15. AN IMPLEMENTATION MODEL FOR THE INTERNET OF FAIR DATA & SERVICES

Luís Fernando Sayão[162]
Luana Farias Sales[163]

15.1 INTRODUCTION

There is nothing new in saying that the vast and growing amount of data and information that spreads throughout contemporary society is profoundly reshaping its modus vivendi in all its dimensions, activated by techno-social systems: leisure, education, public administration, business, health care, cultural expression and, above all, personal interlocutions. This intensively connected and planetary distributed "infosphere" becomes possible through the vertiginous advance of computer and network Technologies – and their digital materiality – that provide the creation, capture, copying, transmission, sharing and massive storage of information in a massive, easy and low--cost way. (National Research Council, 2015). It is to be expected that these transformations overlap in a forceful way with the processes of construction of Science knowledge.

Regardless of the point of observation, what can be seen is that scientific research – due to this global trend allied to its immanent connections with technical systems – is producing an enormous and growing flow of digital data. Countless sensors installed in the most diverse devices ranging from distant satellites, particle accelerators, automatic DNA sequencers, even in unpretentious medical implants, allow data to be captured in an unprecedented amount in all scientific domains, from the exact sciences to humanities, art, and culture.

In light of these findings, research data management is currently a focus of interest and one of the greatest challenges for research organization. As a development, data management and curation, on a planetary scale, stand out with prominence in the 21st century research scenario, as well as the ubiquity of digital Technologies for data collection, analysis and archiving in almost all disciplinary domains (Mayermik, 2012). Therefore, research institutions, in different gradations, are reconceptualizing data management and identifying it as an integral part of research processes, reconsidering or expanding their data treatment strategies, implementing management and curation platforms, acquiring analysis and developing training programs for their teams.

There are many motivations for the implementation of new modalities of information services that support data management in academic and research environments, among them is the need to support research activities, accelerate scientific progress and innovation through intense national and international sharing and collaboration (Mushi, 2020). However, we can identify reuse and reproducibility as the main objectives of data management

[162] PhD in Information Science by PPGCI IBICT-UFRJ, researcher at CNEN, professor at PPGCI IBICT-UFRJ, luis.sayao@cnen.gov.br

[163] PhD in Information Science PPGCI IBICT-UFRJ, professor at PPGCI IBICT-UFRJ, luanasales@ibict.br

and important parameters, from which other benefits are constituted. There are many motivations for storing and preserving data, but the primary reason is reuse and reproducibility, emphasizes Borgman (2007).

The planning, development and implementation of research data management platforms, due to the number of variables that need to be addressed, are complex and multifaceted problems. They need to be articulated around workflows, specific disciplinary domains, informational, technological, political, ethical and legal parameters, sustainability, and expertise in an Odyssey marked by constant changes, whose sign is heterogeneity.

This complex environment can be a suitable terrain for the adoption of FAIR Principles as a horizon for the implementation of management services that make research data findable, accessible, interoperable, so that they can be reused for the long term, thus creating conditions for the transition from a self-contained research to a more open, networked and cooperative research, which, at the same time, meets disciplinary requirements that benefit communities of specific cultures and constraints. "FAIR Principles are not magic and do not represent a panacea, but they guide the development of infrastructures and tools that make all research objects optimally reusable for machines and people" emphasize Barend Mons *et al.* (2017, p. 55), founders of this movement. However, the alignment and implementation of FAIR principles in a research institution requires financial investments, cultural changes, training and building technical infrastructure (Graaf; Waaijers, 2011), factors that can be put together around the concept of "Platform of research data management". This type of platform has the potential of operationalize the several layers of management and establish an increasing infrastructure of informational, scientific and computational services towards applying FAIR Principles in research objects, which is called FAIRfication process, whether data as such or algorithms, codes, procedures, workflows or other physical or conceptual devices that lead to data.

In the attempt to equate this diversity, the work herein aims to present a generic architecture to support data service platform project by defining, realigning, aggregating and articulating the several conceptual models – guidelines, policies, services, tools, infrastructures, among others – around a layer model that, as building blocks, can be adjusted according to the depth, extent, and philosophy of each institution or discipline. The model aims to build a possible scale for measuring the level of maturity of management service projects. The proposal architecture aims at making data adherent to FAIR principles, opening the prospect for a growing number of applications and services can link and process FAIR data, making real the idea of "Internet of FAIR & Data Services" – IFDS– which unfold into several benefits for the various stakeholders involved.

To outline the elements of the proposed architecture, the analysis of the literature in the area was taken as a methodology, with special emphasis on articles, reports, annuals, and data infrastructure projects developed by researchers and research institutions.

15.2 SOME CONSIDERATIONS ON FAIR PRINCIPLES AND THEIR IMPLEMENTATION

The notion of proper research data management, idealized in a way that it can maximize the opportunities of finding data and the efficient reuse of research results by humans and machines, is not exactly new as it has been present for decades in scientific research domains, especially by semantic web and ontology engineering communities. Along this path, many options of implementation have already been carried out by pioneer communities to associate data management to the notion of "machine actionability". FAIR principles can be considered a synthesis of these previous efforts and emerged from the materialization of a view, from multiple stakeholders, of an infrastructure to support the reuse of data that can be processes by computers (Wilkinson *et al.,* 2016), which was later coined "Internet of FAIR & Data Services" (Jacobsen *et al.,* 2020; Mons *et al.*, 2017).

