the public. Over the years, open data activities have evolved to the extent where governments are now scored on how well their activity is going and how sustainable it is.

The Open Data Barometer (ODB) of the World Wide Web Foundation scores governments on three aspects: readiness, implementation and impact. The implementation score is measured by looking for key open datasets to be present, accessible and up to date. In the 2016 ODB, the implementation score looks for the availability of fifteen types of data.

1. Map data
2. Land ownership data
3. Detailed census data
4. Detailed government budget
5. Detailed government spend
6. Companies register
7. Legislation
8. Public transport timetables
9. International trade data
10. Health sector performance
11. Primary and/or secondary education performance data
12. Crime statistics
13. National environment statistics
14. National election results
15. Public contracts

Datasets such as election results, government spend, and education performance are records of historical significance. Such datasets are very static in nature and can be made easily available in spreadsheet form to download via a portal. At the same time, it is these records that are linked more with access to information laws and have less economic potential for wide reuse.

Conversely datasets, such as map, companies and trade data, are much more dynamic, and as such are more suited to a different level of service from a simple file download. This is especially true as map data takes the form of many complex formats, and trade data can exceed sensible file sizes for access via download.

Agriculture is an area which cross cuts many of these areas and, as such, many different datasets exist in portals. For example, mapping data is critical to inform agriculture beyond land ownership: aspects such as water catchments and runoff, land use as well as protected/restricted areas can all help inform understanding.

Many governments now run a data.gov.XX portal (or variants of in the local language, e.g. datos.gov.mx) that provide a central catalogue for government data. Many portals have dedicated agriculture sections, often closely tied in with the relevant government department, that offer large amounts of links and access to downloadable data.

©Adobe Stock/JackF

As mentioned previously, finding data in these portals relies heavily on metadata search. As such, the title and description of each dataset is critical to enable discovery. However, as the portal will often be organized by department or activity, finding a dataset requires either specialist domain knowledge and/or knowledge of how the local government organizes and describes its data in the portal. A good starting point is for a user to browse the portal to gain clues on how it is organized and discover how data is described, before targeting their search using this new knowledge.

3.1.5 Obtaining data from 'on the web'

As mentioned previously, much of the available open data out there is only available 'on the web', either via a download button or contained within the web pages themselves. Section 3.1. Using open data looks at the techniques to start discovering and unlocking this data ready to be used.

Finding downloadable data files

Many suppliers are helpful and provide human-readable download links in order to obtain their data. Most of the datasets on government data portals work in this way. The examples below can be explored to provide further insight as to how they work:

- United Kingdom National Statistics – Latest cereal stock statistics (Department for Environment, Food & Rural Affairs, 2012)
- Tanzania Agricultural Census (Agriculture Department, 2020)

Many search engines, including Google, provide the ability to use advanced searches and prefixes in order to dig out data from sources, such as the ones above. Advanced searches make use of filters and prefixes in order to limit the type of search and results. A list of prefixes and linked examples can be seen in Table 9.

Each of these can help refine a user's search for data. While the top two help narrow the search, the bottom two broaden the search again once a relevant resource has been found.

The 'related' search can help find other relevant content, related to a known page. Sites that link to a specific dataset might have used the data, which can help provide context over the existing usage. Given that the majority of openly licensed dataset require attribution, the 'link' specific search should return (at least) a number of results if the dataset has been used.

Data aggregators

One of the main challenges facing data 'on the web' is the lack of ability to search within the data itself. Existing search engines only enable metadata search, regardless of the format of the file (the same is true for audio and video, but Google does allow a user to search using an existing image).

Many data portals act as aggregators and allow some exploration of data, either from a single data provider or a set. In relation to agriculture, the World Bank aggregates key statistical data from many countries and allows the exploration and download of this data. One example from the World Bank is the Agriculture and Rural Development indicators (The World Bank, 2020). Aggregating the data together allows the exploration of indicators such as agricultural land area versus rural population. This example looks at the comparison between a number of countries in East and West Africa. The DataBank service from the World Bank provides easy access to explore and then download the data ready for use (The World Bank, 2020). Such aggregators can be a crucial source of open data when country portals are either out of date or simply not available.

Table 9. Prefixes for advanced search

Prefix	Description	Example search
filetype:	Search for specific file types only	filetype: xls cereal stocks
site:	Search with a specific domain or site only	site: nbs.go.tz agriculture
related:	Search for content related to a known page	related: **www.gov.uk/government/statistics/cereal-stocks**
link:	List only pages that link to the one given	link: **www.gov.uk/government/statistics/cereal-stocks**

Table 10. Common formats for data 'in the web'

Extension	Description
.csv	*Comma Separated Values.* Tabular data format like excel but stripped back to just contain data in a simple structure.
.json	*JavaScript Object Notation.* A hierarchical data format native to the JavaScript language which is used widely on the web as it forms part of the HTML5 specification.
.xml	*eXtensible Markup Language.* A markup specification that has a wide range of uses. Has been criticised for its complexity and verbosity in comparison to JSON.
.rdf	Although RDF should not be a data format (not covered here). RDF defines a formal data structure, which can be applied in xml, json and csv formats. Use of the extension implies that the structure is used and most commonly the data itself is in XML format.
.rss	Another specific XML structure that is often used for data feeds that regularly update, such as news and weather.

Enigma.io, winner of Techcrunch Disrupt in 2013 (Dickie, 2013), brings together data from a multitude of open data sources and enables fine-grained search within the data itself. This is, in effect, a reverse search on data: rather than searching the metadata to find the data, enigma.io searches the data and shows datasets where the search term can be found.

Scrapers

Sometimes data will not be available for download in a usable format. Sometimes, the data will be available only from within the webpage as a table or a list. In other cases, the data may be available in a document format (such as PDF) rather than a data format. In both cases, the use of data scrapers can help extract this visible data.

Web data scrapers

Web scrapers allow the automatic extraction of structured data from a web page. Tools like grepsr [**grepsr.com**] allow the automatic extraction of data from structured websites in seconds, including the ability to handle pagination and infinite scrolling results. Currently, such tools usually involve a per-record cost for extraction with a limited number of free credits per month.

PDF data scrapers

Another place where useful data is often embedded is within PDF reports produced by statistics agencies. These will often contain long appendices of tabular data which can be extracted using tools like **PDFTables.com**. Try it for with some agricultural statistics from Tanzania (Agriculture Department, 2020). Check where the data starts in a PDF report. It is advisable to reduce the PDF to the exact pages that contain the data required before uploading to a PDF extractor like **PDFTables.com**.

3.1.6 Obtaining data from 'in the web'

The evolution of the web has led to the requirement to separate back-end infrastructure and data from the presentation layer, such as websites and mobile applications that all use this same data. Shopping, weather and travel applications all offer various options for users to interact with essentially the same data.

These applications all use dedicated data services to access and query the data. Many of these data services are documented for anyone to use, while many remain hidden for various commercial or budgetary reasons. Section 3.1. Using open data looks at the different techniques that can be tried to access data that is 'in the web', which helps service such applications.

Filetype extensions

Some websites have been built to offer a way to extract data by adding a file extension to the URL of the web page being viewed. For such websites, usually maintained by organizations who also publish downloadable open data, adding the correct extension will trigger a download of that page in a data format, as opposed to a document format.

A good example of this is the United Kingdom Government website [**www.gov.uk**], which provides any page in a data format simply by adding the relevant extension like '.json', for example **www.gov.uk/browse/business.json**. To view the data in a more human-readable form, copy it into **jsonlint.com**.

The United Kingdom Trade Tariff also has the same functionality and contains details on the international trade codes that can be linked to the trade data available from Revenue and Customs (**Build Your Own Tables, 2019**).

Unfortunately, not many websites make it clearly display that alternative formats (such as JSON) are available. A good indicator is to find modern-looking websites where pages clearly contain data, such as records about individual companies, where such extensions can be tried. Table 10 lists common data formats available for data 'in the web'.

Application programming interfaces (APIs)

APIs are one of the best ways to access data. APIs are a service best described as a 'promise' by one system to constantly and consistently provide a service to another that allows the two to interact. For this reason, APIs have many advantages over any other form of data access, as listed below.

1. Service agreements. As an API is a service, this guarantees access to data and can often be accompanied with service level agreements for those who wish to use them.
2. Live access. APIs provide a mechanism whereby data can be included live within an application. The most common example of a data API is live transport times. On the back of a single API, many hundreds of applications can be created.
3. Designed for data. Perhaps the biggest advantage of an API is how they are designed for data and machines rather than for humans. This means that data availability is no longer constrained by the paradigms of how humans use the web; however, this does create challenges when searching data that might be within an API.

The major disadvantage of APIs is that data is not as easily accessible to download and use straight away. Some third-party applications, such as enigma.io already use APIs to access data from other services to allow easy access, while others like OpenCorporates allow downloads by file extension as part of their API.

Examples of services that have APIs include: OpenCorporates, OpenStreetMap, Twitter, Flickr, and LinkedIn, which provide direct access to the raw data, as well as broad queries to allow faceted search.

Many of the open data platforms provide APIs to access the data including Socrata and OpenDataSoft. Such platforms are used by a number of governments and departments; Socrata is mainly in the US, while OpenDataSoft is throughout Europe. Comprehensive Knowledge Archive Network (CKAN), an open source alternative, also has an API although this API only gives access to the metadata records in many instances.

Table 11 contains some examples of each platform and some of the available agricultural datasets, some of which were mentioned earlier.

Using APIs

Many web APIs take the form of REpresentational State Transfer (REST) APIs, which is an API designed specifically for the web. It has a specific set of guidelines and rules that control if something is a RESTful API.

Broadly speaking, a REST API requires the use of resource identifiers which are then interacted with to upload/download the required resource. In the case of the Socrata example, the API location is the web-based identifier of the resource.

A REST API specifies that a machine should be able to change the request in order to ask for different representations of the same resource. This is a bit like adding file extensions, except where the requested resource does not change any part of its location on the web (as adding '.csv' effectively changes the URL).

Table 11. Examples of open data platforms and agricultural datasets available there

CKAN	Web page: https://data.gov.uk/dataset/cereal_stocks_england_and_wales API: https://data.gov.uk/api/3/action/package_show?id=cereal_stocks_england_and_wales
Socrata	Web Page: https://agtransport.usda.gov API: https://dev.socrata.com/foundry/agtransport.usda.gov/sruw-w49i
OpenDataSoft	*No agricultural examples found.*

REST APIs are simply extensions of the web's existing HyperText Transfer Protocol (HTTP) except used for data. Thus, it is possible to change the type of request from a 'GET' to a 'PUT' and then send structured data to the server to replace the existing data with new data (using authentication, obviously). The City of Chicago uses the POST method to send updated crime statistics to their data portal and have been doing so daily since 2001 (Chicago Data Portal, 2020).

APIs not only allow users access to data; they also form a key part of the provider's data infrastructure allowing data to be managed and kept up to date.

Hidden APIs

Not all websites that dynamically load data make their API known publicly, even if one exists. However, it is possible to discover them. Doing so requires a fair amount of technical knowledge; however, a good Google search often turns up communities of people who may have already built something for the particular service needed.

As many APIs are based upon the REST API design, in many cases, it is fairly straightforward for someone familiar with REST APIs to quickly find if a service has one and how it works. This can be done by trialling out some REST requests with browser extensions like Postman for Google Chrome.

The ODI's experimental Hidden Data Extractor tool has been built to automatically look for REST APIs that exchange JSON data when a web page is loaded.

3.1.7 Checking rights to use data

There are many ways to obtain data from the web, be it clearly visible via a download button or available through a public or hidden API. Regardless of the method of data acquisition, it is critical to check the rights both to use that method and to use the subsequent data.

Just like the data itself, some rights statements will be only human-readable, some only machine-readable and some a combination of both. Commonly, however, service providers will have a human-readable version of their terms of use and/or data licence that will cover both the terms of use of the service and rights to use data once acquired.

Many government data portals will have the data licence listed as a piece of metadata against the record being viewed. For example, in **data.gov.uk** all licences are listed directly under the title of the dataset as a clickable link. The CKAN platform (which data.gov.uk is a version of) is particularly good at exposing rights statements, which helps users ensure the data they are viewing is open data.

Services like Flickr also have licences against each photo. Each Flickr user is able to specify licenses for their own photos. Flickr even provides a search that allows others to find photos with specific licenses.

If licences are also machine-readable (as is the case with CKAN and Flickr), then search engines can use this as a piece of metadata, meaning that search results can be instantly filtered to contain only openly licensed content.

If using a REST API, the rights statement might be returned as part of a Link header (ODI, 2013b), which separates the rights statement from the content, allowing the response to still be the pure data, e.g. CSV file.

If none of these options exist, then it might be necessary to read the terms and conditions of the providers to ensure that method of access, and rights to use the data, are permitted as just because something is accessible on the web does not give everyone the right to use it.

It is still early in the evolution of a 'data age' following the 'information age', and services that specialise in providing fast access to data are evolving.

At the same time, the number of services providing data is also growing, mirroring the early days of the web. There are still lessons to be learnt; however, methods to access data are beginning to stabilise with the emergence of common APIs such as REST.

Data formats have also evolved and thus so too have methods to discover and access data. Search engines are becoming much more intelligent and can be customised to perform highly targeted queries. At the same time, tools to help extract and work with data have evolved such that it is very easy to start working with data regardless of the format.

The evolution of mobile applications that demand instant access to data has also increased the number of available APIs, even if some of them remain hidden.

It is clear that we live in the age of data, however we need to be careful over our rights to use such data; having clear open data licenses is critical to the future of our data infrastructure.

3.2 Quality and provenance

Quality and provenance are two important aspects that determine the usability of a dataset. Section 3.2. Quality and provenance take a broad look at the different aspects that make a quality dataset and a number of best practice guidelines that aid the publication of high-quality usable data.

Part of the quality of a dataset is dictated by the history, or provenance, of that dataset. Knowing that the data is from a reliable source and that it was collected using reliable methods (or via methods with known constraints) can often be more important than having a dataset with a well-controlled vocabulary or schema.

While data is technical in nature, not all quality measures are technical; ensuring a dataset is machine readable does not always mean it is usable; not all quality requirements are technical. The evolution of best practice guidelines reflects this. In Section 3.2. Quality and provenance, we look at the technical and non-technical aspects that make a high-quality usable dataset.

3.2.1 Quality marks and measures

Assessing the quality of open data cannot be done quickly. There are a number of community-based standards and quality marks that can help to assess the quality and usability of data.

One of the first quality marks to emerge for open data is the '5-stars of linked open data' (Berners-Lee, 2012). Awarding of each star is sequential and starts with the requirement to apply an open license to data. The remaining stars are split into two that focus on open data being available 'on the web' to download and two that focus on data being 'in the web' to use via an API, which can instantly retrieve resources. Other than the aspect regarding licensing, the 5-star guideline is focused on the technical availability of data and lacks the non-technical aspects that make up quality usable data.

Similarly, the FAIR principles provide a similar guideline for publishers. According to these principles data should be: Findable, Accessible, Interoperable and Reusable (FAIR) (FORCE11, 2014). While these largely focus again on technical aspects, the principles do, however, set out some terms relating to provenance, stating that 'Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation'.

Both the 5-star schema and the FAIR principles are outlined in more detail in Chapter 4 Exposing data. Section 3.2. Quality and provenance is going to take a look at the other aspects of usability and quality as defined by the ODI's Open Data Certificates.

3.2.2 Open Data Certificates

The Open Data Certificate is a free online tool developed and maintained by ODI to assess and recognise the sustainable publication of quality open data. It addresses the legal, practical, technical and social aspects of publishing open data using best practice guidance.

Like FAIR, the Open Data Certificates (The Open Data Institute (ODI), 2020) process takes an alternative but complementary view to the 5-star scheme. A certificate measures how effectively someone is sharing a dataset for ease of reuse. The scope covers more than just technical issues including rights and licensing, documentation, and guarantees about availability. A certificate therefore offers a more rounded assessment of the quality of publication of a dataset.

For data publishers, the process of assessing a dataset provides insight into how they might improve their publishing process. The assessment process is therefore valuable in itself, but the certificate that is produced is also of value to reusers.

Being a data reuser

For data users, the technical quality at the point of use might be enough for their particular use. However, for reusers sustainability and support are likely to be more dominant in the decision-making process than file format.

Reusers need to be offered assurance that their access to the data will be consistent and reliable. An open data certificate challenges publishers to think beyond the data to address key policy considerations in the support of the data. These considerations are broken down into four categories crucial for reusers.

A reuser's checklist

Using the open data certificates as a guide, the following presents a reuser's checklist for open data. The checklist is split into four categories reflecting the sections of the Open Data Certificate.

©Adobe Stock/MIA Studio

Legal

- Is the data openly licensed and legally usable?
- Is the data model, format or structure also openly licensed and legally usable?
- Are copyright statements clear?
- Are any data or parts of the data that are not openly licensed described?
- Are any privacy and potential ethical constraints to the data use outlined?

Practical

- Is the data well described?
- Is the reason the data is collected clear?
- Is the publisher's use for the data clear?
- Are any other existing uses of the data outlined?
- Is the data accessible?
- Is the data timestamped or up to date?
- Will the data be available for at least a year?
- Will the data be updated regularly?
- Is there a quality control process?

Technical

- Is the data available in a format appropriate for the content?
- Is the data available from a consistent location?
- Is the data well-structured and machine readable?
- Are complex terms and acronyms in the data defined?
- Does the data use a schema or data standard?
- Is there an API available for accessing the data?

Social

- Is there an existing community of users of the data?
- Is the data already relied upon by large numbers of people?
- Is the data officially supported?
- Are service level agreements available for the data?
- Is it clear who maintains and can be contacted about the data?

If a publisher has completed an open data certificate and applied the quality mark to their data, then a reuser can quickly find out the answers to all the questions on the reuser's checklist above. Alternatively, a complete list of all certificates is available in the certified datasets registry (The Open Data Institute, 2013c).

A reuser's guide to provenance

All of the best practice guidelines introduced so far are somewhat focused on data producers, who already own and manage the source data. However, not all open data originates with the publisher; a vast amount of data is derived from other data. For example, a weather forecast is derived by applying complex models to meteorological data. With there being many sources of meteorological data and many organizations that use different models to create a forecast, this can lead to situations where even forecasts based upon the same input data can be vastly different, with potentially devastating consequences (BBC, 2013).

Provenance checklist

The checklist below will help to establish the provenance of a dataset and help establish the level of trust in that dataset.

- Is the data wholly owned and produced by the data provider?
- Does anyone else produce comparable data for cross checking?
- Is it clear if the data has been derived from other sources of data?
- Are the other sources of data clear?
- Are the other sources of data trustworthy and comparable with other data providers?
- Is it clear if and how any data has changed (from any source) prior to being made available as open data at point of access?

Following these points will help establish how trustworthy different open data sources are. It may even reveal potential to bypass the current data source and follow the trail of provenance back to an original source that may be more trustworthy or offer a more completed and/or supported data service.

3.2.3 Post-access quality checking

Establishing trust in a data provider, potentially obtaining a service agreement and then accessing the data is not the end of the quality checking process. Once access to the data has been obtained, it is essential to verify the data.

This stage of checking and preparing data to be ready for use has many aspects, most of them based on technical processes. It is more than likely that problems or inconsistencies will be found in the data at this stage. This is when it is crucial to be connected to the community to provide and/or have access to quality control procedures to help with the understanding of the data or to fix these problems.

Data validation

One of the first things to do with the data is to verify it against any available data schema and description of the data structure. This will reveal potential omissions from the documentation for terms which have emerged in the data. Verifying the data against the schema will also help verify that the correct data exists within the dataset.

If data does not have a schema, it might be necessary to transform the data into a format where one can be easily applied, or to make a schema for the dataset. A schema is a blueprint for data that defines a set of integrity constraints and rules relating to the structure and contents of a data resource.

Designing schemas

Schemas play a key role in enabling the wide, automated reuse of data. A schema defines three key things:

Column/key titles in the data: Defining a consistent set of column titles (or keys) for a dataset is essential to ensure that datasets of the same type can be merged and analysed easily. Often column titles will change or be abbreviated to save time; however, this causes a lot of problems when analysing data over long periods of time. Adding column titles is less of a problem but has to be taken into account when analysing data.

Value types: With the column title/keys defined, it is important to define the valid data type for the values, e.g. number, text, date, coordinate etc.

Value constraints: With the value type defined, valid constraints - such as being required, needing to be unique, being in a certain unit (e.g. gallons [United Kingdom) or within a certain range - should be defined. For example, a column might be entitled 'Cost (GBP million)'; thus, any values should be numbers (without commas). Setting a valid range also helps avoid and explain any errors in the data: for example, setting a range of 0.001–100 on the 'Cost (GBP million)' (if it is known that cost cannot exceed GBP 100 million (USD 124 million). Range validation stops people accidentally misreading the column title/units and entering 100 000 000 instead of 100 for GBP 100 million (USD 124 million).

Using schemas

All spreadsheet packages allow validation rules to be created for data relating to each column title. However, very few packages allow the exporting of schemas alongside the data for others to use without adding complex developer extensions.

One of the main reasons for this is the connection between schemas and hierarchical databases dating back to earlier database design by E F Codd (1970).

©Adobe Stock/Jacob Lund

Schemas are a key part of a database, where multiple tables are linked with pre-defined relationships. Since this time, implementations have been led by this theoretical model and applied technically in relational database packages like MySQL.

The development of the eXtensible Markup Language as a mechanism for sharing data emerged much later in 1996. Five years later, in 2001, the now popular XML Schema specification was released to help formalise the sharing of consistent and verifiable data. During the mid-1990s to late 2000s, and still to some extent today, XML was the standard of choice for the representation of exchangeable data, designed to be both machine and human readable.

Table 12 (shown here in tabular form) is an example of some data taken from a dataset.

The same data is shown in Figure 13 in XML with associated extracts from the schema.

Table 12. Example data

FirstName	LastName	Instrument	Date Of Birth
John	Lennon	Vocal	1940-10-09
Paul	McCartney	Bass Guitar	1942-06-18
George	Harrison	Guitar	1943-02-25
Ringo	Starr	Drums	1940-07-07

Figure 13. Example data in XML format with associated extracts from schema

XML

```
<People>
  <Person>
    <FirstName>John</FirstName>
    <LastName>Lennon</LastName>
    <Instrument>Vocal</Instrument>
    <DateOfBirth>1940-10-09</DateOfBirth>
  </Person>
  <Person>
    <FirstName>Paul</FirstName>
    <LastName>McCartney</LastName>
    <Instrument>Bass Guitar</Instrument>
    <DateOfBirth>1942-06-18</DateOfBirth>
  </Person>
  <Person>
    <FirstName>George</FirstName>
    <LastName>Harrison</LastName>
    <Instrument>Guitar</Instrument>
    <DateOfBirth>1943-02-25</DateOfBirth>
  </Person>
  <Person>
    <FirstName>Ringo</FirstName>
    <LastName>Starr</LastName>
    <Instrument>Drums</Instrument>
    <DateOfBirth>1940-07-07</DateOfBirth>
  </Person>
  </People>
```

XML schema extract

```
...
<xs:simpleType name=”birthsDate”>
 <xs:restriction base=”xs:date”>
  <xs:minInclusive value=”1800-01-01”/>
  <xs:maxInclusive value=”2017-07-31”/>
  <xs:pattern value=”.{10}”/>
 </xs:restriction>
</xs:simpleType>
<xs:simpleType name=”instrument”>
 <xs:restriction base=”xs:token”>
  <xs:enumeration value=”Vocal”/>
  <xs:enumeration value=”Guitar”/>
  <xs:enumeration value=”Bass Guitar”/>
  <xs:enumeration value=”Drums”/>
 </xs:restriction>
</xs:simpleType>
...
<xs:complexType>
 <xs:sequence>
  <xs:element name=”FirstName” type=”xs:string”/>
  <xs:element name=”LastName” type=”xs:string”/>
  <xs:element name=”Instrument” type=”instrument”/>
  <xs:element name=”DateOfBirth” type=”birthsDate”/>
 </xs:sequence>
</xs:complexType>
...
```

Source: FAO, 2020.

Figure 14. Example data in JSON format with associated extracts from schema

XML

```
[
 {
  “FirstName”: “John”,
  “LastName”: “Lennon”,
  “Instrument”: “Vocal”,
  “DateOfBirth”: “1940-10-09”
 },
 {
  “FirstName”: “Paul”,
  “LastName”: “McCartney”,
  “Instrument”: “Bass Guitar”,
  “DateOfBirth”: “1942-06-18”
 },
 {
  “FirstName”: “George”,
  “LastName”: “Harrison”,
  “Instrument”: “Guitar”,
  “DateOfBirth”: “1943-02-25”
 },
 {
  “FirstName”: “Ringo”,
  “LastName”: “Starr”,
  “Instrument”: “Drums”,
  “DateOfBirth”: “1940-07-07”
 }
]
```

XML schema extract

```
{
 “fields”: [
  {
          “name”: “FirstName”,
          “type”: “string”,
          “constraints”: {
            “required”: true
          }
  },
  {
          “name”: “LastName”,
          “type”: “string”,
          “constraints”: {
            “required”: true
          }
  },
  {
          “name”: “Instrument”,
          “enum”: [“Vocal”, “Guitar”, “Bass Guitar”, “Drums”],
          “constraints”: {
           “required”: true
          }
  },
  {
          “name”: “DateOfBirth”,
          “type”: “string”,
          “format”: “date-time”

  }
 ]
}
```

Source: FAO, 2020.

Such schema can be used by machines to automatically validate the structure and contents of datasets; there is also an online validation tool (Briganti, 2020), which can be used with the two examples above.

While XML has seen wide adoption, 2005 was a significant year for data on the web with the emergence of Ajax (Garrett, 2005). Standing for Asynchronous JavaScript and XML, Ajax is a set of web development techniques that uses JavaScript to dynamically load data into web applications. The initial goal was to allow the dynamic use of XML data in web applications, further encouraging the release of data. In practice, modern applications commonly substitute JSON in place of XML due to the advantages of JSON being native to JavaScript, and, as a result is 21 percent faster to work with, as well as being substantially less verbose.

Similar to XML, json-schema.org provides a JSON Schema specification; however, this is currently not as fully developed as the XML Schema specification and lacks the ability to validate a range of inputs (e.g. minimum and maximum values permitted). Figure 14 shows the same data as the XML example, this time in JSON with the equivalent JSON schema.

JSON schema is shown in this table for tabular data files. One of the challenges with both CSV and JSON formats is the requirement to have two files: the data and the schema, which often leads to the predictable situation where the data is maintained and shared, but the schema gets forgotten and lost. The use of namespaces and linked data as a technical solution goes some way to providing a solution; however, the main problem still lies with the lack of integration of such standards in off-the-shelf data-management packages.

Spreadsheet packages like Excel tend to make formatting data and setting up validation rules overly complex, as proven with the ease of setting them up in tools like Airtable [airtable.com]. Again, however, exporting of the schema is not currently possible, neither is the use of namespaced and linked schemas.

The same is true for database software and the move towards noSQL database structures that are not controlled by tightly defined relational schema. While the move to new flat database structures is good for speed when managing big datasets, the use of namespaces to define schemas is still lacking.

Schemas clearly have their advantages and should be adopted where organizations rely heavily on high quality data.

Cleaning data

One of the biggest challenges when working with any data is dealing with errors. Often errors are not even noticed by data publishers because the data can change over many years. In other cases, errors can be the result of human mistakes in data entry, such as mistyping or incorrect abbreviations.

Even when a schema is available, errors and inconsistencies may exist in the data. When working with any data, it is important to know how to find errors and correct them to make the data more useful.

Section 3.2. Quality and provenance introduce a number of different examples of errors and inconsistencies in data, outlines which can be fixed with schema validation and which need a more advanced tool like Open Refine.

Wrong date formats

Dates can be written in inconsistent ways and according to different standards. One of the biggest confusions exists between the American and the British ways of writing dates. In the United States of America, the month comes first, then the day (e.g. 12/30/2017), whereas it is the other way around in the UK (30/12/2017). This is easy to spot when the day is greater than the 12th, however it can cause confusion otherwise (e.g. 6/7/2017?).

The ISO 8601 standard specifies a series of rules for writing dates and times to solve this and other problems. ISO 8601 specifies that dates must be written year first (e.g. YYYY-MM-DD HH:ii:ss). Not only is this format still easy to read, it also works as a way to sort in date order with the most significant sort factor going first.

As dates are complicated, efforts have been made over the years to hide the management of dates from users. For example, if '8-7' is typed in any general cell in Excel it will automatically translate this into a date and save '08-Jul' in a CSV, which is not ISO standard. Formatting the cells as a date allows the specific formatting, however mixing American and British dates is still possible and only visible with the content is left (incorrect) or right (correct) aligned in the cell. Even more concerning is what happens on import of CSV data into Excel: for example, an ISO standard date (YYYY-MM-DD HH:ii:ss) will be translated to a custom date format (DD/MM/YYYY HH:ii:ss) and this format will be saved back to the file upon completion of editing in Excel, unless the format is changed before it is saved. Most users will be completely unaware that Excel has done this translation prior to displaying the data.

Dates are difficult to manage, especially when software makes assumptions on behalf of the user. Schemas can help, providing the data matches the required format and translations do not happen somewhere in the middle.

Multiple representations

People often try to save time when entering data by abbreviating terms. If these abbreviations are not consistent, it can cause errors in the dataset. Schemas that use enumerated predefined lists of acceptable terms can help here, providing users are not able to easily add to the list and thus recreate the same error.

Other errors that exist in this category include differences in capitalisation, spacing, gender and pluralisation of adjectives (e.g., councilman vs councilmen), which can all cause interesting problems.

Duplicate records

A duplicate record is where the same piece of data has been entered more than once. Duplicate records often occur when datasets have been combined or because it was not known there was already an entry. Additionally, record duplication can occur when one person might be referred to by two names (e.g. Dave and David). This might lead to instances where records need disambiguating to discover if the authors of publications are the same or different. This type of error cannot be caught with a schema validation.

Redundant or combined data

Redundant data is anything that is not relevant to the work with the dataset. Often a dataset has been created for a specific purpose, which requires details that may not needed. Common occurrences of redundant data include rows that represent total amounts. These often appear when a dataset in Excel has been exported into other formats without the 'Total' row first being removed. At other times columns of data have been combined or replicated in order to assist human readability.

Mixed use of numerical scales

Numerical values in datasets often use different scales to make it easier for a human to read. In budget datasets, for example, the units are often in the millions. USD 1 200 000 often becomes USD 1.2 million. However, smaller amounts like USD 800 000 are still written in full. For a machine, this means they read the larger figure as USD 1.2, which causes errors. Alternatively, if the column is meant to be in millions, then the second figure becomes USD 800 000 000 000.

Unfortunately, schemas are not great at catching this type of error. This is because all the numbers could be value and errors could be caused at any level. Setting boundaries on values can help but might not solve the problem. Making the units of measurement clear at the point of collection or use is essential here to guarantee data consistency and to ensure disasters do not happen (Witkin, 1983).

Numerical ranges

Data is sometimes measured in ranges, such as age or salary range. In order for a machine to understand these ranges, it is important to separate the high and low values for easier analysis. It might even be necessary to create new brackets if they have changed over the years (e.g. people's ages or retirement age has risen).

Spelling errors

If there are lots of free text in the data, it is important to check for consistent spelling to ensure analysis can be performed with other datasets. The spelling might not be 'correct' (e.g. colour vs color), however, one option should be chosen, and it should be consistent to allow datasets to be comparable and interoperable.

Data-cleaning toolkit

When looking for errors in data, it may be necessary to download and upload datasets in many tools for cleaning and processing. It is also important to keep a note of what changes have been made and share these openly with others so that everyone can benefit from this work, particularly if the data being cleaned is existing open data that has been published.

We have already looked at a number of schema validation tools, however there are a number of other tools that can help clear messy data.

