Improving data identification and tagging for more effective decision making in agriculture

Pascal Neveu, Romain David, Clement Jonquet

Improving data identification and tagging for more effective decision making in agriculture

Pascal Neveu and Romain David, INRA, France and Clement Jonquet, LIRMM, France

Abstract

Data integration, data analytics and decision support methods can help increase agriculture challenges such as climate change adaptation or food security. In this context, smart data acquisition systems, interoperable information systems and frameworks for data structuring are required. In this chapter we describe methods for data identification and provide some recommendations. We also describe how to enrich data with semantics and a way to tag data with the relevant ontology. We illustrate the proposed approach in a case of high-throughput plant phenotyping.

Keywords: semantic; interoperability; identification; agricultural data sources; identifier

1 Introduction
2 Structuring the data
3 Case study: plant phenotyping
4 Conclusion and future trends
5 Where to look for further information
6 References

1 Introduction

Global demand for food products is increasing sharply, and the current growth rates in agriculture are clearly inadequate and ill-adapted. Today, an urgent and profound redesign of agriculture is crucial to increase production and reduce the environmental impact. In this context, a major challenge is the shift to digitization – entering the Big Data era – to enable a better understanding of the complex mechanisms underlying the sustainable improvement in crop yields and adaptation. This requires studying not only the genotype, phenotype and environment relationships but also the social or health aspects. This highly interdisciplinary challenge (agronomy, genetics, biology, sociology, etc.) requires intensive data integration. In this context, we shall develop and promote new methods and tools for decision makers, researchers and other agricultural actors, especially in relation to:

- the development of the use of sensors with a smart data acquisition system suitable for the areas such as precision farming,

- the advances in the design of interoperable information systems for agricultural (Big) data, and
- providing data structuring frameworks for visualization, data analytics, knowledge discovery and decision support.

However, agriculture researchers face multiple data challenges such as (i) the change of scale in the production and exploitation of data (Big Data), (ii) the need to share and reuse these data with others in an open approach (Open Data) and, finally, (iii) interoperability and the transformation of data into knowledge (Linked Data). Structuring the data is a prerequisite for more effective data exploitation, analysis and decision making. By 'structured', we mean data that are organized into different parts following a specific data model, for example, data contained in databases or spreadsheets or stored in a specific standard format. In Section 2, we discuss two specific aspects related to structuring the data:

- Identification: objects, concepts and data must be clearly identified.
- Semantics and tagging: the meaning of objects and concepts and the relation between them must be clearly formalized.

Additionally, data describing contexts, often from outside producers, are key to interpreting antropic and natural phenomena and effects on agrosystems. Parameters concerning the weather, pedology, hydrology and social environment are produced and banked by different organizations. For instance, social and biodiversity data, which can be essential for developing agro-ecological approaches, are produced and hosted by a multitude of actors (institutes, associations, environmental agencies, etc.). For efficient decision support, these heterogeneous contextual data must enrich and complement agricultural data. In the meantime, both agricultural data and context data are now provided by thousands of various data sources (data repositories, registries and knowledge bases), requiring scientists and stakeholders to develop international recommendations and standards to improve interoperability while ensuring data traceability and ownership. Better semantically described data have proved a source for better decision support systems, including in agriculture (Lousteau-Cazalet et al., 2016; Guillard et al., 2015). Stakeholders are now embracing the FAIR data principles (Findable, Accessible, Interoperable, and Reusable) to allow the implementation of integrative analyses and multisource decision support systems, based on better data structuring, analysis and curation (Wilkinson et al., 2016).

High-throughput phenotyping (phenomics), the plant selection process that aims to identify the most adapted genotypes, is a good illustration of the data challenges faced by the agricultural research community. For example, in plant sciences, phenomics platforms produce huge complex datasets (images, spectrum, human readings, soil analysis) from different scales (molecular to plant population) in various contexts of strongly instrumented installations (field, greenhouse). Phenomics datasets must be accessible to the scientific communities (genetician, bioinformatician, ecophysiologist, agronomist, statistician, sociologist, etc.) who have intensive data integration needs in order to help them in their selection. This case study will be further detailed in Section 3.

2 Structuring the data

In agriculture, observation and management systems, developed and used in many settings, produce a large volume of heterogeneous data, which are difficult to aggregate since they focus on specific issues. There are various data sources in agriculture that require miscellaneous knowledge and skills to be used together accordingly. For instance, agricultural data sources can be related to agricultural production, farm practices, transformation, distribution and so on. Since a few years, another important sources of data are not only connected objects in agriculture (Tzounis et al., 2017) – weather stations, insect traps, soil moisture sensors and water meters connected to irrigation – but also various sensors installed on animals to evaluate their conditions (health measures, temperature, movement), milking robots (quantity and quality of milk) or feeding automata. Agro-equipment is increasingly enriched with sensors, for precision farming (e.g., provide the plant exactly what it needs) and predictive maintenance. Satellite images are another example: the Sentinel constellation delivers free images at a very high temporal frequency (every 5 days), which opens up new research and business opportunities. Agricultural production traceability requirements are now supported, in part, by automated reading systems, with radio frequency identification (RFID) and NFC chips, or by the manual input of agricultural interventions from smartphones with direct transmission to the applications software. The challenge is to automate data acquisition so that it has virtually no cost and is not an additional charge for farmers or scientists (Wolfert et al., 2017). Finally, high-throughput phenotyping methods, essential for shortening the production cycle of new seeds, are also sources of massive data (e.g., phenotype-monitoring platforms produce thousands of images per day) to link with genotypic data (Halewood et al., 2018).

