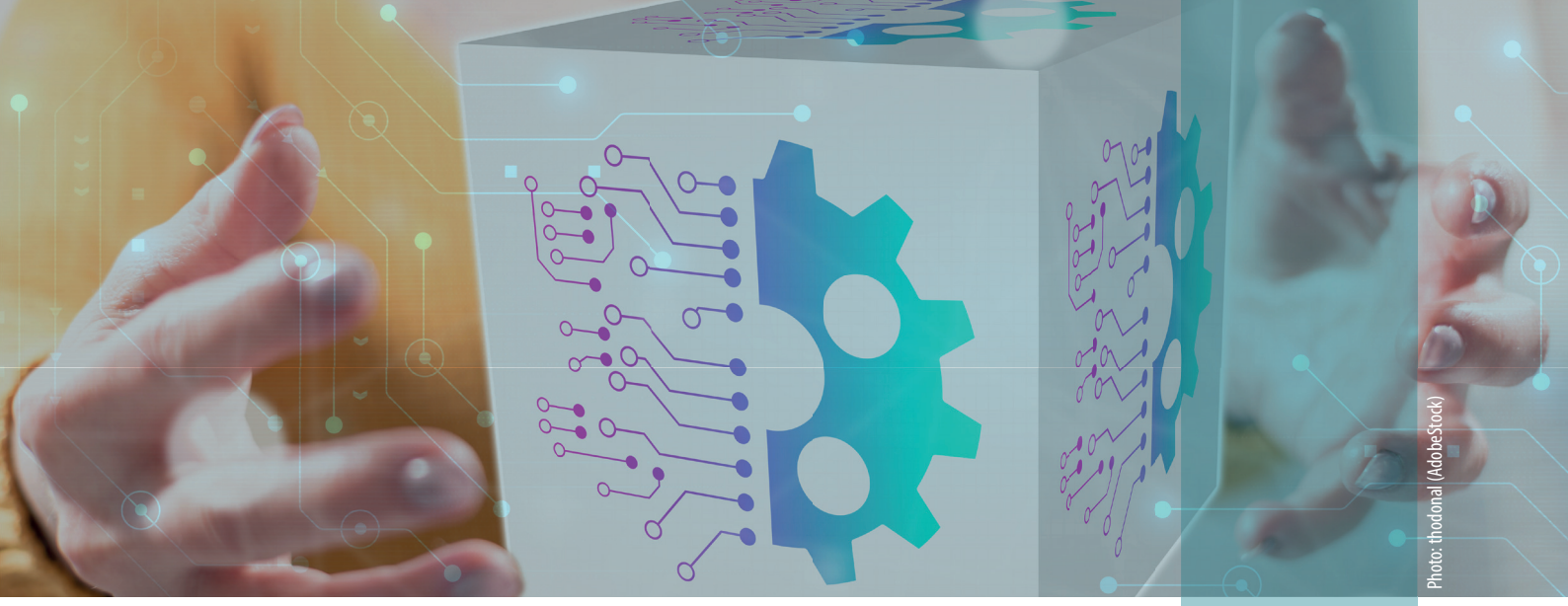


Photo: thodonal (AdobeStock)

# Information engineering
## Contributions to digital agriculture

8

Ivo Pierozzi Júnior | Marcos Cezar Visoli | Marcia Izabel Fugisawa Souza | Luiz Manoel Silva Cunha | Isaque Vacari | Tércia Zavaglia Torres

## Introduction

Digital agriculture comprises the modeling of agricultural phenomena and processes in the environmental, economic and social dimensions through computational artifacts and Information and Communication Technologies (ICT) to bring to the agricultural sector organization, access, use, sharing, and dissemination facilities, as well as the application of scientific knowledge.

> There are multiple challenges in a globalized world, and it is complicated to even reach a consensus on what the priorities should be. However, one of them stands out as the most important due to its structuring effects – making knowledge accessible to everyone. At no other time in history has the production of knowledge been as intense as it is today, and at no other time has its application assumed such a preeminent role. Hence the importance of knowledge management, because between its production and its use there is a chain of complex procedures that may or may not determine its operational success. For some experts like Manuel Castells, the application of knowledge is at the center of the conceptual and operational revolution driven by science and technology advances which are operating in contemporary societies, and that reach all sectors of human life at unprecedented speed. It is therefore important to think about the use of knowledge, pave the way for its various uses and ensure its social and ethical dimension. (Defourny, 2006, p. 7).

The term Information Engineering, as presented by Martin and Finkelstein (1989), is expressed in three different definitions, but converges conceptually. In two of them, the word "automated" stands out:

1) The application of an interconnected set of formal techniques for planning, analyzing, designing and building information systems about an organization as a whole or in one of its main sectors.

2) An interconnected set of automated techniques in which organization models, data models and process models are built into a comprehensive knowledge base used to create and maintain data processing systems.

3) A set of automated disciplines at the organization level used to provide the right information, to the right people, at the right time.

It is based on this perspective and this context that Information Engineering is presented in this chapter, according to how this subject is understood, as a means of mapping, organizing and representing agricultural knowledge and in the context of digital agriculture.

From the perspective of Knowledge Management (KM), the delivery itinerary of scientific knowledge and the guarantee of its effectiveness and efficiency in response to society's demands, including those affecting agriculture (mainly food, energy and fibers) are made possible by Information Engineering. Embrapa has already addressed the relationship between data, information and knowledge (Pierozzi et al., 2017) to enable the use of theoretical and conceptual concepts that align these three levels of organization of human perception about the real world and its consequent technological transformation.

The application of knowledge, that is, to apprehend and use it as a solution to problems and challenges, goes through the decision-making process. There is no best decision to make. There is a possible decision regarding the ability to identify, gather, process and combine the greatest possible amount of information on a given subject. Thus, as a research, development and innovation discipline, Information Engineering is positioned at the central point to combine data, originated from the practice of agricultural research, knowledge, which represents the offer and use of Embrapa's performance results, modeled on ICT. In addition to the aforementioned challenges, the dynamic and massive pace of knowledge production and supply, accelerated by the advancement and support of ICTs.

Therefore, another challenge emerges, simultaneously, in which the quality of the knowledge offered is also configured as a social demand: the knowledge that is sought begins to be demanded as environmentally sustainable, economically viable and socially fair. It is no different in the context of the pragmatics of scientific knowledge and, in particular, in the context of agricultural knowledge.

Embrapa has expressed, in its constant strategic planning efforts, in which it periodically reviews its mission, vision, objectives and goals (Embrapa, 2018), a constant concern with the delivery of technological knowledge to society, using premises of quality and effectiveness. Its recent incursion into the implementation of innovation-oriented research management models corroborates the convergence and coherence of this intention, especially as it is a company that generates knowledge and competences and, therefore, a company that learns and evolves scientifically, technologically, and organizationally (Garcia; Salles Filho, 2009).

In light of this reality, Embrapa Digital Agriculture inserts its contribution, given the efforts it has invested to develop and innovate methodologies and technologies and produce knowledge within its competences in Computing and ICT.

It is in this context that the paths of digital agriculture are aligned and explored, a concept and term that in the scope of Research, Development & Innovation (RD&I) and Science and Technology (S&T) is not only as a trend but, in particular, a new socioeconomic paradigm of the agricultural sector, since it uses ICT to move research data and transform it into information, knowledge and technology for the producer, through the Internet of Things (IoT), Big Data, Cloud Computing, Machine Learning, etc. This paradigm, inherent to digital agriculture, is combined with other contemporary paradigms, such as the Information Economy or Knowledge Economy, Data Science and Open Science (Porat; Rubin, 1977; Powell; Snellman, 2004; Pordes et al., 2007; Aalst, 2016).

The word "engineering" has been associated with computing (software, data, knowledge engineering, etc.), as a way to express the processes of "construction" of computational artifacts that represent things (entities, phenomena, processes) from the real world to machine language. A possible explanation for this linguistic phenomenon of interdisciplinary conceptual and terminological recombinations is the understanding that the practical use of knowledge is an endless, continuous and dynamic process of conceptual recombination, analysis, synthesis and re-signification in permanently emerging contexts. Hence the metaphorical meaning of the word "engineering".