OpenRefine

OpenRefine is a software tool designed to deal with uncleaned data. The tool is a column-based browser that allows errors to be fixed across an entire open dataset in a single action. The errors that can be fixed include:

- date formats
- multiple representations
- duplicate records
- redundant data
- mixed numerical scales
- mixed ranges

Spreadsheet programmes

OpenRefine is a key tool for cleaning data. However, it is sometimes easier to fix some errors in a spreadsheet programme:

- spelling errors
- redundant data
- numerical verification
- fixing shifted data

Other tools

- Drake
- Data Wrangler (jointly developed tool from Stanford and Berkeley Universities in the United States of America)
- Data Cleaner
- WinPure

For data to be usable, there is a lot more needed than it simply being technically great data. Many guidelines try to simplify the requirements for data to be truly usable and each has its merits. The checklist created from the Open Data Certificates work goes to show the extent of the challenge of producing high-quality usable data. The Open Data Certificate is designed to assess the legal, practical, technical and social aspects of publishing open data using best practice guidance. However, even this does not consider the cross checking of data sources to ensure that the right data is used. Provenance of data is also key but again hard to follow fully on a dataset.

Once trust is established in a dataset, verifying its contents is the next challenge. To some extent, the creation and adoption of schemas and data validators helps to some extent but the need to clean and validate the data (potentially by hand) is not going away any time soon. In fact, as it is explored in Section 3.3. Data analysis and visualisation, the process of preparing data ready for analysis could take up to 80 percent of time.

3.3 Data analysis and visualisation

Section 3.3. Data analysis and visualisation looks at the next stage of data analysis and visualisation. Like the other sections, the knowledge required will vary considerably depending on the exact object of the analysis and visualisation. It aims to take a pragmatic approach to both subjects and offer a broad theory of analysis that leads to clear visualisations that generate impact. Practical examples focus on quantitative data in a spreadsheet and feature extraction-based analysis of qualitative data.

3.3.1 Purpose of data analysis

Raw, unprocessed data is often messy and potentially still not ready for visualisation. Section 3.3. Data analysis and visualisation looks at a number of techniques that can be used to turn data into information including:

- derivation and feature extraction;
- dataset combining;
- dataset enrichment.

Derivation and feature extraction

Derivation and feature extraction are similar in that they are designed to add to existing data without requiring external datasets. This results in additional columns (or features) being added to data, which are based upon the existing data only.

Derivation

A derived data element is one that is derived from other data elements using a mathematical, logical or other type of transformation, e.g. arithmetic formula, composition, aggregation.

For example, source data might contain a series of columns of monthly expenditure. A sum total could be made over all months to add another column. This column is thus derived from the others.

Feature extraction

Feature extraction is very similar to derived data but does not necessarily have to involve a function. For example, the city could be extracted from a list of unstructured address data, making the city a distinct feature of the dataset. Likewise, the colour of the centre pixel of an image could be extracted without the use of a mathematical function. A short tutorial on feature extraction is available for Excel (How to Extract Part of Text String from Cell in Excel?, 2020).

Combining datasets

In combining datasets, we are not referring to adding additional data onto the end of an existing dataset (a process known as consolidation), but rather the process of adding to the existing data.

If using a spreadsheet, the result of combining two datasets with 10 columns each will be one dataset with 19 columns. Combining data can only be done in this way if there is a column in each dataset with a shared value. This column becomes the key upon which the combination can be performed.

For example: take a dataset of roads and the numbers of cars during each hour of the day. Each hour period is a column and each road is a row. If there were many sources of the data, e.g. sources with different hours in all individually collected, then it could be keyed on the road name and combine all the data into a single dataset.

Find out how to combine two datasets in Excel (Merge Two Excel Files Using a Common Column, 2020) or in Open Refine (Hirst, 2011).

Enriching geographic data

Much like combining data, where two datasets can be combined based upon a common feature (e.g. road name), the same can be achieved with geographic data. One big difference is that it is possible to combine geographic data based upon location and place a geographic point inside a boundary. This process is known as a spatial join.

A spatial join can be useful when looking at features of different buildings or services and mapping them onto jurisdiction regions to see if any patterns emerge. For example, this could be used to see if schools in different council-controlled areas perform significantly differently.

Find out more about enriching and joining map data using the excellent guides provided by CartoDB (CARTO, 2020).

Another essential stage in preparing geographic data for analysis is that of geocoding. Geocoding is the process of taking any reference or description of a physical location (like a street address) and adding the actual physical location coordinates (like latitude and longitude) to the data. Reverse geocoding is thus the opposite, extracting the description (e.g. London) from the coordinates. Geocoding can also refer to the process of transforming from one coordinate representation system (e.g. easting–northing) to another (e.g. latitude–longitude). Geocoding is essential when trying to analyse geographic data and perform other operations such as spatial joins.

3.3.2 Qualitative and quantitative data analysis

There are two main types of statistical analysis talked about in data: qualitative and quantitative. Put simply, qualitative research deals with open-ended, often text-based data, while quantitative research tries to focus on objective, measurable data in the form of numbers or other structured data. Table 12 (from the Open University) shows some of the main differences between the two types of research.

Although Table 13 illustrates qualitative and quantitative research as distinct and opposite, in practice they are often combined or draw elements from each other. For example, a survey question could be answered on a scale of 1–10, however, the question could be subject to personal bias.

Even something as simple as counting sheep can be fraught with danger: if the field also contains lambs, are these sheep? When do they become sheep?

Analysing quantitative data

Good quantitative researchers will seek to maintain a level of control of the different variables and carefully define both the scope and size of the sample being measured. They will also attempt to eliminate or accept the influence of other factors on the sample and outline this clearly in the research.

One of the most important aspects in research is that of obtaining a statistically significant result. Statistical significance is essentially a measure which states that there is less than a 5 percent chance that the result of the analysis is down to chance. This can be explained best with coin tosses and testing if a coin is biased, which can be tested with a simple null hypothesis test. The null hypothesis is a statement about the world which can plausibly account for the data observed, for example 'the coin is fair'. It is possible to flip the coin 100 times and if heads came up only once, then it is safe to say the null hypothesis can be rejected and that the coin is biased.

But what if 51 of the 100 tosses come out heads? Or flip the coin 100 000 times and 51 000 times it comes out heads. Can either of these be random chance, or is the coin biased?

Calculating whether a result is significant can be done in two ways, either with a statistical significance calculation or a z-score calculation. Either way, what we are trying to calculate is if the result falls outside of the 95 percent of observations where the null hypothesis is true, thus disproving it. As the majority of quantitative data can be mapped to a normal distribution it is this that dictates whether a result is statistically significant.

Table 13. Qualitative vs quantitative research

	Qualitative research	Quantitative
Type of knowledge	Subjective	Objective
Aim	Exploratory and observational	Generalisable and testing
Characteristics	Flexible	Fixed and controlled
	Contextual portrayal	Independent and dependent variables
	Dynamic, continuous view of change	Pre- and post-measurement of change
Sampling	Purposeful	Random
Data collection	Semi-structured or unstructured	Structured
Nature of data	Narratives, quotations, descriptions	Numbers, statistics
	Value uniqueness, particularity	Replication
Analysis		Statistical

Source: Save the Children and Open University, 2014.

A 95 percent level of confidence means that the null hypothesis is rejected if the calculation of statistical significance falls outside of the 95 percent of the area of the normal curve. The z-score is a calculation of how many standard deviations away from the mean the sample is.

So, with a toss of a coin with a null hypothesis that the coin is fair, based upon different amounts of flips and results is as follows, see Table 14.

From Table 13, it can be seen that as the sample size increases, so the significance can be observed with the same percentage of coin tosses that come out heads.

Sample size is one key factor and it helps eliminate the chance of other factors having an effect, such as the coin always being tossed the same way, with the same force and from the same position. However, some tests might require this kind of absolute control when the sample size is small, for example comparing the shatter resistance of expensive phones (Smith, 2017).

With the majority of quantitative analysis focusing on averages, sample size is important to get rid of the effect of outliers in the data. Additionally, it is important to pick the most representative average for the dataset, as the mean is not always an accurate reflection of the data, if there are outliers in that data. For example, in President Bush's 2008 State of the Union Address he attacked his opposition, stating that their tax plans would mean an average tax hike of USD 1 800 per person. However, this was the mean value and people's earnings are not evenly distributed; the proposals actually hit high earners badly, while putting money back in the pockets of those who needed it.

It is not just sample sizes that affect the significance of a result, however, and there are other factors that need to be eliminated first before a result can be declared significant:

Fluctuation

Random events cluster, thus fluctuation is also important to bear in mind. Flipping a whole run of heads on a coin does not mean there will be more to follow (unless the coin is biased). The same is the case for road accidents. Installing speed cameras in accident black spots might not necessarily be correlated to the subsequent reduction in accidents.

Targets

While most data is normally distributed, targets have a profound effect on people's behaviour. Once a target is introduced, people will game their outcome to match this target. For example, one hospital in the United States of America operated only on healthy patients, so they could obtain funding for being the most successful hospital.

Correlation

Just because there appears to be a correlation does not mean there is one. A famous example from spurious correlations was that banana pricing strongly correlates with the number of people who died becoming tangled in their bed sheets. Or the 99 percent correlation between spending on space, science and technology and suicides by hanging, strangulation and suffocation (Vigen, 2020).

Percentages

Percentages can make small numbers look big and are often used this way. For example, last year there was a 100 percent increase in cancer cases linked to mobile phones: it rose from 1 to 2. The other problem with percentages is that many confuse percentage points and percentages. For example, consider when Value Added Tax in the UK rose by 2.5 percent (i.e. 2.5 percentage points) from 17.5 percent to reach 20 percent; a 2.5 percent rise would actually have made it 17.9375 percent.

Table 14. Sample size

Coin	Flips	Percent heads	Z-score	Biased?
#1	100	51%	0.2	No
#2	100	60%	2.00	Yes
#3	50	60%	1.41	Not significant
#4	50	65%	2.12	Yes

Un-normalised data

Another common pitfall is to not adjust figures for floating variables; for example, saying global cancer incidence is predicted to increase by 75 percent by 2030 is pretty shocking (ecancer, 2012). However, this percentage has been calculated from the raw numbers of people and has not been adjusted for population growth. Another common type of headline is saying that spending on a sector or service has risen every year, but not normalising this to inflation and thus revealing that the real value of this spending is actually falling.

Whichever technique is used to analyse quantitative data, being truly objective can take a huge amount of effort, even when the data subject is a simple coin.

Analysing qualitative data

Not all data comes in the form of structured tables or accurately located geographic data. Often it can be the qualitative data that can be hardest to analyse and work with. Qualitative data is information that cannot be measured and is very subjective as a result (even colour can be subjective (Wikipedia, 2020g).

The main aim of qualitative data analysis is to reduce and make sense of vast amounts of information, often from different sources. The result of such analysis is to offer an explanation, interpretation or thematic summary of the data. Inputs to qualitative analysis can take many forms including interview transcripts, documents, blogs, surveys, pictures, videos, etc.

Qualitative data analysis is a more natural process for humans who naturally seek to distil inputs into themes and key outcomes, as is especially true of meetings or focus groups. People will often use mind-mapping or 'Post-it' based thought maps to help group together and categorise wide-ranging discussion into key themes.

Qualitative data analysis ought to pay attention to the spoken word context, consistency and contradictions of views, frequency and intensity of comments, and their specificity, as well as emerging themes and trends.

There are two main ways of analysing qualitative data, framework analysis and thematic network analysis.

Framework analysis involves building a predefined set of criteria that clearly reflect aims, objectives and interests. Using this set of criteria, the relevant pieces of information can be extracted from the data and compared with other inputs in the framework. Using a framework allows many researchers to do the extraction while minimising the chances of qualitative analysis bias. A bias can still be introduced at the framework design stage, which may result in key information being missed.

An alternative approach is to not construct a framework, but rather to apply thematic network analysis. *Thematic network analysis* is a more exploratory approach which encourages the analysis of all of the input data, which can shape the output in unexpected directions. In reality, the majority of qualitative data analysis will involve a combination of the two approaches.

Whichever approach is chosen, the first step in any qualitative data analysis involves familiarisation with the data; reading and re-reading responses. At the same time, it is a good idea to start codifying the data by writing down keywords and topics which attempt to reduce and interpret the data. The result of the coding process could be the thematic network analysis or a framework by which all responses need to be coded. Either way, coding can be a long, slow and repetitive process, however there are a number of tools that can help with thematic network analysis.

Thematic network analysis tools

Entity recognition tools provide one such technique that can help analyse and enrich qualitative data. Essentially entity recognition seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

More widely, such entity recognition is used by services such as TheyWorkForYou [**theyworkforyou.com**] in order to track the activities of politicians and provide this in an easy way to the public.

Calais is a good example of an online tool that can perform entity recognition and provide links into additional data on each topic. Other techniques can be much simpler but just as effective such as word cloud generators.

3.3.3 Data Visualisation

Another way to quickly interpret data is to visualise it. The human brain is much more adept at consuming and understanding data presented visually than as text.

Most charts used in modern data visualisation derive from the original designs of William Playfair (1759–1823), a political economist. Playfair invented several types of diagrams: in 1786, the line, area and bar chart of economic data; and in 1801, the pie chart and circle graph, used to show part–whole relations.

The choice of which visualisation technique to use depends on both the objective of the visualisation and the type of data to be visualised. Section 3.3. Data analysis and visualisation explores the different objectives and the types of visualisation suited to each with examples.

The goals of visualising data

The goal of data visualisation is to communicate information clearly and efficiently to users.

According to Friedman (2008), the *"main goal of data visualization is to communicate information clearly and effectively through graphical means. It does not mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way.*

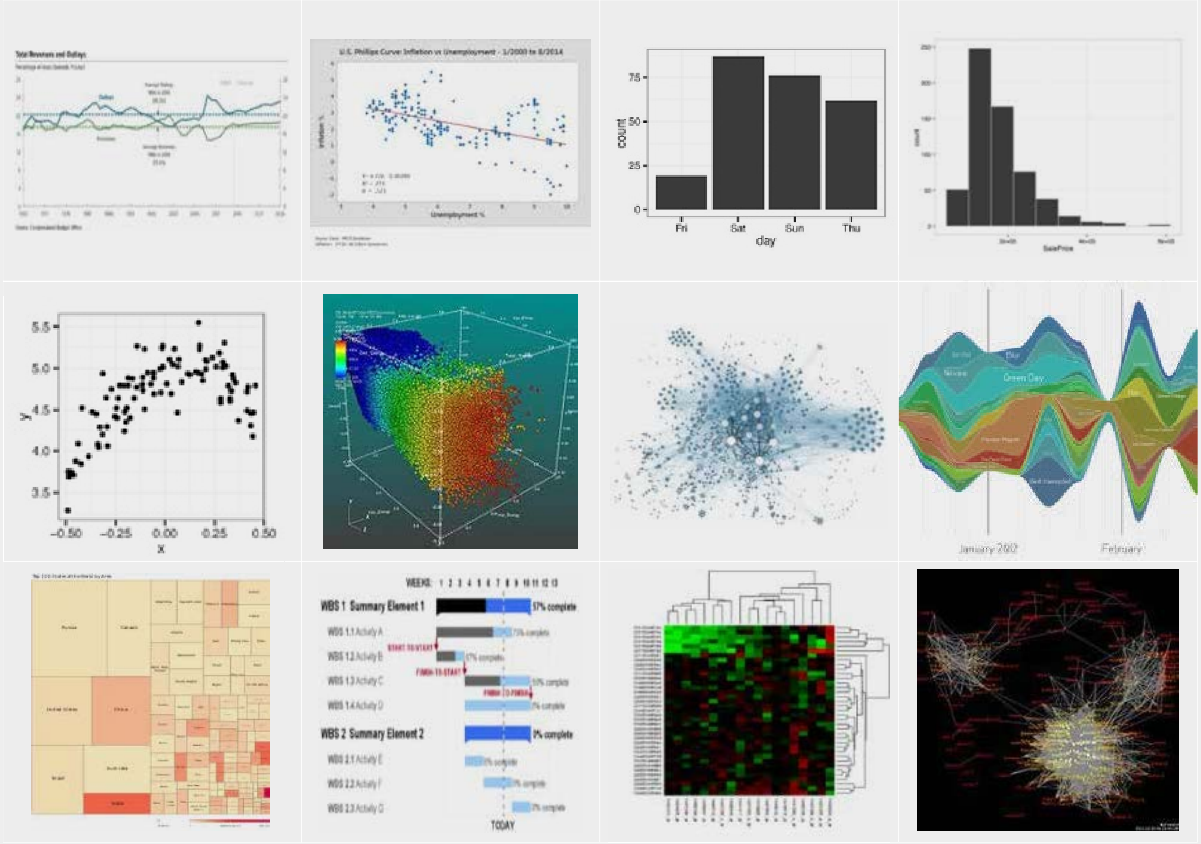
Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualizations which fail to serve their main purpose – to communicate information."

One of the main problems with data visualisation is that it is a broad topic to cover a wide range of different visualisations that are designed for different purposes. Communication is only one goal of a data visualisation. Data visualisations can also be used during the data analysis stage in order to help make sense of the data and guide analysis.

If the main goal of a data visualisation is to communicate information, then the visualisation should be able to do this without the need for any explanatory text or additional knowledge by the reader.

Figure 15 shows examples from Wikipedia's data visualisation page (Wikipedia, 2020h): arguably only

Figure 15. Examples of data visualisation



Source: Wikipedia, 2020h.

Table 15. Types of data

Data type	Description	Example visualisation
Time series	Observations of the same objects over time	Line chart, motion chart, polar area diagram, Gantt chart, bar chart
Population	Observation of different objects at a single point in time	Bar chart, map, treemap, pie chart
Multivariate	Observations of different objects at different points in time	Multidimensional motion chart, bar chart, treemap

the bar charts (top right) and treemap (bottom left) are good visualisations for instant communication of data. This is because both of these charts use a visualisation trick called 'pop out'.

Pop out helps direct the eye to the correct place, instantly and the human eye is drawn towards brighter colours, larger items and things which stand out through difference. It is a feature that is programmed into the visual cortex.

Inside the visual cortex there are two streams, the ventral (what) stream and the dorsal (where/how) stream. It is the dorsal stream that processes information from the eye about our surroundings in real time, so we can instantly react in situations of risk.

It is the dorsal stream which is looking where things are and how they relate to other things that makes pop out work so quickly. Conversely, the ventral stream is responsible for working out 'what' the thing is. This is a much slower process and is the reason why a person might recognise another person's face but can't name the person.

The best data visualisations for communication appeal to the dorsal stream and make information pop. If a visualisation requires use of the ventral stream to help contextualise the information, then there is a high probability that individuals will interpret the information differently to each other.

In any visualisation, it is essential to get rid of any extraneous clutter that detracts from the message being conveyed.

Choosing the correct visualisation for data

Choosing the correct data visualisation depends on two key aspects:

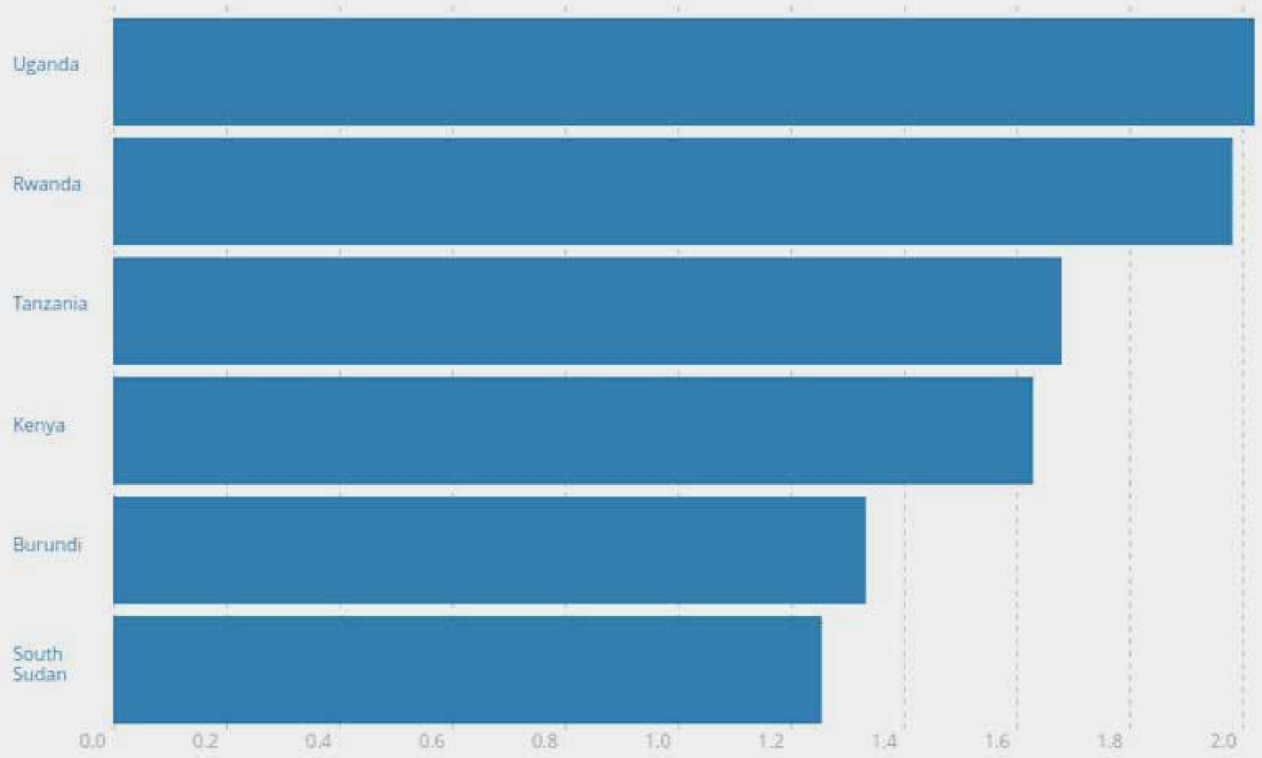
1. the type of data;
2. the message to be conveyed.

There are three main types of data outlined in Table 15.

For example, take the data about cereal yields in the East African Community of Burundi, Kenya, Rwanda, South Sudan Tanzania, and Uganda, which is available as open data from the World Bank (World Bank, 2017). This data can be visualised in many ways due to its multivariate nature.

The data available from the World Bank is available for every country dating back to 1961. If the data was only about a single country, then we would have a time series dataset. If the data was only about one year, then this would be a population dataset; however, both are available, thus the dataset is a multivariate dataset.

Figure 16. Population bar chart visualisation of cereal yield in the East African Community in 2014



Source: World Bank, 2017.

Conversely the bar chart (Figure 16) compares the countries at a single point in time (2014 in this case). Using the same colour here means the eye is not drawn to a single country but is focused on the sizes of the bars. Perhaps what stands out here is the three groupings of Uganda with Rwanda, Tanzania with Kenya, and Burundi with South Sudan. However, the time-series line chart tells us that these groupings may not be significant over time.

Data analysis and visualisation is a difficult and complex topic. This is especially the case given the multivariate nature of data and the fact that many datasets are now 'big', meaning that multivariate problems are multiplied.

Choosing the right analysis technique is also heavily tied to the data collection technique to ensure that bias is limited wherever possible. Careful consideration should be given to the data collection and how analysis might be influenced by external factors such as policy targets.

Data visualisation should also avoid introducing bias or any aspects that mislead the viewer. Given that the eye will process the visualisation in a fraction of a second, the visualisation should be designed carefully to ensure that the key message is being conveyed accurately.

3.4 Open data in policy cycles

The policy cycle is a tool used for designing and delivering policies. Whilst it has its roots in the public sector, it can also be used by private-sector organizations when implementing a company- or department-wide policy.

There are different variations of the policy cycle model which largely depend on the granularity of the breakdown of the stages, as well as the needs of the organization that is using the cycle.

Broadly, however, there are five key steps in a policy cycle, as identified by Anderson (1974):

- agenda setting;
- policy formulation;
- decision-making;
- policy implementation;
- monitoring and evaluation.

Agenda setting is the first step in the policy cycle. This key focus of this stage is the identification of a problem or vision that requires policy intervention. Following the setting of an agenda is policy formulation. This can often be the biggest stage of the policy cycle and involve everything from data analysis to in-depth studies and consultation with a broad range of stakeholders.

Next in the cycle is decision-making where all the various options are discussed, and external factors considered prior to policy implementation.

The step in one complete policy cycle is that of monitoring and evaluation. This is a critical last step to help evaluate the success of a policy and set the agenda for improving or changing the policy to suit the new environment the policy has created. Anderson (1974) offers the definition of the stages as set out in Table 16.

It is also worth noting that the United Kingdom Institute for Government has written a report, *Policy Making in the Real World* (Hallsworth, Parker and Rutter, 2011), detailing how a policy cycle "*reduces policy making to a structured, logical methodical process that does not reflect reality*," as policy making will often not fit neatly into isolated stages. Whilst a policy cycle can be a useful tool when designing a policy and thinking where open data can assist in policy making, it is fine if the actions needed do not fit into neat stages. Additionally, it may be necessary to repeat steps, or complete the cycle more than once with one policy.

Table 16. Stages in the policy cycle as identified by Anderson (1974)

Problem identification	The recognition of certain subject as a problem demanding further government attention.
Policy formulation	Involves exploring a variation of options or alternative courses of action available for addressing the problem. (appraisal, dialogue, formulation, and consolidation)
Decision-making	Government decides on an ultimate course of action, whether to perpetuate the policy status quo or alter it. (Decision could be 'positive', 'negative', or 'no-action')
Implementation	The ultimate decision made earlier will be put into practice.
Evaluation	Assesses the effectiveness of a public policy in terms of its perceived intentions and results. Policy actors attempt to determine whether the course of action is a success or failure by examining its impact and outcomes.

3.4.1 Applying open data to the policy cycle

Agenda setting

Within the policy cycle, data can help identify the problem, by giving a realistic view of what is currently happening and helping decision-making that is informed by the current situation. Without data as an input, it would be merely making a best guess at what is happening. With open data, anyone is able to access, use and share the data, so more people can make more informed decisions from a wider range of available sources.

Using open data to help identify the problem is also known as using open data as an input, such as the example with GroenMonitor in the Netherlands.

Protecting crops from pest outbreaks with vegetation maps: GroenMonitor. Farm productivity is often hit by crop damage caused by pests. Mice and other pests are difficult to detect on large farms through manual inspection alone. The GroenMonitor (GreenMonitor) is a tool that shows a current vegetation map of the Netherlands, based on satellite images and maps made publicly available through the European Space Agency (ESA), which makes pest outbreaks easy to identify and mitigate relatively quickly. In 2014, the GroenMonitor helped to identify 12 000 ha (29 652 acres) of fields affected by mice. The tool is now being exploited for various other applications, including plant phenology, crop identification and yield, identification of agricultural activities (e.g. mowing, ploughing and harvesting), nature and water management.

When using data, whether open or not, to help make informed decisions, having confidence in the quality of that data is essential. Poor data as an input will be reflected in the poor quality of a policy agenda. However, having good-quality data does not just mean data that is accurate or free from errors.

Quality data should also be timely: that might be using the most recent dataset or using data that is still relevant to a decision or policy. For example, crop prices in the United States of America from 2005 may be irrelevant in today's policy-making, unless data is needed to measure the impact of Hurricane Katrina on American crop prices and comparing that to make predictions about the effects of 2017's Hurricane Irma, or future hurricanes in the United States of America. This could then identify a need for a policy on crop insurance, regulated prices or government subsidies.

Also essential is having confidence in the way the data has been collected, especially when the data relates to people. A non-representative sample can skew the data, as has been seen repeatedly in British election predictions in recent years (Sturgis *et al.*, 2016). In the agricultural sector, open data about farm incomes could be skewed by surveying a non-representative sample of farmers. Most open data sources will provide details on how the data was collected (Department for Environment, Food & Rural Affairs, 2010).

Using different sources of data will help to gain a holistic overview of the problem.

Policy formulation

The next stage of the policy cycle involves designing the policy. Having identified the problem, the idea here is to carry out a number of research studies and/or data analysis processes in order to formulate a potential solution to the identified problem. In the case of the GroenMonitor, potential solutions included plant phenology, mowing and even culling.

This stage of a policy cycle can take the longest and be the most challenging due to the nature of statistics and the difficulties in obtaining accurate and controlled scientific data. As discussed in Section 3.3. Data analysis and visualisation, the analysis of data does not always mean the correlation equates to causation and sometimes can lead to the wrong (or no) conclusion.

Once such example in the United Kingdom is how to control the outbreak of bovine tuberculosis in cattle. The government decided that one necessary course of action is to cull tens of thousands of badgers (Carrington, 2017), which have been linked to outbreaks. However, scientific studies of the effects of culling (available from the government's own website) find that *"badger culling cannot meaningfully contribute to the future control of cattle TB in Britain"* (Bourne, 2007).There are many tools that can help those in agriculture make meaningful decisions to protect their crops and react to changing conditions. Plantwise [plantwise.org] offers farmers open access to over 10 000 factsheets regarding crop pest prevalence and best practices to help manage and prevent potential crop loss from pests and diseases. Farmers are thus able to make informed decisions about crop protection.

Decision-making

While the policy formulation stage may have come up with what appears to be an ideal solution to the particular problem identified, the final decision will need to take into account any adverse effects the policy might have on the sector as a whole.

To help guide the decision-making process, it may be necessary to have a set of objectives that the policy has to fulfil overall. A good example of such a process is the European Common Agricultural Policy (CAP) which has five key objectives:

1. to increase productivity, by promoting technical progress and ensuring the optimum use of the factors of production, in particular labour;
2. to ensure a fair standard of living for the agricultural community;
3. to stabilise markets;
4. to secure availability of supplies;
5. to provide consumers with food at reasonable prices.

Balancing the requirements of these five objectives can be a challenging task, especially as some countries have access to more advanced technologies which could create unstable monopolies within Europe. Securing fair standards or living, with stable markets and availability of supplies could be challenging as Europe grows to the east, while potentially losing the UK on the west.

Once again, open data can play a key role in the decision-making process. In relation to the CAP, each of the objectives is either already monitored or can be evaluated using open data on trade, markets and census data. Even multinational companies, such as Syngenta, are now releasing key indicator data they hold regarding productivity, biodiversity and smallholder reach.

Policy implementation

This stage is about taking or setting an expected course of action, either by changing the law, distribution of money or something else. As well as helping inform the policy decision, open data can also play an important role in policy implementation. The publication of open data may be all that is required to help enact the policy. For example, open data plays a key role in tackling obesity and increasing the health of a population:

- Empowering consumers to make smart food choices: USDA National Nutrient Database. Consumers have clearly indicated that they want to be better informed on the quality and ingredients of the food that they are consuming. Although basic information already exists on food packaging, more detailed information on food nutrients could allow people to make better decisions regarding food selection based on their individual needs (e.g. following the advice of a dietitian).
- The USDA National Nutrient Database for Standard Reference (SR25) is the major source of food composition data in the United States of America and provides data sources for most public and private sector databases. SR25 contains nutrient data for more than 8 500 food items and about 150 food components, such as vitamins, minerals, amino acids, and fatty acids. The use of this data is not limited to commercial applications (e.g. smartphone apps). It provides the basis for new services like ChooseMyPlate.gov, an initiative launched by First Lady of the United States of America, Michelle Obama and USDA Secretary Tom Vilsack to provide *"practical information to individuals, health professionals, nutrition educators, and the food industry to help consumers build healthier diets with resources and tools for dietary assessment, nutrition education, and other user-friendly nutrition information."*

Many countries around the world now make it mandatory to include nutritional information on all their food products. Food standards agencies also publish open rating data and mandate that restaurants and food-processing companies must also publicly display their rating data for everyone to access (See UK food hygiene rating data API at **ratings.food.gov.uk/open-data**. This open publication is designed to force behaviour change and help improve the overall quality of the food production supply chain, from farm to plate.

Data is not just an input to the policy-making process. It can also be the output. It may even be sufficient on its own to help solve the policy problem or to create markets in which others solve the policy problem.

Monitoring and evaluation

Depending on the policy implementation, monitoring and evaluation can take many forms, from simple monitoring of continued open data publication (as is the case with food hygiene ratings) to the re-analysis of cereal crop yields and cases of Bovine TB. In nearly all cases, monitoring and evaluation requires the continued collection and analysis of data, which represents a completing of the policy cycle as this data can then be used to set a new agenda.