Organized and structured access to primary agricultural data is a sine qua non condition for building efficient decision support systems to achieve the conservation of biodiversity and sustainable development. Organizing, managing and storing of various data require new approaches. Proper data structuring enables to organize data to suit a specific purpose so that they can be accessed and worked with in appropriate ways. The better the data structure, the better we will be able to group them with other data and learn from them.

2.1 Identification

An identifier is a sort of name that identifies a specific object (digital or not) in a set of objects. In an ideal world, identifier should be unique for each object (bijection); in practice this is rarely the case. In most cases a resource (object) can have several (not all unambiguous) identifiers depending on the context. An identifier is unambiguous if it makes it possible to identify an individual in a specific context in a safe way (McMurry et al., 2017). An unambiguous identifier, which cannot refer two different objects, is called GUID (globally unique identifier) or UUID (universally unique identifier); irrespective of whatever the database or source, all disciplines taken together, no other object will be designated identically, for example, ISBN for books. For software objects, GUIDs are typically randomly generated 128-bit codes. There are several specifications for identifiers, for example, UUID, LSID, ARK, DOI, URI, RFID, XRI. The relevance of these different mechanisms depends on the context and, of course, of the characteristics of objects to identify. Data identification also depends on the range of the use of the resource. If the resource shall be referenced only within a limited range or system, it could be assigned a local identifier. But if it shall move to another system (e.g., for the purposes of expert measures such as chemistry of soil and water quality) or if it shall be reused and aggregated with data of different provenances or contexts, a ‘reliable’ global and long-term identification mechanism is necessary.

Long-term structuring of data requires to reliably identify all the concepts, objects and their properties described in the information systems. A persistent identifier is an identifier that is permanently assigned to an object (ideally usable in several decades). For example, once an ISBN is assigned to a particular book, that number is always associated with that book and no other book will ever receive the same number. Likewise, identifiers must be persistent and shall not change. The problem is that during periods of decades, many changes can occur not only within databases but also in institutions or organizations in charge of the data. It is thus necessary to preserve and

recover dependencies between these elements, in time and in localization. Persistent identifiers play a key role in adopting Open Science (Dappert et al., 2017). The reliability of this identification depends on some essential qualities described, for instance, in W3C Recommendations (https://www.w3.org/TR/cooluris/) and must assure persistent security, traceability and reusability of data. The key to rich integration is a commitment to deploy and reuse globally unique, shared identifiers and to implement services that link those identifiers (Page, 2008). The major persistent identification system appears in chronological order: Handle (1994), Persistent URL (1995), Uniform Resource Name (URN; 1997), Archival Resource Keys (ARKs; 2001) and eXtensible Resource Identifier (XRI; 2005).

For instance, persistent GUIDs are usually generated as groups of dash-separated hexadecimal characters, for example, 120a-e29f-a861-12f5-5a52. Their three main qualities are: 1) to be generated in a non-centralized way, 2) to make extremely improbable the random generation of two identical identifiers and 3) to be completely opaque and not sensitive to the changes of authorities or names of authorities. These automatically generated GUIDs can be used as a basis for the construction of other identifiers, for example, by adding a prefix (URI, URL, domain name, authority name). A GUID, when it is integrated in an URI, can be dereferenceable as explained hereafter.

Uniform resource identifier (URI) is defined by the RFC 3986 standard[1] provided by the W3C, which specifies: '*An URI is a compact sequence of characters that identifies an abstract or physical resource. This specification defines the generic URI syntax and a process for resolving URI references that might be in relative form, along with guidelines and security considerations for the use of URIs on the Internet. The URI syntax defines a grammar that is a superset of all valid URIs, allowing an implementation to parse the common components of a URI reference without knowing the scheme-specific requirements of every possible identifier. This specification does not define a generative grammar for URIs; that task is performed by the individual specifications of each URI scheme*'. All Web hyperlinks (URLs) are expressed as URIs. Dereferencing is also an important aspect: an URI is said to be dereferenceable if it is possible to obtain all the digital contents describing the referenced resource (e.g., URL). The act of retrieving

[1] https://tools.ietf.org/html/rfc3986

an information of a resource identified by a URI is known as dereferencing that URI. To summarize on URI/URL one can say that:[2]

- URL identifies what exists on the Web;
- URI identifies, on the Web, what exists; and
- IRI identifies, on the Web, in any language, what exists.

Life science identifiers (LSIDs) are represented as an URN with the following format: urn:lsid:<Authority>:<Namespace>:<ObjectID>[:<Version>]. But LSIDs are not strictly URIs, and so are not always dereferenceable. Bioinformatics and biodiversity communities use them as a way of identifying species in global catalogs. LSIDs have been criticized as violating the Web architecture's good practice of reusing existing URI schemes.

Digital Object Identifier (DOI), initially used in bibliographic databases, allows the identification of digital resources, such as a report, scientific articles or any other type of digital object objects. The purpose of the DOI is to associate metadata describing the object, for example, in bibliography, to produce more reliable, unambiguous and longer-lasting citations. DOIs are issued by DOI agencies, part of the DataCite consortium. A DOI is a special case of Handle ID[3] with the following format: doi:10.<Naming Authority>:<Registry_Number>, and it contains a link to the metadata (restrictions of use or copyright and naming authority among others), described by a data model common to all DOIs, the indecs Data Dictionary, an address or physical location for the digital object (usually a URL) that the DOI translator will use to redirect. For instance, prefixing a DOI with https://doi.org/ allows to dereference the identifier into a landing page storing or describing the object identified. DOI provides a good frame for a persistent identification of agricultural datasets.