At the same time, in the same itinerary of knowledge construction and development, another conceptual reflection has associated the words "data", "information" and "knowledge" (D-I-K), creating several representation models of this relationship and thus giving new meaning to the term "engineering". Recently, in the proposition of a Knowledge Data and Information Governance model at Embrapa, a conception of a model related to this relationship, different from the conventional ones, was presented to facilitate its organizational and operational implementation in order to provide support to corporate processes of data, information and knowledge management (Pierozzi Junior et al., 2017). The model has as theoretical and conceptual references, in addition to the notion of the D-I-K relationship, the life cycles of data, information and knowledge, conceived in a conjugated and aligned way and represented as a mandala.

This model also supports an ontological approach (Mol, 2008), which serves as an itinerary for the construction of the ontic, that is, the entity itself. The term "entity" is used as a reference to the computational objects or artifacts to be engineered (software, applications, information systems, etc.), since these are the objects that operationally implement scientific and multidisciplinary knowledge, enabling its application in the solution of problems or in response to demands from the agricultural sector.

Based on this general conception, Information Engineering, as an area of knowledge, discipline or proposal for a productive process of technologies and innovation, was configured as an attractive conceptual and terminological option to bring together competences, technologies and solutions executed and produced by Embrapa Digital Agriculture throughout its history and, mainly, as an appropriate option to systematize the process of transforming scientific data into pragmatic knowledge.

Therefore, a conceptual metamodel is being elaborated to organize worldviews in the environmental, agricultural, social and economic spheres (Figure 1), for the development of computational products in response to challenges and opportunities in digital agriculture. Another metamodel (Figure 2) brings together conceptual, methodological and technological approaches that are aligned, based on the concept of Information Engineering, as an integrative construct of knowledge pragmatics that are inherent to various sciences such as Cognition, Information and Computing.

In the following sections, research actions and results will be presented, discussed and contextualized, which in the context of Information Engineering, are being developed at Embrapa Digital Agriculture.

# Knowledge organization and representation systems

To make knowledge accessible and usable, whether by human agents or technological agents, it has to be organized (Soergel, 2009). Thus, including technological perception, Information Engineering can be understood as an area or discipline of knowledge that allows the construction of an operational itinerary, supported by computing and ICT, so that knowledge becomes accessible and usable.