As with all stages of the policy cycle, wider involvement in the monitoring will also be key to ensuring policy success. This is especially the case where there might be evidence of corruption anywhere in the policy chain:

- Exposing misspent farm subsidies in Mexico: FUNDAR. PROCAMPO is the largest federal farm subsidy programme in Mexico supporting the poorest farmers. Since 2007, there have been concerns that its subsidies were not received by those meeting the requirements, who were in dire need of support. To better understand the situation, the FUNDAR Centre of Analysis and Research, a Mexican NGO, called for information related to the distribution of subsidies from the Mexican Ministry of Agriculture. After initial requests resulted in incomplete data in non-machine-readable formats, the agency in charge finally published the data.

  Analysis showed that 57 percent of the benefits were distributed among the wealthiest 10 percent of recipients, confirming initial fears. An important outcome of the data release was the development of a database (Subsidios al Campo en México) by FUNDAR and other NGOs, which publishes ongoing information about the farm subsidies to ensure more transparency over the process. A series of resignations followed the revelations and the government-imposed limits on the eligibility of subsidies.

Open data can play an even more important role when there are many international players involved in the policy process. This is particularly relevant in agriculture in the area of international aid funding. The United Kingdom **DFID DevTracker** allows anyone to find what money is being spent by the United Kingdom government on international aid and in which countries. Challenges are faced with monitoring the impact of such funds, especially on a global scale. Community groups such as Follow the Money (Nigeria) are important to ensure that funds reach their designated projects and that policies are effectively implemented.

Being open and transparent throughout the policy cycle is critically important to allow anyone to join in the conversation and assist with the monitoring and evaluation, be that as a member of overnment, a private organization, the press, or a community member.

©Adobe Stock/luchschenF

3.4.2 The impact of open on policies

An open approach to policy-making has many benefits beyond simply being able to use open data as an input to inform decision-making.

The policy cycle relies on the constant use of data to inform the correct decision, as well as monitor the effectiveness of that decision. If the data was open at the point of agenda identification but is closed at the point of monitoring and evaluation, then the policy cycle has been broken and a second cycle is not possible following the same process.

As seen in the sub-section, Policy implementation, in 3.4.2. Applying open data to the policy cycle, open data could also be sufficient to solve a policy problem without further intervention. For example, open transport data can help ease congestion by informing people about public transport options, possibly avoiding the need to spend millions on new roads.

There are many other benefits open data brings to the policy-making process including transparency and inclusion, which is important if there is a high number of stakeholders to reach. In order to obtain 'buy in', it is important to include as many stakeholders as possible. Open data not only allows stakeholders to engage but to also perform their own analysis to discover any potential policy solutions.

Another big benefit of publishing open data as part of the policy implementation process is the way it allows markets to evolve around the data and take on the challenge of solving policy problems on one behalf, such as with food hygiene ratings in San Francisco:

- Highlighting restaurant inspection scores and improving food safety: Open data is being used to help consumers choose where to dine, while incentivising improvements in food safety. Local Inspector Value-Entry Specification (LIVES) aims to *"normalise restaurant inspection scores across jurisdictions, allowing consumers to get a sense for restaurant food safety compliance across municipalities and within their home town"* (Socrata, 2015).
- LIVES was launched in 2013 as a project between the city and county of San Francisco, city of New York, and Yelp, and is providing the standard for publishing open data on restaurant inspections. By allowing citizens to make better use of inspection results, LIVES facilitates food transparency and decision-making on approved restaurants. When the City of Los Angeles began to require that restaurants displayed hygiene grade cards on their entrances, studies found it was associated with a 13 percent decrease in hospitalisations due to foodborne illness (Dyson, 2013).

While most think of data as simply an input in policy-making, the benefits can be multiplied, and policy realised quicker if open data is considered throughout.

As the world grows in population, the role of open data in agriculture and nutrition will be key to driving effective policy and efficiency to ensure that everyone has an equal chance. Agricultural policies will be of critical importance in the next fifty years, especially as the world's climate also changes.

4

Exposing data

©Adobe Stock/Darren Baker

4.1 Managing data for reuse

This chapter focuses on how to expose data so that other actors can reuse it. Data cannot be exposed in an effective way and be of real use if it is not managed appropriately. For example:

- If weather data is not collected and stored according to specific temporal and spatial criteria, or is not annotated with source and methodology information, it is difficult to expose it selectively based on users' needs.
- If a database of farmer profiles is not regularly updated, users will not trust it; and if it is not regularly checked for consistency (e.g. of values entered by different enumerators), the data will not be sufficiently harmonised for reuse.

This is why Section 4.1. Managing data for reuse is dedicated to data management.

Since this book is targeted at farmers and professionals working for farmers, data management is a relevant topic, as far as specific data services are concerned. This includes on-farm information systems/precision agriculture, data services for farmers, farmer profiling systems and the like. It is not expected that the building and maintenance of big public open data repositories is of high interest to readers of this book, so Section 4.1. Managing data for reuse does not go into the details of open data repository management practices or open data lifecycles.

Many recommendations in Section 4.1. Managing data for reuse will be useful to all types of services provided to farmers. Some recommendations are particularly tailored to those services that collect large amounts of data (either directly or from other sources) and need to manage challenging datasets with dynamicity, multiple versions, licensing and usage rights issues etc.

Good data management is essential to enable users to expose quality data to: (a) other actors for reuse, for instance farm soil data, weather data or price data exposed to farm management information systems; (b) other parts of the system in complex distributed systems, for instance where weather data is used in precision agriculture equipment.

After an overview of good practices in data management, including an indication of software tools that can help in managing and exposing data, Section 4.1. Managing data for reuse illustrates the importance of having a data management plan and a data sharing policy.

Finally, some suggestions on how to maximise the use of a data service, in particular how to engage users will be provided.

4.1.1 Good data management

In order to build and maintain trust in data, it is necessary to have stable data management principles and practices in place. Good data management principles help to ensure that data produced or used are registered, stored, made accessible for use and reuse, managed over time and/or disposed of, according to legal, ethical and funder requirements and good practice. For open data consumers, trust depends on numerous factors:

- *Knowing the source.* Trust in data begins with knowledge of its source.
- *Trusting the source.* If the data comes from a trusted source, it can be trusted, as can the conclusions drawn from it.
- *Timeliness of the data.* Even when from a trusted source, data is not useful if it is outdated.
- *Data quality.* Trusted data must accurately and precisely reflect what it measures.
- *Sustainability.* A trusted dataset must have some guarantee of availability.
- *Discoverability.* Like documents, data is only useful if it is straightforward to find.
- *Documentation and support.* Consumers should be able to access support for data if needed.
- *Interaction.* If there is a problem with data, consumers should be able to provide feedback.

Data management, therefore, is a process involving a broad range of activities from administrative to technical aspects of handling data in a manner that addresses the factors listed above. A sound data management policy will define strategic long-term goals for data management across all aspects of a project or enterprise.

A data management policy is a set of high-level principles that establish a guiding framework for data management. A data management policy can be used to address strategic issues such as data access, relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other issues. As a high-level framework, the data management policy should be flexible and dynamic, which allows it to readily adapt to unanticipated challenges, different types of projects and potentially opportunistic partnerships while still maintaining its guiding strategic focus. The data management policy will help inform and develop a data management plan, which will be discussed in more detail later in Section 4.1. Managing data for reuse.

In order to meet data management goals and standards, all involved parties must understand their associated roles and responsibilities. The objectives of delineating data management roles and responsibilities are to:

- clearly define roles associated with functions;
- establish data ownership throughout all project phases;
- instil data accountability;
- ensure that adequate, agreed-upon data quality and metadata metrics are maintained on a continuous basis.

Data quality and consistency

Quality, as applied to data, has been defined as fitness for use or potential use. Many data quality principles apply when dealing with species data, for example, and with the spatial aspects of those data. These principles are involved at all stages of the data management process, beginning with data collection and capture. A loss of data quality at any one of these stages reduces the applicability and uses to which the data can be adequately put. Data quality standards may be available for accuracy, precision, resolution, reliability, repeatability, reproducibility, currency, relevance, ability to audit, completeness and timeliness.

Data quality is assessed by applying verification and validation procedures as part of the quality control process. Verification and validation are important components of data management that help ensure data is valid and reliable.

Consistency

Consistent data is data that is technically correct and fit for statistical analysis. This is data that has been checked for missing values, special values, (obvious) errors and outliers, which are either removed, corrected or replaced. The data is consistent with constraints based on real-world knowledge about the subject that the data describes.

Data quality is assessed by applying verification and validation procedures as part of the quality control process.

Cataloguing: metadata and content standardisation

All datasets should be identified and documented to facilitate their subsequent identification, proper management and effective use, and to avoid collecting the same data more than once.

To provide an accurate list of datasets held by an organization, a catalogue of data should be compiled. This is a collection of discovery-level metadata for each dataset, in a form suitable for users to reference. These metadata should provide information about the content, geographic extent, currency and accessibility of the data, together with contact details for further information about the data.

All datasets, once catalogued, should also be documented in a detailed form suitable for users to reference when using the data. These detailed metadata should describe the content, characteristics and use of the dataset, using a standard detailed metadata template.

Metadata

Metadata, or 'data about data' explains the dataset and allows the data owner to document important information for:

- finding the data later;
- understanding what the data is later;
- sharing the data (both with collaborators and future secondary data users).

It should be considered an investment of time that will save trouble later several-fold.

There are several useful vocabularies that have been created in order to use standardised common terms for metadata: examples are:

- Dublin Core (for any information resource, including datasets)
- Darwin Core (for biological resources)
- FGDC (Federal Geographic Data Committee, for geographic metadata)
- DDI (Data Documentation Initiative, for datasets)
- DCAT (Data Catalog Vocabulary, for datasets)
- ABCD (Access to Biological Collections Data, for germplasm databases)
- CSDGM (Content Standard for Digital Geospatial Metadata, for geographic metadata).

File content standardisation

In order for others to use the data, they must understand the contents of the dataset, including the parameter names, units of measure, formats, and definitions of coded values. At the top of the file, include several header rows containing descriptors that link the data file to the dataset (for example, the data file name, dataset title, author, today's date, the date the data within the file was last modified, and companion file names). Other header rows should describe the content of each column, including one row for parameter names and one for parameter units. For those datasets that are large and complex and may require a lot of descriptive information about dataset contents, that information may be provided in a separate linked document rather than as headers in the data file itself.

- *Parameters.* The parameters reported in datasets need to have names that describe their contents, and their units need to be defined so that others understand what is being reported. Use commonly accepted parameter names. A good name is short, unique (at least within a given dataset), and descriptive of the parameter contents. Column headings should be constructed for easy importing by various data systems. Use consistent capitalisation and use only letters, numerals, and underscores – no spaces or decimal characters – in the parameter name. Choose a consistent format for each parameter and use that format throughout the dataset. When possible, try to use standardised formats, such as those used for dates, times, and spatial coordinates.

- *Data type.* All cells within each column should contain only one type of information (e.g., text, numbers, etc.). Common data types include text, numeric, date/time, Boolean (Yes/No or True/False), and comments (used for storing large quantities of text).

- *Coded fields.* Coded fields, as opposed to free text fields, often have standardised lists of predefined values from which the data provider may choose. Data collectors may establish their own coded fields with defined values to be consistently used across several data files. Coded fields are more efficient for the storage and retrieval of data than free text fields.

- *Missing values.* There are several options for dealing with a missing value. One is to leave the value blank, but this poses a problem as some software does not differentiate a blank from a zero, or a user might wonder if the data provider accidentally skipped a column. Another option is to put a period where the number would go. This makes it clear that a value should be there, although it says nothing about why the data is missing. One more option is to use different codes to indicate different reasons why the data is missing.

Version control

It is important to consistently identify and distinguish versions of datasets. This ensures that a clear audit trail exists for tracking the development of a dataset and identifying earlier versions when needed. Thus, it is necessary to establish a method that indicates the version of the dataset:

- A common form for expressing data file versions is to use ordinal numbers (1, 2, 3, etc.) for major version changes and decimals for minor changes (e.g., v1, v1.1, v2.6);
- Confusing labels should be avoided e.g. revision, final, final2, definitive copy;
- Record ALL changes (minor and major);
- Discard or delete obsolete versions (whilst retaining the original 'raw' copy)
- Use an auto-backup facility (if available) rather than saving or archiving multiple versions;
- Turn on versioning or tracking in collaborative documents or storage utilities such as Wikis, GoogleDocs etc.;
- Consider using version control software e.g. Subversion, TortoiseSVN.

Security and storage

Effective data sharing depends on a strong network of trust between data providers and consumers. Infrastructure for data sharing will not be used if the parties who provide and use the data do not trust the infrastructure or one another. If sensitive data is to be shared, there must be provisions in the platform to ensure security of that data.

Whether data is closed or shared with specific individuals or organizations, it will need to be hosted in a controlled way. Depending on data sensitivity, this will include some guarantee of security, e.g. against hacking.

Data access and dissemination

Data production is one thing, its dissemination is another. Open data is useful when it can be delivered into the right hands (or the right machine) and within a context where it can be most valuable. In typical use cases of data for farmers, data is reused and delivered (through intermediaries) into the field, so it can be used to help farmers make informed decisions on which crop varieties to grow or which treatments to apply.

There must also be a variety of data delivery channels, fine-tuned to each case for data delivery. The 'fine-tuning' of data delivery channels can become a business opportunity for data intermediaries in the case where the data is fully 'open'. An intermediary can provide services to customise data delivery for the vast range of customers that might exist for the data. Open data creates the possibility of a marketplace, where alternative sources of relevant data are available.

To be made available, data has to be stored in a way that makes it accessible. Even in the modern era of cloud deployment, the data and applications are stored on some hardware somewhere. A strategy for sharing data on a global scale must specify where it will be stored, and what service level agreements will be maintained (up time, throughput, access controls, etc).

The following should strongly be considered when deciding how to disseminate data:

- access to the data should be provided in line with the organization's data policy and the national laws/acts on access to information;
- access to data should be granted without infringing the copyright or intellectual property rights of the data or any statutory/departmental obligations.

Dynamic vs. static datasets

Dynamic data denotes data that is periodically updated and asynchronously changed as further updates become available. The opposite of this is static data, also referred to as persistent data, which is infrequently accessed and not likely to be modified. Dynamic data is different from streaming data in that there is not a constant flow of information; rather, updates may come at any time, with periods of inactivity in between.

The rise of digital agriculture is changing the way farmers manage their land and livestock, such as with satellite-driven geo-positioning systems and sensors that detect nutrients and water in soil. These technologies ultimately result in the collection of more dynamic data, which is processed automatically.

Static datasets are normally stored and catalogued as described in the previous chapters and are exposed/published as downloads from an FTP server or a website. This can be done using specialised dataset management tools that combine storage, cataloguing and publication, but normally entails the manual upload of the dataset and its subsequent versions.

Dynamic datasets instead require some more effective form of data automation. Data automation is the process of updating data on a data portal programmatically.

Data automation includes the automation of the three steps that Extract, Transform, and Load the data (Extract–Transform–Load, ETL (Wikipedia, 2020i):

- Extract: the process of extracting data from one or many source systems;
- Transform: the process of transforming data into the necessary structure, such as a flat file format like a CSV; this could also include things like normalising values or applying schemas;
- Load: the process of loading the data into the final system, be it a data portal or the backend of an API for other machines.

Dynamic datasets are usually exposed through APIs, which allow to take data and functionality that may already be available on a website and make them available through a programmatic API that both web and mobile applications can use by just calling a URL. Then, instead of returning HTML to represent the information like a website would, an API returns data in a machine-readable format, like XML or JSON (see Section 3.1. Using open data on file type extensions and API).

Developers can then take this data and use it in web and mobile applications. However, XML and JSON are easily consumed by spreadsheets and other tools non-developers can use as well, making APIs accessible by potentially anyone.

The reason why dynamic datasets are more accessible via APIs is that downloading a static file is not needed. Instead, it is possible to query the data directly and get a file that is created 'on the fly'.

ODI suggests (The Open Data Institute (ODI), 2013a) that the technical documentation with an API should include:

- documentation about the data formats that being providing, possibly including schemas for any vocabularies that used;
- code lists that provide more details about each of the codes used within the data. One way to provide this information is to have a URL that provides documentation about each code and to link to that URL within the data;
- service documentation that describes the way any API provided works; this might include links to machine-readable service descriptions if applicable.

Equipped with this information, reusers should be able to understand the data to be published and how to create applications that use it.

Examples of existing APIs relevant for agriculture:

- The IFPRI Food Security Portal contains over 40 indicators related to food security, commodity prices, economics, and human wellbeing (IFPRI, 2020). Much of this data is available for every country in the world and goes back over 50 years. IFPRI draws data from public, authoritative data sources like the World Bank, the FAO, UNICEF, and others, as well as their own data.
- The OpenWeatherMap collects data from professional and private weather stations and publishes it through an API for developers (OpenWeather, 2020). Today, it has more than 40 000 weather stations; most are professional stations which are installed in airports, large cities, etc.

Data management plans

The planning process for data management begins with a data planning checklist. A checklist later assists in the development of a data management plan. That checklist might include some or all of the following questions.

- What data will be collected or created, how will it be created, and for what purpose?
- How will any ethical issues be managed? How will copyright and intellectual property rights issues be managed?
- What file formats will be used? Are they non-proprietary, transparent and sustainable? What directory and file naming conventions will be used? Are there any formal standards that will be adopted? What documentation and metadata will accompany the data?
- How will the data be stored and backed up? How will access and security be managed? Who will be responsible for data management?
- Are there existing procedures on which to base the approach? For example, are there institutional data protection or security policies to follow, or guidelines/best practices/codes of conduct?
- What is the long-term preservation plan for the dataset? For example, which data should be retained, shared, and/or preserved? How will the data be shared, and are any restrictions on data sharing required?
- What resources are required to deliver the plan? For example, are there tools or software needed to create, process or visualise the data?

It is important to note that data management plans must be continuously maintained and kept up-to-date throughout the course of a project or research.

4.1.2 Data management catalog software

Generally, a data repository serves as a central location to find data, a venue for standardising practices, and a showpiece of the use of that data. In a practical sense, a repository serves as a central, searchable place for people to find data. Some repository software will automatically convert data from one format to others, so even if data is provided in only one format (e.g., CSV), it will generate data in XML, JSON, Excel, etc.

Some repository software will visualise datasets in the browser, letting people map, sort, search, and combine datasets, without requiring any knowledge of programming. Others allow syndication, permitting other organizations to automatically incorporate data (e.g. a state transportation agency could gather up all localities' transportation data and republish it). Some tools also comply with existing metadata standards although, in most cases, only with generic vocabularies like Dublin Core but rarely with published standards for data catalogs.

Generally, repository software supports either uploading files to be stored in the repository or pointing the repository to an existing website address where the file lives. If the tool has been designed to also facilitate collection of data in the field, it will probably have a data collection module, preferably optimised for mobiles.

Data catalog software

To make a broad division, there are two types of data catalog software: third-party/cloud hosted or self-hosted (to be deployed on a server), and both can be either paid or free.

Self-hosted, free, open source

There are some excellent open source data repository programmes that are solid options for technically savvy organizations, for organizations with a commitment to use the open source software, or for organizations with the budget to hire a consultant to deploy the software.

- *CKAN:* CKAN is nominally an acronym for 'Comprehensive Knowledge Archive Network,' but it is only ever referred to as CKAN. A creation of UK-based Open Knowledge, CKAN is the most commonly used open source data repository software. It is written in Python and is the standard-bearer for repository software. Lamentably, it is also known for being difficult to install, although Docker images have simplified this substantially. CKAN users include Data.gov and the National Oceanic and Atmospheric Administration, among many others. CKAN consultants include Open Knowledge, Ontodia, and Accela, in addition to many independent consultants. Paid CKAN hosts include Open Knowledge and Ontodia. Please check a CKAN demo site at **demo.ckan.org**.
- *DKAN:* DKAN is a clone of CKAN, although it shares no code with CKAN – it has been rewritten in PHP, as a Drupal module. For an organization that uses the Drupal content management system and also wants a data repository, DKAN is an especially good option. DKAN users include the USDA, among others. Please check a DKAN demo site at **demo.getdkan.org**.
- *JKAN:* JKAN is nominally based on CKAN, although it shares no code with it. JKAN was created by Tim Wisniewski, Philadelphia's Chief Data Officer, as a data catalog powered by Jekyll. Note that JKAN is a data catalogue, not a repository, which is to say that it stores links to data and metadata about that data, but not the data itself. The data could be hosted on an FTP server, in-place on agency websites, in Amazon S3, in Dropbox, or anywhere else one might store a file for public access. Setting up a site takes just a few minutes. Please check a JKAN demo site at **https://demo.jkan.io**.
- *Open Data Kit (ODK):* Free and open source software to collect, manage, and use data.
- *QGIS:* Free and open source geographic information system specialized for geospatial data.

Cloud-hosted, commercial

For some organizations, commercial hosting is going to be a viable option. Paying somebody to host data requires little to no technical knowledge on the part of an organization, and the host will provide support through the process. An organization will not have to provide any technical infrastructure (e.g., servers) or know how to program (although some of these platforms also have a self-hosted version). It is important, however, to carefully consider the service level agreements.

- *ArcGIS Open Data:* ArcGIS Open Data is a new entrant in the field, having been released in late 2014. ArcGIS Open Data is included with an ArcGIS Online contract – because of the universality of that service among municipalities and states, it is effectively free for those existing customers. This makes it a very attractive option for governments with low levels of buy-in to an open data program, because it eliminates the cost of a data catalogue. ArcGIS Open Data is only available as hosted software – it is not possible to run an instance of it on another server.

- *Junar:* Junar provides platforms and packages for businesses, governments, NGOs, and academia with a focus on data collection and analysis. Junar is bilingual, supporting English and Spanish audiences. Their pricing is targeted at small- to medium-sized organizations, starting at around USD 10 000. Junar's demo site is available upon request.
- *NuCivic Data:* NuCivic Data is based on DKAN, which was created and is maintained by nücivic. They are a mid-range provider, in terms of pricing – their rates are much lower than Socrata, but more expensive than, for example, Junar.
- *CivicDashboards:* Open data consulting firm Ontodia provides hosted CKAN under the CivicDashboards banner. They offer a free tier for storing a small number of datasets. Their pricing is comparable with Junar's.
- *OpenDataSoft:* OpenDataSoft is a French company that has moved into the market in the United States of America recently. They offer a free tier (up to 5 datasets, each of up to 20 000 records).
- *Socrata Open Data:* Socrata is the major vendor in the open data repository space, with their Socrata Open Data platform. Socrata only offers hosted options – there is no way to run Socrata's software on other servers. It is both the most feature-rich and the most expensive option, with plans running into hundreds of thousands of dollars a year.

More oriented towards distributed data collection, especially through mobile phones:

- *ONA,* with special features for data collection from smartphones and data visualisations. They also have a free plan.

Cloud-hosted, free

There are some options available for free hosting of open data repositories. (Note that the above-listed open source options are also free, but require set-up, a server, and maintenance time.) Besides, there is often a free option for the lowest tier of service provided by paid hosted services like the ones mentioned above.

- *DataHub:* The Open Knowledge Foundation provides DataHub, a free, CKAN-based data host. It is a large, collective repository – users do not get their own site, although it is possible to list only one's own data and share a URL that only lists those datasets.
- *GitHub:* GitHub is not really meant as a data repository, but it can serve as one. It has none of the niceties of proper repository software (conversion of formats, retrieving data from remote URLs, etc.), but it does offer previews of some types of data, tracks changes publicly, and is a reasonable place to store datasets. It does offer one significant advantage, which is that GitHub – unlike any other repository software – provides a mechanism for people to propose changes to datasets, which can be accepted or declined, if they spot mistakes or areas for enhancement.
- *JKAN on GitHub:* JKAN is designed to be deployed on GitHub, where the resulting data catalogue can be hosted for free. In this way, GitHub can serve as a free host without sacrificing the niceties of a data catalogue.

More oriented towards distributed data collection, through flexible web forms:

- *Kobo toolbox,* a suite of tools for field data collection for use in challenging environments. Besides the cloud service, the tool can also be self-hosted using Docker.

In the most extreme cases, the security requirements for shared data in agriculture could be as severe as for shared data in the military. These principles are not unique to agricultural data and have been studied in depth.

The basic concepts behind these principles are that: services should be hard to compromise; a compromise should be easy to detect; and the impact of a compromise can be contained. For open data, this is much less of a concern but, to build trust among data providers, some support for data security must be in place.

Data access and dissemination

Data production is one thing, its dissemination is another. Open data is useful when it can be delivered into the right hands (or the right machine) and within a context where it can be most valuable. In typical use cases of data for farmers, data is reused and delivered (through intermediaries) into the field, so it can be used to help farmers make informed decisions on which crop varieties to grow or which treatments to apply.

There must also be a variety of data delivery channels, fine-tuned to each case for data delivery. The 'fine-tuning' of data delivery channels can become a business opportunity for data intermediaries in the case where the data is fully 'open'. An intermediary can provide services to customise data delivery for the vast range of customers that might exist for the data. Open data creates the possibility of a marketplace, where alternative sources of relevant data are available.

To be made available, data has to be stored in a way that makes it accessible. Even in the modern era of cloud deployment, the data and applications are stored on some hardware somewhere. A strategy for sharing data on a global scale must specify where it will be stored, and what service level agreements will be maintained (up time, throughput, access controls, etc).

The following should strongly be considered when deciding how to disseminate data:

- access to the data should be provided in line with the organization's data policy and the national laws/ acts on access to information;
- access to data should be granted without infringing the copyright or intellectual property rights of the data or any statutory/departmental obligations.

Dynamic vs. static datasets

Dynamic data denotes data that is periodically updated and asynchronously changed as further updates become available. The opposite of this is static data, also referred to as persistent data, which is infrequently accessed and not likely to be modified.

Dynamic data is different from streaming data in that there is not a constant flow of information; rather, updates may come at any time, with periods of inactivity in between.

The rise of digital agriculture is changing the way farmers manage their land and livestock, such as with satellite-driven geo-positioning systems and sensors that detect nutrients and water in soil. These technologies ultimately result in the collection of more dynamic data, which is processed automatically.

Static datasets are normally stored and catalogued as described in the previous chapters and are exposed/ published as downloads from an FTP server or a website. This can be done using specialised dataset management tools that combine storage, cataloguing and publication, but normally entails the manual upload of the dataset and its subsequent versions.

Dynamic datasets instead require some more effective form of data automation. Data automation is the process of updating data on a data portal programmatically.

Data automation includes the automation of the three steps that Extract, Transform, and Load the data (Extract–Transform–Load, ETL (Wikipedia, 2020i):

- *Extract:* the process of extracting data from one or many source systems;
- *Transform:* the process of transforming data into the necessary structure, such as a flat file format like a CSV; this could also include things like normalising values or applying schemas;
- *Load:* the process of loading the data into the final system, be it a data portal or the backend of an API for other machines.

Dynamic datasets are usually exposed through APIs, which allow to take data and functionality that may already be available on a website and make them available through a programmatic API that both web and mobile applications can use by just calling a URL. Then, instead of returning HTML to represent the information like a website would, an API returns data in a machine-readable format, like XML or JSON (see Section 3.1. Using open data on file type extensions and API).

Developers can then take this data and use it in web and mobile applications. However, XML and JSON are easily consumed by spreadsheets and other tools non-developers can use as well, making APIs accessible by potentially anyone.

The reason why dynamic datasets are more accessible via APIs is that downloading a static file is not needed. Instead, it is possible to query the data directly and get a file that is created 'on the fly'.

ODI suggests (The Open Data Institute (ODI), 2013a) that the technical documentation with an API should include:

- documentation about the data formats that being providing, possibly including schemas for any vocabularies that used;
- code lists that provide more details about each of the codes used within the data. One way to provide this information is to have a URL that provides documentation about each code and to link to that URL within the data;
- service documentation that describes the way any API provided works; this might include links to machine-readable service descriptions if applicable.

Equipped with this information, reusers should be able to understand the data to be published and how to create applications that use it.

Examples of existing APIs relevant for agriculture:

- The IFPRI Food Security Portal contains over 40 indicators related to food security, commodity prices, economics, and human wellbeing (IFPRI, 2020). Much of this data is available for every country in the world and goes back over 50 years. IFPRI draws data from public, authoritative data sources like the World Bank, the FAO, UNICEF, and others, as well as their own data.
- The OpenWeatherMap collects data from professional and private weather stations and publishes it through an API for developers (OpenWeather, 2020). Today, it has more than 40 000 weather stations; most are professional stations which are installed in airports, large cities, etc.

Data management plans

The planning process for data management begins with a data planning checklist. A checklist later assists in the development of a data management plan. That checklist might include some or all of the following questions.

- What data will be collected or created, how will it be created, and for what purpose?
- How will any ethical issues be managed? How will copyright and intellectual property rights issues be managed?
- What file formats will be used? Are they non-proprietary, transparent and sustainable? What directory and file naming conventions will be used? Are there any formal standards that will be adopted? What documentation and metadata will accompany the data?
- How will the data be stored and backed up? How will access and security be managed? Who will be responsible for data management?
- Are there existing procedures on which to base the approach? For example, are there institutional data protection or security policies to follow, or guidelines/best practices/codes of conduct?
- What is the long-term preservation plan for the dataset? For example, which data should be retained, shared, and/or preserved? How will the data be shared, and are any restrictions on data sharing required?
- What resources are required to deliver the plan? For example, are there tools or software needed to create, process or visualise the data?

It is important to note that data management plans must be continuously maintained and kept up-to-date throughout the course of a project or research.

4.1.2 Data management catalog software

Generally, a data repository serves as a central location to find data, a venue for standardising practices, and a showpiece of the use of that data. In a practical sense, a repository serves as a central, searchable place for people to find data. Some repository software will automatically convert data from one format to others, so even if data is provided in only one format (e.g., CSV), it will generate data in XML, JSON, Excel, etc.

Some repository software will visualise datasets in the browser, letting people map, sort, search, and combine datasets, without requiring any knowledge of programming. Others allow syndication, permitting other organizations to automatically incorporate data (e.g. a state transportation agency could gather up all localities' transportation data and republish it). Some tools also comply with existing metadata standards although, in most cases, only with generic vocabularies like Dublin Core but rarely with published standards for data catalogs.

Generally, repository software supports either uploading files to be stored in the repository or pointing the repository to an existing website address where the file lives. If the tool has been designed to also facilitate collection of data in the field, it will probably have a data collection module, preferably optimised for mobiles.

Data catalog software

To make a broad division, there are two types of data catalog software: third-party/cloud hosted or self-hosted (to be deployed on a server), and both can be either paid or free.

Self-hosted, free, open source

There are some excellent open source data repository programmes that are solid options for technically savvy organizations, for organizations with a commitment to use the open source software, or for organizations with the budget to hire a consultant to deploy the software.

- *CKAN:* CKAN is nominally an acronym for 'Comprehensive Knowledge Archive Network,' but it is only ever referred to as CKAN. A creation of UK-based Open Knowledge, CKAN is the most commonly used open source data repository software. It is written in Python and is the standard-bearer for repository software. Lamentably, it is also known for being difficult to install, although Docker images have simplified this substantially. CKAN users include Data.gov and the National Oceanic and Atmospheric Administration, among many others. CKAN consultants include Open Knowledge, Ontodia, and Accela, in addition to many independent consultants. Paid CKAN hosts include Open Knowledge and Ontodia. Please check a CKAN demo site at **demo.ckan.org**.
- *DKAN:* DKAN is a clone of CKAN, although it shares no code with CKAN – it has been rewritten in PHP, as a Drupal module. For an organization that uses the Drupal content management system and also wants a data repository, DKAN is an especially good option. DKAN users include the USDA, among others. Please check a DKAN demo site at **demo.getdkan.org**.
- *JKAN:* JKAN is nominally based on CKAN, although it shares no code with it. JKAN was created by Tim Wisniewski, Philadelphia's Chief Data Officer, as a data catalog powered by Jekyll. Note that JKAN is a data catalogue, not a repository, which is to say that it stores links to data and metadata about that data, but not the data itself. The data could be hosted on an FTP server, in-place on agency websites, in Amazon S3, in Dropbox, or anywhere else one might store a file for public access. Setting up a site takes just a few minutes. Please check a JKAN demo site at **https://demo.jkan.io**.
- *Open Data Kit (ODK):* Free and open source software to collect, manage, and use data.
- *QGIS:* Free and open source geographic information system specialized for geospatial data.