ARK is a perennial identifier system based on the URI standard. ARK is designed to ensure long term identification of a resource, scalability and independence. An ARK contains a portion impervious to changes and a flexible portion, which designates a shape of the object or a mode of access thereto. An ARK URL is subdivided into two URLs: the first, optional, gives the addressing authority NMA (Name Mapping Authority), while the second is the ARK URL, fixed and proper, which includes a NAAN (Name Assigning Authority Number) and the name given to the object.

2 Credit to Fabien Gandon's (INRIA).

3 The Handle System is a technical specification for assigning, managing and resolving persistent identifiers assigned to digital objects and other internet resources.

An XRI is a schema and resolution protocol for abstract identifiers compatible with URIs. The goal of XRI is to provide a universal format for abstracts, structured identifiers that are independent of domains, locations and transport applications, so that they can be shared across a large number of domains, directory and protocols.

Identifying samples and real objects with a persistent identifier is possible with several standardized methods that can be linked with previous persistent identifiers (e.g., bar code that is a visual, machine-readable representation that describes something about the object that carries the barcode). It can have one or two dimensions and represents a numerical identifier. For instance, Universal Product Code from industrial sector is a worldwide retail, GS1-approved international standard (ISO/IEC 15420).

The identification of real objects has been increasing since the appearance of the internet of things (IoT). Between RFID chips, naming solutions and middlewares, the IoT is composed of many complementary elements, each having its own specificities.

For real objects, RFIDs are based on radio tags that can be pasted or embedded in objects or products and even implanted into living organisms (animals, human body). This identification method can be used to identify objects, such as those with a barcode (electronic label), people (being integrated in passports, transport card, payment card or domestic carnivores by implantation under the skin), cats, dogs and so on. The RFID identification of pets is mandatory in many countries. For traceability purposes, this is often the case for farm animals. The International Geo Sample Number (IGSN) retains the identity of a sample even if it is transferred from one laboratory to another, and the data appear in different publications, thus eliminating any ambiguity from similar names of other terrestrial samples. It allows researchers to reconstruct the analytical history of a sample. The IGSN, developed as part of SESAR (System for Earth Sample Registration), is a nine-character identifier. It is designed to ensure backward compatibility with previously collected data as new techniques are developed. The IGSN network allows to link data generated by researchers and published in different scientific articles. Research Resource Identifier (RRID) aims to authenticate the key resources: antibodies, model organisms and tools (software, databases, services). But it is dedicated to the medical domain, and it could be relevant to extend it to agriculture.

The persistence of an identifier relies on the durability of the system to provide the identifier and its capacity to dereference. It is clear that this sustainability is not always a strong priority of

institutions. Durability also depends on the durability of the organizations themselves. The missions and perimeters of national and international identifying organizations are regularly re-evaluated and modified. Several global database organizations were created, to catalog and monitor research organizations worldwide, such as GRID (Global Research Identifier Database), Ringgold IDs, ISNIs (International Standard Name Identifiers) or the Research Organization Registry. For instance, most URIs, regardless of their type, include the name of the institute that generates or host them. These changes needed to be tracked so that identifiers stay valid; this is the role of so-called data authorities and a prerequisite to the consensual adoption of a truly perennial and shared global identification system. Today, many self-established or defecto reference identifier generators, some of them proprietary, co-exist (e.g., Pensoft, Zenodo, PubMed, ResearchGate, ResearcherID, HAL-ID). However, the existence of systems established and promoted by economic actors is questionable, either for ethical or economical reasons, for example, what if Google-Bing-Yahoo-launched Schema.org reference system was stopped because it is unprofitable? Therefore, identifier governance and management should be based on a system equivalent to the management of domain names and supported by a standard Web organization like W3C.[4]

We clearly established identifiers that are used in schemas or standard vocabularies and ontologies (cf. Section 2.2) to provide information (properties, relations) about the object (e.g., responsible organization, type of object, definition, labels). As ontologies are changing with digital objects, the persistent identification method must support different versions of an object. Versioning becomes then an important aspect when building identifiers, for example, predefined period and important update releases (curation of data, campaign of collection). On the other hand, some versioning processes must trace all the transformations made on the data for history management. Services such as B2HANDLE can allow to support this.

Today, the URI system is a standard used in a large variety of domains: genetic, chemistry, IoT, life sciences and so on. As an identifier, the URI must have some properties: non-ambiguousness, unicity, persistence, stability and resolvability.

[4] The World Wide Web Consortium is led by three organizations: the MIT Computer Science and Artificial Intelligence Laboratory (the United States), Keio University (Japan), the National Institute of Research in Computer Science and Automation (France). Its role is only advisory.

Here is an example of an URI: http://www.phenome-fppn.fr/m3p/arch/2017/c17000915 (which is not dereferenceable). It uses the following pattern: http://subdomain.yourdomain.topdomain/path/identifier

Properties of URIs are:

- Non-ambiguousness: the URI must be associated with only one resource.
- Unicity: only one URI for one resource.
- Persistence: once a resource is given an URI, one should not replace or delete the URI.
- Stability: URI has to remain the longer possible (at least 20 years) and should not be reassigned to another resource. The definition is close to the persistence; stability is persistence over a long time.
- Resolvability: URI should be used through internet browser to find information about the resource or the resource itself (also called dereferenceable).

When these principles are not respected, one may encounter several issues. Usually non-ambiguousness and unicity are usually not a problem as everyone understands their importances. However, stability and persistence are much more difficult to get: typical case is when part of this URI is changing. For example, the domain name www.phenome-fppn.fr later becomes phenotyping.fr. Thus, the unicity of phenotyping.fr/m3p/arch/2017/c17000915 is not guaranteed; there could be two different resources identified with the same URI, the phenotyping.fr ID and another with the phenome.fppn.fr ID.