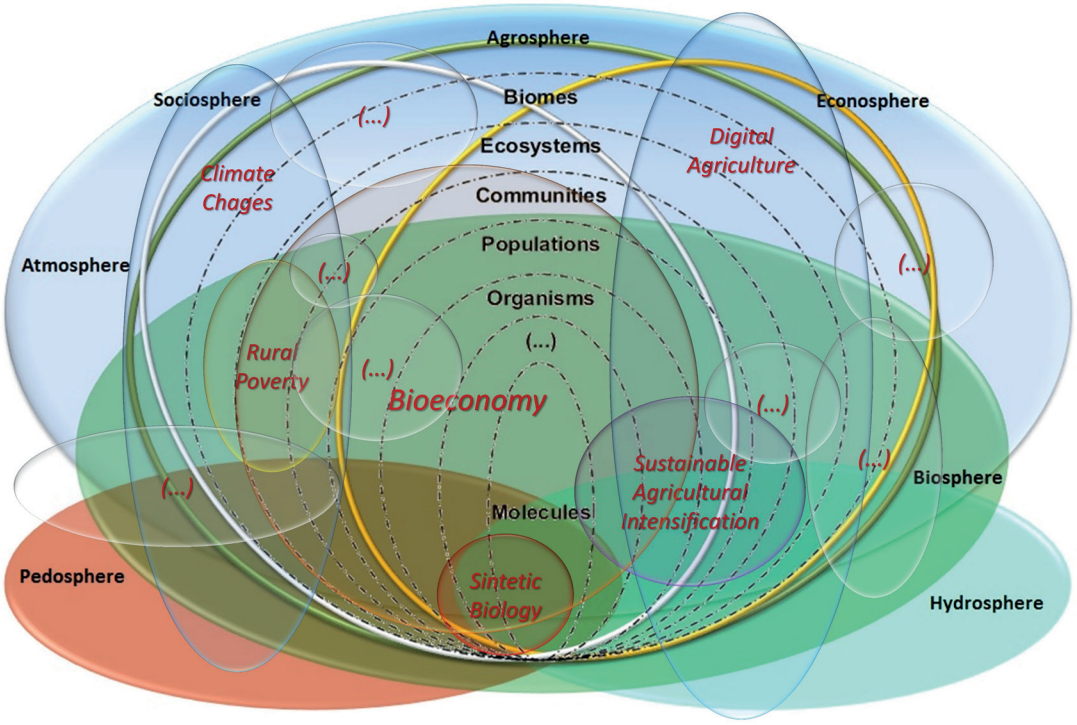



**Figure 1.** Multi, inter and transdisciplinary conceptual representation of agriculture.

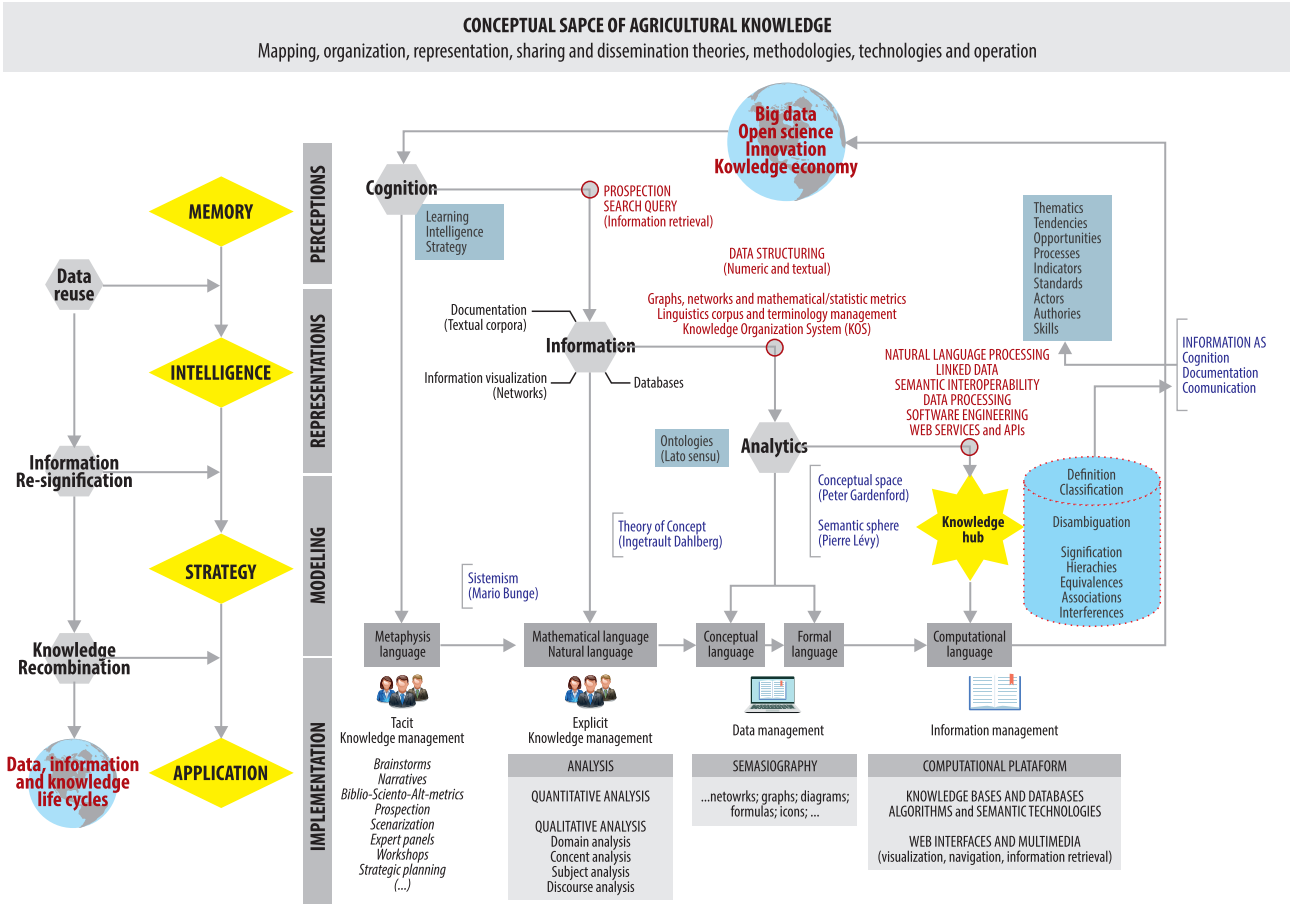


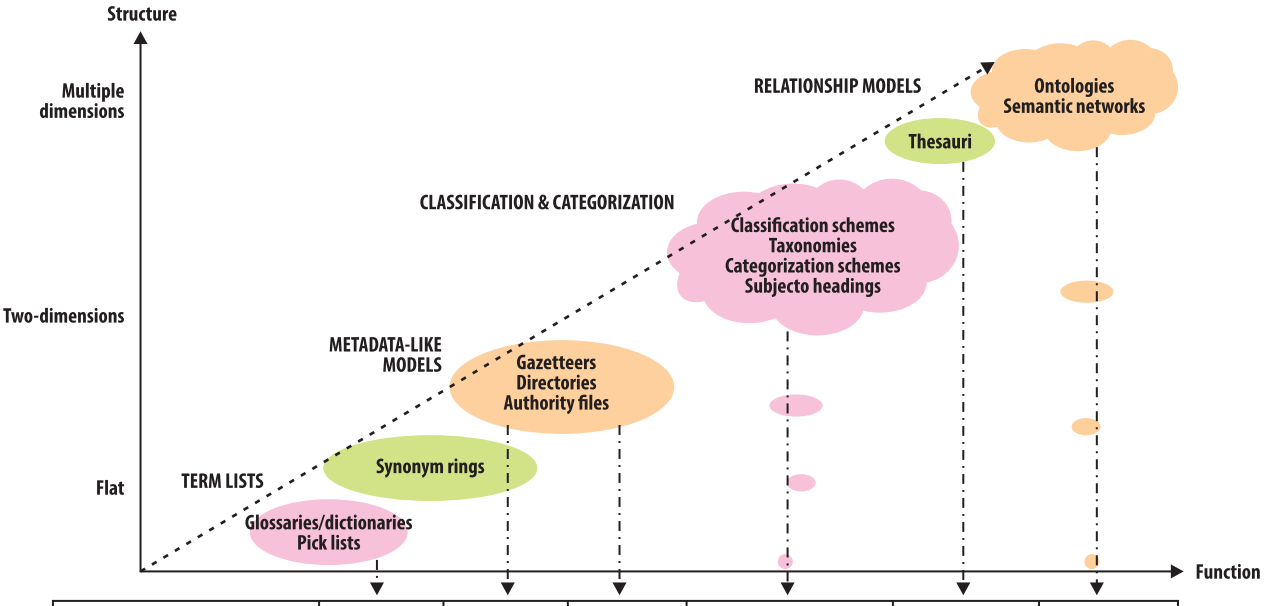

**Figure 2.** Agricultural Information Engineering Metamodel.

The Knowledge Organization Encyclopedia defines Knowledge Organization Systems (KOS) as:

> [...] a generic term used to refer to a wide range of items (e.g., subject titles, thesauruses, classification schemes, and ontologies) that were conceived with regard to different purposes at different historical times. They are characterized by different structures and specific functions, different ways of relating to technology, and used in a plurality of contexts by various communities. However, what they all have in common is that they are designed to support the organization of knowledge and information to facilitate management and retrieval (Mazzocchi, 2019, p. 1).

They can also be defined as "[...] semantically structured conceptual systems that include terms, definitions, relationships and properties of the concepts" (Carlan; Medeiros, 2011, p. 54, own translation). The term Knowledge Organization System (KOS), was proposed by the Networked Knowledge Organization Systems Working Group, at the 1st Conference of the ACM Digital Libraries, in 1998, in Pittsburgh, Pennsylvania (Carlan; Medeiros, 2011 p. 54).

Figures 3 and 4 illustrate how KOS can be understood and indicate how they can be learned and used in the context of Information Engineering.




| Major functions | | | | | | | |
|---|---|---|---|---|---|---|---|
| Eliminating ambiguity | xxx | | xxx | xx | xxxx | xx |
| Controlling synonyms | | xxxx | xxx | xx | xxxx | xx |
| Establishing relationships: hierachical | | | x | xxxx | xxx | xxx |
| Establishing relationships: associative | | | | | xxxx | xxxxx |
| Presenting properties | | | | | | xxxxx |

**Figure 3.** Types of Knowledge Organization System.

Source: Zeng (2008).

The process of organizing and representing knowledge has always been part of human history (Martins; Moraes, 2015). This process is configured as a multi and interdisciplinary action inherent to all sciences and cultures. Thus, when Aristotle[1] defined the ten categories of 'being' as metaphysical categories, which classify words in relation to our knowledge of 'being', and today in the plural modernity in which

[1] 1 The ten categories of knowledge are used in the classification and representation of human thought, that is, thoughts became the starting point for representations.

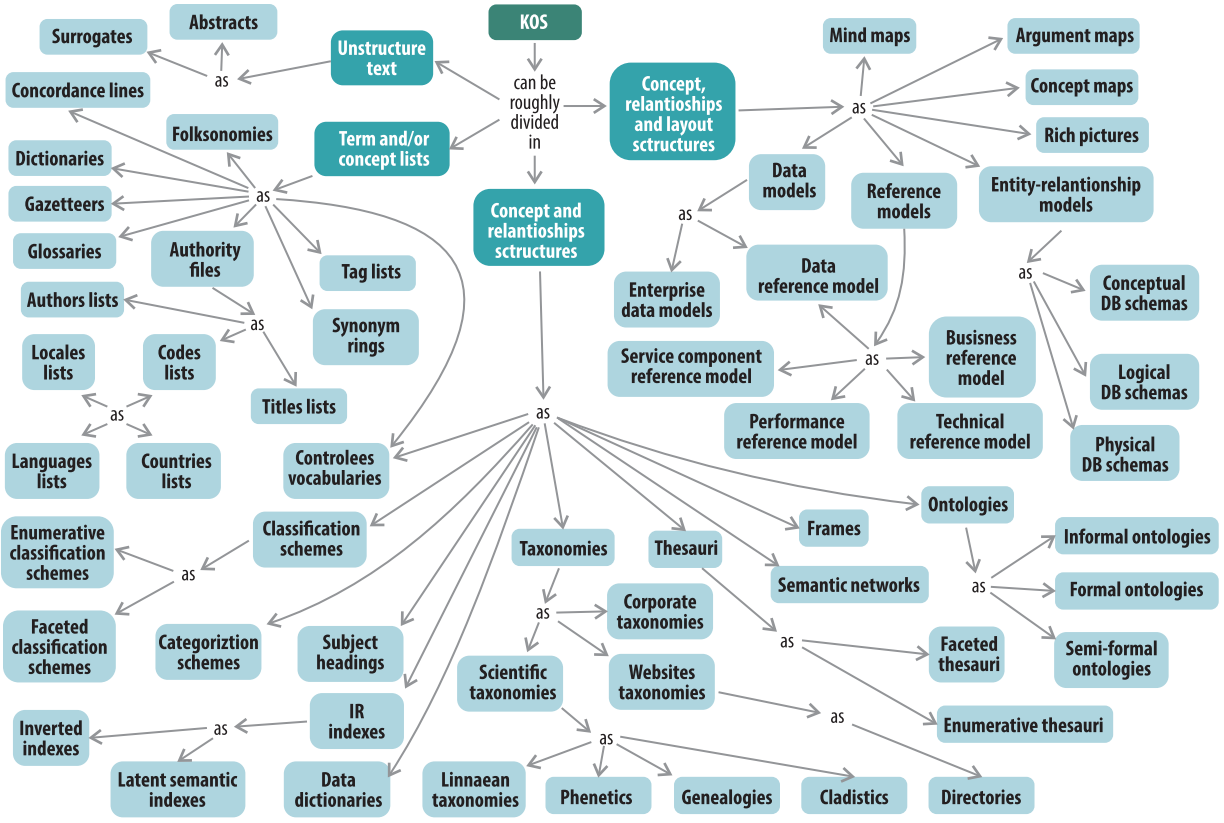



**Figure 4.** Classification of Knowledge Organization System.

Source: Souza et al. (2012).

we are in, we need methods of organization and representation of knowledge to understand the world around us (Mendonça, 2005). It so happens that to represent knowledge, before taking any action, it has to be classified, in other words, we need to organize it in a system so that we can understand it (Martins; Moraes, 2015).