Cloud-hosted, commercial

For some organizations, commercial hosting is going to be a viable option. Paying somebody to host data requires little to no technical knowledge on the part of an organization, and the host will provide support through the process. An organization will not have to provide any technical infrastructure (e.g., servers) or know how to program (although some of these platforms also have a self-hosted version). It is important, however, to carefully consider the service level agreements.

- *ArcGIS Open Data:* ArcGIS Open Data is a new entrant in the field, having been released in late 2014. ArcGIS Open Data is included with an ArcGIS Online contract – because of the universality of that service among municipalities and states, it is effectively free for those existing customers. This makes it a very attractive option for governments with low levels of buy-in to an open data program, because it eliminates the cost of a data catalogue. ArcGIS Open Data is only available as hosted software – it is not possible to run an instance of it on another server.
- *Junar:* Junar provides platforms and packages for businesses, governments, NGOs, and academia with a focus on data collection and analysis. Junar is bilingual, supporting English and Spanish audiences. Their pricing is targeted at small- to medium-sized organizations, starting at around USD 10 000. Junar's demo site is available upon request.
- *NuCivic Data:* NuCivic Data is based on DKAN, which was created and is maintained by nücivic. They are a mid-range provider, in terms of pricing – their rates are much lower than Socrata, but more expensive than, for example, Junar.
- *CivicDashboards:* Open data consulting firm Ontodia provides hosted CKAN under the CivicDashboards banner. They offer a free tier for storing a small number of datasets. Their pricing is comparable with Junar's.
- *OpenDataSoft:* OpenDataSoft is a French company that has moved into the market in the United States of America recently. They offer a free tier (up to 5 datasets, each of up to 20 000 records).
- *Socrata Open Data:* Socrata is the major vendor in the open data repository space, with their Socrata Open Data platform. Socrata only offers hosted options – there is no way to run Socrata's software on other servers. It is both the most feature-rich and the most expensive option, with plans running into hundreds of thousands of dollars a year.

More oriented towards distributed data collection, especially through mobile phones:

- *ONA*, with special features for data collection from smartphones and data visualisations. They also have a free plan.

Cloud-hosted, free

There are some options available for free hosting of open data repositories. (Note that the above-listed open source options are also free, but require set-up, a server, and maintenance time.) Besides, there is often a free option for the lowest tier of service provided by paid hosted services like the ones mentioned above.

- *DataHub:* The Open Knowledge Foundation provides DataHub, a free, CKAN-based data host. It is a large, collective repository – users do not get their own site, although it is possible to list only one's own data and share a URL that only lists those datasets.
- *GitHub:* GitHub is not really meant as a data repository, but it can serve as one. It has none of the niceties of proper repository software (conversion of formats, retrieving data from remote URLs, etc.), but it does offer previews of some types of data, tracks changes publicly, and is a reasonable place to store datasets. It does offer one significant advantage, which is that GitHub – unlike any other repository software – provides a mechanism for people to propose changes to datasets, which can be accepted or declined, if they spot mistakes or areas for enhancement.
- *JKAN on GitHub:* JKAN is designed to be deployed on GitHub, where the resulting data catalogue can be hosted for free. In this way, GitHub can serve as a free host without sacrificing the niceties of a data catalogue.

More oriented towards distributed data collection, through flexible web forms:

- *Kobo toolbox,* a suite of tools for field data collection for use in challenging environments. Besides the cloud service, the tool can also be self-hosted using Docker.

4.1.3 Data sharing policies

Data exists on a spectrum: it can be closed, shared, or open. Closed data can only be accessed by the owners or the contracting parties. Shared data can be accessed on specific conditions (authentication, payment). Open data is data that anyone can access, use and share. In the spectrum, there are intermediate solutions. Whatever the position in the spectrum, it is important that data sharing conditions are clarified in a data policy.

A growing number of public and private-sector organizations are adopting the open data approach and are drafting open data policies. However, in the agricultural value chain, data cannot always be fully open: in Chapter 2 the sensitivities around various types of data shared in the value chain was covered. In such cases, a data sharing policy is equally, if not more important.

A good data sharing policy will increase the transparency of an organization and ensure the best use of its data. The translation of a data strategy into a solid policy is of great importance to ensure its successful implementation. Some elements that will be in an open data policy may be different from the elements in a non-open data sharing policy, but the elements highlighted below can be a good starting point for any data sharing policy.

A good (open) data policy will include some general context that helps to define its scope, for example:

- a definition of the general data strategy (whether it's fully open or not) – why it is important to the organization and the reasons for defining a policy;
- a general declaration of principles that should guide the release and reuse of the data;
- an outline of the types of data collected by the organization and whether they are covered by the policy
- references to any relevant legislation, policies or other guidance, which also apply to the management and sharing of information with third parties.

And it will then consider the following elements:

- the approach to identifying and prioritising data for release: how will data be inventoried, reviewed and then released?
- privacy considerations: ensuring that personal information is not shared and recommending steps to mitigate, e.g. by undertaking privacy impact assessments or approaches to anonymisation;
- data ownership, copyright, consent to reuse: especially important when reusing data collected from other actors;
- data licensing and reuse rights: this will include not only the licence under which data will be released, but also the importance of clearing rights during data collection;
- data publishing standards: ensuring that data is shared in well-structured, machine-readable formats, with clear metadata and documentation;
- engaging with reusers: how the organization will work with external stakeholders to help guide release of data and ensure it can be easily used;
- measuring success: what metrics the organization will use to measure whether the policy is successful and how these measures will be shared;
- approach to consuming open data: for organizations that are reusing open data, guidance on how to identify high-quality datasets and ensure reuse rights are clear;
- concrete commitments: what the organization is committing to do, in concrete terms, over the timespan of the policy;
- policy transparency: how the policy and the processes it describes will be reviewed based on feedback from stakeholders and lessons learned.

A policy document will not necessarily include detailed information on each of these areas, e.g. specific standards or release processes. It will instead focus on general principles that should be followed and which may inform the drafting of more detailed guidance for practitioners.

4.1.4 Ensuring reuse of exposed data

Data and services built on data will only have an impact if they are used by actors in society to achieve some of their goals. In particular, data shared in agricultural value chains has to be reused by actors, ultimately the farmer, to achieve their business objectives to improve the value chain itself and to ensure transparency for the public good.

For commercial actors sharing (selling) data and services, acquiring users for their services is essential, but also public providers need to demonstrate that their data makes an impact.

Some recommendations on how to make data useful and reusable have been given already in Section 1.1. Data for agriculture, especially regarding data quality. Those recommendations were inherent to the data, while other actions can be taken once the data is published to maximise its reuse.

The first recommendation is to engage with users to make sure the data meets their needs and they have all the information on how to use it. Building on Tim Berners-Lee's Five Stars for Open Data (Berners-Lee, 2012), Tim Davis has developed a five-star open data engagement model. The model suggests five steps to engage with users that will ensure that shared data is reused:

"★ Be demand driven

★★ Put data in context

★★★ Support conversation around data

★★★★ Build capacity, skills and networks

★★★★★ Collaborate on data as a common resource" (Davies, 2012)

Other useful suggestions for engaging with data users are:

- providing tools, such as plugins, visualisations, software libraries and services that enable reusers to build on others' work with the data (these tools are sometimes built by third parties);
- blogging to showcase good examples of use of data;
- arranging hackdays or competitions to encourage the use of data;
- Whilst not worthwhile for all datasets, in some cases explicit community building, engagement and outreach work can help to maximise the value that data brings to an organization and to others.

4.2 Guiding frameworks for data sharing

Guiding frameworks for data sharing introduces the most important conceptual frameworks on data sharing. Chapter 2 presented ethical and legal considerations on data sharing and related policy instruments, while the frameworks presented in Section 4.2. Guiding frameworks for data sharing cover mainly technical aspects of data sharing. They define what is meant by open or accessible data, how data should be made technically accessible, and when it comes to legal conditions and data rights, they focus more on how to formalise such conditions and rights in the data itself.

In the context of agricultural value chains and, in particular, regarding data from and for farmers, awareness of these frameworks is useful for a number of reasons:

- Analysing the existing frameworks clarifies the distinction between *'fully open'* or *'open by default'* – mostly applicable to public data like government data and publicly-funded research data, or private-sector data that is deemed suitable for public reuse, like Google Earth data, Twitter feeds etc., and *'shared data'*, which is conditionally shared, e.g. behind authorisation or payment or contract, which is a more common in the agricultural value chain. For professionals working for farmers, fully open data methodologies are important mostly as open data users, in cases when they have to reuse open data in their services, while shared data practices are relevant for any service that needs to share data and wants to do it following best practices, independent of access conditions.
- These frameworks recommend methodologies to formulate access rights in the data, so they describe useful techniques to make access rights clear and understandable by machines, as well as the most suitable access protocols.
- Independent of the access conditions, these frameworks recommend technical methodologies for making data reusable for those who can access it. Data authorisation layer apart, the 'rules' to make data reusable are the same for open and shared data.

The frameworks that are presented are: Tim Berners-Lee's Five Star open data deployment scheme, the FAIR principles, the World Wide Web Consortium (W3C) Data on the Web Best Practices, and the general Open Access paradigm.

4.2.3. Practical recommendations for applying the data sharing principles describes in more detail a few selected practical recommendations from the described frameworks, namely on licensing, repository infrastructure (protocols, persistence and identifiers), data formats, metadata and linking.

©Adobe Stock/Kara

4.2.1 Guiding frameworks for data: from open to FAIR

Data sharing is something that has always happened, in different ways over the centuries. In agriculture, as in other areas, data exchange is much needed for innovation and business (e.g. market prices, weather information, soil data, crop growth data), as well as for policy and regulations (e.g. tracking of food products, use of pesticides). The stakeholders involved are therefore varied and their interests are sometimes aligned and sometimes not but, in the end, they all need data from other actors or from other parts of the data value chain.

A key factor that has changed the game of data sharing in the last decades is, of course, the advent of the internet and, more recently, the cloud and big-data technologies that have multiplied the potential of data processing power. At the beginning of the web its inventor Tim Berners-Lee saw it as a 'web of data' and, over the last decade, everybody has come to acknowledge that the primary consumers of data and the actual intermediaries in the data sharing process are machines and therefore, in order to be shared, data must be machine-readable.

On the other hand, the ease of sharing data on the web brings with it legal concerns about the use that can be made of those data: the fact that they are on the web only means that they can be read, but nothing is said about permission to reuse, modify or re-distribute. If permissions are not made explicit, data cannot legally be reused (see Chapter 2 Data sharing principles for more on this topic).

Sharing data therefore requires agreements on how data should be 'written' for machines because machines have to be programmed to read the data and have to know the rules. It also requires agreement on how data can be used and whether they can be modified or redistributed. Agreeing on this can be just a matter of ad hoc agreements or technical/legal specifications or, with a more ambitious and long-term view, it can lead to the creation of general authoritative guiding frameworks that become widely endorsed. Section 4.2. Guiding frameworks for data sharing will introduce the two major frameworks that have been set out and endorsed for sharing data.

Tim Berners-Lee's 5-star deployment scheme for open data

The first guiding framework for sharing data was designed by the inventor of the web, Sir Tim Berners-Lee (TBL). The technical framework he designed for the web of data is the 'Linked Open Data' (LOD) or simply Linked Data (Berners-Lee, 2006) good practice, which was fully formalised in 2006; it consists of very technical guidelines to make data fully linked and it basically recommends:

- use uniform resource identifiers (URIs)[15] as names for things;
- use HTTP URIs so that people can look them up;
- when someone looks up a URI, provide useful information; and
- include links to other URIs, so that they can discover more things.

However, given the high entry barrier to such technological solutions, around 2010, TBL published the more comprehensive 5-star open data deployment scheme 'in order to encourage people -- especially government data owners -- along the road to good linked data' (Berners-Lee, 2013). The 5-star scheme illustrates the 'continuum' in data publishing that leads to the final steps of fully linked open data. The LOD framework is only a subset (last two starts) of the full scheme.

[15] Uniform Resource Identifier (URI) is a string of characters that unambiguously identifies a particular resource (Wikipedia, 2020s).

Table 17. Tim Berners-Lee's 5-star open data deployment scheme

★	Make your stuff available on the web (whatever format) under an open licence
★★	Make it available as structured data (e.g. Excel instead of image scan of a tab)
★★★	Non-proprietary format (e.g. CSV instead of Excel)
★★★★	Use URIs to identify things, so that people can point at your stuff
★★★★★	Link your data to other people's data to provide context

TBL's 5-stars are still a reference framework for everybody working on open data. They have been interpreted as 'cumulative', in that 'each additional star presumes the data meets the criteria of the previous step(s)'. This puts enormous weight on the first step, an open licence, as a requirement, as in its absence, implementing the other four stars would not result in really open data. All the other stars relate to the interoperability of data, while the first star is all about openness for reuse. TBL's 5-stars are really a framework for openness and hence they are also called 'the 5-stars of openness'.

[To be more precise, to distinguish between Linked Data and Linked Open Data, TBL said: *'Linked Data does not of course in general have to be open -- there is a lot of important use of linked data internally, and for personal and group-wide data. You can have 5-star Linked Data without it being open. However, if it claims to be Linked Open Data then it does have to be open to get any star at all'* (Berners-Lee, 2006).]

The concept of 'open' has been the cornerstone of all initiatives on knowledge and data sharing for decades. However, in recent years some aspects of this concept have been potentially discouraging the sharing of data, especially in the context of data-intensive research and data transmitted across the various steps of the data value chain, because they are so strict:

- The definition of 'open licence' is very strict: in their basic course on open data the ODI says that *"this licence must permit people to use the data in any way they want, including transforming, combining and sharing it with others, even commercially"* (The Open Data Institute (ODI), 2019b). The requirement of an open licence could prevent data sharing that has some access restrictions but can still be reused in large communities or through simple agreements and can therefore still have a big impact.
- The fourth and fifth stars, especially in versions that describe each star further, are sometimes seen as too much tied to the Resource Description Framework (RDF) technical approach, instead of being generic principles that can be implemented with any technology. See 4.3.4. Structural interoperability: formats and structures for more information on RDF.

This does not mean that TBL's framework has been superseded: it is still the reference framework for high data interoperability and for a loosely coupled, bottom-up open web of data.

The FAIR principles

Recently, since the reuse of data is unanimously recognised as a big driver for innovation, and the way data is shared is key to its reuse, there has been new interest around a definition of a more formal and more coordinated framework that could cater more for data-intensive research and data sharing across the data value chain.

In 2014, the need to better define the 'rules' for more effective sharing of data led a group of different stakeholders – academia, industry, publishers, funding agencies – to meet in Leiden, The Netherlands and discuss a *"minimal set of community-agreed guiding principles and practices"* (FORCE11, 2014).

What came out of these discussions was a set of principles called the FAIR principles, according to which data must be: Findable, Accessible, Interoperable and Reusable (FAIR).

By just reading the four principles, this framework seems very much in line with the concept of open and TBL's 5-star scheme (indeed there is no discontinuity between the two, and the FAIR framework is not contradicting or replacing the TBL 5-star framework). However, when reading the details about each of the principles, there are some key differences and a higher level or generalisation, which make the FAIR principle more of a formal 'framework':

1. The focus of the FAIR principles is on clear access rights rather than openness. This in line with the need for more flexibility and more precision in data sharing in order to facilitate access restrictions on data. It can still be reused under specific conditions.
2. There is a special attention for provenance and attribution and for persistence, in line with the fact that the FAIR principles have been agreed by a community that wants to work together and share data and needs a trusted environment with some basic rules. From this perspective, the FAIR principles need more rigorous rules when it comes to trust than the more bottom-up open web of data designed by TBL.

3. The principles do not want to go to the level of technical specifications; they are: *"a general 'guide to FAIRness of data', not a 'specification'. In compiling the FAIR guiding principles for this document, technical implementation choices have been consciously avoided. The minimal [FAIR Guiding Principles] are meant to guide implementers of FAIR data environments in checking whether their particular implementation choices are indeed rendering the resulting data FAIR."*

4. In contrast to TBL's 5-stars, the FAIR principles are not 'cumulative': *"These FAIR facets are obviously related, but technically somewhat independent from one another, and may be implemented in any combination, incrementally, as data providers […] evolve to increasing degrees of FAIR-ness."*

Here are some of the key details about each principle that exemplify the features described above:

1) Some aspects of the Findable principle:

a) Data Objects should be persistent, with emphasis on their metadata.
b) Identifiers for any concept used in Data Objects should therefore be unique and persistent.

2) Some aspects of the Accessible principle:

a) 'upon appropriate authorisation' – this means that even data that require special permission are considered FAIR if they apply the other principles, so data does not have to necessarily be usable 'by anyone'.
b) 'through a well-defined protocol' – this means that interoperability is not tied to one protocol (e.g. HTTP and more specifically a REST API or a SPARQL endpoint, which underlie the Linked Data framework).

3) Some aspects of the Interoperable principle:

a) '(Meta) data is machine-actionable.'
b)'(Meta) data formats utilise shared vocabularies and/or ontologies.'
c) '(Meta) data within the Data Object should thus be both syntactically parsable and semantically machine-accessible' – these two indications are, of course, perfectly implemented using RDF, but no specific technology is mentioned. In general, the whole interoperable principle is very much in line with TBL's stars and all the principles are based on the key assumption that data have to be *"fair for machines as well as people"* and that "metadata being machine-readable is a *conditio sine qua non* for FAIRness".

4) Among the practical indications that detail the Reusable principle, an important one is:

a) 'Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation' – this is in line with the objective of building an infrastructure of trusted data repositories where authorship and attribution are particularly important.

The FAIR principles are being rapidly adopted by many stakeholders, especially research funders. They have been recently adopted by the European Commission Guidelines on FAIR Data Management in Horizon 2020, European Union Framework for Research and Innovation.

Other general frameworks

There are other frameworks to be considered when publishing data, one that is more generic than the two previously mentioned (Open Access), one that is more strict on licences (Open Content) and one that is a very detailed technical 'best practice' with clear implementation choices (W3C Data on the Web Best Practices).

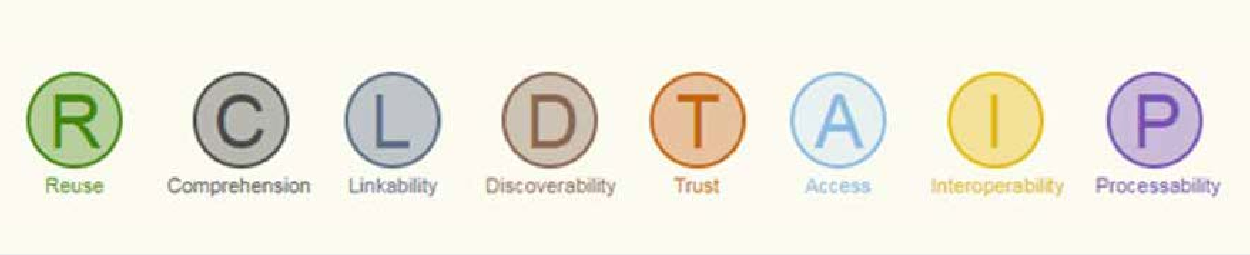
Open Access

The Open Access and Open Science movements advocate, respectively, for publication of resources (from journal articles to theses to datasets) in accessible ways (either via self-archiving or via open access journals) and for making 'scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional'.

Open access can be achieved by following either the green route (self-archiving) or the gold route (publishing in open access journals, some of which now also publish datasets) (Wikipedia, 2020j).

In new European Union-funded projects, open access is a requirement and datasets can be either self-archived in an open-access catalog or published in public catalogs like OpenAIRE [openaire.eu] or Dryad [**datadryad.org**].

Figure 17. The W3C benefits for applying the data on the web best practices



Source: W3C, 2017.

W3C Data on the Web

The W3C Data on the Web Best Practices are very detailed guidelines related to the publication and usage of data on the Web designed to help support 'a self-sustaining ecosystem' (W3C, 2017). Compared to the FAIR principles, these best practices go much deeper into the actual technical implementation and recommend specific solutions for publication and usage of data.

The approach to data sharing is very similar to that of the FAIR principles, highlighting the need to cater also for the publishing of data with controlled access, the need for the reliability and persistence of the data, and the need to agree on a set of common rules.

The Best Practices are all linked to a set of 'benefits': each benefit represents an improvement in the way datasets are available on the Web, see Figure 17.

It can be noted that the FAIR principles are all covered, but the W3C framework also covers one step back (human comprehension) and one step further (data processability). Some of the technical solutions recommended by the W3C Best Practices are presented in Section 4.3. Introduction to data interoperability.

Open Content

An even more open definition of what 'open' means is provided by the 5Rs for Open Content. Open content is *"any copyrightable work (traditionally excluding software, which is described by other terms like 'open source') that is licensed in a manner that provides users with free and perpetual permission to engage in the 5R activities: Retain, Reuse, Revise, Remix, Redistribute"* (Wiley, 2020).

An open data framework for agriculture

Besides these general frameworks for sharing data, there have been initiatives dedicated to advocacy for open data in specific sectors. For food and agriculture, after an agreement at the G8 International Conference on Open Data for Agriculture in 2012, the Global Open Data for Agriculture and Nutrition (GODAN) initiative was launched at the Open Government Partnership Conference in October 2013. The initiative focuses on building high-level support among governments, policy-makers, international organizations and business.

GODAN has a Statement of Purpose to which more than 500 partners have adhered so far, and the Statement is oriented towards fully open data:

"The Global Open Data for Agriculture and Nutrition (GODAN) initiative seeks to support global efforts to make agricultural and nutritionally relevant data available, accessible, and usable for unrestricted use worldwide." (GODAN Secretariat, 2020)

Related data evaluation tools

Some tools have been developed to assess the openness and/or fairness of data: two examples are the ODI certificates developed by ODI (The Open Data Institute (ODI), 2020), which is a tool to assess and recognise the sustainable publication of quality open data standards, building on frameworks such as **opendefinition.org**, 5-star Open Data, Sunlight principles[16], W3C DCAT[17] and the Data Seal of Approval (Core Trust Seal, 2020) by the Dutch Data Archiving and Networked Services (very much in line with the FAIR principles but more related to the quality of digital repositories, not individual datasets).

[16] This is a framework to assess the openness of government data.

[17] The W3C Data Catalog vocabulary, designed for data catalog and dataset metadata.

4.2.2 Practical recommendations for applying the data sharing principles

Some of the principles described above have implications in terms of data policy, but most of them have heavy implications in terms of technical implementation choices. Especially for the Interoperability principle, many of the implications are very technical and they will be discussed in more detail in Section 4.3. Introduction to data interoperability and Section 4.4. Interoperability of farm data. Section 4.2. Guiding frameworks for data sharing only provides some general recommendations that can help the data manager to either select appropriate tools or guide developers in the choice of technological solutions for applying the most important data sharing principles.

When it is useful to mention data catalog tools, examples with CKAN and Dataverse are used.

While the FAIR principles do not go into explicit technical implementation choices, TBL's 5-star scheme provides some examples and indicates specific technologies. But most of all, the document that can help in identifying technological solutions for publishing data is the W3C Data on the Web Best Practices (DWBP) mentioned above, which is often referenced in Section 4.2. Guiding frameworks for data sharing. It is a very technical document, but it contains advice for virtually everything that is needed for publishing data on the web, and it is in line with both TBL's 5-stars and the FAIR principles.

Licensing

As mentioned above, prescriptions for licensing are different in TBL's scheme (open licence) and in the FAIR framework (any licence that clarifies usage rights).

Of course, this is primarily a data policy choice – several factors have to be considered including: if the organization has a mandate to publish open data; if a dataset includes sensitive data; if the data has commercial value (or if it could fall within a pre-competitive space) etc. Different parts of datasets can be published under different licences or apply licences to specific data elements (following the FAIR 'modular' and 'recurrent' Data Object model).

Both approaches to licensing can, in any case, be implemented following DWBP Best Practice 4: Provide data licence information:

"In the context of data on the Web, the licence of a dataset can be specified within the data, or outside of it, in a separate document to which it is linked. […] It should be possible for machines to automatically detect the data licence of a distribution." (W3C, 2017)

DWBP also provides implementation examples: vocabularies[18] that can be used to express the licence metadata (Dublin Core Terms, schema.org) and even machine-readable rights languages such as the Creative Commons Rights Expression Language, the Open Data Rights Language, and the Open Data Rights Statement Vocabulary.

Ideally, in order to be precise, licence metadata would point to a URI or at least a URL of a published licence. For more on open data licensing, see Section 4.5. for data.

Data service infrastructure: protocols, persistence and identifiers

In this sub-section, Data service infrastructure: protocols, persistence and identifiers. readers are introduced to a number of architectural requirements that concern data service infrastructure[19] e.g. repository, access layer. These requirements are significant for the choice of technical solutions to implement a data service. Readers can use these requirements as evaluation criteria for selecting a third-party software platform.

The TBL 4th star simply recommends: *"use URIs to denote things, so that people can point at your stuff."*

The FAIR framework has much more demanding requirements for the data service:

- (meta)data are assigned a globally unique and eternally persistent identifier;
- (meta)data are retrievable by their identifier using a standardised communications protocol;
- the protocol is open, free, and universally implementable;
- the protocol allows for an authentication and authorisation procedure, where necessary;

- metadata are accessible, even when the data are no longer available.

[18] In the context of interoperability technologies, vocabularies are sets of terms that are prescribed to describe or classify resources. Vocabularies can be represented by a simple textual specification or, more appropriately, for interoperability purposes, in machine-readable schemas. They are used to improve semantic interoperability of data. See Section 4.3. Introduction to data interoperability for more information on semantic interoperability.

[19] "Data infrastructure consists of data assets supported by people, processes and technology." **https://theodi.org/topic/data-infrastructure**

©Adobe Stock/bugarskipavle3

Regarding protocols, DWBP has Best Practice 21: Use Web Standardised Interfaces – *"It is recommended to use URIs, HTTP verbs, HTTP response codes, MIME types, typed HTTP links and content negotiation when designing APIs."* The best practice explicitly recommends RESTful APIs.[20]

Regarding persistent identifiers, the most common practices are to use:

- Unique Resource Identifiers (URIs) that resolve to URLs
- Digital Object Identifiers (DOIs).

DWBP Best Practice 9: *"Use persistent URIs as identifiers of datasets"* is heavily based on the Linked Data framework and provides links to many technical documents on how to build URIs and how to ensure URI persistence. However, it also considers DOIs and suggests a way of merging the two approaches by appending the DOI to a URI pattern. *"Digital Object Identifiers (DOIs) offer a similar alternative. These identifiers are defined independently of any web technology but can be appended to a 'URI stub.' DOIs are an important part of the digital infrastructure for research data and libraries."*

While implementing all of the above is technically challenging, most existing data repository/catalog tools create either URIs (e.g. CKAN) or DOIs (e.g. Dataverse) for uploaded datasets and use open protocols. Besides HTTP, they use REST APIs like SPARQL or OAI-PMH.

However, the URI domain is a responsibility of the data service owner and applying the persistence principle will require some policy commitment to maintain the URI domain (or the DOIs). For URIs, DWBP suggests an alternative solution: *"Where a data publisher is unable or unwilling to manage its URI space directly for persistence, an alternative approach is to use a redirection service such as purl.org. This provides persistent URIs that can be redirected as required so that the eventual location can be ephemeral."* (W3C, 2017)

[20] "Representational state transfer (REST) or RESTful web services is a way of providing interoperability between computer systems on the Internet. REST-compliant web services allow requesting systems to access and manipulate textual representations of web resources using a uniform and predefined set of stateless operations." (Wikipedia, 2020g).

Data formats

Regarding data formats, there are easy recommendations that will comply with the principles of all the described frameworks:

- TBL third star recommends: "make [content] available in a non-proprietary open format (e.g., CSV instead of Excel)" (Berners-Lee, 2012).
- FAIR: Interoperability point 1 recommends: *"(meta) data use a formal, accessible, shared, and broadly applicable language for knowledge representation."* (FORCE11, 2014)

Basically, data have to be 'serialised', exposed, in a machine-readable format, possibly without the need for proprietary software. There are formats that are technically machine-readable (HTML, Excel), but they are not necessarily easy to parse, either because they are not rigorously structured or because the algorithms to parse them are proprietary. Note, parsing means splitting a file or other input into pieces of data that can be easily stored or manipulated.

The most complete recommendation in this sense is the W3C DWBP Best Practice 12: Use machine-readable standardised data formats: *"Make data available in a machine-readable standardised data format that is easily parsable including but not limited to CSV, XML, Turtle, NetCDF, JSON and RDF."* (W3C, 2017)

Most existing data catalog tools expose data in RDF (e.g. CKAN) or in JSON or XML (e.g. Dataverse).

Metadata

TBL's 5-stars do not mention metadata. The FAIR principles lay out a fundamental role for metadata, for findability:

- data are described with rich metadata;
- metadata specify the data identifier.

and for reusability:

- meta(data) have a plurality of accurate and relevant attributes;
- (meta)data are released with a clear and accessible data usage licence;
- (meta)data are associated with their provenance;
- (meta)data meet domain-relevant community standards.

One thing to note about metadata is that if a third-party software tool is used, these tools may come with their metadata models and their metadata exposure layer. This means that in most cases, modifying them means hacking the programming code, so modifying metadata can be a challenge. It is true that these tools tend to comply with existing metadata standards (e.g. CKAN exposes metadata using the W3C DCAT vocabulary, although not in full compliance), but this is not always true. In some cases, it can be helpful to add domain-specific standards (e.g. to comply with '(meta) data meet domain-relevant community standards').

So, on the one hand, it is very important to evaluate the metadata model of a third-party tool before adopting it; on the other hand, it would be desirable that tools allow data managers to easily adjust the metadata model. Even if the metadata model is managed by the data software, it may be interesting to check the DWBP 1: 'Provide metadata', especially regarding machine-readable metadata. First, in line with the best practice of using easily parsable formats, it recommends serialising metadata in formats like Turtle or JSON, then it recommends the use of existing vocabularies: *"When defining machine-readable metadata, reusing existing standard terms and popular vocabularies are strongly recommended. For example, Dublin Core Metadata (DCMI) terms and Data Catalog Vocabulary should be used to provide descriptive metadata."*

Using existing metadata standards helps implement the FAIR requirement that *"(eta)data meet domain-relevant community standards"* and hopefully the requirements about licence and provenance metadata as well, considering that appropriate dataset vocabularies should include those metadata.

Metadata and catalogs

Ideally, metadata about the dataset should be in the dataset itself (many structured formats and vocabularies allow for hierarchical or relational models that include metadata about the dataset and the actual data) for self-discovery.

However, self-discovery would assume an existing infrastructure of distributed datasets and dataset repositories that expose standardised metadata and that are crawled or federated through distributed searches, which is not the case. The FAIR principles foresee that dataset metadata are registered in dataset catalogs where they can be more easily found: *"(meta)data are registered or indexed in a searchable resource"*. (FORCE11, 2014)

Many data repository/data service tools normally also provide good functions of data catalogs, providing metadata search functionalities and exposing all the metadata through APIs.

Even TBL added a note on this to the Linked Data design page: *"Now in 2010, people have been pressing me, for government data, to add a new requirement, and that is there should be metadata about the data itself, and that that metadata should be available from a major catalogue. […] Yes, there should be metadata about your dataset"* (Berners-Lee, 2006).

Linking

TBL's 5th star recommends: 'link your data to other data to provide context'.

The FAIR principles recommend, for interoperability: '(meta)data include qualified references to other (meta)data'.

This means that for instance, instead of indicating a category, a country or a unit of measure by a conventional local name, there should be a reference (a link) to that concept from an authoritative source to identify the concept precisely and unambiguously across systems. A typical way to do this is to identify the concept by the URI assigned to it in the authoritative system (for example, the URI of a country from GeoNames). This is the core of the Linked Data architecture and the basic mechanism of the Semantic Web.