In summary, few rules must be followed to create good URIs in agriculture: i) use minimal information and do not use everything that may change, ii) use persistent URL, iii) provide multiple output format – content negotiation – and link them together, iv) request on the external identifier (identifiers.org, n2t.net, w3id.org, ePIC[5]), v) integrate and reuse already-existing identifiers. Things to avoid include: i) avoid file extension in the URI, ii) avoid query-specific characters (e.g., '?' or '&'), iii) use auto-incrementation carefully (not for versioning purpose). To conclude, identification is a first step to go further in order to improve decision support. Once the objects have been identified, it is necessary to specify how they are used and the role of each in relation to the others, in other words, to allow the implementation of methods related to semantics and tagging.

5 https://www.pidconsortium.eu/

2.2 Semantics and tagging

2.2.1 Interoperability with ontologies and the Semantic Web

Data structuring and understanding need metadata for their description and use. Several categories of metadata must be provided such as descriptive, administrative, technical and provenance. Too often, metadata are poor and incomplete, hampering effective data reuse. In some virtuous situations, these data and associated contexts can be informed, very precisely, but not in a machine-readable format. Metadata are often simple (wording, date of creation, contact point, cartographic projection, size, etc.) or very detailed (data quality measurements for each data element, provenance, versioning or historical maintenance service of the measurement instruments, constraints of use, etc.). Structuring the data and providing metadata are essential for the understanding and good use of data in decision processes. It is therefore important to pay special attention to the semantics of the data. Earlier we get metadata better they are.

Semantic interoperability enables data integration and fosters new scientific discoveries by exploiting various data acquired from different perspectives (e.g., agricultural and context data). For instance, a scientist experimentally measures the sensitivity of a plant to a disease (agronomy vision), whereas a farmer concretely observes the leaves of the plant turning brown (agriculture vision). Both are phenotypes, or traits, information, but they come from two different worlds that must yet be more connected. This shall be possible only through lifting the data into meaningful knowledge for humans, yet exploitable by machines.

A researcher studying a certain plant trait (e.g., resistance to a disease) is interested in the gene that controls this phenotype, the expression of this trait in different crop varieties observed in different environments and, of course, its effect on the crop yield or for associated needs such as the use of pesticides. The information we need to answer such questions is available in multiple datasets expressed using various ontologies (crop ontology, plant ontology, trait ontology, etc.) and at various levels (e.g., population, individual, organ); the issue is finding that information and combining it in a meaningful way for researchers, breeders and ultimately farmers, consumers or any stakeholders of the value chain.

Ontology engineering is a sub-domain of knowledge engineering that deals with knowledge representation and reasoning. An ontology is described as a ‘formal specification of conceptualization’ (Gruber, 1993); it ‘defines the terms to describe and represent an area of

knowledge'. Ontologies are composed of concepts, relations and instances. For example, if you want to define a car, you should say: 'a car is a transportation object, with four wheels, and one needs a licence to drive it. My blue Ford Mustang is a car'. 'Car' is a concept, 'is a' is a relation and 'My blue Ford Mustang' is an instance.

The Semantic Web is the area in which ontologies are used to structure data into formal knowledge. The Semantic Web provides the necessary techniques and technologies to build a Web of data (or Linked Open Data (LOD)) as well as reasoning on ontology concepts and mapping between ontologies. The Semantic Web relies on a set of core technologies such as RDF, RDF-S, OWL, SPARQL and SKOS; all of them are built on top of the notion of URIs, which are employed to formally identify objects and remove ambiguity. The Resource Description Framework (RDF) is the W3C language to describe data. It is the backbone of the Semantic Web. SPARQL is the corresponding query language. Complementary, RDF Schema (RDF-S), the Web Ontology Language (OWL) and the Simple Knowledge Organization System (SKOS) are languages to build schemas, ontologies and vocabularies/thesaurus. Figure X (credit to http://lod-cloud.net) illustrates, as of the beginning of 2017 and previously, part of the amount available as LOD and the importance of ontologies/vocabularies (most of the life sciences section in pink are ontologies listed in the NCBO BioPortal; Noy et al., 2009).[6,7]

The Semantic Web offers the methods and technologies to extract/transform Big Data into actionable knowledge (Grigoris and Van Harmelen, 2004).[8] It relies on standard vocabularies and ontologies to formally capture the knowledge of a domain into semantic resources that computers use to index, search or reason on the data.

Tim Berners-Lee, the inventor of the Web and initiator of the Linked Data project, suggested a five-star deployment scheme for Linked Data. The five-star Linked Data system is cumulative. Each additional star presumes the data meet the criteria of the previous step(s).[9]

☆*Available on the Web, in whatever format*

☆☆ *Available as machine-readable structured data (i.e., not a scanned image)*

☆☆☆ *Available in a non-proprietary format (i.e., CSV, not Microsoft Excel)*

6 Linking Open Data cloud diagram 2017-08-22, CC-BY-SA by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

7 https://commons.wikimedia.org/wiki/File:Graphe_Web_des_donn%C3%A9es_depuis_4_ans.png

8 See the MOOC Web of Data for a quick introduction: https://www.coursera.org/learn/web-data

9 https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

☆☆☆☆ *Published using open standards from the W3C (RDF and SPARQL)*

☆☆☆☆☆ *All of the above and links to other LOD*

The purpose of the Web of data is not to create another Web, since it is based on its current architecture (the URI system and the HTTP protocol), but to create an extension. RDF is to structured data what HTML is to documents, an interoperability framework that ensures consistency in the handling and processing of these data by machines.