The process of organizing objects is a classificatory and relational process, which requires our cognitive ability to associate ideas, create order and meaning in our experiences, using the interpretation of the world, the attribution of meanings and the structuring of ideas. The representation of knowledge is, therefore, part of a totality that, through perception and reason, looks to formulate abstract concepts about the reality to which it belongs (Martins; Moraes, 2015).

It can be argued that the world as an object of human knowledge exists as an interpreted world that is completely infused with meaning. Human cognition cannot see simple facts without these being part of its structure of meaning (Tuomi, 1999). Hence the importance of knowledge organization and representation systems, especially when they fall within the scope of the developing Information and Communication Technologies (ICT) solutions for digital agriculture.

In the context of the Information Engineering Research Group of Embrapa Digital Agriculture, these systems are understood as "performed" objects, that is, constructed and executed in line with their contexts, as objects that insert multiple realities, responding directly to the peculiarities of modeling complex systems. In the context of Brazilian agricultural knowledge, the KOS are presented as organizational and representative systems of knowledge in this domain, describing associations between the multiple interdependent elements that act in the production of knowledge. (Latour, 2012).

They must be perceived as spaces that integrate the social, cultural, environmental and economic contexts of Brazilian agriculture in interdependent relations that, in addition to action, consider the process of knowledge production within the scope of an articulated and hybrid reality, with complex capillarity involving human and non-human agents.

In this sense, the system of representation and organization of knowledge at Embrapa is an open system, which is configured more as a constitutive map of a network of actors that influence themselves by being in permanent interaction – redesigning new routes – than as a closed, limited and static territory.

As open systems that have agents (human and non-human) that articulate and transform each other, the KOS at Embrapa have been constructed and employed in different dimensions of knowledge domains, resulting in fragmented and dispersed ontologies in time and in space, whose representation potential is impaired due to the absence of higher-level perceptions. Thus, the use of these conceptual artifacts as they have been conceived and used, to support computational modeling of agricultural knowledge, still faces difficulties of coherence and convergence, as KOS should represent multiple ontologies resulting from various methods and practices used by researchers, who move the knowledge of Brazilian agriculture at Embrapa (Baum et al., 2020). Therefore, KOS conceived from the perspective of ontological policies (Mol, 2008) are understood as performative objects with greater adherence to the paradigm of complexity related to Brazilian agriculture.

Thus, agriculture as an object of knowledge is not a passive subject waiting to be perceived from the point of view of an endless series of perspectives. On the contrary, it is a living and open object constituted through scientific practices, through which it is manipulated to be better understood (Baum et al., 2020).

This rationality justified the need to engineer a system of knowledge organization and representation for Embrapa that had a pluralist ontological conception, which implies that it is an oriented system and "[...] favorable to the coexistence of a variety of explanations, assumptions, methods, methodologies, approaches, theories" (Baum et al., 2020, p. 14), which together perform agriculture as an open object, exposed to its multiple relationships, interactions, capillaries and, consequently, to its own entirety.

## Agroterms: controlled vocabulary at Embrapa

Knowledge is a personal intellectual experience and, thus, the term "transfer of knowledge" may have a conceptual meaning, but in practice it has no operational meaning. Such transfer, in fact, happens through a process of encoding brain energy in natural language, and is manifested via communication between a "sending" agent and a "receiving" agent. Humanity has performed this process so naturally that at times it is not even considered that there are other possible means of encoding knowledge, such as symbols, sounds, smells, textures, etc. The truth is that, fundamentally, almost all human knowledge has been encoded in spoken or written natural language, and more recently, digitally, which still retains its original nature. This human naturalness of representation is indeed based on the preponderance of visual perception over other senses.

Thus, a good part of KOS is conceived, built and executed in this manner: based on the lexicon, which technologically, can be modeled through Natural Language Processing (NLP) methods and tools, and therefore, graphically encoded.

Controlled vocabularies are KOS that collect and organize words or terms, in the field of scientific specialties. Based on Concept Theory, terms denote concepts, but they are only one of the vertices of the triangle that represents a given concept. Another vertex is the referent, that is, what the human mind perceives in the real world. Finally, the third vertex refers to the properties that can be attributed to the

referent and that, finally, guide the choice of a lexical element in natural language that best synthesizes the perception of the referent (Dahlberg, 1978). According to the same author, concepts are considered units of knowledge. Thus, the role of controlled vocabularies is contextualized as facilitating resources in the management of institutions, which are confronted with the production, access and sharing of D-I-K, including volume scales and Big Data flow. And in the same logic, currently controlled vocabularies greatly benefit from Information Engineering to be conceived, managed and maintained as open and dynamic systems, in accordance with the premises of best practices of knowledge representation and their pragmatic applications.

Until recently, Embrapa used externally controlled vocabularies in its corporate processes for managing D-I-K, conceived and managed in contexts not perfectly aligned with its own range of contents related to tropical agriculture, which greatly hindered the alignment of this collection of knowledge at various stages of the D-I-K lifecycles, and even more so when deployed on global scales. At the level of cataloging, indexing, retrieving, accessing and disseminating D-I-K processes, problems of consistency, ambiguity and lack of interoperability between information systems accumulated in the same proportion this collection grew exponentially.

The Agroterms was then built, through a combination of Portuguese language terminologies found in national and international agricultural thesauruses, with a methodological and technological basis in the Global Agricultural Concept Space (GACS) initiative (Research Data Alliance, 2020). As a result, Embrapa was recognized as content curator in Portuguese, for the Brazilian variant, by the editor group of Agrovoc (FAO, 2020) – the controlled vocabulary of the Food and Agriculture Organization (FAO) of the United Nations (UN).

Currently, Agroterms is composed of approximately 245,000 terms. Through Information Engineering, using NLP methodologies and tools, Corpus Linguistics and semantic modeling, it is being prepared to expand its technological functionality as a terminological resource to a level of conceptual space for Brazilian agricultural knowledge. In order to achieve this potential, Agroterms is being addressed organizationally by a permanent working group from Embrapa, the Gtermos, responsible for its conception, curation and management, within the context of Data Governance and Information for Knowledge Policy, already implemented at Embrapa and which contributes to the intention and efforts in order to bring digital agriculture to the reality of the Brazilian agricultural sector.

# Research data management

The first decades of the 21st century have been characterized by an explosive growth in human capacity to acquire, store and communicate digital data. From a scientific perspective, the concept of "data-intensive science" or "e-Science" (Borgman, 2007; Gray, 2009), has been consolidated as a reality in numerous fields of knowledge, many of them relevant to agricultural research.

Technological advances have allowed for greater accuracy and coverage in data acquisition. Some examples are Internet of Things (IoT) applications, which are making the use of sensors a reality in the field; the growing possibilities of imaging rural areas using drones (also known as Unmanned Aerial Vehicles - UAVs); other geotechnology applications; and the evolution of bioinformatics and nanotechnology areas.

The current situation has also been called the "Big Data" Era, characterized by the 5 Vs: Volume, Velocity, Variety, Veracity and Value (Mcafee; Brynjolfsson, 2012). For this large amount of data to be useful, it must be well managed, retrievable, and accessible, understandable, and integrated. This scenario has led

to transformations in how data, information and knowledge are created and used: to intelligently and quickly deal with the data "flood", new skills are required to ensure the preservation, integration and reuse of the data.

Research data management (RDM) is a discipline that brings together a set of activities that are essential to the planning, implementation and execution of strategies, procedures and practices aimed at effective data management. There are several approaches to understanding RDM; one of them refers to different conceptions of data life cycle management (DLM) models, which provide a view of the dynamics of data, from its generation to its reuse.

Data lifecycle is defined by DataONE (2020b) as a high-level representation of the stages involved in management and preservation of data for use and reuse (Figure 5). In this chapter, this definition of DataONE is taken as a reference, due to the suitability for adaptation and the versatility of interpretation of its definitions and concepts. This life cycle is composed of eight stages: planning, collecting, ensuring quality, describing, preserving, discovering, integrating and analyzing, which are briefly described, based on Sayão and Sales (2015), Strasser et al. (2015), Araújo et al. (2019) and DataONE (2020b). These steps involve cyclical actions that enable: a) to construct a data management plan aimed at meeting the data policy of the research institution; b) data collection for guaranteeing its usability and long-term reuse; c) ensured guarantee to data sets so they can be used and are reproducible; d) precise and detailed description of the data, adopting a standard of metadata, taxonomies and controlled vocabularies; e) data preservation through proper storage in data centers in order to ensure interoperability, recovery and search; f) discovery of potentially useful data, which, as described by metadata, can be easily found; g) integration of data from different sources, which after combined, generate new sets of data that can be used; h) data analysis provided by the new datasets created in the integration, in order to provide relevant information for future research.