DWBP Best Practice 10: Use persistent URIs as identifiers within datasets recommends that *"datasets should use and reuse other people's URIs as identifiers where possible."* This is another requirement that should be used as a criterion when choosing a third-party tool (and as a requirement if developing a new tool): so far, it seems that the most popular tools don't allow either to link an internal concept (like a category) to an external one or to use a URI as a value for a metadata element (except as a string, without considering it a resource). Therefore, complying with this requirement using tools like Dataverse or CKAN is – at the moment – difficult.

In conclusion, it is important for data managers to analyse the practical implications of implementing the requirements of the major data sharing frameworks and make informed decisions about their data repository/service accordingly.

4.3 Introduction to data interoperability

Data interoperability is the ability of a data set to be reused by any system without special effort. There are different layers to data interoperability: data can be technically interoperable thanks to a machine-readable format (e.g. CSV) and an easy-to-parse structure (e.g. JSON) but, in order for an external system to perform more operations on a dataset, the 'meaning' of data in the structure has to be explicit, and this is achieved through semantic interoperability by using metadata and values that have been previously assigned an unambiguous meaning and an identifier that can be used across systems.

Section 4.3. Introduction to data interoperability introduces the different layers of data interoperability: foundational or infrastructural (exposure protocols), structural (file formats and content structure) and semantic (use of metadata and values from agreed vocabularies). It only provides an overview and some examples: mastering the technologies for data interoperability would require a dedicated course and is something that is expected only of developers and data engineers.

However, a good understanding of the possible technological choices is useful for data reusers, who should be able to assess the level of data interoperability they want to reuse, and for agricultural data/service providers, who want to make sure their data is interoperable with other systems.

4.3.5. Semantic interoperability: vocabularies is devoted to semantic interoperability, as it is the most sector-specific aspect of data interoperability: while protocols and formats/structures are sector-agnostic, semantics are created and agreed upon in specific domains and, in the case of agricultural data, it is very important to know which semantic structures have been published in the domain and for which sub-domains (agronomy, food and nutrition, natural resources etc.) or types of data (soil data, weather data, crop growth data etc.).

Section 4.4. Interoperability of farm data focuses on examples of interoperable data using vocabularies created for farm data.

4.3.1 Data interoperability

The most used definition of 'interoperability' on the web is: *"the ability of a system or a product to work with other systems or products without special effort on the part of the user."* Wikipedia defines it as *"a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, at present or future, in either implementation or access, without any restrictions."*

When it comes to data interoperability, considering that data are serialised in datasets, the definitions above can be applied easily to a dataset as a product.

In the proceedings of a conference organized by the Coherence in Information for Agricultural Research for Development (CIARD) Open Agricultural Knowledge for Development community on data interoperability for agriculture, data interoperability was defined as *"a feature of datasets … whereby data can be easily retrieved, processed, re-used, and re-packaged ("operated") by other systems"* (CIARD, 2011).

There are some definitions that define it as the interoperability 'between two systems', however, it is a common view that something is really interoperable (or more interoperable) when as many systems as possible can interoperate it. Even more, by using certain data formats and applying certain data standards, data can be made 'interoperable by design' without necessarily knowing with which system they will be interoperable: planned interoperability with specific systems means that the data will be 'tightly coupled' with those systems, while maximised interoperability aims at loose coupling with as many systems as possible.

However, there will never be something like universal or perfect interoperability, a way of exposing data that will be suitable for all possible cases. Interoperability is always relative to a system of shared standards and common ways of using data that are in some cases very broad and all-purpose (like Dublin Core or schema.org) and, in other cases, very specific to scientific or interest communities (like data specifications and visualisations of gene sequences).

Indeed, the definitions above define interoperability as a feature of data(sets) alone, which is correct because they are the object of the interoperation, but the ecosystem of actors and products that have to co-operate for achieving full interoperability is broader: an interesting definition of interoperability that highlights the importance of 'shared expectations' is the one from the Data Interoperability Standards Consortium (DISC): *"Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data."* (Data Interoperability Standards Consortium, 2020)

Levels of interoperability

Interoperability can be achieved at different levels. While Wikipedia distinguishes between syntactic interoperability and semantic interoperability, a more granular distinction is made in the classification of types of interoperability by the Healthcare Information and Management Systems Society (HIMSS):

"'Foundational' interoperability allows data exchange from one information technology system to be received by another and does not require the ability for the receiving information technology system to interpret the data." (HIMSS, 2019)

This level is about infrastructural interoperability and is mostly about transmission protocols, which will not be addressed here as it is not of interest to the data manager. However, foundational interoperability also covers some higher-level exchange protocols, mostly working on top of the common HTTP protocol (for instance special REST APIs like OAI-PMH, SPARQL, Linked Data API, or content negotiation based on HTTP content-type requests). These may be of interest to data managers because, before being read and understood, data has to be transmitted, and there are different and more convenient ways to do this besides FTP downloads.

'Structural' interoperability is an intermediate level that defines the structure or format of data exchange i.e. the message format standards. Structural interoperability defines the syntax of the data exchange. It ensures that data exchanges between information technology systems can be interpreted at the data field level.

Structural interoperability is also the level where file formats and data formats play the most important role and it is the level where (meta)data become 'machine-readable' and can be parsed by machines: the easier it is for machine to parse the (meta)data format/syntax (XML, Json, CSV etc.), the more structurally interoperable data are.

'Semantic' interoperability provides interoperability at the highest level, which is the ability of two or more systems or elements to exchange information and to use the information that has been exchanged.

Semantic interoperability takes advantage of both the structuring of the data exchange and the codification of the data, including vocabulary (see 4.3.5. Semantic interoperability: vocabularies) so that the receiving information technology systems can interpret the data.

While with structural interoperability, machines understand what the different elements are (and their reciprocal structural relation), with semantic interoperability they also understand the meaning of those elements and can process them with semantic-aware tools and do some advanced reasoning. Data formats are not enough for this: semantic technologies allow embedding machine-readable semantic elements from agreed vocabularies in (meta)data serialised in most of the existing data formats, although the most suitable formats for this so far are the various formats of RDF (RDF/XML, Turtle, N-Triples) and JSON-LD.

For more detailed recommendations on how to implement data interoperability, the W3C DWBP is probably the best reference document. It is heavily based on the 'Linked Data' approach, but many of the recommendations help implement interoperability at different levels, even if not aiming at 100 percent linked data.

Interoperability of data and metadata
Data without metadata cannot be understood by machines. Data usually come with metadata.

The definition of data in Wikipedia is: *"Data are values of qualitative or quantitative variables, belonging to a set of items."* (Wikipedia, 2020k) Data are always part of a collection (a set of items) and, in the individual item (row, record) in a set, data are the values of some variables. They will always be encapsulated in some form of key-value pair where the key (the variable) is a metadata element that gives meaning to the data. This key-value pair is what in FAIR principles is called *"a single association between two concepts"* (FORCE11, 2014), which is 'one of the smallest possible 'data elements'.

Both parts in the key-value pair can present interoperability issues, so one can deal with interoperability issues of the metadata (e.g. agreed schemas for metadata elements, agreed variable names) and interoperability issues of the data/value (e.g. agreed controlled lists/ranges from which the value has to be taken, or syntax issues). However, since it is also common to consider controlled values and specification of syntax or unit of measure as 'metadata' (especially because they should be ideally defined by separate metadata elements and not within the value itself), one can also say that interoperability is mostly about metadata. The interoperability of data is achieved through the interoperability of metadata. For instance, the metadata 'air temperature' has to be interoperable to then get the actual value (the data) that one needs. The number that expresses the value will never be findable or understandable per se without the associated metadata.

Section 4.3. Introduction to data interoperability will follow the same convention adopted in the FAIR guiding principles mentioned above, using the term (meta)data when something applies indifferently to data and metadata.

4.3.2 Infrastructural interoperability: exchange protocols

The most popular protocols for exposing data as a service are the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and SPARQL. OAI-PMH and SPARQL are technically RESTful[21] APIs and, like all APIs, they expose 'methods' which accept a number of parameters and these methods can be called via an HTTP request and return a machine-readable response.

Besides OAI-PMH and SPARQL, it is very common to have data exposed through custom APIs, with standardised documentation about parameters, data types and return values.

OAI-PMH was born in the library environment and it was conceived mainly for metadata, but it can be used to expose any type of records.

The OAI-PMH API methods are designed to enable a series of calls that allow the caller (an application) to preliminarily check some metadata about the repository (the name, what metadata schemas are supported, which subsets can be retrieved etc.) and then retrieve records filtered according to metadata format, date ranges, subsets. Records can be retrieved in smaller batches using a 'resumption token'.

SPARQL (Wikipedia, 2020l) is a language for querying RDF data, but it is also the name of the protocol that enables the query to be sent via an HTTP request and get the response in several RDF-enabled formats (RDF/XML, Turtle, JSON etc.).

Compared to OAI-PMH, SPARQL allows for much more complex queries built using the powerful RDF triple grammar and can expose triples with any type of subject (a dataset, an organization, a place, a topic etc.) all from the same interface. This makes it possible to filter data very granularly and look up other properties (e.g. labels) of resources referenced by URIs. SPARQL clients can also send the same query to several SPARQL engines and combine the responses: the use of URIs allows the client to merge duplicates and consolidate properties for the same entity coming from different systems.

[21] 'Representational state transfer (REST) or RESTful web services is a way of providing interoperability between computer systems on the Internet. REST-compliant Web services allow requesting systems to access and manipulate textual representations of Web resources using a uniform and predefined set of stateless operations' (Wikipedia, 2020m).

Other protocols have been developed in scientific communities, building on common scientific data exchange practices, focusing on efficiency of data transfer (catering for large amounts of data) and leveraging traditional exchange formats like NetCDF and HDF5 (see Section 4.4. Interoperability of farm data). An example is OPeNDAP.

OPeNDAP ('Open-source Project for a Network Data Access Protocol') is *"a data transport architecture and protocol widely used by earth scientists. The protocol is based on HTTP and […] includes standards for encapsulating structured data, annotating the data with attributes and adding semantics that describe the data. […] An OPeNDAP client sends requests to an OPeNDAP server and receives various types of documents or binary data as a response. […] Data on the server is often in HDF or NetCDF format but can be in any format including a user-defined format. Compared to ordinary file transfer protocols (e.g. FTP), a major advantage using OPeNDAP is the ability to retrieve subsets of files, and also the ability to aggregate data from several files in one transfer operation"* (Wikipedia, 2020n).

The choice of protocols to implement to share one's data should be definitely influenced primarily by the community with which data are expected to be shared and then by technical considerations, such as ease of implementation, support for specific formats and general support for the data sharing principles.

For example, OPeNDAP is widely used by governmental agencies, such as NASA and the National Oceanic and Atmospheric Administration (NOAA), to serve satellite, weather and other observed earth science data, so it could be a good choice for institutions sharing these or similar types of data. On the other hand, if wider sharing across communities is desired, a custom API or known APIs like OAI-PMH or SPARQL can be used.

4.3.3 Structural interoperability: formats and structures

This is the level where (meta)data become machine-readable. However, 'machine-readable' in its most literal sense is not enough for structural interoperability: in order for machines to understand the structure of the (meta)data, i.e. which are the labels/metadata elements/column names/variables and which are the values, and whether there's a hierarchy, the (meta)data have to be structured enough for a machine to extract individual values.

In other words, (meta)data have to be not just readable but 'parsable' by machines. Since 'parsing' means 'splitting a file or other input into pieces of data that can be easily stored or manipulated', the more regular and rigorous a format is, the easier it is to parse it. Ideally, parsable formats have a published specification so that developers know how the structure is built and can write parsers accordingly.

There are many file formats that are parsable by machines. What makes a file format more easily parsable and therefore more interoperable is:

- The simplicity of the structure: the fewer the elements of the structure and the fewer the possible constructs are, the easier it is for a machine to parse the file without too complex reasoning. On the other hand, the format should also cater for complex data structures: the most suitable file format is the one that combines simplicity with enough flexibility to represent complex data structures.
- The rigorousness and 'regularity' of the structure: the fewer the options to serialise the same thing in a different way, the easier it is for a machine to parse the file and identify the role of each element.
- The existence of a clear and open specification for the format, which makes it easy for any developer to write parsers. (Many formats are tied to a software product and can be correctly and fully parsed only by that product. This makes their interoperability level very low.)
- An additional help is the existence of software libraries and APIs that can parse the format.

The formats that are considered the most interoperable against the criteria above are CSV, XML and JSON. These formats were explored in Section 3.1. Using open data in Chapter 3.

Binary array-based formats like NetCDF and HDF5 retain a special place for use by researchers. They will not be described here for a few reasons: they are more tied to specific software libraries (however many tools can read them); they are still more oriented towards compactness and efficiency of data transmission than towards broader interoperability, especially semantic interoperability; and some work has already been done to represent NetCDF in CSV (the CEDA BADC-CSV format) and in XML (the UNIDATA NcML). Other formats that are extremely well interoperable are the main native RDF formats, Turtle and N-Triples.

The RDF is *"a general method for conceptual description or modeling of information"* (Wikipedia, 2020s). As such, it is not a format and not tied to a specific format: any format that can represent the basic RDF 'grammar' ("triples") can implement RDF.

The RDF grammar is based on statements made of subject – predicate – object; each statement is a 'triple' and the assumption is that combinations of such triples can represent and describe everything. Ideally all three components of the statement should be represented by URIs that always refer precisely to the same entity.

The triples syntax is very simple:

> *Subject (node) - Predicate / property (arc) - Object (node)*
> <book_war&peace_URI> – <has_title> – 'War and Peace'
> <book_war&peace_URI> – <is_of_type> – <book_type_URI>
> <book_war&peace_URI> – <has_author> – <tolstoj_URI>
> <tolstoj_URI> – <is_of_type> – <person_type_URI>
> <tolstoj_URI> – <has_name> – 'Lev Tolstoj'

The RDF grammar has been successfully applied to: (a) XML: XML/RDF is not really a new 'format' (it's still formally XML), but rather an XML that uses specific constraints that enforce the triple logic; and (b) JSON: the JSON-LD (JSON for Linking Data) specification provides a method of encoding Linked Data as JSON.

4.3.4 Semantic interoperability: vocabularies

Human beings can interpret data through human-readable semantics that have always been used in (meta)data in different ways. For instance, a string to identify the topic or the colour of an object (e.g. in germplasm phenotypical descriptions) can be taken from a list of authoritative values (e.g. type of soil) or conventional variable names. As mentioned, interoperability is all about being understood by computer software: strings can be different in each dataset and in different languages. Even codes without a reference system behind them do not mean anything to computers as codes do not allow computers to make decisions on how to treat the values.

Metadata should contain information about each variable and each value comes from a reference system (a 'semantic structure' like a thesaurus or a code list). That semantic structure must be machine-readable and provide some stable identifiers that computer programs could use as stable values to design their behaviour (e.g. using the values as common search values across different datasets) to achieve semantic interoperability.

So, on the one hand the metadata have to embed information on the reference semantic structures and point to the exact elements they are using from that structure; on the other hand, these semantic structures, like the data, have to be 'serialised' in such a way that machines can read and process them, and use them to interpret the data.

Details on how to publish a semantic structure, or a 'vocabulary', in machine-readable format are beyond the scope of Section 4.3. Introduction to data interoperability. In short, such vocabularies are published as datasets, whose records are terms/concepts and their related descriptions, codes and ideally URIs in a machine-readable format – for the moment, let us assume XML or RDF/XML.

Semantic structures or 'vocabularies'

Vocabularies are agreed sets of terms, possibly with defined relationships between them. This includes both terms used for description metadata, like metadata element names, properties, predicates (so terms in description vocabularies: metadata schemas, ontologies) and terms used to categorise, annotate, classify (so terms in value vocabularies: thesauri, code lists, classifications, authority lists, also sometimes called 'Knowledge Organization Systems' (KOS)).

There is no formal classification of types of vocabularies which, in itself, could be a useful example of a value vocabulary.

Limited to what was defined above as 'value vocabularies', the exercise of creating a taxonomy of vocabulary types has been partially done by the Dublin Core initiative: their 'KOS Types Vocabulary' is quite useful to give an idea of the great variety of KOS and of the mixture of features that are combined in their definition:

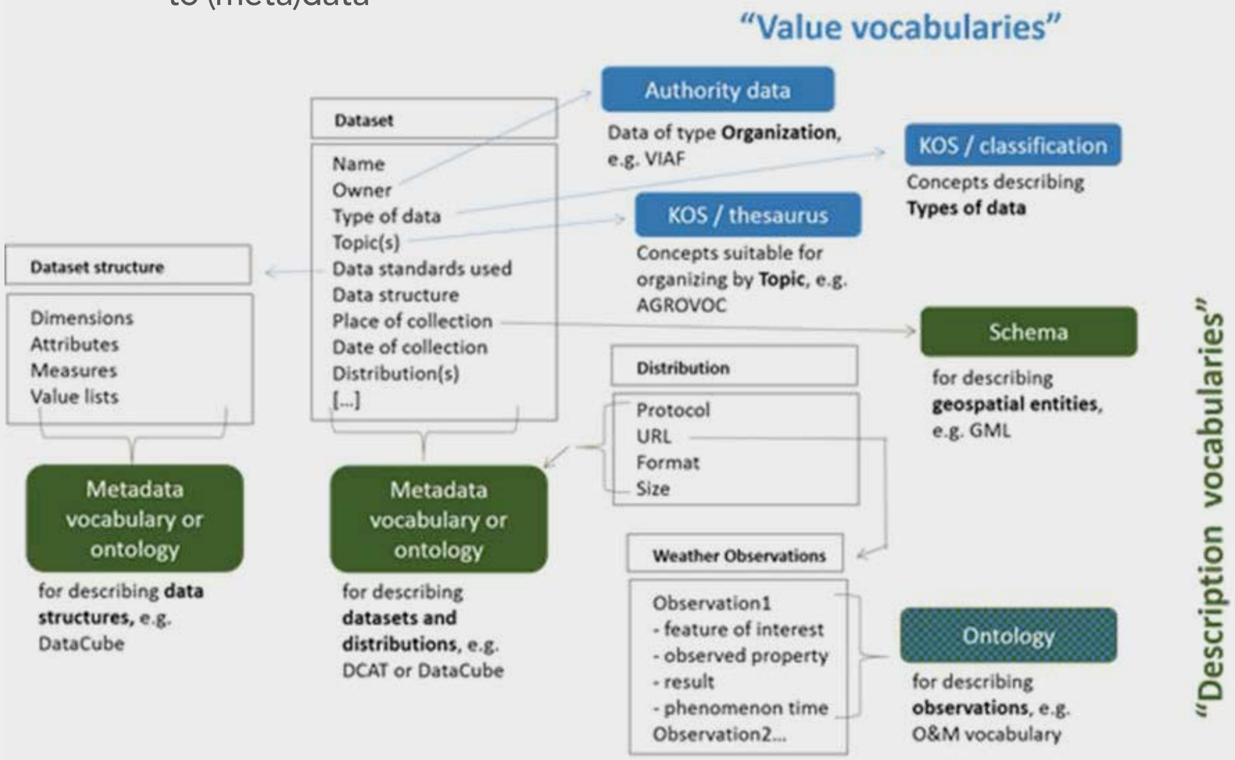
- categorisation scheme: loosely formed grouping scheme;
- classification scheme: schedule of concepts and pre-coordinated combinations of concepts, arranged by classification;
- dictionary: reference source containing words usually alphabetically arranged along with information about their forms, pronunciations, functions, etymologies, meanings, and syntactical and idiomatic uses;
- gazetteer: geospatial dictionary of named and typed places;
- glossary: collection of textual glosses or of specialised terms with their meanings;list: a limited set of terms arranged as a simple alphabetical list or in some other logically evident way; containing no relationships of any kind;
- name authority list (or authority file): controlled vocabulary for use in naming particular entities consistently;
- ontology: formal model that allows knowledge to be represented for a specific domain; an ontology describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms) *[see below a note on how an ontology can be seen as a KOS but also as a description vocabulary, an extended schema]*;
- semantic network: set of terms representing concepts, modelled as the nodes in a network of variable relationship types;
- subject heading scheme: structured vocabulary comprising terms available for subject indexing, plus rules for combining them into pre-coordinated strings of terms where necessary;
- synonym ring: set of synonymous or almost synonymous terms, any of which can be used to refer to a particular concept;
- taxonomy: scheme of categories and subcategories that can be used to sort and otherwise organize items of knowledge or information;
- terminology: set of designations belonging to one special language;
- thesaurus: controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.

As for description/modelling vocabularies, there is no authoritative list, but the most commonly used types are:

- schema (or metadata element set): any set of metadata elements, like XML schemas, RDF schemas or less formalised set of descriptors;
- application profile: a schema which consist of metadata elements drawn from one or more namespaces, combined together by implementers, and optimised for a particular local application;
- messaging standard: standards which describe how to format syntactically (and sometimes semantically) a message usually describing some event- or time-related information; messages are triggered by an event and transmitted in some way;
- ontology: this can define complex schemas with constraints and rules for reasoning.

As can be seen from the two lists above, ontologies are a special case: *"In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse"* (Wikipedia, 2020p). As such, it can be used for multiple purposes: it can be used as a description vocabulary, using the relations or even the classes defined by the ontology as metadata elements/properties describing the data (e.g. 'extreme temperature resistance' or 'frost resistance' in the Wheat Trait Ontology), or as a value vocabulary using classes or entities as terms for controlled values (e.g. wheat illnesses like *Puccinia striiformis* from the Wheat Trait Ontology, or countries from the FAO Geopolitical Ontology).

Figure 18. Example of use of different types of vocabularies to add semantics to (meta)data


"Value vocabularies"
Authority data
Data of type Organization, e.g. VIAF
KOS / classification
Concepts describing Types of data
KOS / thesaurus
Concepts suitable for organizing by Topic, e.g. AGROVOC
Dataset
Name
Owner
Type of data
Topic(s)
Data standards used
Data structure
Place of collection
Date of collection
Distribution(s)
[...]
Dataset structure
Dimensions
Attributes
Measures
Value lists
Schema
for describing geospatial entities, e.g. GML
Distribution
Protocol
URL
Format
Size
Metadata vocabulary or ontology
for describing data structures, e.g. DataCube
Metadata vocabulary or ontology
for describing datasets and distributions, e.g. DCAT or DataCube
Weather Observations
Observation1
- feature of interest
- observed property
- result
- phenomenon time
Observation2...
Ontology
for describing observations, e.g. O&M vocabulary
"Description vocabularies"

Source: Valeria Pesce, 2017.

Sometimes the boundaries between a schema and an ontology are blurred, but perhaps what can be considered typical of an ontology is the 'functional' more than descriptive design: classes, properties and especially relationships are designed as a model that is 'actionable' and can be used by applications for reasoning. However, the tendency nowadays is to use just the word 'vocabulary' and not delve too much into the definition of the different types (W3C, 2015). See Figure 18 for the use of different types of vocabularies to add semantics to (meta)data (Valeria Pesce, 2017).

Examples of metadata vocabularies are:

- Dublin Core [**www.dublincore.org/specifications/dublin-core/dces**], a metadata set providing metadata elements to describe almost any resource (title, description, identifier, creator, date, etc.);
- The W3C Data Catalog Vocabulary [**www.w3.org/TR/vocab-dcat**], a metadata model to describe different entities relevant to data sharing (catalog, dataset, distribution) and their relations;
- The Multi-Crop Passport Descriptors (Wikipedia, 2020r), a widely used international standard to facilitate germplasm passport information exchange developed jointly by IPGRI and FAO.

Examples of value vocabularies, or KOS, are:

- AGROVOC [**www.fao.org/agrovoc**], a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment, etc.
- GeoNames [**www.geonames.org**], an authoritative database of geographic entities. It covers 645 geographic features, among which countries, regions, administrative divisions, physical features like lakes, rivers, mountains etc.

A combination of metadata vocabulary and value vocabulary are ontologies that define the complete set of metadata elements and controlled values/entities that should be used to describe and classify specific things. Examples of ontologies are:

- The Crop Ontology [**www.cropontology.org**], which compiles validated concepts along with their inter-relationships on anatomy, structure and phenotype of crops, on trait measurement and methods, as well as on germplasm with the multi-crop passport terms.

- The Semantic Sensor Network (SSN) ontology [**www.w3.org/TR/vocab-ssn**], an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators.

Besides 'vocabularies', other terms that are used for defining these resources are 'semantic resources' or 'semantic structures'.

How to identify the most suitable published vocabularies

To discover what published vocabularies are available, the most useful source of information is catalogues that are dedicated to the agri-food domain. However, general catalogues searchable by domain can also be of help. Registries are conceived as metadata catalogues, which provide descriptions and categorisation of vocabularies and link to the original website and original serialisation of the standard. Repositories host the full content of the vocabulary, so that the terms themselves can be browsed. Below is an overview of some existing catalogues/repositories of data standards and vocabularies.

Agri-food domain

- GODAN Action Map of data standards [**vest.agrisemantics.org**] – a catalogue of data standards of different types and formats for the agri-food domain, categorised according to sub-domain, types of data, format and other criteria.
- AgroPortal [**agroportal.lirmm.fr**] – a repository of ontologies and value vocabularies, specialised in agronomy and food.
- Planteome [**browser.planteome.org/amigo**] – a repository of ontologies for plant biology.

General

- FAIRsharing [**fairsharing.org**] - this evolved from the Biosharing directory of standards for life sciences, it is now a general directory of data standards of different types. It has a good tagging system, but the coverage of agri-food standards is still poor.
- Linked Open Vocabularies (LOV) [**lov.okfn.org/dataset/lov**] – directory of RDF vocabularies spanning across all disciplines; it is not organized by domain or discipline and vocabularies can only be browsed through a small number of free tags.
- The Basel Register of Thesauri, Ontologies and Classifications (BARTOC) [**bartoc.org**] - BARTOC includes all types of KOS in any format, across all subject areas. The categorisation of vocabularies is quite generic (food and agriculture would fall partly under pure science and partly under technology without further sub-categorisations).

Besides choosing a vocabulary suitable for the data managed, some other criteria are useful for selecting the most appropriate vocabulary:

- The ideal implementation of the Semantic Web is through namespaces, URIs and the Linked Data approach: in order to be able to exploit these technologies, it is preferable to choose vocabularies that have been published as RDF (XML can be a good option too) and follow the Linked Data approach (use URIs, link to other vocabularies).
- It is preferable to choose vocabularies that are widely used, so that the data is interoperable with more datasets and more systems.

If existing vocabularies do not meet current needs, it is possible to create a new vocabulary and publish it (ideally, in collaboration with other partners in the community so as to ensure wide adoption).

Embedding semantics in the (meta)data

Semantic interoperability can be achieved at different levels and the implementation is different depending on the data format being used and the planned use of the vocabulary.

Using a schema for metadata

By identifying a metadata vocabulary/schema that has the classes and properties needed to describe the data, it can be reused to model and represent the data.

By using an existing published schema, the data will be already more semantically interoperable. This is because instead of using local metadata element names, which are meaningless for a computer, element names from a published vocabulary were used. Software tools can recognise that vocabulary to do something with it, e.g. match the values with values from other datasets that use the same schema. The adopted schema becomes the 'language' of the data structure. The element name of the selected schema will be used instead of a local one, with a prefix that indicates from which schema the element comes.

The agreed way to show in the metadata that an element is coming from a published vocabulary is to add a prefix that identifies the vocabulary.

In XML and in RDF-compliant formats like JSON-LD, prefixes can be associated with the URI of the vocabulary, so that the reference to the vocabulary is explicit. In other formats, like CSV, there may be no other way than just using conventional prefixes before column names (or the full URI of the vocabulary before the property name, but this is not very practical for column names).

In the example below, an XML file defines the prefixes for all the vocabularies (schemes) it will use; the om: prefix refers to the ISO data standard for Observations and Measurements (O&M):

```
<om:OM_Observation
xmlns:om="http://www.opengis.net/om/2.0"
xmlns:gml="http://www.opengis.net/gml/3.2"
xsi:schemaLocation="http://www.opengis.net/
om/2.0 http://schemas.opengis.net/om/2.0/
observation.xsd">
```

After the definition of the prefixes, every time the om: prefix is used, machines know that the following property name comes from the O&M vocabulary.

Using value vocabularies for data

A slightly different case is when the intention is to use values from an existing vocabulary as values (not metadata/properties) in the dataset in order to use unambiguous values across systems. Examples might be using the AGROVOC term for Oryza sativa or the term identifying a country from the FAO Geopolitical Ontology.

As a minimum, once a suitable vocabulary has been identified, if the vocabulary does not use URIs, and/or the URIs cannot be used in the dataset, at least the literal values of the terms can be used; systems can recognise the vocabulary and can match the literal against the URI. Preferably, the URI of the term should be used.

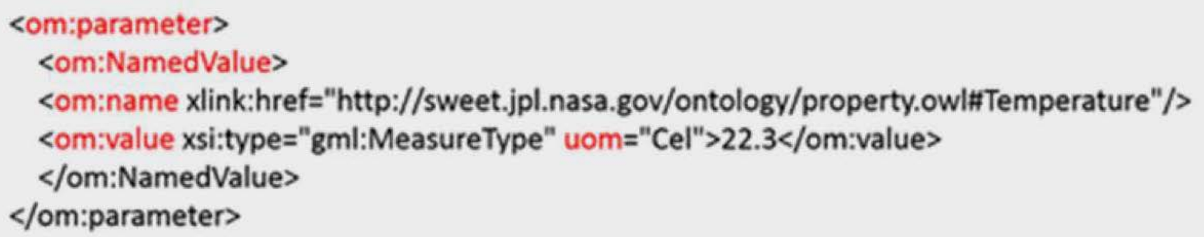
Depending on the data format used, value vocabularies can be used in different ways. In non-RDF XML or in CSV or in JSON, the URI can be used as the value of the element/column/label (e.g. in XML, the URI of the Geopolitical Ontology country can be used as the value of the dc:spatial element, perhaps specifying the scheme='URI' attribute to make it clear it's a URI).

Ideally, semantic interoperability is fully achieved using an RDF-enabled serialisation format (XML/RDF, Turtle, N3, JSON-LD). The advantage of using RDF is that RDF parsers and crawlers would normally look up additional properties from the URI address so including redundant values in the data may not be needed.

Even if there is not an ideal vocabulary that meets the needs and using custom terms is needed, one can link the local term to some similar or broader term in existing vocabularies. See an XML example below.

The name of the dimension (Temperature) is taken from a NASA-controlled vocabulary of property names using the full URI, while the fact that the value is a measure is indicated using the MeasureType value from the Open Geospatial Consortium GML controlled vocabulary:

Figure 19. Example of use of prefixes or URIs to refer to values defined in external vocabularies



```
<om:parameter>
  <om:NamedValue>
  <om:name xlink:href="http://sweet.jpl.nasa.gov/ontology/property.owl#Temperature"/>
  <om:value xsi:type="gml:MeasureType" uom="Cel">22.3</om:value>
  </om:NamedValue>
</om:parameter>
```

Source: FAO, 2020.

4.4 Interoperability of farm data

This book addresses the management of data from and for farmers, so primarily farmer profiles and farm data management systems or Farm Management Information Systems, FMIS. FMIS normally handle more specific, often agronomic or technical information, for farm decision-making. Some of this information is, in some cases, also used in farmer profiling platforms, especially in order to provide data-driven advice to farmers, but this is not common.

Regarding farmer profiles, strictly, there does not seem to be any initiatives to develop exchange standards for this type of data. However, the methodologies used for agricultural censuses can give some guidance (FAO, 2015).

FMIS are a relatively new area. These are added-value services, in which primary data (either from the farm or from external data services) of very different types (crop data, data on nutrients, pesticides, soil, weather etc.) is integrated, processed, and sometimes run against models, visualised and made actionable. They have been, so far, mostly the domain of the private sector: big companies as well as smaller start-ups have created all sorts of farm management services.