2.2.2 Ontologies and semantic tagging in agriculture

In recent years, we have seen an explosion in the number of semantic resources (thesauri, terminologies, vocabularies and ontologies) being developed in agronomy and agriculture, for instance, the plant ontology, environment ontology, crop ontology or agronomy ontology, which opened the space from various types of semantic applications to data integration or decision support. Ontologies in agriculture are spread out around the Web (or even unshared), in many different formats and artifact types, and with different structures. Agronomy (and its related domains such as food, plant sciences and biodiversity) needs a one-stop shop, allowing users to identify and select ontologies for specific tasks, as well as offering generic services to exploit them in search, annotation or other scientific data management processes. The need is also for a community-oriented platform that will enable ontology developers and users to meet and discuss their respective opinions and wishes. And, with such a number of ontologies, new problems have raised such as describing, selecting, evaluating, trusting and interconnecting ontologies. For this reason, ontology repositories, such as AgroPortal (Jonquet et al., 2018), are offering a reference point of entry for interconnected vocabularies and ontologies in agronomy and agriculture. AgroPortal offers a robust and reliable service to the community that provides ontology hosting, search, versioning, visualization, comment and recommendation; enables semantic annotation; stores and exploits ontology alignments; and enables interoperation with the Semantic Web.

One important use of ontologies is for annotating and indexing text data. Indeed, ontologies allow representing data with clear semantics that can be leveraged by computing algorithms to search, query or reason on the data. One way of using ontologies is by means of creating semantic annotations or semantic tags. An annotation is a link from an ontology term to a data element, indicating that the data element (e.g., article, experiment, observation, medical record) refers to the term. When doing ontology-based indexing, we use these annotations to ‘bring together’ the

data elements from these resources. However, explicitly annotating data is still not a common practice for several reasons (Jonquet et al., 2009):

- Annotation often needs to be done either manually by expert curators or directly by the authors of the data.
- The number and format of ontologies available for use are large, and ontologies change often and frequently overlap.
- Users do not always know the structure of an ontology's content or how to use the ontology to do the annotation themselves.
- Annotation is often a boring additional task without immediate reward for the author.

Semantic annotation is an important research topic in the Semantic Web community. Tools vary along with the types of documents that they annotate (e.g., image annotation). For an overview and comparison of semantic annotation tools, the reader may refer to the study by Uren et al. (2006).

Previous work has encouraged and exalted the use of ontologies for annotation at various levels (Rhee et al., 2008). For a while, the prevalent paradigm in the use of ontologies was that of manual annotation and curation. However, several researchers have shown that such manual annotation, though highly desirable, will not scale to the large amounts of data being generated (e.g., in the life sciences; Baumgartner et al., 2007). If one examines the reasons for the low adoption of ontology-based annotation methods among database providers, the high cost of manual data curation remains the main obstacle. In light of this situation, researchers have called for the need of automated annotation methods (Dowell et al., 2009) and for leveraging natural language processing tools in the curation process (Altman et al., 2008).

An example of such service for agronomic/agricultural data is the AgroPortal Annotator (http://agroportal.lirmm.fr/annotator), a Web service that provides a mechanism to employ ontology-based annotation in curation, data integration and indexing workflows, using any of the ontologies in the AgroPortal repository.[10] The Annotator tags raw text descriptions with relevant ontology concepts and returns the annotations to end users. Those annotations are mainly made of an URI identifying the annotating concept. Services like the AgroPortal Annotator can be used to

10 This service is AgroPortal's version of the NCBO Annotator described in (Jonquet et al., 2009).

structure data into unambiguous and semantically identified parts, hence contributing to the process of transformation/extraction of data into knowledge.

- *x*

3 Case study: plant phenotyping

In this section, we provide an example of the use of some of the technologies and methods presented in Section 2 in a case study of plant phenotyping and agriculture (Neveu et al., 2019).

Plant-derived products are at the center of challenges posed by increasing requirements for food, feed and raw materials. Integrating approaches across all scales from molecular to field applications is necessary to develop sustainable plant production with a higher yield and using limited resources. While significant progress has been made in molecular and genetic approaches in recent years, the quantitative analysis of plant phenotypes – the structure and function of the plant – has become the major bottleneck. Plant phenomics is an interdisciplinary science that links genomics with plant ecophysiology and agronomy. The functional plant body (phenotype) is formed during plant growth and development from the dynamic interaction between the genetic background (genotype) and the physical world in which plants develop (environment). These interactions determine plant performance and productivity, measured as accumulated biomass and commercial yield and resource use efficiency.

Phenomics platforms produce huge complex datasets (images, spectrum, human readings) from different scales (genetic to plant population). Phenomics datasets need to be accessible to the large scientific community (genetician, bioinformatician, ecophysiologist, agronomist, etc.). Their reanalysis requires tracing relevant information on thousands of plants, sensors and events. The open-source Phenotyping Hybrid Information System (PHIS – http://www.phis.inra.fr) is proposed for plant phenotyping experiments in various categories of installations (field, glasshouse).

3.1 Identification in PHIS

Tracking all objects involved in a phenotyping experiment (e.g., plants, pots, sensors) and representing relationships between them are essential in a high-throughput context where

thousands of plots, plants or sensors are involved. This requires a proper strategy that allows to individually identify each specific object as well as semantic properties for creating relationships between such objects.

For instance, the replacement of a sensor at a given position (e.g., meteorological sensor or soil tensiometer) is not obvious in the outputs of an environmental database. In greenhouse experiments, a plant can be replaced by another plant at the same position and vector (e.g., pot, cart) during an experiment, potentially generating confusion. All objects therefore need to be identified in order to keep the necessary information associated to them (e.g., positions over time, successive calibration for sensors, origin for plants).