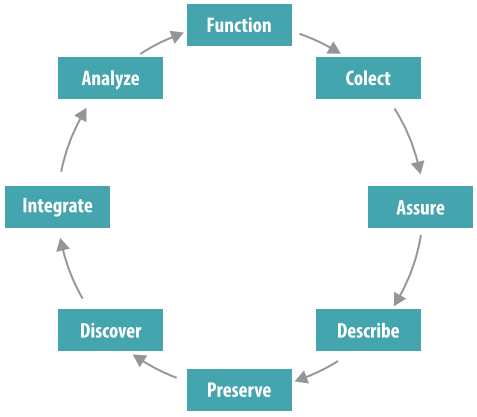



**Figure 5.** Data Lifecycle.
Source: DataOne (2020b).

The RDM based on the data life cycle refers to adopting best practices, which are defined as "[...] methods or approaches that are recognized by a community as being correct or more appropriate for acquisition, management, analysis and data sharing" (Sayão; Sales, 2015, p. 81, our translation). Such practices guide people on how to effectively work with their data at each stage of its life cycle, thus helping to map the processes involved in DML (DataONE, 2020a). Examples of best practices in the planning stage, as those recommended by DataONE it recommends: creation, management and documentation of data; definition of the types and format of data to be produced, etc. For the step to describe, for example, it recommends: to name the files so they describe and reflect their content; describe format for geospatial and temporal location of the data; adopt standard taxonomies for describing any datasets; adopt specialized controlled vocabulary; define set of metadata elements, etc.

However, only the recommendation of best practices does not guarantee efficiency and effectiveness in data management and, according to Veiga (2019, p. 15, own translation), "it is not enough to share data, they need to be FAIR". It is therefore necessary to associate such best practices with the FAIR principles (Findable, Accessible, Interoperable, Reusable), so that research data, in addition to being well managed, can also become findable, accessible, interoperable and reusable (Wilkinson et al., 2016). The four FAIR principles are

made up of 15 elements that contribute to complement and enrich the information essential to the eight stages of data lifecycle, expanding the possibilities of location, access, interoperability and reuse.

The FAIR principles apply to any research objects, so that they become available and understandable to humans and machines, ensuring transparency, reproducibility and reuse, in addition to providing the proper and adequate citation of information generated by data-intensive science (Wilkinson et al., 2016). The FAIR principles also guided the design of Embrapa's Data, Information and Knowledge Governance Policy (Embrapa, 2019) to bring researchers and other research subjects closer to the datasets available in data repositories and platforms.

# Metadata and data cataloging

The objective of cataloging and metadata in this subject is largely data. Data is a word with several meanings, including generic and specialized ones, depending on the context in which it is used. According to Semeler and Pinto (2019, p. 113, own translation), "[...] data means a single piece of information, "while" research data are the result of any systematic investigation that involves research processes of observation, experimentation or simulation of scientific research procedures."

Metadata and data cataloging are necessary so that research data can be "[...] identifiable, citable, visible, retrievable, interpretable, contextualizable, interoperable and reusable when considering consistency and origin" (Semeler; Pinto, 2019, p. 116). Also noteworthy is the need to consider the context of research data management and the data life cycle in which metadata and data cataloging are inserted, which, in addition to being two important sub-steps of the description stage, should be aligned with the FAIR principles.

Given the need to expand data and information representation mechanisms to better manage them, and the consequent complexity involved in defining their attributes, metadata can no longer be defined as only "data about data". Currently, this definition is considered an expression that does not help to understand what exactly metadata means (Sayão; Sales, 2015). What expands this understanding and expands its application domain is the definition given by Riley (2017, p. 1), who considers metadata as the information we create, store and share to describe things, and which allows us to interact with these things to obtain the necessary knowledge.

Metadata allow exploring other dimensions and facets of the data, which, when revealed by cataloging, contribute to improving management and quality, favoring the discovery of data collections for the scientific community. Such dimensions bring to light the need to create new metadata elements that are capable of expanding and enriching the adopted metadata scheme. Metadata are essential so that in the future digital content can be accessed and interpreted. Without metadata, according to Gray (2009 cited by Sayão and Sales 2016, slide 83), the users

> [...] will not know the details of how the data was obtained and prepared: 1) how the instruments were designed and built; 2) when, where and how data were collected; and 3) will not have a description of the processes that led to the derived data, which are typically used for scientific analyses.

Metadata are also essential for technical and semantic interoperability, so that without them, data repositories and platforms will not be able to exchange data and information. Metadata consists of well-defined descriptive elements, for example: author, title, description, subject, keyword, identifier, producer, types of data, access conditions, terms of use of collections, etc., and based on the data cataloging, formulating a body of information capable of contextualizing the data in terms of provenance, history, nature, purpose and other aspects.

The implementation of enriched metadata brings direct benefits to data management, positively impacting archiving and preservation, as well as interoperability and retrieval of research datasets. Data will only be useful for analysis if it has been described by quality metadata, and for that to happen, the best recommendation is to use FAIR principles when cataloging it.

Also with regard to metadata, and according to Veiga (2019, p. 18-22), it is necessary to briefly highlight the key elements that should guide the implementation of FAIR principles, especially regarding the descriptive aspect of metadata: a) metadata elements for single and persistent identifiers for both the data and the dataset; b) dataset using metadata enriched with a wide range of precise and relevant attributes; c) metadata element that clearly and explicitly indicates persistent identifiers, both from the dataset and from the metadata itself in the data repositories and platforms; d) metadata registered or indexed in identification resources that offer search capability; e) metadata using standardized communication protocols to facilitate data retrieval via metadata, including; f) availability of access to metadata, even if the data is not accessible and available; g) metadata element for the representation of knowledge through formal language and the use of taxonomies and controlled vocabularies according to FAIR principles, specialized and standardized by specific area of the domain; h) metadata element for qualified dataset references and other derived research objects, which interconnect, ensuring semantic interconnections between them, and which are linkable to other datasets; i) metadata with a wealth of attributes and high level of detail to allow researcher to evaluate the possibility of reuse and relevance to their needs; j) metadata element with unambiguous information, clearly defining who can have access to the data, for what purpose and under what conditions; k) metadata element that specifies the origin of the data, supporting the researcher when deciding on the usefulness of the data or metadata and when attributing credit to the data producer; m) implementation of metadata must be aligned with relevant and specific standards of the community and the research area.

In the context of digital agriculture, metadata described in accordance with FAIR principles will directly contribute to the discovery and reuse of data by other researchers and research institutions.

# Dataverse platform

The report entitled *Open access to research data in Brazil: technological solutions – 2018 report* (Rocha, 2018) presents the results of the Brazilian Research Data Network (RDP Brazil) research project, which identifies, explores and analyzes in depth three technological solutions (Dataverse, DSpace and CKAN) for building an Open Access to Research Data repository.

Based on the Open Access and Scholarly Information System (OASIS) model – composed of 56 criteria, classified in Repository Environment Representation, Data Set Representation, Data Set Description and Documentation, Data Set Production, Long Storage Deadline and Planning for Preservation, Access and Use of Data Sets and Use, Development and Maintenance of the Software –, it is concluded that the Dataverse and DSpace technologies have resources for configuring various types of data repository, including organizational, thematic and hierarchies distinct data policies for research groups or units, with metadata schemas and support for usage licenses. However, CKAN software is a good alternative when used as a publishing and access service, with submission and digital preservation performed by other repository environments.

The Dataverse Platform is an open-source software application for storing, publishing and sharing data (Dataverse, 2020b). It provides facilities to represent scenarios composed of several hierarchical entities, such as universities, institutions, laboratories, research groups and departments, so as to autonomously

implement the details of data management, such as defining who can create, authorize the publication or access data sets, establish licenses, as well as define that the data can only be used upon request.

The platform uses metadata schemas (compatible with DDI Lite, DDI Codebook, Dublin Core, DataCite, VORResource, ISA-Tab), manages dataset versions, uniquely identifies datasets (considering versions) in a universal and persistent way (DOI or Handle System), provides citation metadata and a citation structure that involves checking the immutability of the cited material. It also enables the storage of complementary documents together with a dataset, adds visualization and data exploration tools, allows the customization of its interface, and provides the collection of metadata from the Open Archives Initiative Protocol for Metadata protocol Harvesting (OAI-PMH).