FMIS have to interface with the machinery that collects the data (e.g. soil moisture) and the machinery that executes the operations (e.g. a sprinkler) and possibly other systems that process the data down the line. In the majority of cases, data in these systems is designed to be interoperable within the system, using formats that are tightly coupled with the suppliers' machinery interface, with no apparent incentive for collaboration and no need to share data widely.

However, there is demand for wider interoperability, primarily for making machines and data from different suppliers work with equipment from other suppliers, but also for making farm data reusable by other FMIS and giving the farmer freedom to switch providers.

Section 4.4. Interoperability of farm data illustrates the current situation of data standards for farm data, listing published standards as well as projects that are working on new standards.

At the end of Section 4.4. Interoperability of farm data, some examples of interoperable farm data fragments are provided, with a commentary to explain the different approaches and how they use formats and vocabularies to make data interoperable.

These descriptions and examples illustrate that there is work underway to develop standards for various types of information in this area. However, it has to be noted that none of these standards is so widely accepted that its adoption should be considered a best practice: they are illustrated here to make learners familiar with interoperable farm data and to provide inspiration to data managers and developers on new ways to make their data interoperable.

4.4.1 Standards for farm observation data

For some types of data normally used in FMIS, like crop basic data, crop growth data, soil data and soil profiles, weather data etc., data standards exist, but they have been developed outside the community of intermediaries that develop tools for farmers:

- Crop basic data (from germplasm descriptors to official names to product classifications) have been standardised by normative bodies (e.g. FAO, International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRA). The European Food Safety Authority (EFSA), the United States Department of Agriculture (USDA)) and their core properties have been modelled in ontologies by research institutions (e.g. Consultative Group on International Agricultural Research (CGIAR), and the Institut national de la recherche agronomique (INRA)).
- Some data standards for crop growth data and crop growth models have been developed by research institutions that wanted to share or reuse models (e.g. the Agricultural Model Intercomparison and Improvement Project (AgMIP) and the Global Agricultural Trial Repository and Database (AgTrials)).
- Data standards for soil observations, soil profiles and soil properties (chemical properties, physical properties) exist (from USDA and FAO classifications to Infrastructure for Spatial Data in Europe (INSPIRE) data specifications).
- Weather data standards, as seen in the previous chapters, have been created by meteorology agencies.

Other data used in FMIS (data about machinery, sensors, agricultural input like fertilisers and pesticides, in some cases sales management data) partly follow industry standards and partly are just encoded in closed proprietary formats.

While FMIS developers may need to be aware of data standards when they import external data (crop core data, historical observations, climate data, predictive models), they have no reason to apply them in the storage or further re-packaging of data, at least as long as farm management data is meant to be used only locally or within the service network. And in case data is exchanged between pieces of machinery or within a network, it is normally in proprietary and closed formats established by suppliers (suppliers of agricultural input, machinery and software).

The only standards that are normally applied in the interface between FMIS and machinery are ISO standards, especially ISO 11783 *"Tractors and machinery for agriculture and forestry -- Serial control and communications data network"* (known as ISOBUS), because it is a standard that allows pieces of machinery to communicate.

However, there are a few trends that indicate that there is a growing need for more standardisation of data in these services:

- Farmers and farmers associations create or join consortia with the intent of sharing data or at least being able to transfer data across software packages. Some intermediaries are intercepting these data sharing needs. For instance, the AgGateway and the Open Ag Data Alliance consortia are leading efforts towards standardising all farm management data, or at least providing crosswalks to be able to transfer and share data.
- The need for more efficient pluggability of machinery components of different brands and their communication with FMIS. The AgGateway consortium is intercepting this need and working with machinery providers, data providers and FMIS providers to standardise data formats, at least in the crucial parts of the workflow where different pieces of machinery have to communicate between themselves or with the FMIS.
- While competition and patents lead manufacturers to keep part of the data and messages in their machinery operation in proprietary formats, they have been collaborating for a long time on agreed interface standards, especially ISOBUS. (There is, of course, the exception of some bigger players not adhering to ISOBUS and counting more on their monopoly position.)
- Intermediaries may want to be able to reuse predictive models coming from research (e.g. crop growth models, climate models etc.) and therefore need to model their data in a compatible way (*"If observed and simulated data are to be compared, it will be helpful if their metadata describes them in the same way"* (Eaton *et al.*, 2011)). Examples may be the models available from initiatives like AgMIP or crop management data standards like the International Consortium for Agricultural Systems Applications (ICASA) standards.
- There are international industry standards (ISO, UN etc) that were designed mainly for traceability and food security reasons and are normative standards when it comes to trade. In general, they are recommended for information exchange between farms and suppliers, traders and other partners in the agrifood supply chain.

Existing published data standards

United Nations and ISO standards

These standards are important for FMIS for regulatory reasons. On the one hand, data related to trade - from product traceability to invoicing -, especially when it comes to export, has to comply with these standards. On the other hand, in the case of ISO 11783, it is essential for the operation of machines. Since decision support is not the objective of these standards, weather data is not essential, although there are segments for measurements of different types and for event conditions, including weather.

- UNECE standards, in particular UN/EDIFACT Data Plot Sheet: This is a "Data Plot Sheet" (DPLOS) published by the United Nations (UNECE, 2010). The Detail section provides the breakdown of 1 to n plots sheets contained in the exchange:
    - General points on the plot sheet (dates, species, variety, area, contracts, etc.);
    - History of the plot (previous crops, enrichment, etc.);
    - Analysis (details of soil analyses carried out on the plot);
    - Events (i.e. all events such as observations, advice, actions taken, etc.).
- More specialised 'messages' compliant with EDIFACT messages have been created, with related XML schemas. In Europe, Agro EDI Europe is leading these efforts, producing standards like the 'Agronomic observations' (AgroObs) part of the EPIPHYT project.

- ISO standards, in particular ISO 11783 'ISOBUS': ISO 11783 *"Tractors and machinery for agriculture and forestry -- Serial control and communications data network"* is a communication protocol for the agriculture industry.

The ISOBUS Data Dictionary, part 11, is particularly interesting concerning data standardisation. The dictionary lists a huge number of 'entities' (and related definitions, units of measure and symbols) used in the transmission of data from farms, from all sorts of observed treatments and properties of crops (spray application, tillage, seeding depth, yield, crop loss etc.) and devices to (a few) properties read from weather stations, like air humidity and temperature.

Research-based agronomic data models

These schemas were designed for crop management and are quite suitable for decision support. They all cover weather data.

- *DSSAT ICASA2 data standards:* The Decision Support System for Agrotechnology Transfer (DSSAT) is a software tool developed through collaboration between several universities in the United States of America, USDA-Agricultural Research Service, and ICASA (DSSAT, 2020). It combines crop, soil, and weather databases with crop models and application programs to simulate multi-year outcomes of crop management strategies. DSSAT also provides for evaluation of crop model outputs with experimental data, thus allowing users to compare simulated outcomes with observed results. It comprises crop simulation models for over 42 crops. It is supported by data base management programs for soil, weather, and crop management and experimental data.

  DSSAT uses the ICASA2 data standard. Although the software and standards were created for research purposes, more precisely for field experiments, the experiments are about crop growth and the prescribed minimum set of data largely corresponds to the weather and agronomic data normally used in farm management software. As the standard authors say, ICASA is intended *"to allow description of essentially any field experiment or commercial crop production situation"*. The ICASA standard is also used by AgMIP as a format to import data as input to models (indeed the data needed for building the crop growth model are the same data needed to apply the model to crop growth decision-support tools).

  ICASA is explicitly designed to support implementations in a variety of formats, including plain text, spreadsheets or structured formats (it has an XML schema). The core of the ICASA standard is the master list of variables (ICASA, 2013), a naming convention for agricultural model variables, which is also used in AgMIP standards that build on the ICASA standards, like ACMO (see below). There are plans to render ICASA in RDF in the Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (TERRA-REF) project and the ICASA data dictionary is also being mapped to various ontologies as part of the Agronomy Ontology project).

- Agronomy Ontology: *"AGRO, the AGRonomy Ontology, describes agronomic practices, agronomic techniques, and agronomic variables used in agronomic experiments. AGRO is being built using traits identified by agronomists, the ICASA variables, and other existing ontologies such as Environmental Ontology (ENVO), Units of measurement ontology (UO), and Phenotype and Trait Ontology (PATO), The Information Artifact Ontology (IAO), and Chemical Entities of Biologica Interest (CHEBI). Further, AGRO will power an Agronomy Management System and fieldbook modelled on a CGIAR Breeding Management System to capture agronomic data"* (Aubert *et al*., 2017).

- Crop Ontology *(CO)*: The CO [**cropontology.org**] is managed by the CGIAR. It describes experimental design, environmental conditions and methods associated with the crop study/experiment/trial and their evaluation. The CO concepts are used to curate agronomic databases and describe the data.

Geospatial and observations data standards

Besides strictly agronomical data standards, standards that support observations and measurements, with the related geospatial dimensions, are very relevant. The most commonly used are:

- The ISO 19101 Domain Reference Standard model (ISO, 2014) for geospatial data infrastructures, defining the relations between dataset, metadata, feature instances, application schemas and services, and all related standards developed by ISO TC211. This model defines the concept of feature as an 'abstraction of real world phenomena', which can occur as a type or an instance) and 'feature types' (classes of features having common characteristics).

- The spatial data models and web services defined by the Open Geospatial Consortium (OGC), built along the lines of the above ISO specifications (almost always as joint ISO/OGC standards) and around the main OGC Geography Markup Language (GML): Web Feature Service, Web Coverage Service, Web Map Service (OGC, 2020a). These services, rather than sharing geographic information at the file level using FTP, offer direct fine-grained access to geographic information at different levels: the feature and feature property level, the coverage level, and map level.

- The ISO 19156:2011/OGC O&M model. This is a conceptual model (and a schema) for observations, and for features involved in sampling when making observations (observations commonly involve sampling of an ultimate feature-of-interest); ISO 19156:2011 defines a common set of sampling feature types classified primarily by topological dimension and therefore embedded in geospatial features.

  Since one of the key types of data in meteorology are weather observations, this ISO/OGC approach to geospatial and observational data is very relevant to weather data. Also the OGC Timeseries Profile of Observations and Measurements (TSML) and the OGC Sensor Model Language (SensorML) are part of the same framework and relevant to farm and weather observations.

- In the direction of a more semantic web of geospatial data, the W3C is now working on taking stock of geospatial ontologies following the ISO/OGC feature types approach. They are considering a Geospatial Features ontology and a 'Feature Types' ontology. For sensors data and observations, they have already developed, together with OGC, the OGC W3C Semantic Sensor Network Ontology (OGC and W3C, 2017).

- Among researchers, common data formats used for generic observations are very popular, like Network Common Data Form (NetCDF) (preferably following the Climate and Forecast Metadata Conventions, see below), HDF5, or in general formats that are understood by widely used services (like OPeNDAP). NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. It was developed by Unidata, one of the University Corporation for Atmospheric Research (UCAR)'s Community Programs (UCP).

- Weather observations: In formats like NetCDF, variable names are either arbitrary and only locally defined or they are coded with implicit reference to a code list or table which is not machine readable; recommended syntaxes and units of measures are sometimes indicated in the metadata in a human-readable way and sometimes described in attached guidelines or even just assumed (e.g. common scientific practices). It is recommended that *"the names of variables and dimensions should be meaningful and conform to any relevant conventions."*

  For NetCDF, a specific naming 'convention' for climate and forecast data was developed, called 'Climate and Forecast (CF) Metadata Conventions' or simply 'CF Conventions', which *"define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities"* (CF Metadata, 2020).

- Dataset metadata: In Section 4.3. Introduction to data interoperability, the distinction between data and dataset metadata standardisation was introduced. Many of the existing standards for observations described above (ISO 19101 OGC geospatial standards, NetCDF and CF Conventions) prescribe metadata conventions for the dataset. Dataset metadata should also include descriptions of the content and structure of the dataset, e.g. the phenomenon observed, the dataset dimensions, the units of measure etc. An RDF vocabulary that aims to provide a semantic web approach to describing the structure of a dataset together with its content is the W3C DataCube vocabulary (W3C, 2014).

Schemas for farm management data

As far as we know, except for the standards under development in initiatives like AgGateway, AgroRDF is the only published data standard for representing and describing farm work.

- *Association for Technology and Structures in Agriculture (KTBL)'s AGROXML/AGRORDF* agroXML is an XML/RDF schema developed in the iGreen project for representing and describing farm work. agroXML has been developed by a team consisting of members from makers of agricultural software systems, machinery companies, service providers and research organizations. It provides elements and XML data types for representing data on work processes on the farm including accompanying operating supplies like fertilisers, pesticides, crops etc. It can be used within FMIS as a file format for documentation purposes but also within web services and interfaces between the farm and external stakeholders as a means to exchange data in a structured, standardised and easy to use way. agroXML covers topics relevant to on farm activity including: 'crop', 'cropSpecies', 'chemical substance', 'harvestDate', 'enginePower'.

Projects that use some data standards and are developing new standards

Projects from public research

These are normally more for field experiments and crop growth models, but with similar data needs.

- *APSIM [*www.apsim.info*]:* The Agricultural Production Systems sIMulator (APSIM) is internationally recognised as a highly advanced simulator of agricultural systems. The APSIM initiative (AI) was established in 2007 to promote the development and use of the science modules and infrastructure software of APSIM. The founding members of the AI are CSIRO, the State of Queensland and The University of Queensland. AgResearch Ltd., New Zealand became a party in 2015 and other organizations may apply to join at any time. APSIM contains a suite of modules which enable the simulation of systems that cover a range of plant, animal, soil, climate and management interactions.

  In terms of data standards, for the formatting of messages, APSIM uses an XML schema called Data Description Markup Language (DDML) for data types, units, scales, and the Simulation Description Markup Language (SDML) for the simulation data.

  These formats, as far as we could see, are not meant as exchange standards and are not used outside of the APSIM software. However, data can be exported into a CSV that can be read for instance by AgMIP tools (and there is an R package available for direct importing into the R platform).

- *AgMIP [*tools.agmip.org*]:* The Agricultural Model Intercomparison and Improvement Project (AgMIP) aims to utilise intercomparisons of various types of methods to improve crop and economic models and ensemble projections and to produce enhanced assessments by the crop and economic modeling communities researching climate change agricultural impacts and adaptation.

  AgMIP also collaborates with the CCAFS AgTrials project on metadata standards. AgMIP uses some of its own standards (like ACMO and ACE) but heavily reuses and extends the ICASA2 standards. AgMIP provides QuadUI, a simple desktop application for converting crop modelling data to standard AgMIP format (JSON) and then translating to compatible model-ready formats for multiple crop models. Currently, the application reads weather, soil and field management information in either DSSAT format or a harmonised AgMIP CSV format. Output formats currently supported are DSSAT (ICASA2) and APSIM.

  They also provide the AgMIP Crop Model Output (ACMO) desktop utility to help generate ACMO files from model outputs (ACMO, 2013). The AgMIP team established the AgMIP Crop Experiment (ACE) harmonised data format to overcome incompatible file organization and structural complexity. Definitions of data elements are based on the ICASA standards which provide a comprehensive and extensible ontology for the description and definition of agricultural practices. Data are managed in ACE using JSON key-value structures. The key in each key-value pair corresponds to an ICASA parameter definition and units.

- *AgTrials [*www.agtrials.org*]:* AgTrials is managed by the CGIAR Research Program on Climate Change and Agricultural Food Security (CCAFS). The alpha version of a web application to compile and store information on the performance of agricultural technology, so far, AgTrials allows the collection, organization and uploading of raw data and their associated metadata from more than 800 trials carried out in the last three decades covering more than 20 countries across Africa, South Asia and Latin America and 16 crops and 7 livestock species. It is run in collaboration with the crop modelling initiatives AgMIP and Global Futures. The results are available in AgMIP formats as well as modelled with the Crop Ontology.

4.4.2 Projects from/for industry

There are hundreds of FMIS that, of course, use some internal data model and store data in some format but, in most cases, they focus on the end user and do not make their data formats explicit nor do they engage with the creation or negotiation of data standards. On the other hand, there are a few initiatives that work at a broader level, involving actors from different sectors, which focus on the interoperability of FMIS data.

- *AgGateway:* It is a non-profit consortium of businesses serving the agriculture industry. It currently has more than 230 member companies working on eConnectivity activities within eight major segments: agricultural retail; systems and software developers and service providers; crop nutrition; crop protection; grain and feed; precision agriculture; seed; specialty chemicals. Their work on standardisation covers all aspects of farm management.

  Currently, AgGateway is working on three major projects: (a) the Standardised Precision Ag Data Exchange (SPADE) project, which aims to establish a framework of standards to simplify mixed-fleet field operations and regulatory compliance and to allow seamless data exchange between hardware systems and software applications that collect field data across farming operations; (b) the Precision Ag Irrigation Language (PAIL) project to provide an industry-wide format that will enable the exchange and use of data to and from irrigation management systems, which are currently stored in a variety of proprietary formats; and, (c) the ADAPT toolkit to enable interoperability between different precision agriculture software and hardware applications.

  AgGateway publishes their guidelines and standards in different forms: some are public and some are available only to members (AgGateway, 2020). These standards are meant to be used by intermediaries who create FMIS but also by machinery manufacturers. The standards are data models, expressed as XML schemas but translatable to XML and Json according to specific guidelines, and controlled vocabularies.

- *Open Ag Data Alliance* (OADA): It was formed in early 2014 as an open source project with widespread industry support and is headed by the Open Ag Technology and Systems Group (OATS) at Purdue University. Its goal is *"to help the industry get data flowing automatically for farmers in agriculture so they can reap the benefits of making data-driven decisions and stop wrangling data and incompatible systems. The alliance has over 25 commercial partners worldwide."*(Ault, 2016)

  The alliance aims to pursue their objective through the development of open data sharing standards (APIs) and open source software libraries that will serve as a conduit between data generation and data consumption. They also aim to facilitate easy transferability of a farmer's data among solution providers and to allow farmers and actors (suppliers, advisors) to collaborate on the same platform based on specific and very granular access permissions: rather than focusing on ownership. The alliance focuses on access rights by facilitating the protocols for giving and revoking access to data.

  OADA's objectives are similar to AgGateway's, only the method is different: while AgGateway focuses on common data models, data formats and controlled vocabularies, OADA focuses more on the web service side like their real-time connections API for weather, soil moisture data (Ault, 2016).

- *AgroConnect:* This is the only initiative focusing on data standards for FMIS that our research revealed in Europe. It has very similar objectives to those of AgGateway and they collaborate closely. The members of the AgroConnect association are companies and organizations that trade goods, services, products and produce with farmers. A group of stakeholders are the FMIS providers: *"All these companies and organisations share a common goal and that is to enable easy data exchange in the agricultural supply chain between all parties involved."* (AgroConnect, 2020)

  AgroConnect promotes standard data models; standard interface definitions (EDI-messages, API's) for data interchange; standards for identifying farms, persons, crop fields, animal; all type of standard code lists, e.g. for crop types, soil types, animal types, etc.; and standard protocols for data exchange.

There do not seem to be other initiatives of this type, focusing on data sharing through FMIS and involving different stakeholders. Apparently, standardisation of FMIS in Europe is high around the ISOBUS standard but has not extended to the integration of components of the FMIS workflow that are not tractors and typical machinery and do not aim to transport data across different FMIS. There are efforts trying to attract the industry towards a data exchange platform around the FIWARE platform [www.fiware.org] through the SmartAgriFood accelerator [smartagrifood.com] and projects like AGICOLUS [www.fiware.org/success-stories/agricolus], but there does not seem to be much work on data standards in addition to ISOBUS.

4.4.3 Examples of interoperable observation data

This section does not aim to make learners experts of XML, JSON or RDF data serialisations. It does not even aim to provide a full understanding of the fragments of data serialisations that are presented. The objective of presenting short examples of data expressed in different ways is to make learners familiar with how interoperable farm data may look and to provide inspiration for data managers, service providers and developers who want to make their data more interoperable.

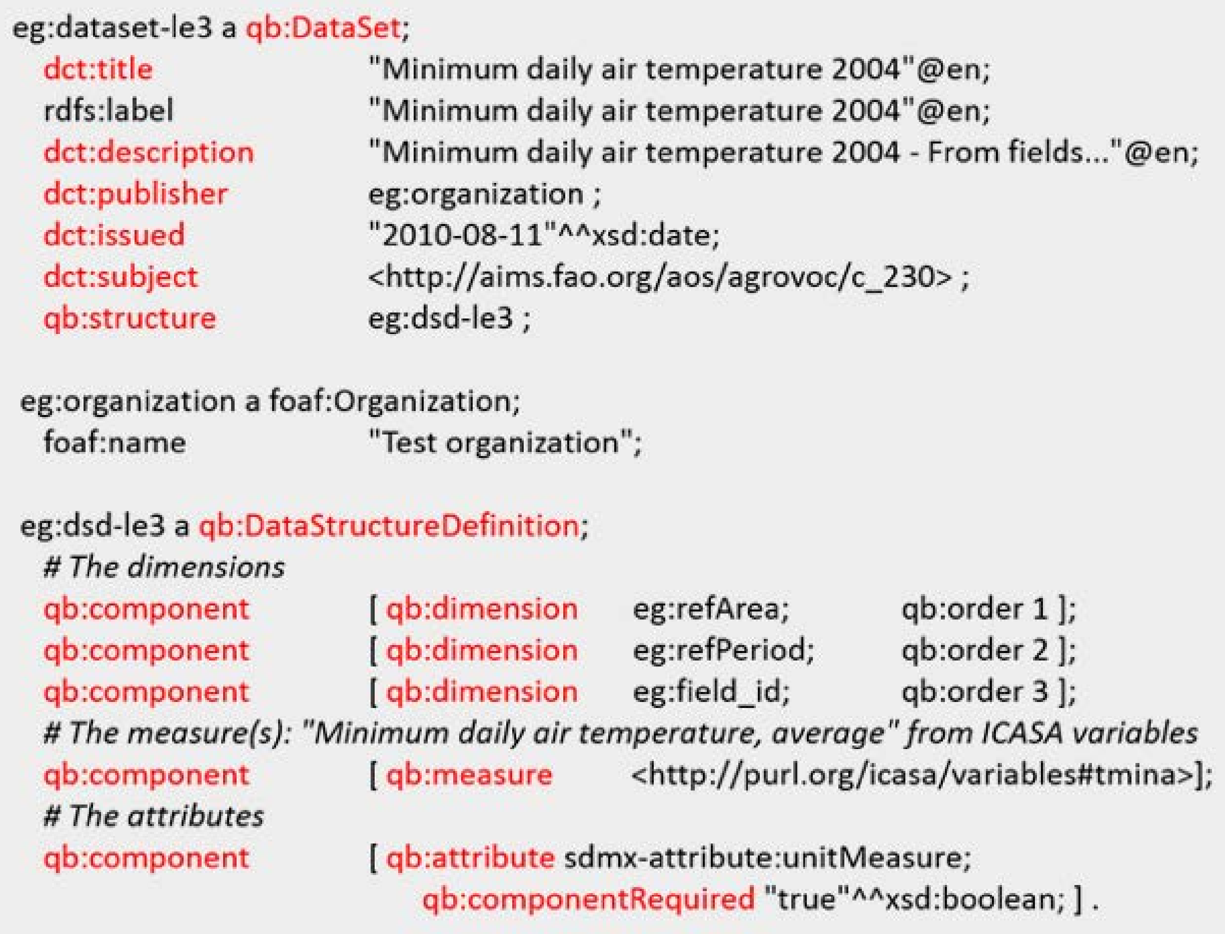
Dataset metadata describing the structure of the observations dataset using DataCube RDF

The example given in Figure 20. describes a data structure designed to contain one measurement ('minimum daily air temperature, average', indicated with a conventional URI from the ICASA Master Variables List, not yet published) and three dimensions: area, period and identifier of the field where the measurement is taken. This means that these will be the data in that dataset and the names and attributes used in the data will be those indicated here.

The example uses the RDF Turtle syntax and, among other vocabularies, Dublin Core (identified by the dct: prefix) and DataCube (qb: prefix):

ISO 19115 'Geographic information – Metadata', in the ISO 19101 series mentioned above, is also a suitable vocabulary covering dozens of metadata elements for a dataset. While ISO/TS 19139 provides an XML schema for ISO 19115, an RDF (OWL) representation of ISO 19115 has been developed by CSIRO Australia.

Figure 20. Example of RDF encoding of dataset metadata and data structure using Data Cube



```
eg:dataset-le3 a qb:DataSet;
   dct:title              "Minimum daily air temperature 2004"@en;
   rdfs:label             "Minimum daily air temperature 2004"@en;
   dct:description        "Minimum daily air temperature 2004 - From fields..."@en;
   dct:publisher          eg:organization ;
   dct:issued             "2010-08-11"^^xsd:date;
   dct:subject            <http://aims.fao.org/aos/agrovoc/c_230> ;
   qb:structure           eg:dsd-le3 ;

 eg:organization a foaf:Organization;
   foaf:name              "Test organization";

 eg:dsd-le3 a qb:DataStructureDefinition;
   # The dimensions
   qb:component           [ qb:dimension    eg:refArea;       qb:order 1 ];
   qb:component           [ qb:dimension    eg:refPeriod;     qb:order 2 ];
   qb:component           [ qb:dimension    eg:field_id;      qb:order 3 ];
   # The measure(s): "Minimum daily air temperature, average" from ICASA variables
   qb:component           [ qb:measure      <http://purl.org/icasa/variables#tmina>];
   # The attributes
   qb:component           [ qb:attribute sdmx-attribute:unitMeasure;
                               qb:componentRequired "true"^^xsd:boolean; ] .
```

Source: FAO, 2020.

Record of measurement of fruit mass at the temperature of 22.3 °C using O&M XML

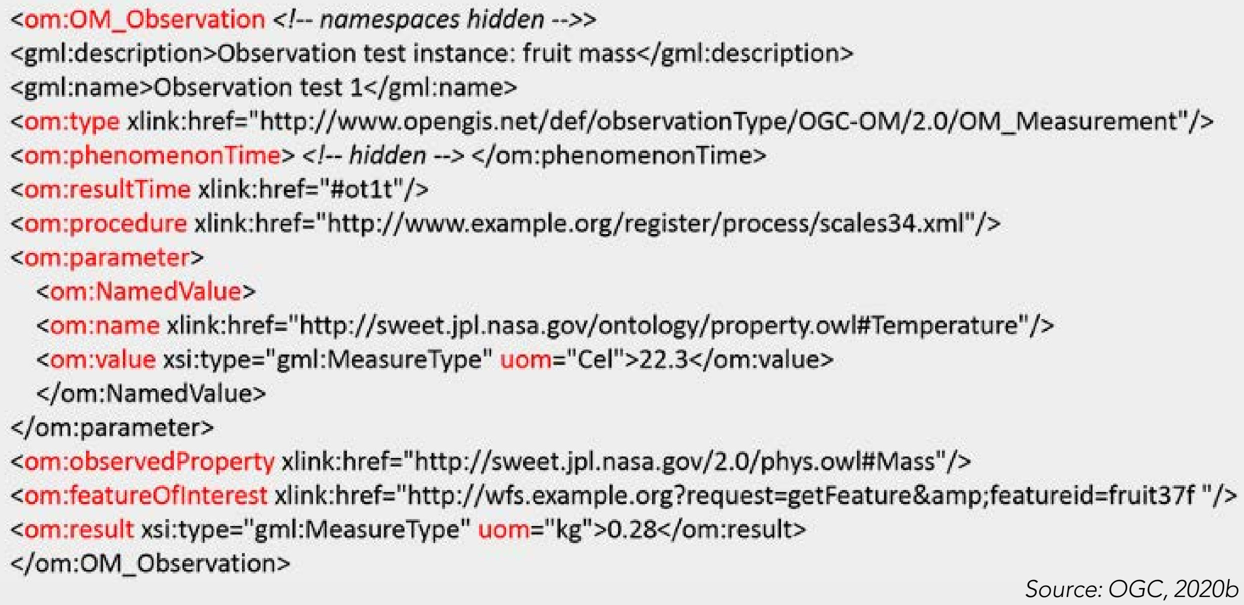
The example given in Figure 21 is from a dataset of observations from agricultural experiments. It uses the XML schema of the ISO data standard for O&M. It shows how to use elements from the O&M XML schema to describe a simple observation: the measurement of fruit mass at 22.3 °C.

Measurement of air temperature at a specific point using O&M JSON

The example given in Figure 22 shows how to use the O&M JSON schema to describe a simple observation: the measurement of air temperature at a specific point. Although no prefix is used, the fact that the JSON follows the O&M JSON schema tells us that the properties used are from O&M. The labels used ('observedProperty', 'featureOfInterest') and the nesting structure ('uom' under 'result') clearly show that, even if in a different format, the schema is the same as the one used in the previous XML example.

In the XML file, the prefix om: would be mapped to the namespace of the XML schema; in the JSON file, the @context would be the URL of the JSON schema. In both cases, any software parsing the datasets will interpret the elements/labels in the same way.