In the following text, we illustrate PHIS's identification system. PHIS object identification is based on URIs. This ensures traceability in space and time, while a typical identification by numbers (e.g., 'plant 736') refers to different plants in different experiments and installations. URIs are generated automatically for each object via the user interface and implemented by QR codes, creating a set of connected objects that can be accessed, along with all their properties, from any terminal (e.g., mobile device, barcode reader).

What are the things to identify? Ideally, we want to identify everything, but we have very different resources – do we identify them the same way? Are URIs the best option to identify every resource? Those are questions one should ask before designing an URI scheme. For example, measures collected by a sensor can be gathered in a dataset and require only one URI for the dataset, or even be aggregated in a database. Then, the measures per day are identified with a primary key or an incremental ID.

3.1.1 How to make non-ambiguous URI

PHIS's non-ambiguous identifiers use an incremental number (the number of plants), prefixed with a letter that helps human manipulate the URI and real objects.

In PHIS, the semantic implementation is realized by a set of standardized ontologies written in OWL2. Based on these ontologies, the first step is to organize objects and concepts with a specialization hierarchy (sort of). For instance, corncob is a sort of a plant organ, that is, corncob is subClassOf plantOrgan. The description of this object (metadata) is formalized as properties. These properties can be values (dataProperty) or objects (objectProperty). Semantic links between objects, between events and between traits used in PHIS are realized through the annotation

ontology and some specific application ontologies (such as Ontology of Experimental Events (OEEV)).[11]

In order to integrate data, the relations between objects need to be represented adequately in a high-throughput context. For instance, if thousands of sample tissues have been collected on different leaves of different plants, the information 'sample 884 belongs to the leaf 7 of plant 736' may be lost if kept in a spreadsheet. The links between objects are based on two OWL application ontologies. The Ontology for Experimental Phenotypic Objects (OEPO)[12] describes objects involved in phenotyping experiments (e.g., infrastructure, devices, germplasm, scientific objects) and defines specialization hierarchy between them according to the specificities of the installations and experiments. OEEV characterizes events that occur during an experiment, for example, moving of plants, dates of sowing, application of a given treatment, harvesting, measurements or sampling for -omics measurements or any category of technical problems. For instance, the Trouble concept distinguishes Breakdown (sensor or conveyor), Dysfunction (sensor fault, irrigation trouble) and Incident (a pot falls down, a leaf is blocked in an imaging cabin, lodging of a plot, human error, etc.). As described in the associated semantic graph, an event can be associated with objects (e.g., plant, plot, sensor) and with the user who has annotated the event, and the occurrence data can be tracked along with every relevant detail.

The use of ontologies allows to deal with the complexity of phenotyping data in order to link a large number of different data sources. Data integration process can be done automatically:

- Concept mapping is one of the approaches for data integration from different sources. Ontologies will help for concept mapping. For instance, the 'field' is equivalent to 'cultivated land'.
- Data-linking approach is based on the use of common standardized RDF properties in several data sources. It allows to identify common individuals in different sources. For instance, GPS coordinate values and the plant species name allow to know common plots of different datasets.

Ontology-driven approach for data management allows to deal with the same system data from greenhouse or fields, thanks to a precise formalization of agricultural objects. This approach makes

11 http://agroportal.lirmm.fr/ontologies/OEEV
12 http://agroportal.lirmm.fr/ontologies/OEPO

easier the data integration process. In other words, by connecting greenhouse and field experiments, the decision-making process is strongly improved.

Other uses are made. Indeed, this approach based on ontology-driven information systems can facilitate decision making for many agricultural applications such as agroecological system design, precision agriculture and breeding. For instance, in agroecology we formalized bioagressors, lifecycles and impacts. All these applications require interdisciplinary work and intensive data integration. The formalization of concepts, the links between concepts and tagging are fundamental and constitute a crucial step. This information system generation encourages the production of FAIR data that can be used across disciplines.

4 Conclusion and future trends

As we have seen earlier, the structuring of data in order to make them reusable is based on their identification in the long term (beyond the decade) and on the reuse of ontologies and interdisciplinary standards. In practice, organization evolutions and staff turnover have important effects for the long-term data management. In too many cases, data are often produced and designed for 'immediate consumption'. Reusing ontologies is the way that we must choose, but efficient tools for improving reuse are needed. Data come from various devices; simulation, observation or crowdsourcing and too often data repeatability/reproducibility is not well known or impossible. Structuring data will be a significant advance. For projects, institutions and companies, Data Management Plans (DMPs) are a sine qua non condition for the evaluation of produced data in agriculture. DMPs will allow the development and improvements of methods for the identification of agricultural objects and the associated data semantics. An interesting example is the world of software where many developers do not hesitate and are very active to share their production. Data papers improve the process of data sharing and data indexing.[14] A citation mechanism is designed to reward the efforts of people and institutes that collect and manage data. But recognizing data sharing is still in its infancy, and the generalization of persistent identifiers, data papers and the Web of data could help change things. Part of the answer is also in the availability of integrative data tools for visualization, analysis, prediction and decision support.

[14] https://freshwaterblog.net/2012/06/29/what-does-a-data-paper-look-like/

Access to a new generation of tools can motivate communities of agriculture. It will support agriculture to raise challenges.