Additional functionality can be incorporated, such as support for storing large volumes of data; support for data visualization and exploration; improvement in the indexing and search engine and in user authentication systems.

## Metadata on the Dataverse platform: the experience of Embrapa Digital Agriculture

As mentioned before, the Dataverse Platform is a web application dedicated to the sharing, preservation, citation, exploration and analysis of research data, which hosts several dataverses composed of datasets, which are processed through metadata (Dataverse, 2020b). The Dataverse Platform was chosen by Embrapa Digital Agriculture to support the management of research datasets, with the Embrapa Bioinformatic Multi-user Laboratory (LMB) as a pilot project (Embrapa Digital Agriculture, 2020). Therefore, it is in this environment of the Dataverse Platform dedicated to the LMB that the experience reported for metadata takes place. (Dataverse, 2020a).

As in several platforms and data repositories, the datasets in the Dataverse Platform are inserted using a registration form, which contains numerous fields, corresponding to metadata elements. Some of these fields are mandatory, but most are optional. In addition, there are additional metadata sets that can be added for specific data domains. Metadata sets follow established standards, ensuring interoperability with other platforms.

Based on the FAIR principles, studies and tests were conducted that involved users and producers of the LMB pilot project data, searching to define metadata elements (terms) to meet the specificities of the pilot project users. Based on this work, definitions were obtained regarding: a) basic metadata, some mandatory, as well as complementary, non-mandatory metadata; b) tools for metadata description: description norms, taxonomy, thesaurus and controlled vocabularies. These definitions are important for generating enriched metadata, which are those that use domain-specific description tools, such as taxonomy, thesaurus and controlled vocabulary, associated with provenance information, bringing semantic clarity when compared to basic (original) metadata. According to Lira (2014, p. 43):

> A set of enriched metadata must present features such as: (i) greater number of semantic attributes...; (ii) ease of interpretation and processing of dataset content; (iii) associated standard vocabulary terms [...].

Metadata on the Dataverse Platform, based on experience with the creation of dataverses and datasets for managing research data in the LMB, are mirrored in the FAIR principles, especially to make them rich in highlights and with numerous precise and relevant attributes.

## Data cataloging on the Dataverse platform

The process of cataloging the datasets of the LMB pilot project begins with the self-deposit activity, which is carried out by the researcher, the dataset owner or another designated person. The time of self-deposit of the dataverse or dataset on the Dataverse Platform corresponds to the pre-cataloging phase, filling in the fields corresponding to the basic metadata elements and mandatory information: title, author, contact, description and subject.

The cataloging takes place in the next phase, which begins by reviewing the previous filling in of the contents related to the basic metadata elements. The catalog description of the dataverse and dataset essentially fills in the complementary metadata elements, which will be done by the domain specialist, preferably by the dataset owner. However, this activity requires knowing librarianship norms and standards, giving greater meaning and quality to datasets.

The use of cataloging techniques to describe and represent data or any research objects, based on a structured set of FAIR metadata elements, is essential to ensure technical and semantic interoperability, sharing, use and reuse research data, with the technical responsibility of the librarian and/or the information scientist.

# Data quality

The high standard of quality in data archiving and maintenance is widely recognized by different segments of society. In digital agriculture, data quality is particularly important for a high level of assertiveness in the decision-making process, planning activities, and others. Based on this finding, and on the impact on different types of business, the data quality (DQ) theme is seen as a strong pillar in the data management process. This has increasingly called the attention of researchers from different areas of knowledge to investigate and expand studies on the subject.

As shown in the literature, there are different definitions for DQ. These variations are related to the context that discusses DQ and the degree of demand the user addresses quality.

It is important to highlight the definition of Data Management Body Knowledge (DMBOK) for DQ:

> "[...] the planning, implementation and control of activities that apply data quality management techniques, in order to ensure they are suitable for the consumer and meet the needs of data consumers" (Knight, 2017, p. 1).

For an organization, whether public or private, focused on business or research, the archiving, maintenance and recovery of high quality data bring opportunities to formulate better business strategies, decision-making facilities for high level of success, and for achieving better competitive advantage conditions. The lack of this quality level, in addition to making it difficult to achieve the aforementioned advantages, contributes to added data processing costs. In addition, the customer's level of satisfaction decreases when he/she receives the result of a service or a required information as a response to a research action. (Jaya et al., 2017).

The DQ theme should not be treated independently, as poor-quality data leads to misleading and costly conclusions, which can lead to the data team's distrust or loss of credibility. Problems found in the organizational culture of an institution also affect the data collection and quality process.

DQ can be developed under qualitative and quantitative approaches (Vancauwenbergh, 2019); in the qualitative approach, categories (for example, measurement, contextualization, representation and access) are determined, and dimensions/characteristics are associated with each one of them. In turn, in the quantitative approach, DQ is guided by the adequacy of data to serve a purpose in a given context, that is, in decision-making and/or planning operations.

# Measures for evaluating data quality

Data quality assessment is a crucial process within data quality management, as it comprises different stages that involve several groups of people in an organization. The purpose of this type of assessment is to identify data containing some type of error and measure the impact of various data-driven business processes.

The process can be started in the data collection phase, including minimum guidelines for quality assurance, helping to reduce the amount of work performed in the data qualification/preparation phase (Pyle, 1999) and thus providing better conditions for carrying out data analysis.

DQ can be evaluated using subjective measures and/or determined by computational calculations. In practice, when a data quality assessment process is initiated, basic measures are used, such as the number of missing data, amount of data with typographical error, amount of non-standard data, basic statistical measures and visual analyses. Adopting these measures provides a preliminary notion of how good the level of data quality is. In some situations, the use of a single measure is not enough to achieve the desired result, hence using other measures together.

In addition to the measures mentioned for evaluating data quality, others are described by Cichy and Rass (2019). They are related to frameworks available for establishing this type of evaluation. These measures are used when there is an interest in expanding the assessment level. The higher the quality level, the greater the accuracy of the results generated, which brings greater possibilities for more assertive decision-making. Another important contribution of having quality data is contributing to the discovery of knowledge in a database, knowledge that is inserted in the data and not visualized, at first, by the user (Fayyad et al., 1996).

# Data quality management

DQ is one of the crucial problems to correctly measure and analyze science, technology and innovation, which allows the adequate monitoring of research efficiency, productivity and also strategic decision-making. (Fan; Geerts, 2017).

Typically, data has some kind of inconsistency – some are duplicated, incomplete, inaccurate and obsolete. To enable them to produce quality results, after being processed and analyzed, the Data Quality Management (DQM) process is used.

The main objective of DQM is to remove any and all problems found, raising the quality of the data and enabling them to contribute to the addition of value to the business processes and/or produce qualified answers to the questions addressed (Vancauwenbergh, 2019).

DQM involves performing various tasks, defining parameters and assigning values to them, and establishing a workflow. All this must be easily recorded, updated and retrieved. To support all this work, different frameworks are available, with emphasis on: DAMA DMBOK´s Data Governance Model (Barbieri, 2013); EWSolutions' EIM Maturity Model (Smith, 2009); and Oracle's Data Quality Management Process (Oracle, 2009).

These models are centered on three basic elements, which are the metadata associated with the data, the processes for recording, organizing and (re)using data, and the organizational context in relation to the data. The quality of each element and the interaction between them, will ultimately determine the quality and therefore the true value of an organization's data assets.

Permission to describe metadata that is understandable by the entire organization and aligned with the processes, strategies and business objectives of that organization is a resource available in these models. In addition, these models provide the means to report critical success factors, which are useful elements for developing effective DQ management strategies.