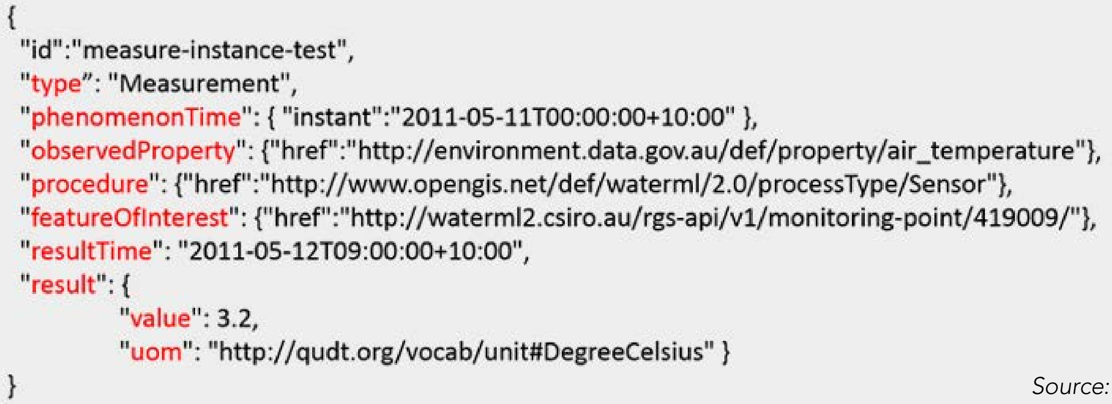
Figure 21. Example of observation in XML format using the O&M schema



```
<om:OM_Observation <!-- namespaces hidden -->>
<gml:description>Observation test instance: fruit mass</gml:description>
<gml:name>Observation test 1</gml:name>
<om:type xlink:href="http://www.opengis.net/def/observationType/OGC-OM/2.0/OM_Measurement"/>
<om:phenomenonTime> <!-- hidden --> </om:phenomenonTime>
<om:resultTime xlink:href="#ot1t"/>
<om:procedure xlink:href="http://www.example.org/register/process/scales34.xml"/>
<om:parameter>
  <om:NamedValue>
  <om:name xlink:href="http://sweet.jpl.nasa.gov/ontology/property.owl#Temperature"/>
  <om:value xsi:type="gml:MeasureType" uom="Cel">22.3</om:value>
  </om:NamedValue>
</om:parameter>
<om:observedProperty xlink:href="http://sweet.jpl.nasa.gov/2.0/phys.owl#Mass"/>
<om:featureOfInterest xlink:href="http://wfs.example.org?request=getFeature&amp;featureid=fruit37f "/>
<om:result xsi:type="gml:MeasureType" uom="kg">0.28</om:result>
</om:OM_Observation>
```

Source: OGC, 2020b

Figure 22. Example of observation in JSON format using the O&M model



```
{
 "id":"measure-instance-test",
 "type": "Measurement",
 "phenomenonTime": { "instant":"2011-05-11T00:00:00+10:00" },
 "observedProperty": {"href":"http://environment.data.gov.au/def/property/air_temperature"},
 "procedure": {"href":"http://www.opengis.net/def/waterml/2.0/processType/Sensor"},
 "featureOfInterest": {"href":"http://waterml2.csiro.au/rgs-api/v1/monitoring-point/419009/"},
 "resultTime": "2011-05-12T09:00:00+10:00",
 "result": {
          "value": 3.2,
          "uom": "http://qudt.org/vocab/unit#DegreeCelsius" }
}
```

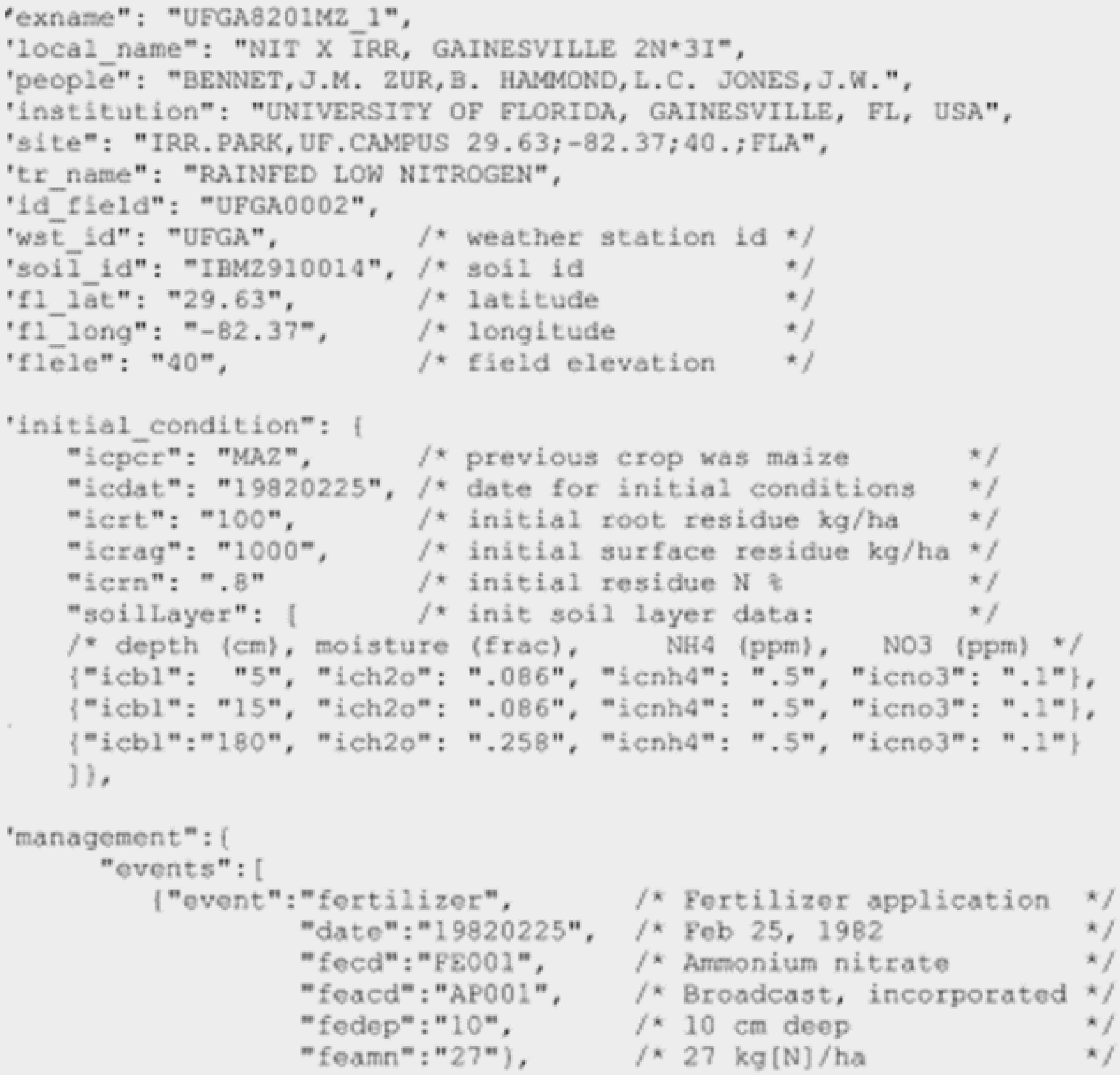
Source: OGC, 2020b.

Field experiment data using ICASA variable names in JSON

The previously mentioned ICASA data standard (White *et al.*, 2013) has been serialised into JSON. The example given in Figure 23 shows an experiment encoded in JSON using the ICASA data model and variables. Instead of URIs, ICASA uses short coded variable names. All the codes are in the ICASA Master Variables list, which defines the meaning of all variables and constitutes the ICASA semantic resource, which indicates that 'fielele' means 'field elevation' and 'icpcr' means 'residue, crop code for previous crop' (ICASA, 2013).

Since variables are identified by codes and not by URIs, and codes are not even associated with definitions in a machine-readable file, software tools cannot look up the meaning and cannot infer the reference semantics behind the code. Therefore, even if the ICASA variables are probably the most complete list for agricultural experiments and are used in other systems, at the moment using them does not ensure full semantic interoperability. There is a work to do to express the ICASA variables in an ontology.

Figure 23. Example of an experiment described in JSON format using the ICASA standard



```
'exname": "UFGA8201MZ_1",
'local_name": "NIT X IRR, GAINESVILLE 2N*3I",
'people": "BENNET,J.M. ZUR,B. HAMMOND,L.C. JONES,J.W.",
'institution": "UNIVERSITY OF FLORIDA, GAINESVILLE, FL, USA",
'site": "IRR.PARK,UF.CAMPUS 29.63;-82.37;40.;FLA",
'tr_name": "RAINFED LOW NITROGEN",
'id_field": "UFGA0002",
'wst_id": "UFGA",         /* weather station id */
'soil_id": "IBM2910014", /* soil id            */
'fl_lat": "29.63",       /* latitude           */
'fl_long": "-82.37",     /* longitude          */
'flele": "40",           /* field elevation    */

'initial_condition": {
    "icpcr": "MAZ",       /* previous crop was maize       */
    "icdat": "19820225", /* date for initial conditions   */
    "icrt": "100",        /* initial root residue kg/ha    */
    "icrag": "1000",      /* initial surface residue kg/ha */
    "icrn": ".8"          /* initial residue N %           */
    "soilLayer": [        /* init soil layer data:         */
    /* depth (cm), moisture (frac),      NH4 (ppm),    NO3 (ppm) */
    {"icbl":  "5", "ich2o": ".086", "icnh4": ".5", "icno3": ".1"},
    {"icbl": "15", "ich2o": ".086", "icnh4": ".5", "icno3": ".1"},
    {"icbl":"180", "ich2o": ".258", "icnh4": ".5", "icno3": ".1"}
    ]},

'management":{
      "events":[
         {"event":"fertilizer",         /* Fertilizer application  */
                  "date":"19820225",    /* Feb 25, 1982            */
                  "fecd":"FE001",       /* Ammonium nitrate        */
                  "feacd":"AP001",      /* Broadcast, incorporated */
                  "fedep":"10",         /* 10 cm deep              */
                  "feamn":"27"},        /* 27 kg[N]/ha             */
```

Source: White et al., 2013.

Weather observations using ICASA variables in tabular format

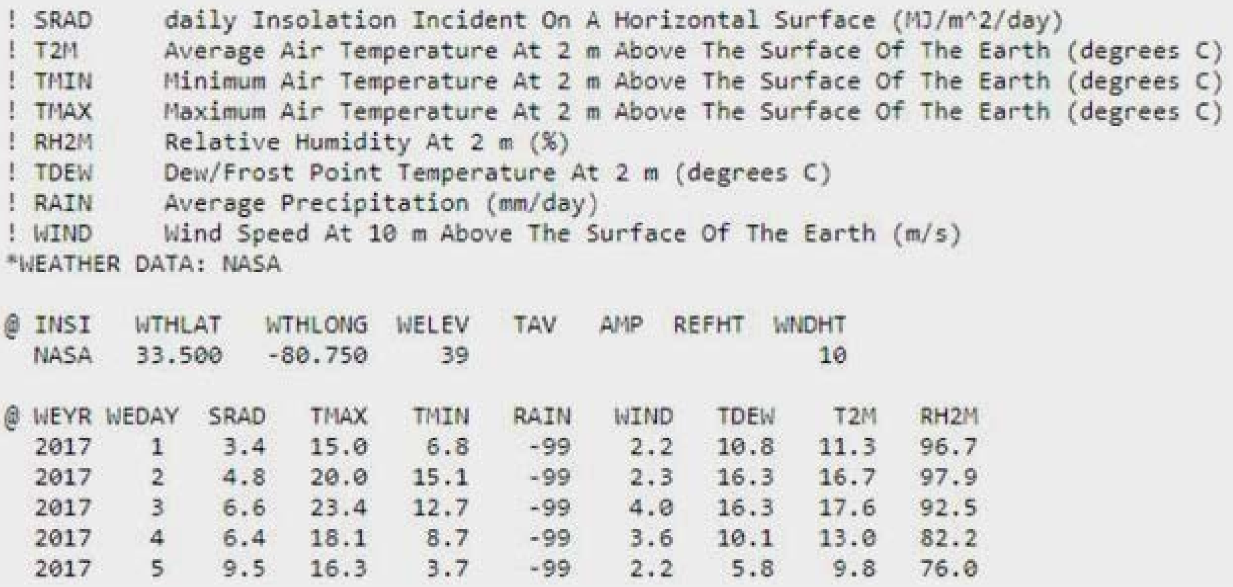
ICASA variables can be used in tabular format as well as in column names. In the example given in Figure 24, variables definition are provided at the beginning of the file, mainly for human reading. Although the ICASA variables are used as simple strings, the fact that they come from a controlled vocabulary can allow for a good degree of interoperability, for example to aggregate the data with other datasets using the ICASA variables.

Observation of tree height using the SSN/SOSA Ontology in RDF

The W3C 'Semantic Sensor Network Ontology' (SSN) is an ontology built based on OGC SensorML and O&M standards. The classes and properties that are most relevant for observations are under the namespace at w3.org/ns/sosa, Sensor, Observation, Sample, and Actuator (SOSA). The vocabulary follows the OGC Observation – FeatureOfInterest - ObservedProperty model. Note that this description also uses other vocabularies, in particular a vocabulary for units of measure:

@prefix qudt-1-1: <http://qudt.org/1.1/schema/qudt#>.

Figure 24. Example of tabular data using the ICASA variables



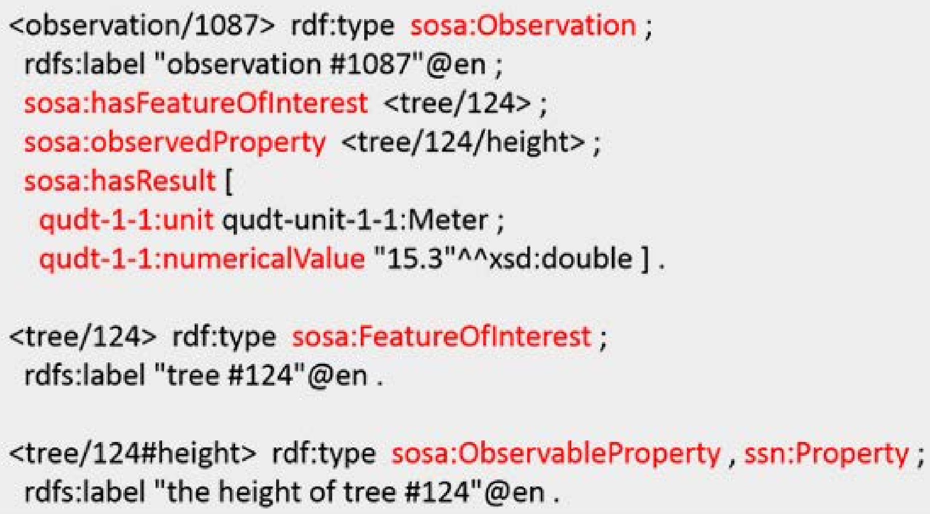
```
! SRAD    daily Insolation Incident On A Horizontal Surface (MJ/m^2/day)
! T2M     Average Air Temperature At 2 m Above The Surface Of The Earth (degrees C)
! TMIN    Minimum Air Temperature At 2 m Above The Surface Of The Earth (degrees C)
! TMAX    Maximum Air Temperature At 2 m Above The Surface Of The Earth (degrees C)
! RH2M    Relative Humidity At 2 m (%)
! TDEW    Dew/Frost Point Temperature At 2 m (degrees C)
! RAIN    Average Precipitation (mm/day)
! WIND    Wind Speed At 10 m Above The Surface Of The Earth (m/s)
*WEATHER DATA: NASA

@ INSI   WTHLAT   WTHLONG  WELEV   TAV   AMP  REFHT  WNDHT
  NASA   33.500   -80.750     39                      10

@ WEYR WEDAY  SRAD  TMAX  TMIN  RAIN  WIND  TDEW   T2M  RH2M
  2017     1   3.4  15.0   6.8   -99   2.2  10.8  11.3  96.7
  2017     2   4.8  20.0  15.1   -99   2.3  16.3  16.7  97.9
  2017     3   6.6  23.4  12.7   -99   4.0  16.3  17.6  92.5
  2017     4   6.4  18.1   8.7   -99   3.6  10.1  13.0  82.2
  2017     5   9.5  16.3   3.7   -99   2.2   5.8   9.8  76.0
```

Source: White et al., 2013.

Figure 25. Example of observation in RDF Turtle using the W3C SSN ontology



```
<observation/1087> rdf:type sosa:Observation ;
  rdfs:label "observation #1087"@en ;
  sosa:hasFeatureOfInterest <tree/124> ;
  sosa:observedProperty <tree/124/height> ;
  sosa:hasResult [
    qudt-1-1:unit qudt-unit-1-1:Meter ;
    qudt-1-1:numericalValue "15.3"^^xsd:double ] .

<tree/124> rdf:type sosa:FeatureOfInterest ;
  rdfs:label "tree #124"@en .

<tree/124#height> rdf:type sosa:ObservableProperty , ssn:Property ;
  rdfs:label "the height of tree #124"@en .
```

Source: OGC and W3C, 2017.

Observation of minimum air temperature using data cube and ICASA variables

Observations can also be encoded using data cube, although inside the observation entity other vocabularies are needed. The example given in Figure 26 encodes the measurement of daily minimum temperature. Besides the namespaces used in the previous example, here we have two additional namespaces for the metadata elements, one for statistical attributes (SDMX) and one for variables (the ICASA variables):

@prefix sdmx-attribute: <**http://purl.org/linked-data/sdmx/2009/attribute#**>

@prefix icasa-var: <**http://purl.org/icasa/variables#**>

The use of the icasa-var: prefix before the tmina property tells the machine that it should look up the tmina variable in the list of variables published at http://purl.org/icasa/variables#.

4.4.4 Controlled variable names

It is clear from the examples above that there is a strong need for standardised variable names and that this need is currently being addressed mainly in two ways, either URIs in published vocabularies (like O&M) or prescribed strings/codes published as controlled lists (like ICASA). However, the second approach is probably also going to adopt the URI method.

The issue of variable names is not only just an issue of URI or string: a variable name can be a combination of several dimensions, usually a feature of interest + observed property + observation methods + parameters (e.g. hourly + average + wind speed + at 10 m).

The AgGateway PAIL standard is taking an interesting approach, which is to reduce an observation to a key-value pair, with the key expressing all the meaning and the value. There is a controlled vocabulary for each of the aspects of a variable (time window, aggregation level, feature of interest, observed property, observation methods, parameters etc) and the observation key is a new concept (in turn making up another vocabulary), which is the orthogonal combination of concepts from these vocabularies. The idea is to have a registry for all the orthogonal keys. The PAIL team is considering mapping the final valid keys in their orthogonal vocabulary to existing standardised variable lists. Those interested in new developments on the PAIL standard can join the project community.

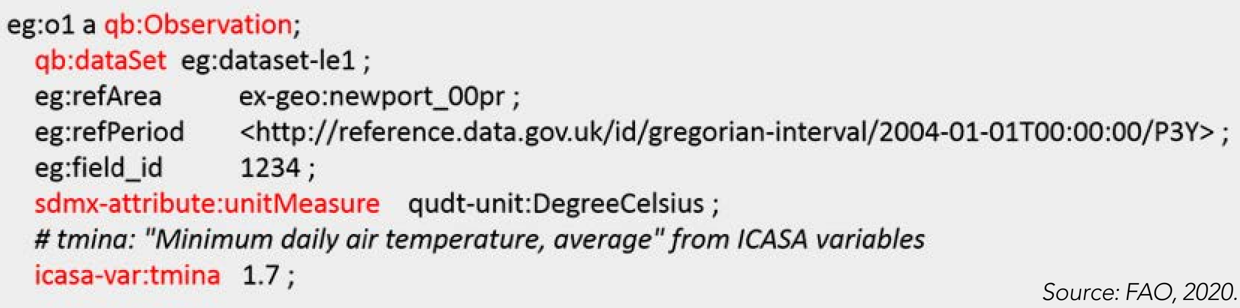
4.5 Open licensing for data

The importance of licensing data was highlighted in Section 4.2. Guiding frameworks for data sharing. The recommendation was that data should always be accompanied by a licence, no matter how openly or limitedly it is being shared: licences indicate if and how the data can be reused, at any point in the data spectrum (The Open Data Institute (ODI), 2019a).

Considering all the different permissions that licences can give and the legal aspects involved, licenses can be very complex and tailored to each specific project, especially in the case of shared or closed data. Therefore, it is difficult to provide a comprehensive guide to licensing non-open data.

In contrast, since open data has been quite clearly defined and there is broad agreement on its main requirements, a lot of work has been done on defining standard open data licences that can be adopted as they are. Section 4.5. for data focuses on open data licensing.

Figure 26. Observation encoded in RDF (Turtle) using data cube and ICASA



```
eg:o1 a qb:Observation;
   qb:dataSet  eg:dataset-le1 ;
   eg:refArea        ex-geo:newport_00pr ;
   eg:refPeriod      <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
   eg:field_id       1234 ;
   sdmx-attribute:unitMeasure    qudt-unit:DegreeCelsius ;
   # tmina: "Minimum daily air temperature, average" from ICASA variables
   icasa-var:tmina   1.7 ;
```

Source: FAO, 2020.

Sharing data on the web, making it publicly available to everyone means that data can only be viewed and read, but cannot be reused or modified legally unless permissions are explicitly given to do so on the source. Going back to the definition given in Section 2.1.

What is shared and open data, open data is data that can be freely used, reused (modified) and redistributed (shared) by anyone. For reuse and redistribution, data must be provided under terms that permit reuse and redistribution, including intermixing with other datasets. This is where the (open) licensing is applied to the data and, in Section 4.5. for data, readers will be introduced to licensing terminology and existing licensing tools.

4.5.1 Licensing and reuse

Licensing means that the copyright owner retains ownership but authorises a third party to carry out certain acts covered by the economic rights, generally for a specific period of time and for a specific purpose (WIPO, 2016). In order to facilitate the reuse of data, it is crucial that others know the terms of use for the database and the data content. To ensure that happens, the rights holder should mark the data with associated permissions.

There are two ways of communicating permissions to potential data reusers. The rights holder can license a second party to do things that would otherwise infringe on the rights held; alternatively, the rights holder can give up the rights to a resource so that infringement becomes a non-issue. In both cases, only the rights holder can grant permissions or waive the rights with a licence (Ball, 2014).

An 'open licence' may sound a contradiction. In general, a licence on a certain piece of content is an agreement between two parties: the licensor and the licensee. It usually comes with provisions for terms, territory and renewal conditions.

- The *terms* lay down what the licensor allows the licensee to do with the content. For example, the licensee may be granted the right to use software that the licensor owns.
- The *territory* is the geographical area where the licence is valid. For example, a distributor may have the right to distribute the data in Europe, but not in the United States of America.
- A *renewal* clause is customary because an agreement usually has a duration and can (or cannot) be renewed after licence expiration.

In the early days of the web, some people mistakenly assumed that they could do anything with the content that they found there. That isa misunderstanding. Without a licence, it is not allowed to do anything with data or other content beyond what is considered as 'fair use'. If a provider wants data to be open, to be used, redistributed and mixed with other content, it should come with an appropriate licence. Such open licences are different from many other content licences:

- to achieve universal participation, no *licensee* is specified;
- to make all uses possible the rights holder waives most or all rights, so no specific terms apply;
- open data is distributed via the internet, so the licence is not limited to a specific *territory*;
- the duration is the same as the duration of the rights that are being waived (we have seen that copyrights expire after a certain time), so there are no *renewal* clauses.

The following open licences are defined as complying with principlesset by the Open Definition (Open Knowledge Foundation, 2020b):

- public domain licence which has no restrictions at all;
- attribution licence which requires credit to the rights holder.
- attribution and share-alike licence which requires attribution and share any derived content or data under the same licence (ODI, 2013c).

4.5.2 Standard open licences

Theoretically, providers could choose to make up their own bespoke open licence. But that is quite complex because the data can be reused anywhere in the world, and so the licence should be valid in many different legislations. Fortunately, there are numerous standard open licences that exist in many languages and for many different legislations. These licences come with statements on different levels:

- a machine-readable version;
- the 'commons deed', a text that is meant to be understandable for everyone, not just legal experts;
- the 'legal code', a text that contains the legal statements that are formulated in such a way that they can be used in court proceedings; there are legal code documents for different national legislations.

Standard open licences are:

- Creative Commons (CC) [**creativecommons.org**]
- Open Data Commons (ODC) [**opendatacommons.org**]
- Government licences, such as the UK Open Government Licence or the French Licence Ouverte.

There are debates about the differences between Creative Commons and Open Data Commons. Creative Commons licences can be applied to many different things that creators want to make available in the public domain, like music and music recordings, pictures, or texts (Creative Commons, 2020). Open Data Commons licences deal with collections held in databases, and the structure of databases, but not the individual content items in the database.

Both CC and ODC licences are used for open data. Government licences are often used to deal with legal requirements that should be met for government organizations, such as a Freedom of Information Act. But a CC or ODC is often used for government data.

Creative Commons licences for creative content

CC is a non-profit organization established in 2001. CC helps to avoid the time and effort to granting/ obtaining permission by providing tools to have the relevant licence on the work in a digital environment. CC licences are available in English by default, but they are also translated into other languages in other national legal systems. CC licences consist of four conditions and six main combinations.

The four main conditions for CC open licences are:

- Attribution (BY): All CC licences require credit to the rights holder in the way it was requested.
- ShareAlike (SA): It is allowed to copy, distribute, display, perform, and modify the work, as long as any modified work is distributed on the same terms.
- NonCommercial (NC): It is allowed to copy, distribute, display, perform, and (unless NoDerivatives is chosen) modify and use the work for any purpose other than commercially.
- NoDerivatives (ND): It is allowed to copy, distribute, display and perform only original copies of the work.

In addition to these four conditions, CC also provides public domain tools for which copyright interests and database rights are waived, allowing the data to be used as freely as possible:

- CC Zero (CC0): The author waives all of his/her copyright and neighbouring and related rights on the work; the rights waived include database rights, so CC0 is suitable to use for data.
- CC Public Domain Mark (PDM): CC provides a public domain mark to generate a licence and anyone can use it to assert that a work is already in the public domain.

Six main combination of licences and their details are given below:

1. Attribution (CC BY): This licence lets others distribute, remix, tweak, and build upon the work, even commercially, as long as they credit the author for the original creation.
2. Attribution-ShareAlike (CC BY-SA): This licence lets others remix, tweak, and build upon the work even for commercial purposes, as long as they credit the author and license their new creations under the identical terms.
3. Attribution-NonCommercial (CC BY-NC): This licence lets others remix, tweak, and build upon the work non-commercially, and although their new works must also acknowledge the author and be non-commercial, they do not have to license their derivative works on the same terms.
4. Attribution-NoDerivs (CC BY-ND): This licence allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.
5. Attribution-NonCommercial-ShareAlike (CC BY-NC-SA): This licence lets others remix, tweak, and build upon the work non-commercially, as long as they credit the author and license their new creations under the identical terms.
6. Attribution-NonCommercial-NoDerivs (CC BY-NC-ND): This licence is the most restrictive of the six main licences, only allowing others to download the works and share them with others as long as they credit you, but they cannot change them in any way or use them commercially.

'Non-commercial' and 'no derivative works' rights are seldom or never reserved for open data. If no derivative works would be allowed, combinations with other datasets or their use in apps would be blocked. There is also a grey area between commercial and non-commercial distribution and, if commercial use is excluded, there is no universal participation.

It is recommended that the latest version of the CC licences is used which are international. The versions of the licences prior to CC version 4.0 International were not specifically aimed at data. Version 4.0 licenses explicitly cover sui generis database rights such as the one in force in the European Union. "All versions of the licences treat datasets and databases as a whole: they do not treat the individual data themselves differently from the collection/database" (Ball, 2014). Therefore, they should be carefully applied in certain complex cases such as collections of variously copyrighted works. The degree of openness in CC licences also matters. Some of the CC licences are more 'free' than the others which are CC0, PDM, CC BY, and CC BY-SA and described as free culture licences (Creative Commons, 2020).

Open data licences for databases

The Open Data Commons, which started in 2007, is an Open Knowledge Foundation project. It offers similar licences to Creative Commons but designed specifically for databases. Open Data Commons has three licences as follows:

1. Public Domain Dedication and Licence (PDDL) – 'Public domain for data/databases': this allows a user to copy, distribute and use the database (share); to produce works from the database (create); and to modify, transform and build upon the database (adapt). The PDDL imposes no restrictions on the use of the PDDL licensed database. It accomplishes the same thing in the same way as CC0 but is worded specifically in database terms.
2. Attribution Licence (ODC-By) – 'Attribution for data/databases': it allows a user to copy, distribute and use the database (share); to produce works from the database (create); and to modify, transform and build upon the database (adapt), as long as the user attributes any public use of the database, or works produced from the database, in the manner specified in the licence.
3. Open Database Licence (ODC-ODbL) – 'Attribution Share-Alike for data/databases': It gives the same permissions as ODC-By. In addition, (i) any adapted version of this database or works produced from an adapted database should also be offered under the ODbL; (ii) a licensor can apply technical restrictions to new work as long as an alternative copy without the restrictions is made equally available.

©Adobe Stock/jackfrog

4.5.3 How to use open licences?

Open licences usually come with layers including human-readable and machine-readable versions. The human-readable layer is for people to acknowledge and the machine-readable layer is for machines to read and process the licences. They both should clearly indicate which licence applies to the content or data and how it can be reused by others. Creative Commons and Open Data Commons define what statements and marks should be used for each of their licences on their web sites.

Creative Commons offers a web-based tool, the license chooser, to help select the right licence for specific needs. Open Data Commons also provides instructions on how to apply licences.

Having a machine-readable licence, including a complete description of the metadata, is important for the content and data to be correctly harvested by machines, e.g. search engines and web APIs. ODI's *Publisher's Guide to the Open Data Rights Statement Vocabulary* provides excellent insight on the topic (The Open Data Institute (ODI), 2013b). This is equally important for the licensed work to be searched, browsed or filtered correctly on search engines. This topic was discussed widely in Section 3.1. Using open data and Section 3.2. Quality and provenance.

Table 18. Standard open licenses compliant with Open Definition

Name	Licence	Attribution (BY)	Share Alike (SA)	Remarks
CC Zero (CC)	CC	No	No	All rights waived. Recommended for scientific data to make data mining and meta analyses possible
Public Domain Dedication and Licence (PDDL)	ODC	No	No	All rights waived. Recommended for scientific data to make data mining and meta analyses possible
Creative Commons Attribution 4.0 (CC BY)	CC	Yes	No	
Open Data Commons Attribution Licence (ODC BY)	ODC	Yes	No	
Creative Commons Attribution Share Alike (CC BY SA)	CC	Yes	Yes	
Open Database Licence (ODbL)	ODC	Yes	Yes	

Source: Open Knowledge Foundation, 2020c.

References

Abalobi – A mobile app suite for small-scale fisheries governance. 2020 [online]. the Republic of South Africa. [Cited 5 May 2020]. http://abalobi.info/about/#graphic-harvest

ACMO. 2013. ACMO (AgMIP Crop Model Output) Data definitions - In GODAN Agrisemantics Map of Standards [online]. [Cited 27 September 2020].

Addison, C., Figuères, C., Oweyesiga, H., Muwonge, D., Nsimidala, E., Sezibera, A., Boyera, S., Besemer, H., Pesce, V., Birba, A. & Muyiramye, D. 2020. Data-driven opportunities for farmer organisations. CTA. (also available at https://cgspace.cgiar.org/handle/10568/108356).

African Union. 2014. African Union Convention on Cyber Security and Personal Data Protection. Addis Ababa, African Union. 37 pp. (also available at https://au.int/sites/default/files/treaties/29560-treaty-0048_-_african_union_convention_on_cyber_security_and_personal_data_protection_e.pdf).

Agbox. 2020. In Canada Digital Agri-Food. [online]. Canada. [Cited 24 September 2020]. www.digitalag.ca

AgGateway. 2020. Standards & Guidelines. In: AgGateway [online]. [Cited 27 September 2020]. www.aggateway.org/GetConnected/StandardsGuidelines.aspx

Agriculture Department. 2020. Agriculture Statistics. In Tanzania National Bureau of Statistics. [online]. Tanzania. [Cited 24 September 2020]. www.nbs.go.tz/index.php/en/census-surveys/agriculture-statistics

Anannya, A.M.A. 2018. Messages for safe fishing in Senegal. In: NRC [online]. [Cited 27 September 2020]. www.nrc.no/expert-deployment/2016/2018/messages-for-safe-fishing-in-senegal

Anderson, J.E. 1974. Public Policymaking. New York, Praeger.

ArcGIS Desktop. 2020. How Aspect works. In: ArcMap [online]. [Cited 1 October 2020]. https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-aspect-works.htm

Aubert, C., Buttigieg, P.L., Laporte, M.A., Devare, M. & Arnaud, E. 2017. CGIAR Agronomy Ontology. In: The OBO Foundry [online]. [Cited 27 September 2020]. https://github.com/AgriculturalSemantics/agro

Ault, A. 2016. Open Ag Data Alliance, Servi-Tech launch Real-Time Connections API for weather, soil moisture data. In: Purdue University Agriculture News [online]. [Cited 27 September 2020]. www.purdue.edu/newsroom/releases/2016/Q3/open-ag-data-alliance,-servi-tech-launch-real-time-connections-api-for-weather,-soil-moisture-data.html

Ball, A. 2014. How to License Research Data: Licensing concepts. In: Digital Curation Centre [online]. [Cited 27 September 2020]. www.dcc.ac.uk/guidance/how-guides/license-research-data#top

Ball, A. 2014. How to Licence Research Data. Digital Curation Centre. 16 pp. (also available at www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf)

Banisar, D. 2019. National Comprehensive Data Protection/Privacy Laws and Bills 2019. SSRN [online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1951416

Berners-Lee, T. 2012. 5-star Open Data. [online]. United Kingdom. [Cited 24 September 2020]. http://5stardata.info/en

BBC. 2010. Farmers win privacy fight over EU funding [online]. [Cited 15 September 2020]. www.bbc.co.uk/news/world-europe-11724893

BBC. 2013. Michael Fish's denial of a hurricane in 1987 [online]. [Cited 15 September 2020]. www.bbc.co.uk/news/av/world-24713504/michael-fish-s-denial-of-a-hurricane-in-1987

de Beer, J. 2017. Ownership of Open Data: Governance Options for Agriculture and Nutrition. Wallingford, Global Open Data for Agriculture & Nutrition (GODAN). 24 pp. (also available at www.godan.info/sites/default/files/documents/Godan_Ownership_of_Open_Data_Publication_lowres.pdf).

Berners-Lee, T. 2006. Linked Data. World Wide Web Consortium [online]. [Cited 15 September 2020]. www.w3.org/DesignIssues/LinkedData.html

Bloch, S. 2018. If farmers sold their data instead of giving it away, would anybody buy? [online]. [Cited 15 September 2020]. https://newfoodeconomy.org/farmobile-farm-data/

Bourne, J., ed. 2007. Bovine TB: The Scientific Evidence. London, Independent Scientific Group on Cattle TB. 287 pp. (also available at www.bovinetb.info/docs/final_report.pdf).