The Web of data – built out of LOD – is the concrete and most salient outcome of 20 years of Semantic Web research. Ontologies and vocabularies are its backbone as they are used to semantically annotate and interlink datasets. Methods and techniques have recently been developed, allowing the massive publication of structured data on the Web. Yet, that vision has not been fully applied to agronomy and agriculture for which data have some specificities that require new models (e.g., spatio/temporal dimension, complex and multi-scale data (from gene to environment), data streaming from IoT devices in precision agriculture). And agronomy and agriculture mix data from different disciplinary fields and scientific perspectives, making the integration even more challenging. Despite recent initiatives like the Agronomic Linked Data RDF knowledge base (AgroLD – www.agrold.org) (Venkatesan et al., 2018), we have not seen integrated semantic resources that have had a major impact in agronomy and agriculture such as the ones that have been developed in biomedical and health sciences (e.g., Bio2RDF.org, EBI RDF). Indeed, despite the large adoption of some semantic resources, like AGROVOC – the most widely adopted vocabulary to index, retrieve and organize agricultural system data – we cannot yet measure their impact in terms of LOD produced and made available to the rest of the world. Some may ask: where are agronomy and agriculture in the famous LOD cloud diagram previously illustrated? This question is at the center of a the D2KAB project (hereafter referenced).

Data curation needs to be developed and to go further than ‘cleaning’ its imperfections. The curation of data, from Latin *curare*, which means ‘to take care’, is essential before any process of analysis or decision. It consists of improving the capacity of the data to describe a system in an unambiguous and explicit way. It is essential to prepare a dataset for a large set of analysis methods, given the opportunity to aggregate different datasets of different provenances, structures and semantics.

To meet the agricultural challenges, well-structured and described data are essential, but how to use them better? Ideally, we shall have powerful tools to automatically select and integrate huge datasets from various sources (agriculture, environment, social, health, etc.). A first stage is more reasonably to use semi-automatic tools, in order to produce the most complete knowledge that constitutes the decision support material.

The structuring of data must be accompanied and allow the construction of different kinds of decision support tools. The main goal is to promote the adoption of increasingly decision-making and ‘smart’ decision support tools in the agricultural domain. These systems will use not only more data but also better data, updated, cheaper to produce, more standardized and more efficient for decision making.

Produced data should meet FAIR principles; if widely adopted, the connections they enable will result in improved access to information, opportunities for collaboration, reduced administrative overhead and, ultimately, increased trust in studies and research (Meadows et al., 2019).

However, ensuring long-term persistence of identification is a challenge: it is theoretically easy to install and practically very difficult to maintain. Reuse, obsolescence and updating standards (semantic resources, formats, access protocols, etc.) are another main challenge to the sustainable interoperability of information systems, especially in areas handling heterogeneous data such as agriculture. For these reasons, interoperability level is strongly linked to the quantity of work insured at a long-term scale and to the capacity of different data authorities to build together community-approved ontologies and data schemes. Interoperability at a human level as well as a machine level is a key to integratively analyzing environmental data and building decision support systems that are relevant to address the future challenges of agriculture.

5 Where to look for further information

Further reading and references on identification:

- ARK: https://tools.ietf.org/html/draft-kunze-ark-18
- B2HANDLE: https://eudat.eu/catalogue/B2HANDLE
- DOI: https://www.doi.org/
- Datacite: https://datacite.org/
- ePIC: https://www.pidconsortium.eu/
- GRID: https://www.grid.ac/
- Handle System Namespace and Service Definition: http://www.ietf.org/rfc/rfc3651.txt
- Handle System Protocol: http://www.ietf.org/rfc/rfc3652.txt
- IGSN | SESARSystem for Earth Sample Registration: www.geosamples.org/igsnabout
- IRI: https://www.ietf.org/rfc/rfc3987.txt
- ISNI: http://www.isni.org/
- LSID: http://www.lsid.info/

- ORCID: https://orcid.org/
- RINGGOLD: https://www.ringgold.com/ringgold-identifier/
- RRID: https://scicrunch.org/resources
- UUID: https://www.w3.org/wiki/UriSchemes/uuid
- XRI (OASIS): https://www.oasis-open.org/committees/xri/x

Research data organizations such as Research Data Alliance[17] (RDA) or Force11 are developing. They coordinate actions, research or communication focusing on the structuring of the data for the next tens. Force11[18] is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. The RDA was started in 2013 by the European Commission and several governments with the goal of building the social and technical infrastructure to enable open sharing and reuse of research data. RDA supports interest groups and working groups all along the range of research data issues from DMPs to data repository, identification, standardization and more. Agriculture is especially well represented at the RDA with the Interest Group on Agricultural Data (IGAD). This group gathers several working-groups such as on wheat data interoperability (https://ist.blogs.inra.fr/wdi/), rice data interoperability, and Agrisemantics (https://agrisemantics.org/) which is focus on data management and interoperability with adopting semantic resources and tools.

Additionally, there are a number of current research projects designed to support data management issues in agronomy or agriculture such as:

- French ANR project (*Data to Knowledge in Agronomy and Biodiversity* – www.d2kab.org), which goals is to create a framework to turn agronomy and biodiversity data into knowledge –semantically described, interoperable, actionable, open– and investigate scientific methods and tools to exploit this knowledge for applications in science & agriculture.
- Big Data Grapes H2020 project (*Big Data to Enable Global Disruption of the Grapevine-powered industries* – http://www.bigdatagrapes.eu) which aims to help European companies in the wine and natural cosmetics industries become more competitive in the

17 https://rd-alliance.org/
18 https://www.force11.org/

international markets. This project helps companies across the grapevine-powered value chain ride the big data wave, supporting business decisions with real time and cross-stream analysis of very large, diverse and multimodal data sources.

- EMPHASIS ESFRI and the EPPN H2020 (*European Plant Phenotyping Network* – https://www.plant-phenotyping-network.eu) research infrastructure projects aims to address the technological and organizational limits of European plant phenotyping to make the most of genetic and genomic resources available and essential for crop improvement in times of a changing climate.