## Critical success factors for implementing Data quality management

According to Milosevic and Patanakul (2005, p. 183), critical success factors (CFS) are "[...] characteristics, conditions or variables that can have a significant impact on the success of an organization or a project when appropriately sustained, maintained or managed." Santos (2015) presents 20 CFS applicable to the DQM which, according to Milosevic and Patanakul (2005), form four large groups: a) operational; b) management; c) governance; and d) qualification. The operating group focuses on the operational processes involved in data collection, storage, analysis and security, all of which are highly interdependent. The management group brings together the management processes, which originate from the operational group, mainly aligning data quality with the organization's goals in relation to data and the results of data analysis. The third group, governance, involves the governance processes associated with DQM. These processes can be presented by the organization's senior management as a priority commitment for the implementation of DQM, stimulating a culture change throughout the organization focused on this subject. Finally, the qualification group is deemed essential to invest in a DQM program, even if the company has employee training structure for operational, management and governance actions. The main objective of this group is to inform people about the importance of qualitative data for the organization. In addition to training for the systematic implementation of DQ, throughout the organization, it must institute a continuous monitoring action of the qualifications. This will allow quick adjustments, if errors and adjustments in the business rules are identified.

# Final considerations

The chapter presented an account of the research actions and the results that were carried out and being developed at Embrapa Digital Agriculture, in the context of Information Engineering. Thus, the objective is to align this work with the actions of digital agriculture and also to add value to the pragmatics of scientific knowledge, offering technologies more easily perceived and assimilated by its potential users.

These actions, in progress in the Information Engineering Research Group, take place in the domain of three natures of information (cognitive, documentary and communicative) and through computational artifacts that operationalize the processes that constitute the data life cycles of information and knowledge (D-I-K). Computer Science is the main source generating these actions, able to combine and complement methodological and technological contributions originating in other fields of knowledge, producing inter, multi and transdisciplinary developments with Agronomy, Ecology, Mathematics, Economics, Sociology and the full range of imaginable intersections. Inserted and articulated in this universe of interactions between different areas of knowledge, Information Engineering is an alternative

for the operationalization of strategies, aiming at greater alignment between the actions developed in the areas of Research and Development (R&D) and Embrapa's innovation process.

Embrapa Digital Agriculture, by inaugurating the Information Engineering research line, reorganizes, guides and rescues its competences towards a repositioning of its RD&I actions, considering the perspective of the innovation process implemented in the company. In particular, with regard to facing the current research challenges to consolidate the digital transformation in agriculture, Information Engineering is capable of effectively contributing to conceptual, methodological, procedural contributions and, especially, to support the development of quality artifacts, objects and computational tools, according to an engineering process designed within a pluralistic ontological conception. In other words, Information Engineering expands the possibilities of the various actors that circumscribe the phenomenon of "Brazilian agriculture" to perceive it as an object that admits multiple explanations, assumptions, methods, methodologies, approaches, theories, etc. Furthermore, the efforts and initiatives in Information Engineering, thus far, are aligned with the trends and contemporary opportunities for development and computational applications and ICT in digital agriculture.

Based on the heuristics made possible by Information Engineering, the computational artifacts of D-I-K representation can be used beyond their immediate functionalities (Pierozzi Junior et al., 2018). However, the contributions of Information Engineering can be translated and materialized under different perspectives and exemplified in the form of repositories and databases that enable collaborative work; in the cataloging, indexing and intelligent retrieval of information; use, reuse and data management; in the redefinition of information and interoperability with other systems; in discovering knowledge; in facilitated and controlled access, communication, sharing, learning and collective intelligence. Since digital agriculture is fundamentally based on digital content, from data obtained through the Internet of Things, it is envisaged that Information Engineering will promote facilities and improvements in the construction of computational artifacts that meet the interests of users in different segments of Brazilian agriculture.

Another reading is possible: a) when the data is addressed based on the perspectives of classification, meaning and access, it is said that what is being worked on are its cognitive properties; b) when information is worked on based on the perspectives of cataloging, indexing and retrieval, it is said that what is being worked on are its documentary properties; c) when knowledge is worked on based on visualization perspectives or machine languages, it is said that what is being worked on are its communication properties (dissemination). In addition to these properties, those that are worked on and inherited from the preceding levels of data and information, respectively, must be considered. Thus, the result is a continuous, cyclic feedback movement, which occurs when the communicated knowledge returns as insight for a new round of data and information cycles.

# References

AALST, W. van der. **Processing mining**: data science in action. Berlin: Springer-Verlag, 2016. 467 p. DOI: 10.1007/978-3-662-49851-4.

ARAÚJO, D. G. de; ALMEIDA LLARENA, M. A.; SIEBRA, S. de A.; DIAS, G. A. Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. **Liinc em Revista**, v. 15, n. 2, p. 32-51, nov. 2019. DOI: 10.18617/liinc.v15i2.4686.

BARBIERI, C. **Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK)**. Belo Horizonte: Fumsoft, 2013. 46 p.

BAUM, C.; MARASCHIN, C.; MARKUART, E. N. Política ontológica como abordagem para as relações intercientíficas. **Psicología, Conocimiento y Sociedad**, v. 9, n. 2, p. 8-30, nov. 2019; abr. 2020. Available at: http://www.scielo.edu.uy/pdf/pcs/v9n2/1688-7026-pcs-9-02-6.pdf. Accessed on: May 2020.

BORGMAN, C. L. **Scholarship in the digital age**: information, infrastructure and internet. Cambridge, MA: MIT Press, 2007. DOI: 10.7551/mitpress/7434.001.0001.

CARLAN, E.; MEDEIROS, M. B. B. Sistemas de organização do conhecimento na visão da Ciência da Informação. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, p. 53-73, ago./dez. 2011. DOI: 10.26512/rici.v4.n2.2011.1675.

CICHY, C.; RASS, S. An overview of data quality frameworks. **IEEE Access**, v. 7, p. 24634-24648, Feb. 2019. DOI: 10.1109/ACCESS.2019.2899751.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, v. 7, n. 2, p. 101-107, dez. 1978. Available at: http://revista.ibict.br/ciinf/article/view/115/115. Accessed on: 19 May 2020.

DATAONE. **Best practices**. Albuquerque, NM: University of New Mexico, 2020a. Available at: https://www.dataone.org/best-practices. Accessed on: 27 May 2020.

DATAONE. **Data life cycle**. Albuquerque, NM: University of New Mexico, 2020b. Available at: https://www.dataone.org/data-life-cycle. Accessed on: 2 May 2020.

DATAVERSE. **GenClima**. Campinas: Embrapa Informática Agropecuária, 2020a. Available at: https://www.dataverse-h.cnptia.embrapa.br/dataverse/umip. Accessed on: 2 May 2020.

DATAVERSE. **Harvard Dataverse**. Cambridge, MA: Harvard College, 2020b. Available at: https://dataverse.harvard.edu. Accessed on: 2 May 2020.

DEFOURNY, V. Apresentação. In: TARAPANOFF, K. (org.). **Inteligência, informação e conhecimento em corporações**. Brasília, DF: Ibict: Unesco, 2006. p. 7.

EMBRAPA DIGITAL AGRICULTURE. **Laboratório Multiusuário de Bioinformática**. Campinas, 2020. Available at: https://www.embrapa.br/en/agricultura-digital/lmb. Accessed on: 2 June 2020.

EMBRAPA. Política de Governança de Dados, Informação e Conhecimento da Embrapa. **Boletim de Comunicações Administrativas**, v. 45, n. 16, p. 1-19, abr. 2019. 19 p. (Manual de normas da Embrapa).

EMBRAPA. **Visão 2030**: o futuro da agricultura brasileira. Brasília, DF, 2018. 212 p. Available at: https://www.embrapa.br/visao/o-futuro-da-agricultura-brasileira. Accessed on: 15 May 2020.

FAN, W.; GEERTS, F. Foundations of data quality management. **Synthesis Lectures on Data Management**, v. 4, n. 5. p. 1-227, July 2017. DOI: 10.2200/S00439ED1V01Y201207DTM030.

FAO. **AGROVOC**. Rome: FAO-AIMS, 2020. Available at: http://aims.fao.org/ vest-registry/vocabularies/agrovoc. Accessed on: 20 May 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, Fall 1996.

GARCIA, A. E. B.; SALLES FILHO, S. L. M. Trajetória institucional de um instituto público de pesquisa: o caso do Ital após 1995. **Revista de Administração Pública**, v. 43, n. 3, p. 661-693, maio/jun. 2009. DOI: 10.1590/S0034-76122009000300007.

GRAY, J. Jim Gray on eScience: a transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm**: data-intensive scientific discovery. Redmond, WA: Microsoft Research, 2009. p. xvii-xxxi.

JAYA, I.; SIDI, F.; ISHAK, I.; AFFENDEY, L. S.; JABAR, M. A. A review of data quality research in achieving high data quality within organization. **Journal of Theoretical and Applied Information Technology**, v. 95, n. 12, p. 2647-2657, 2017.