Boyera, S., Addison, C. & Msengezi, C. 2017. Farmer profiling: Making data work for smallholder farmers. Wageningen, Technical Centre for Agricultural and Rural Cooperation (CTA). 76 pp. (also available at https://publications.cta.int/media/publications/downloads/2014_PDF.pdf).

Boyera, S. van Schalkwyk, F. & Grewal, A. 2017. Open Data in Ethiopia: Report on the Open Data Landscape in Ethiopia. Prepared in Support of the Development of the National Policy & Guidelines for the Implementation of Open Data in Ethiopia. Addis Ababa, Ministry of Communication & Information Technology of The Federal Democratic Republic of Ethiopia. 38 pp. (also available at www.academia.edu/38416197/Open_Data_in_Ethiopia).

Briganti, D. 2020. W3C XML Schema (XSD) Validation online. In Utilities-Online.Info. [online]. [Cited 15 September 2020]. www.utilities-online.info/xsdvalidation

Brussels Briefings. 2018a. Projet AGRITIC FEPAB. In: SlideShare [online]. [Cited 15 September 2020]. www.slideshare.net/brusselsbriefings/fepab-agritic-project-west-africa.

Brussels Briefings. 2018b. The Impact of Improved Technologies at Igara Growers Tea Factory Ltd (IGTF), Uganda. In: SlideShare [online]. [Cited 15 September 2020]. https://www.slideshare.net/brusselsbriefings/the-impact-of-of-improved-technologies-at-igara-growers-tea-factory-ltd-uganda.

Build your own tables. 2019. In HM Revenue & Customs. [online]. [Cited 15 September 2020]. www.uktradeinfo.com/trade-data/help-with-data-tables

Carbonell, I.M. 2016. The ethics of big data in big agriculture. Internet Policy Review, 5(1) [online]. https://policyreview.info/articles/analysis/ethics-big-data-big-agriculture

Carolan, L., Smith, F., Protonotarios, V., Schaap, B., Broad, E., Hardings, J. & Gerry, W. 2015. How can we improve agriculture, food and nutrition with open data? Wallingford, Global Open Data for Agriculture & Nutrition (GODAN). 34 pp. (also available at www.godan.info/documents/how-can-we-improve-agriculture-food-and-nutrition-open-data).

Carrington, D. 2017. Huge increase in badger culling will see up to 33,500 animals shot [online]. [Cited 15 September 2020]. www.theguardian.com/environment/2017/sep/11/huge-increase-badger-culling-see-up-to-33500-animals-shot

CARTO. 2020. Tutorials. In CARTO. [online]. [Cited 5 May 2020]. https://carto.com/help/tutorials/your-account

CartONG. 2017. Benchmarking of Mobile Data Collection Solutions: What aspects to consider when choosing a tool/platform. Chambéry, CartONG. 67 pp. (also available at https://blog.cartong.org/wordpress/wp-content/uploads/2017/08/Benchmarking_MDC_2017_CartONG_2.pdf).

Cecconi, G. & Radu, C. 2018. Open Data Maturity in Europe. Brussels, European Commission. 113 pp. (also available at www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n4_2018.pdf).

CF Metadata. 2020. NetCDF CF Metadata Conventions. In: CF Metadata [online]. [Cited 27 September 2020]. http://cfconventions.org

Chicago Data Portal. 2020. Crimes – 2001 to Present. In The City of Chicago Data Portal. [online]. United States of America. [Cited 5 May 2020]. https://data.cityofchicago.org/Public-Safety/Crimes-2020/qzdf-xmn8

CNIL. 2019. Data protection around the world. In: CNIL [online]. [Cited 27 September 2020]. www.cnil.fr/en/data-protection-around-the-world

Coherence in Information for Agricultural Research for Development (CIARD). 2011. Interim Proceedings of International Expert Consultation on "Building the CIARD Framework for Data and Information Sharing". Rome, CIARD. 242 pp. (also available at www.fao.org/docs/eims/upload/297074/IECProceedings-main-doc.pdf).

Communauté Economique des Etats de l'Afrique de l'Ouest (CEDEAO) & Economic Community of West African States (ECOWAS). 2010. Trente Septieme Session de la Conference des Chefs d'Etat et de Gouvernement. Lagos, Nigeria. 24 pp. (also available at www.afapdp.org/wp-content/uploads/2018/06/CEDEAO-Acte-2010-01-protection-des-donnees.pdf).

Confédération des Associations des Producteurs Agricoles pour le développement. [online]. the Republic of Burundi. [Cited 5 May 2020]. ww.capad.info/?lang=fr

Copa Cogeca, CEMA, Fertilizers Europe, CEETTAR, CEJA, ECPA, EFFAB, FEFAC, & ESA. 2018. EU Code of conduct on agricultural data sharing by contractual agreement. Copa Cogeca. (also available at www.ecpa.eu/sites/default/files/documents/AgriDataSharingCoC_2018.pdf)

Core Trust Seal. 2020. Data Seal of Approval Synopsis (2008–2018). In: CoreTrustSeal [online]. [Cited 27 September 2020]. www.coretrustseal.org/about/history/data-seal-of-approval-synopsis-2008-2018

CTA. 2018a. Jackson Byaruhanga and Hamlus Owoyesiga from Igara Growers Tea Factory Limited, Uganda. [video]. [Cited 24 September 2020] https://vimeo.com/263101646

CTA. 2018b. Hamlus Owoyesiga, Network and Systems Administrator, Igara. Growers Tea Factory Ltd., Uganda. [video]. [Cited 24 September 2020]. https://vimeo.com/262555014

CTA. 2019. Farmer registration and profiling: How did it go? In: KM4ARD Experience Capitalization [online]. [Cited 27 September 2020]. http://experience-capitalization.cta.int/farmer-registration-and-profiling-how-did-it-go/index.html

Data Collaboratives. 2020. [online]. [Cited 24 September 2020]. http://datacollaboratives.org/

Daunivalu, J. 2018. Empowering Fijian Farmers with Mobile Apps. In CTA. Resilience and productivity in the Pacific. Experience Capitalization Series 7. Wageningen, Technical Centre for Agricultural and Rural Cooperation (CTA). 68 pp. (also available at https://cgspace.cgiar.org/bitstream/handle/10568/98979/Exp_Cap7_Daunivalu.pdf?sequence=1&isAllowed=y).

Data Interoperability Standards Consortium. 2020. What is 'Data Interoperability?' In: The Data Interoperability Standards Consortium [online]. [Cited 27 September 2020]. https://datainteroperability.org

Davies, T. 2012. Five Stars. In: Open Data Engagement [online]. [Cited 27 September 2020]. www.opendataimpacts.net/engagement

Dickie, M.R. 2013. This TechCrunch Disrupt Winner Could Be the Future Of Search [online]. [Cited 15 September 2020]. www.businessinsider.com/techcrunch-disrupt-winner-enigma-2013-5?IR=T

DLG. 2018. Digital Agriculture – Opportunities. Risks. Acceptance: A DLG position paper [online]. Frankfurt. [Cited 15 September 2020]. www.dlg.org/en/agriculture/topics/a-dlg-position-paper/

DSSAT. 2020. Data Standards. In: DSSAT.net [online]. [Cited 27 September 2020]. https://dssat.net/data/standards_v2

Dyson, L. 2013. Foodies and Open Data Enthusiasts Rejoice. Code for America [online]. [Cited 18 May 2015]. www.codeforamerica.org/blog/2013/01/17/foodies-and-open-data-enthusiasts-rejoice

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J. & Signell, R. 2011. NetCDF Climate and Forecast (CF) Metadata Conventions. (also available at http://cfconventions.org/cf-conventions/v1.6.0/cf-conventions.html).

ecancer. 2012. Cancer incidence predicted to increase 75% by 2030 – Ecancer. In Ecancer. [online]. [Cited 24 September 2020]. http://ecancer.org/en/news/2870-cancer-incidence-predicted-to-increase-75-by-2030

Eligibility Criteria & OGP Values Check Assessment. 2020. In Open Government Partnership. [online]. [Cited 24 September 2020]. www.opengovpartnership.org/process/joining-ogp/eligibility-criteria/

European Commission. 2018. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Towards a Common European Data Space. Brussels, European Commission. 15 pp. (also available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0232&from=EN).

European Commission. 2018a. Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union. Official Journal of the European Union (L 303/59). (also available at https://eur-lex.europa.eu/eli/reg/2018/1807/oj).

European Commission. 2018b. Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union (L 127/2). (also available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679R(02)).

European Commission. 2019. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. Official Journal of the European Union (PE/28/2019/REV/1). (also available at https://eur-lex.europa.eu/eli/dir/2019/1024/oj).

European Commission. 2020. Open data: agent of change [online]. [Cited 15 September 2020]. www.europeandataportal.eu/elearning/en/module3/#/id/co-01

European Statistical System. 2017. Position paper on access to privately held data which are of public interest. Opening up new data sources for a new generation of official statistics – in light of the growing European Digital Single Market and the revision of the Public Sector Information Directive. Brussels, European Statistical System. 19 pp. (also available at https://ec.europa.eu/eurostat/documents/7330775/8463599/ESS+Position+Paper+on+Access+to+privately+held+data+final+-+Nov+2017.pdf/6ef6398f-6580-4731-86ab-9d9d015d15ae).

FAO. 2014. The State of Food and Agriculture 2014: Innovation in Family Farming. Rome, FAO. 139 pp. (also available at www.fao.org/publications/sofa/2014/en/).

FAO. 2015. World Programme for the Census of Agriculture 2020: Volume 1 – Programme, concepts and definitions. Rome. 204 pp. (also available at www.fao.org/3/a-i4913e.pdf).

FAO. 2018. Status of Implementation of e-Agriculture in Central and Eastern Europe and Central Asia. Insights from selected countries in Europe and Central Asia. Budapest, FAO. 64 pp. (also available at www.fao.org/3/I8303EN/i8303en.pdf).

FAO. 2019. Family Farming Knowledge Platform: Smallholders Dataportrait [online]. Rome. [Cited 19 September 2019]. www.fao.org/family-farming/data-sources/dataportrait/production/en/.

FAO, IFAD, UNICEF, WFP & WHO. 2019. The State of Food Security and Nutrition in the World: Safeguarding Against Economic Slowdowns and Downturns. Rome, FAO. 214 pp. (also available at www.fao.org/3/ca5162en/ca5162en.pdf).

Farmobile. 2020. [online]. United States of America. [Cited 24 September 2020]. www.farmobile.com/

Federal Ministry of Food and Agriculture (BMEL). 2019. Berlin World Food Conference: Governments agree on communiqué on digital transformation and agriculture [online]. [Cited 15 September 2020]. www.bmel.de/EN/topics/international-affairs/global-forum-for-food-and-agriculture/gffa-2019.html

Ferris, L. & Rahman, Z. 2016a. Responsible Data in Agriculture. In: The Engine Room Library [online]. [Cited 5 May 2020]. https://library.theengineroom.org/agriculture

Ferris, L. & Rahman, Z. 2016b. Responsible Data in Agriculture. Wallingford, Global Open Data for Agriculture & Nutrition (GODAN). 16 pp. (also available at www.godan.info/sites/default/files/documents/Godan_Responsible_Data_in_Agriculture_Publication_lowres.pdf).

FORCE11. 2014. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0. In FORCE11. [online]. [Cited 5 May 2020] www.force11.org/fairprinciples

Garrett, J.J. 2005. Ajax: A New Approach to Web Applications. Adaptive Path. 4 pp. (also available at https://immagic.com/eLibrary/ARCHIVES/GENERAL/ADTVPATH/A050218G.pdf).

GODAN Action. 2018. Open Data Management in Agriculture and Nutrition Course. In: GODAN Global Open Data for Agriculture and Nutrition [online]. [Cited 1 October 2020]. www.godan.info/open-data-management-agriculture-and-nutrition-course-english

OGC. 2020a. Geography Markup Language. In: OGC [online]. [Cited 27 September 2020]. www.ogc.org/standards/gml

OGC. 2020b. Observations and Measurements. In: OGC [online]. [Cited 27 September 2020]. www.ogc.org/standards/om

GiSC. 2020. Grower's Information Services Coop. [online]. United States of America. [Cited 24 September 2020]. www.gisc.coop/

Gray, J. 2014. Towards a Genealogy of Open Data. Presented at General Conference of the European Consortium for Political Research, Glasgow, UK, 3–6 September 2014. (also available at http://dx.doi.org/10.2139/ssrn.2605828).

Gray, B., Babcock, L., Tobias, L., McCord, M., Herrera, A., Osei, C. & Cadavid, R. 2018. Digital Farmer Profiles: Reimagining Smallholder Agriculture. Washington DC, United States Agency for International Development (USAID). 91 pp. (also available at www.usaid.gov/sites/default/files/documents/15396/Data_Driven_Agriculture_Farmer_Profile.pdf).

Global Open Data for Agriculture & Nutrition (GODAN). 2016. Introducing the Agriculture Open Data Package BETA Version. Wallingford, GODAN. 16 pp. (also available at www.godan.info/sites/default/files/GODAN_AG_Open_Data_Package_PRINT_1.pdf).

GODAN Secretariat. 2020. Mission. In: GODAN - Global Open Data for Agriculture and Nutrition [online]. [Cited 27 September 2020]. www.godan.info/pages/mission

GODAN Secretariat. 2017a. Open Climate. [video]. [Cited 24 September 2020]. www.youtube.com/watch?v=y9SwIpHT5A4&feature=youtu.be

GODAN Secretariat. 2017b. Open Skies. [video]. [Cited 24 September 2020]. www.youtube.com/watch?v=MJ0VX7CqmSc&feature=youtu.be

Good Growth Plan Progress Data. 2020. In Syngenta Global. [online]. [Cited 24 September 2020]. http://opendata.syngenta.agroknow.com/the-good-growth-plan-progress-data

Government of the Republic of Macedonia. 2018. Open Data Strategy (2018-2020). Skopje, Ministry of Information Society and Administration. 86 pp. (also available at http://mioa.gov.mk/sites/default/files/pbl_files/documents/strategies/open_data_strategy_en.pdf).

Hallsworth, M., Parker, S. & Rutter, J. 2011. Policy Making in the Real World: Evidence and Impact. London, Institute for Government. 105 pp. (also available at www.instituteforgovernment.org.uk/sites/default/files/publications/Policy%20making%20in%20the%20real%20world.pdf).

HIMSS. 2019. What is Interoperability? HIMSS 2013 Definition. In: HIMSS [online]. [Cited 27 September 2020]. www.himss.org/previous-himss-interoperability-definitions

Hirst, A. T. 2011. Merging Datasets with Common Columns in Google Refine. OUseful.Info, the Blog... [online]. [Cited 24 September 2020]. https://blog.ouseful.info/2011/05/06/merging-datesets-with-common-columns-in-google-refine

How to extract part of text string from cell in Excel? 2020. Extend Office. [online]. [Cited 24 September 2020]. www.extendoffice.com/documents/excel/3639-excel-extract-part-of-string.html

ICASA. 2013. ICASA Master Variable List. In: AgMIP Software Development [online]. [Cited 27 September 2020]. https://vest.agrisemantics.org/content/agmip-icasa-master-variable-list

IFPRI. 2020. Food Security Portal [online]. [Cited 27 September 2020]. www.foodsecurityportal.org

International Food Policy Research Institute (IFPRI). 2016. Global Nutrition Report 2016: From Promise to Impact: Ending Malnutrition by 2030. Washington DC, IFPRI. 180 pp. (also available at https://globalnutritionreport.org/reports/2016-global-nutrition-report/).

Technologies for African Agricultural transformation (TAAT). 2018. Taat policy support retreat. Paper presented at, 15 March 2018. [Cited 20 September 2019]. www.slideshare.net/ifpriwcao/taat-policy-support-retreat

ISO. 2014. ISO 19101-1:2014(en), Geographic information – Reference model – Part 1: Fundamentals. In: ISO Online Browsing Platform (OBP) [online]. [Cited 27 September 2020]. www.iso.org/obp/ui/#iso:std:iso:19101:-1:ed-1:v1:en

Jansen, B.J. & Pooch, U. 2000. A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology [online]. http://onlinelibrary.wiley.com/doi/10.1002/1097-4571(2000)9999:9999%3C::AID-ASI1607%3E3.0.CO%3B2-F/full

Jensen, R. 2007. The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector, The Quarterly Journal of Economics, CXXII(3): 879–924. (also available at https://mmd4d.files.wordpress.com/2009/04/jensen-indian-fisheries.pdf).

Kritikos, M. 2017. Precision agriculture in Europe. Legal, social and ethical considerations. Brussels, European Parliament. 80 pp. (also available at www.europarl.europa.eu/RegData/etudes/STUD/2017/603207/EPRS_STU(2017)603207_EN.pdf).

Maru, A., Berne, D., De Beer, J., Ballantyne, P., Pesce, V., Kalyesubula, S., Addison, C., *et al.* 2018. Digital and Data-Driven Agriculture: Harnessing the Power of Data for Smallholders. Rome, Global Forum on Agricultural Research and Innovation (GFAR); Wallingford, Global Open Data for Agriculture & Nutrition (GODAN); Wageningen, Technical Centre for Agricultural and Rural Cooperation (CTA). 38 pp. (also available at https://hdl.handle.net/10568/92477).

Mazzucato, M. 2018. Let's make private data into a public good [online]. [Cited 15 September 2020]. www.technologyreview.com/2018/06/27/141776/lets-make-private-data-into-a-public-good

Merge two excel files using a common column. 2020. Stack Exchange - Super User. [online]. [Cited 15 September 2020] https://superuser.com/questions/366647/merge-two-excel-files-using-a-common-column

Microsoft. 2015. FarmBeats: AI, Edge & IoT for Agriculture. In: Microsoft Research [online]. [Cited 27 September 2020]. www.microsoft.com/en-us/research/project/farmbeats-iot-agriculture

Muwonge, D. 2018. Value addition through digitalisation for Uganda coffee farmers. ICT Update, Issue 89 [online]. [Cited 15 September 2020]. www.cta.int/en/article/value-addition-through-digitalisation-for-ugandan-coffee-farmers-sid063cfbfb5-a904-4078-a296-4b563772f53c

NUCAFE. 2020. [online]. Uganda. [Cited 5 May 2020]. https://nucafe.org

OGC & W3C. 2017. Semantic Sensor Network Ontology. In: W3C [online]. [Cited 27 September 2020]. https://github.com/NCEAS/oboe/tree/OBOE_1_1

Open Data Institute (ODI). 2013a. Engaging with Reusers [online]. [Cited 15 September 2020]. https://theodi.org/article/engaging-with-reusers

Open Data Institute (ODI). 2013b. Publisher's Guide to the Open Data Rights Statement Vocabulary [online]. [Cited 15 September 2020]. https://theodi.org/article/publishers-guide-to-the-open-data-rights-statement-vocabulary

Open Data Institute (ODI). 2013c. Reuser's Guide to Open Data Licensing [online]. [Cited 15 September 2020]. https://theodi.org/article/reusers-guide-to-open-data-licensing

Open Data Institute (ODI). 2016. We Need to Learn How to Search the Web of Data [online]. [Cited 15 September 2020]. https://theodi.org/article/we-need-to-learn-how-to-search-the-web-of-data

Open Data Institute (ODI). 2019a. The Data Spectrum. In: The Open Data Institute (ODI) [online]. [Cited 27 September 2020]. https://theodi.org/about-the-odi/the-data-spectrum

The Open Data Institute (ODI). 2019b. Open Data in a Day [online]. [Cited 1 October 2020]. http://training.theodi.org/InADay/#/id/co-01

The Open Data Institute (ODI). 2020. ODI Open Data Certificate [online]. [Cited 25 September 2020]. https://certificates.theodi.org/en

Open Knowledge Foundation. 2020a. Why open data? In: Open Knowledge Foundation [online]. [Cited 27 September 2020]. https://okfn.org/opendata/why-open-data

Open Knowledge Foundation. 2020b. Open Definition 2.1 - Open Definition – Defining Open in Open Data, Open Content and Open Knowledge. In: Open Definition [online]. [Cited 27 September 2020]. http://opendefinition.org/od/2.1/en

Open Knowledge Foundation. 2020c. Conformant Licenses [online]. [Cited 27 September 2020]. http://opendefinition.org/licenses

Pesce. V. 2017. Semantic challenges in sharing dataset metadata and creating federated dataset catalogs. The example of the CIARD RING. Paper presented at MACS G20 Workshop, 2017, Berlin. [Cited 27 September 2020]. www.slideshare.net/valeriap/semantic-challenges-in-sharing-dataset-metadata-and-creating-federated-dataset-catalogs-the-example-of-the-ciard-ring

Pesce, V. 2019. Digital agriculture: data-related policy issues and current policy directions in Europe and Central Asia. Background research funded by the World Bank. (Unpublished).

Pesce, V., Tennison, J., Dodds, L. & Zervas, P. 2017. Weather data standards: a gap exploration report. Wallingford, Global Open Data for Agriculture and Nutrition (GODAN). 46 pp. (also available at https://doi.org/10.7490/f1000research.1115856.1).

Rapsomanikis, G. 2015. The Economic Lives of Smallholder Farmers: An Analysis Based on Household Data from Nine Countries. Rome, FAO. 39 pp. (also available at www.fao.org/3/a-i5251e.pdf).

Rasmussen, N. 2016. From Precision Agriculture to Market Manipulation: A New Frontier in the Legal Community. Minnesota Journal of Law, Science & Technology, 17(1): 489–516. (also available at https://scholarship.law.umn.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1008&context=mjlst).

Rivenes, L. 2019. A GDPR Overview for eCommerce Retailers. In. DataFeedWatch Blog [online]. [Cited 30 November 2020]. www.datafeedwatch.com/blog/a-gdpr-overview-for-ecommerce-retailers

Sanderson, J., Wiseman, L. & Poncini, S. 2018. What's behind the ag-data logo? An examination of voluntary agricultural-data codes of practice. International Journal of Rural Law and Policy, 01 [online]. https://epress.lib.uts.edu.au/journals/index.php/ijrlp/article/view/6043

Save the Children and the Open University. 2014. 6 Methods of Data Collection and Analysis. Save the Children. 30 pp. (also available at https://resourcecentre.savethechildren.net/node/12203/pdf/6_methods_of_data_collection.pdf).

Schrijver, R., ed. 2016. Precision agriculture and the future of farming in Europe: Scientific Foresight Study. Brussels, European Parliament. 42 pp. (also available at https://publications.europa.eu/en/publication-detail/-/publication/40fe549e-cb49-11e7-a5d5-01aa75ed71a1/language-en).

Smart Nkunganire System. 2020. [online]. Rwanda. [Cited 24 September 2020]. www.smartnkunganire.rw

Smith, C. 2017. The iPhone 8's glass back might be fragile, but it's still better than the Note 8. In BGR. [online]. [Cited 24 September 2020]. https://bgr.com/2017/09/25/iphone-8-plus-vs-galaxy-note-8-drop-test-glass

Socrata. 2015. Could the LIVES Standard Reduce Food Poisonings? [online]. [Cited 18 May 2015]. www.socrata.com/blog/lives-helps-prevent-food-poisoning-in-restaurants

Stoodley, C. 2019. The Brain andReading [online]. [Cited 1 October 2020]. www.waece.org/cd_morelia2006/ponencias/stoodley.htm

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. & Smith, P. 2016. Report of the Inquiry into the 2015 British general election opinion polls. Project Report. London, British Polling Council, and the Market Research Society. 120 pp. (also available at http://eprints.ncrm.ac.uk/3789/).

Svensson, J. & Yanagizawa, D. 2010. Getting Prices Right: The Impact of the Market Information Service in Uganda, Journal of the European Economic Association, 7(2-3): 435–445. (also available at https://yanagizawadrott.com/wp-content/uploads/2016/02/getting-prices-right.pdf).

Sylvester, G., ed. 2019. E-agriculture in Action: Blockchain for Agriculture – Opportunities and Challenges. Rome, FAO. 72 pp. (also available at www.fao.org/3/CA2906EN/ca2906en.pdf).

Technical Centre for Agricultural and Rural Cooperation (CTA). 2019. Farmers organisations experiences with the data collection for farmer's profiles: Second interim report of CTA. Wageningen. (unpublished).

TechnoServe. 2017. Technical Solutions for Food Security in Africa [online]. [Cited 15 September 2020]. www.technoserve.org/blog/technical-solutions-for-food-security-in-africa

Tennison, J. & Scott, A. 2018. Getting paid for personal data won't make things better [online]. [Cited 15 September 2020]. https://theodi.org/article/jeni-tennison-getting-paid-for-personal-data-wont-make-things-better

Townsend, T., Lampietti, J., Treguer, D., Schroeder, K., Haile, M., Juergenliemk, A., Hasiner, E. *et al.* 2019. Future of Food: Harnessing Digital Technologies to Improve Food System Outcomes. Washington DC, World Bank. 44 pp. (also available at http://hdl.handle.net/10986/31565).

Tsan, M., Totapally, S., Hailu, M. & Addom, B.K. 2019. The Digitalisation of African Agriculture Report, 2018-2019. Wagneingen, CTA. 238 pp. (also available at **www.cta.int/en/digitalisation-agriculture-africa**).

UNECE. 2010. UN/EDIFACT Message DAPLOS Release: 10A. In: UN/EDIFACT [online]. [Cited 27 September 2020]. https://service.unece.org/trade/untdid/d10a/trmd/daplos_c.htm

UK Data Service. 2020. Costing data management [online]. [Cited 27 September 2020]. www.ukdataservice.ac.uk/manage-data/plan/costing

Uganda Tea Development Agency. 2020. [online]. Uganda. [Cited 5 May 2020]. http://www.ugatea.com/

Van der Wees, A., Stefanatou, D., Svorc, J., van den Ham, M., Vermesan, O., Annicchino, P., Ziegler, S. & Scudiero, L. 2017. H2020 – CREATE-IoT Project. Deliverable 05.05. Legal IoT Framework (Initial). Brussels, European Commission. (also available at https://european-iot-pilots.eu/wp-content/uploads/2018/02/D05_05_WP05_H2020_CREATE-IoT_Final.pdf).

Vigen, T. 2020. Spurious Correlations. In Tylervigen. [online]. [Cited 25 September 2020]. http://tylervigen.com/spurious-correlations

W3C. 2014. RDF Data Cube Vocabulary. In: W3C [online]. [Cited 27 September 2020]. www.w3.org/TR/vocab-data-cube

W3C. 2015. Ontologies. In: W3C [online]. [Cited 27 September 2020]. www.w3.org/standards/semanticweb/ontology

W3C. 2017. Data on the Web Best Practices. In: W3C [online]. [Cited 27 September 2020]. www.w3.org/TR/dwbp

Waite, R., Hanson, C., Searchinger, T, & Ranganathan, J. 2018. This Is How to Sustainably Feed 10 Billion People by 2050. In: World Economic Forum [online]. Geneva. [Cited 20 September 2019]. www.weforum.org/agenda/2018/12/how-to-sustainably-feed-10-billion-people-by-2050-in-21-charts/.

What is open data. In European Data Portal. [online]. [Cited 24 September 2020]. www.europeandataportal.eu/en/training/what-open-data

What is Open Data?. In Open Data Handbook [online]. [Cited 24 September 2020]. http://opendatahandbook.org/guide/en/what-is-open-data

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. Computers and Electronics in Agriculture, 96:1–12. (also available at http://dssat.net/wp-content/uploads/2014/02/White2013ICASA_V2_standards.pdf).

Wiebe, K.D., Sulser, T.B., Mason-D'Croz, D. & Rosegrant, M.W. 2017. The effects of climate change on agriculture and food security in Africa. In A. De Pinto & J.M. Ulimwengu, eds. A Thriving Agricultural Sector in a Changing Climate: Meeting Malabo Declaration Goals Through Climate-Smart Agriculture. pp. 5–21. Washington DC, USA, International Food Policy Research Institute. (also available at http://dx.doi.org/10.2499/9780896292949_02).

Wikipedia. 2019. Data localization [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Data_localization

Wikipedia. 2020a. Facebook – Cambridge Analytica data scandal [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

Wikipedia. 2020b. World Wide Web [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/World_Wide_Web

Wikipedia. 2020c. Yahoo! Directory [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Yahoo!_Directory

Wikipedia. 2020d. AltaVista [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/AltaVista

Wikipedia. 2020e. Lycos [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Lycos

Wikipedia. 2020f. PageRank [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/PageRank

Wikipedia. 2020g. The dress [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/The_dress

Wikipedia. 2020h. Data visualization [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Data_visualization

Wikipedia. 2020i. Extract, transform, load [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Extract,_transform,_load

Wikipedia. 2020j. Open access [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Open_access

Wikipedia. 2020k. Data. [online]. [Cited 27 September 2020]. https://en.wikipedia.org/w/index.php?title=Data&oldid=978303903

Wikipedia. 2020l. SPARQL [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/SPARQL

Wikipedia. 2020m. Representational state transfer [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Representational_state_transfer

Wikipedia. 2020n. OPeNDAP [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/OPeNDAP

Wikipedia. 2020o. Resource Description Framework [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Resource_Description_Framework

Wikipedia. 2020p. Ontology (information science) [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Ontology_(information_science)

Wikipedia. 2020r. Multi-Crop Passport Descriptor [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Multi-Crop_Passport_Descriptor

Wikipedia. 2020s. Uniform Resource Identifier [online]. [Cited 15 September 2020]. https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

Wiley, D. 2020. Defining the 'Open' in Open Content and Open Educational Resources [online]. [Cited 27 September 2020]. http://opencontent.org/definition

Witkin, R. 1983. Jet's Fuel Ran Out After Metric Conversion Errors [online]. [Cited 15 September 2020]. www.nytimes.com/1983/07/30/us/jet-s-fuel-ran-out-after-metric-conversion-errors.html?mcubz=1

World Bank. 2017. Cereal yield (kg per hectare) | Data. In: World Bank Open Data [online]. [Cited 27 September 2020]. https://data.worldbank.org/indicator/AG.YLD.CREL.KG

World Bank. 2020. Agriculture & Rural Development. In World Bank Open Data. [online]. [Cited 24 September 2020]. https://data.worldbank.org/topic/agriculture-and-rural-development?locations=KE-TZ-RW-GH-NG-ML-BF

World Bank. 2020. DataBank. [online]. [Cited 24 September 2020] https://databank.worldbank.org/home.aspx

World Intellectual Property Organization (WIPO). 2014. Intellectual Property Handbook: Policy, Law and Use. Geneva, WIPO. 488 pp. (also available at www.wipo.int/about-ip/en/iprm/).

World Intellectual Property Organization (WIPO). 2016. Understanding Copyright and Related Rights. Geneva, WIPO. 40 pp. (also available at http://www.wipo.int/edocs/pubdocs/en/wipo_pub_909_2016.pdf).

World Wide Web Foundation. 2020. Get the data | Open Data Barometer. In: Open Data Barometer [online]. [Cited 27 September 2020]. https://opendatabarometer.org/leadersedition/data

World Wide Web Size. 2020. The size of the World Wide Web (The Internet) [online]. [Cited 15 September 2020]. www.worldwidewebsize.com

©Adobe Stock/pressmaster

This book aims to strengthen the skills of professionals who use, manage data for the benefit of farmers and farmers organizations by exposing them to the topics of importance of data in the agriculture value chain and how new and existing technologies, products and services can leverage farm level and global data to improve yield, reduce loss, add value and increase profitability and resilience.

The areas covered in this book include: value of data, the different types and sources of data and identify the type of services that data enables in agriculture; how data is used and generated in the value chain; the challenges and risks that smallholders face when sharing data; the strategies related to farmer profiling; how and where to find open data; data analysis and visualisation techniques; the legal and policy aspects when dealing with farmers' data sharing; and the basics of licencing, copyright and database rights.

The publication raises awareness about data on and for farmers as well as the products and services that have become a growth area, driving expectations and investments including e-extension, precision agriculture and digital financial services.

Partnerships and UN Collaboration Division

Food and Agriculture Organization of the United Nations
Rome, Italy

ISBN 978-92-5-133837-7

9 789251 338377

CB2840EN/1/02.21