6 References

1. Altman, R. B., Bergman, C. M., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L., Jensen, L. J., Krallinger, M., Mons, B., O'Donoghue, S. I., Peitsch, M. C., Rebholz-Schuhmann, D., Shatkay, H. and Valencia, A. 2008. Text mining for biology – the way forward: opinions from leading scientists . *Genome Biology* 9(Suppl. 2), S7. doi:10.1186/gb-2008-9-s2-s7.

2. Antoniou, G. and Van Harmelen, F. 2004. *A Semantic Web Primer*. MIT press.

4. Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G. and Hunter, L. A. 2007. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 23(13), i41–8. doi:10.1093/bioinformatics/btm229.

6. Dappert, A., Farquhar, A., Kotarski, R. and Hewlett, K. 2017. Connecting the persistent identifier ecosystem: building the technical and human infrastructure for open research. *Data Science Journal* 16, 28. doi:10.5334/dsj-2017-028.

7. Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J. and Blake, J. A. 2009. Integrating text mining into the MGI biocuration workflow. *Database: the Journal of Biological Databases and Curation* 2009, bap019. doi:10.1093/database/bap019.

8. Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2), 199–220. doi:10.1006/knac.1993.1008.

9. Guillard, V., Buche, P., Destercke, S., Tamani, N., Croitoru, M., Menut, L., Guillaume, C. and Gontard, N. 2015. A decision support system to design modified atmosphere packaging for fresh

produce based on a bipolar flexible querying approach. *Computers and Electronics in Agriculture* 111, 131–9. doi:10.1016/j.compag.2014.12.010.

10. Halewood, M., Chiurugwi, T., Sackville Hamilton, R., Kurtz, B., Marden, E., Welch, E., Michiels, F., Mozafari, J., Sabran, M., Patron, N., Kersey, P., Bastow, R., Dorius, S., Dias, S., McCouch, S. and Powell, W. 2018. Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution . *The New Phytologist* 217, 1407–19. doi:10.1111/nph.14993.

5. Jonquet, C., Shah, N. H. and Musen, M. A. 2009. The open biomedical annotator. In: *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09*, San Francisco, CA, March 2009, pp. 56–60.

11. Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., Emonet, V., Graybeal, J., Laporte, M., Musen, M. A., Pesce, V. and Larmande, P. 2018. AgroPortal: a vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture* 144, 126–43. doi:10.1016/j.compag.2017.10.012.

12. Lousteau-Cazalet, C., Barakat, A., Belaud, J., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C. and Vialle, C. 2016. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture* 127, 351–67. doi:10.1016/j.compag.2016.06.020.

13. McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., Courtot, M., Deck, J., Dumontier, M., Fellows, D. K., Gonzalez-Beltran, A., Gormanns, P., Grethe, J., Hastings, J., Hériché, J. K., Hermjakob, H., Ison, J. C., Jimenez, R. C., Jupp, S., Kunze, J., Laibe, C., Le Novère, N., Malone, J., Martin, M. J., McEntyre, J. R., Morris, C., Muilu, J., Müller, W., Rocca-Serra, P., Sansone, S. A., Sariyar, M., Snoep, J. L., Soiland-Reyes, S., Stanford, N. J., Swainston, N., Washington, N., Williams, A. R., Wimalaratne, S. M., Winfree, L. M., Wolstencroft, K., Goble, C., Mungall, C. J., Haendel, M. A. and Parkinson, H. 2017. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biology* 15(6), e2001414. doi:10.1371/journal.pbio.2001414.

14. Meadows, A., Haak, L. L. and Brown, J. 2019. Persistent identifiers: the building blocks of the research information infrastructure. *Insights the UKSG Journal* 32(9), 1–6. doi:10.1629/uksg.457.

16. Neveu, P., Tireau, A., Hilgert, N., Nègre, V., Mineau-Cesari, J., Brichet, N., Chapuis, R., Sanchez, I., Pommier, C., Charnomordic, B., Tardieu, F. and Cabrera-Bosquet, L. 2019. Dealing

with multi-source and multi-scale information in plant phenomics: the ontology-driven phenotyping hybrid information system. *The New Phytologist* 221, 588–601. doi:10.1111/nph.15385.

15. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N. B., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G. and Musen, M. A. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* 37 , W170–3. doi:10.1093/nar/gkp440.

17. Page, R. D. M. 2008. Biodiversity informatics: the challenge of linking data and the role of shared identifiers . *Briefings in Bioinformatics* 9, 345–54. doi:10.1093/bib/bbn022.

18. Rhee, S. Y., Wood, V., Dolinski, K. and Draghici, S. 2008. Use and misuse of the gene ontology annotations. *Nature Reviews. Genetics* 9(7), 509–15. doi:10.1038/nrg2363.

19. Tzounis, A., Katsoulas, N., Bartzanas, T. and Kittas, C. 2017. Internet of things in agriculture, recent advances and future challenges. *Biosystems Engineering* 164, 31–48. doi:10.1016/j.biosystemseng.2017.09.007.

20. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. and Ciravegna, F. 2006. Semantic annotation for knowledge management: requirements and a survey of the state of the art. *Journal of Web Semantics* 4, 14–28. doi:10.1016/j.websem.2005.10.002.

3. Venkatesan, A., Tagny Ngompe, G., Hassouni, N. E., Chentli, I., Guignon, V., Jonquet, C., Ruiz, M. and Larmande, P. 2018. Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy. *PLoS ONE* 13(11), e0198270. doi:10.1371/journal.pone.0198270.

21. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. doi:10.1038/sdata.2016.18.

22. Wolfert, S., Ge, L., Verdouw, C. and Bogaardt, M. 2017. Big data in smart farming – a review. *Agricultural Systems* 153, 69–80. doi:10.1016/j.agsy.2017.01.023.