KNIGHT, M. What is data quality? In: DATAVERSITY. **Dataversity.net**. Studio City, CA; Dataversity Education, 2017. Available at: https://www.dataversity.net/what-is-data-quality. Accessed on: 19 May 2020.

LATOUR, B. Reagregando o social: uma introdução à Teoria do Ator-Rede. Salvador: Edufba; Bauru: Edusc, 2012. 399 p. Resenha de: SEGATA J. **Ilha Revista de Antropologia**, v. 14, n. 2, p. 238-243, jul./dez. 2012. DOI: 10.5007/2175-8034.2012v14n1-2p238.

LIRA, M. A. B. de. **Uma abordagem para enriquecimento semântico de metadados para publicação de dados abertos**. 2014. 95 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco, Centro de Informática, Recife.

MARTIN, J.; FINKELSTEIN, C. **Information engineering**. Englewood Cliffs: Prentice Hall, 1989.

MARTINS, G. K.; MORAES, J. B. E. Organização e representação do conhecimento: institucionalização como disciplina científica no âmbito da Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 16., 2015, João Pessoa. **Anais**... João Pessoa: Ed. UFPb, 2015. Available at: http://www.ufpb.br/evento/index. php/enancib2015/enancib2015/paper/viewFile/3162/1030. Accessed on: 16 May 2020.

MAZZOCCHI, F. Knowledge organization system (KOS). In: HJORLAND, B.; GNOLI, C. (ed.). **Encyclopedia of Knowledge Organization**. Alberta: University of Alberta, 2019. Available at: http://www.isko.org/cyclo/kos. Accessed on: 16 May 2020.

MCAFEE, A.; BRYNJOLFSSON, E. Big data: the management revolution. **Harvard Business Review**, v. 90, n. 10, p. 61-68, Oct. 2012.

MENDONÇA, E. S. A organização e a representação do conhecimento no tempo. **Revista de Ciências Humanas**, n. 38, p. 277-94, out. 2005.

MILOSEVIC, D.; PATANAKUL, P. Standardized project management may increase development projects success. **International Journal of Project Management**, v. 23, n. 3, p. 181-192, Apr. 2005. DOI: 10.1016/j.ijproman.2004.11.002.

MOL, A. Política ontológica: algumas ideias e várias perguntas. In: NUNES, A.; ROQUE, R. (ed.). **Objectos impuros**: experiências em estudos sobre a ciência. Porto: Ed. Afrontamento, 2008. p. 63-106.

ORACLE. **Oracle® Warehouse Builder**: user's guide - 11g release 1(11.1). Redwood City, CA, 2009. 764 p. Available at: https://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223.pdf. Accessed on: 16 May 2020.

PIEROZZI JUNIOR, I.; BERTIN, P. R. B.; MACHADO, C. R. de L.; SILVA, A. R. da. Towards semantic knowledge maps applications: modelling the ontological nature of data and information governance in a R&D organization. In: THOMAS, C. (ed.). **Ontology in information science**. Rijeka: InTech, 2017. p. 83-104. DOI: 10.5772/67978.

PORAT, M. U.; RUBIN, M. R. **The information economy**. Washington, DC: US Govt. Print Off, 1977.

PORDES, R.; PETRAVICK, D.; KRAMER, B.; OLSON, D.; LIVNY, M.; ROY, A.; AVERY, P.; BLACKBURN, K.; WENAUS, T. The open science grid. **Journal of Physics**: conference series, v. 78, p. 1-15, 2007. DOI: 10.1088/1742-6596/78/1/012057.

POWELL, W. W.; SNELLMAN, K. The knowledge economy. **Annual Review of Sociology**, v. 30, p. 199-220, Aug. 2004. DOI: 10.1146/annurev.KOS.29.010202.100037.

PYLE, D. **Data preparation for data mining**. San Francisco, CA: Morgan Kaufmann, 1999. 560 p.

RESEARCH DATA ALLIANCE. **Global Agricultural Concept Space (GACS)**. [S.l.]: European Commission; National Science Foundation, 2020. Available at: https://agrisemantics.org/#GACShome. Accessed on: 20 May 2020.

RILEY, J. **Understanding metadata**: what is metadata, and what is it for? Bethesda, MD: NISO Press, 2017. 49 p. Available at: https://groups.niso.org/apps/group_public/ download.php/17446/Understanding%20Metadata.pdf. Accessed on: 1 May 2020.

ROCHA, R. P. da. (coord.). **Acesso aberto a dados de pesquisa no Brasil:** soluções tecnológicas - relatório 2018. Porto Alegre: UFRGS, 2018. 75 p. Available at: http://hdl.handle.net/10183/185126. Accessed on: 14 May 2020.

SANTOS, M. P. da C. dos. **Fatores críticos de sucesso na gestão da qualidade dos dados**. 2015. 55 f. Dissertação (Mestrado em Gestão de Sistemas de Informação) – Lisbon School of Economics & Management, Lisboa, Portugal.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015. 93 p.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa**: [minicurso]. Rio de Janeiro: CNEN, 2016. 196 slides.

SEMELER, A. R.; PINTO, A. L. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. **Ciência da Informação**, v. 48, n. 1, p. 113-129, jan./abr. 2019.

SMITH, A. Enterprise information management maturity: data governance's role. **EIMInsight Magazine**, v. 3, n. 1, jan. 2009. Available at: http://www.eiminstitute.org/library/eimi-archives/ volume-3-issue-1-january-2009-edition/EIM-Maturity. Accessed on: 16 May 2020.

SOERGEL, D. Digital libraries and knowledge organization. In: KRUK, S. R.; MCDANIEL, B. (ed.). **Semantic digital librarie**s. Berlin: Springer, 2009. p. 9-39. DOI: 10.1007/978-3-540-85434-0_2.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. B. Towards a taxonomy of KOS: dimensions for classifying knowledge organization systems. **Knowledge Organization,** v. 39, n. 3, p. 179-192, 2012. DOI: 10.5771/0943-7444-2012-3-179.

STRASSER, C.; COOK, R.; MICHENER, W.; BUDDEN, A. **Primer on data management**: what you always wanted to know. [S.l.]: California Digital Library, 2015. 12 p. (DataONE best practices primer).

TUOMI, I. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. **Journal of Management Information Systems**, v. 16, n. 3, p. 103-117, 1999. DOI: 10.1080/07421222.1999.11518258.

VANCAUWENBERGH, S. Data quality management. In: KUNOSIC, S.; ZEREM, E. (ed.). **Scientometrics recent advances**. London: Intechopen Limited, 2019. p. 1-15. DOI: 10.5772/intechopen.86819.

VEIGA, V. Gestão de dados de pesquisa FAIR: dando um JUMP em seus dados. In: ENCONTRO DA REDE SUDESTE DE REPOSITÓRIOS INSTITUCIONAIS, 1., 2019, Rio de Janeiro. **Anais** ... Rio de Janeiro: Fiocruz/Icit/ UFRJ, 2019. 59 p. Available at: https://www.arca.fiocruz.br/handle/icict/33343. Accessed on: 27 Apr. 2020.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, J. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. B. da S.; BOURNE, P. E.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; GONZALEZ-BELTRAN, A.; GRAY, A. J. G.; GROTH, P.; GOBLE, C.; GRETHE, J. S.; HERINGA, J.; HOEN, P. A. C T.;

HOOFT, R.; KUHN, T.; KOK, R.; KOK, J.; LUSHERM, S. J.; MARTONE, M. E.; MONS, A.; PACKER, A. L.; PERSSON, B.; ROCCA-SERRA, P.; ROOS, M.; SCHAIK, R. van; SANSONE, S.-A.; SCHULTES, E.; SENGSTAG, T.; SLATER, T.; STRAWN, G.; SWERTZ, M. A.; THOMPSON, M.; LEI, J. van der; MULLIGEN, E. van; VELTEROP, J.; WAAGMEESTER, A.; WITTENBURG, P.; WOLSTENCROFT, K.; ZHAO, J.; MONS, B. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, article number 160018, 2016. DOI: 10.1038/sdata.2016.18.

ZENG, M. L. Knowledge Organization Systems (KOS). **Knowledge Organization**, v. 35, n. 2-3, p. 160-182, 2008. DOI: 10.5771/0943-7444-2008-2-3-160.