FAIR Data Principles advocate that all research products should be findable, accessible, interoperable and thus reusable by humans and machines, expressing the researchers' expectations regarding data resources in current science, and offering a guide for data producers and publishers so they can navigate more securely and objectively around the complexity inherent to research data management. The primary focus of the Principles is to ensure that data can be reusable, by both humans and machines, in subsequent research and transversally reinterpreted accelerating interdisciplinarity and innovation, becoming even more valuable; and also maximizing the added value obtained by the developments of academic publications that have digital and network technologies as their substrate (Wilkinson *et al.*, 2016). In this direction, FAIR Principles outline considerations that are part of contemporary publications of research data and are related to the deposit, exploitation, sharing, and reuse of these resources through manual and automated processes. As such, they describe the characteristics that data resources, tools, vocabularies, and infrastructures must have to support discovery and reuse by other stakeholders in subsequent endeavors, now and in the future.

Unlike different initiatives shaped by disciplinary domains that establish specific practices for managing and archiving data, FAIR "describes high-level, concise, domain-independent principles that can be applied to a broad spectrum of research product" (Wilkinson *et al.*, 2016, p. 2), however, it may be "a basis for the development of flexible community [and disciplinary] standards" (Boeckhout; Zielhuis; Bredenoord, 2018, p. 932). Despite this neutrality, well-known standards, such as WC3 Resource Description Framework (RDF) with formal ontologies, are currently frequently applied solutions for interoperability and information and knowledge sharing that meet FAIR requirements, especially at the metadata level (Mons *et al.*, 2017, p. 51).

As a high-level design, the adoption of FAIR Principles precedes implementation choices, which do not recommend any specific technology or solution, which does not constitute a norm, standard, or specification. However, they offer a set of guidelines for management focused on the reuse of digital research resources. The elements of the four FAIR principles are related, yet independent and separable, and can be implemented in any combination and incrementally as the publishing environment evolves towards higher levels of "FAIRness". The importance and degree of implementation of each principle may depend on the priorities and maturity of each community in the use of certain research objects (Hong *et al.*, 2020). These characteristics contribute to the broad adoption of the principles, as specific communities, including those outside the scientific world, can implement their own FAIR solutions, allowing them to be reconfigured over time to follow the evolution of the underlying technologies (Jacobsen *et al.,* 2020). Therefore, it must be recognized that different disciplines require different types of technical solutions to achieve the same benefits of FAIR data.

It is crucially important to note that the application of FAIR principles extrapolates research data in its most conventional sense. In the narrower scope of scientific and methodological practices, FAIR principles should also be extended to algorithms, codes, tools, methodologies and workflows, objects that lead to the acquisition of data and that, if well documented, allow tracking the provenance of these assets. Thus, they need to be identified, described and reuse, like data.

All digital research objects – from data to analytical pipelines – benefit from the application of these objects, as all the components of the research process must be available to ensure "transparency, reproducibility and reusability" (Wilkinson *et al.*, 2016, p. 1).

This characteristic brings FAIR principles closer to the assumptions of open Science, whose considerations need to go beyond conventional publications.

The primary idea of implementing Internet of FAIR Data & Services is not executed by itself. For such, a data management process is needed that can effectively add value over time. The degree of adherence of research products to FAIR principles is linked to the scope and depth of management to which they are submitted. This assumes the need for a multi-layered framework – scientific, technological, informational and governance, which address the numerous ethical, methodological and organizational problems that lodge between the flows of sharing, integrity, reproducibility, research accountability, as well as new needs and opportunities for large-scale analysis and reanalysis (Wilkinson *et al.*, 2016).

15.3 FAIR PRINCIPLES X SERVICE MANAGEMENT

The implementation of FAIR principles is due to the varying degrees of actions applied to data by the set of data management services made available mainly by the various disciplinary platforms. These sets of FAIRification are captured by the models in three categories: informational, computational and scientific. Boeckhout, Zielhuis and Bredenoord (2018) make this relationship obvious in their analysis for the area of genome that, however, can be generalized.

- The **Findability** principle stipulates that data must be easy to find by humans and machines. In this way, data must be **identified, described and recorded or indexed in a clear and unambiguous way so that they can be located, and their content understood by humans and computational explorers.** In terms of services, this means that a unique and persistent **identifier** must be assigned in a data collection; that the main features are systematically specified, ideally using standardized formats; and that it is deposited or indexed in a public device such as an archive or data center or a disciplinary or institutional repository, which emphasizes the need for information services and management infrastructure. Meaningful and machine-actionable metadata is essential for the automatic findability of relevant datasets and services, and therefore is an essential component of the FAIRfication process (Jacobsen *et al.*, 2020).

- **The Accessibility principal** advocates that research objects are preferably accessible through the implementation, when appropriate, of automated data retrieval protocols; it also recommends that data are available according to clear and well-defined procedures. These conditions involve the establishment of authentication and authorization processes that are aligned to the organization policies and to the disciplinary culture, and also to data specificities – for example, sensibility level. As a FAIR mantra, which must be worked on by data management services, we have that metadata must be unconditionally accessible even if data is not, or is no longer available.

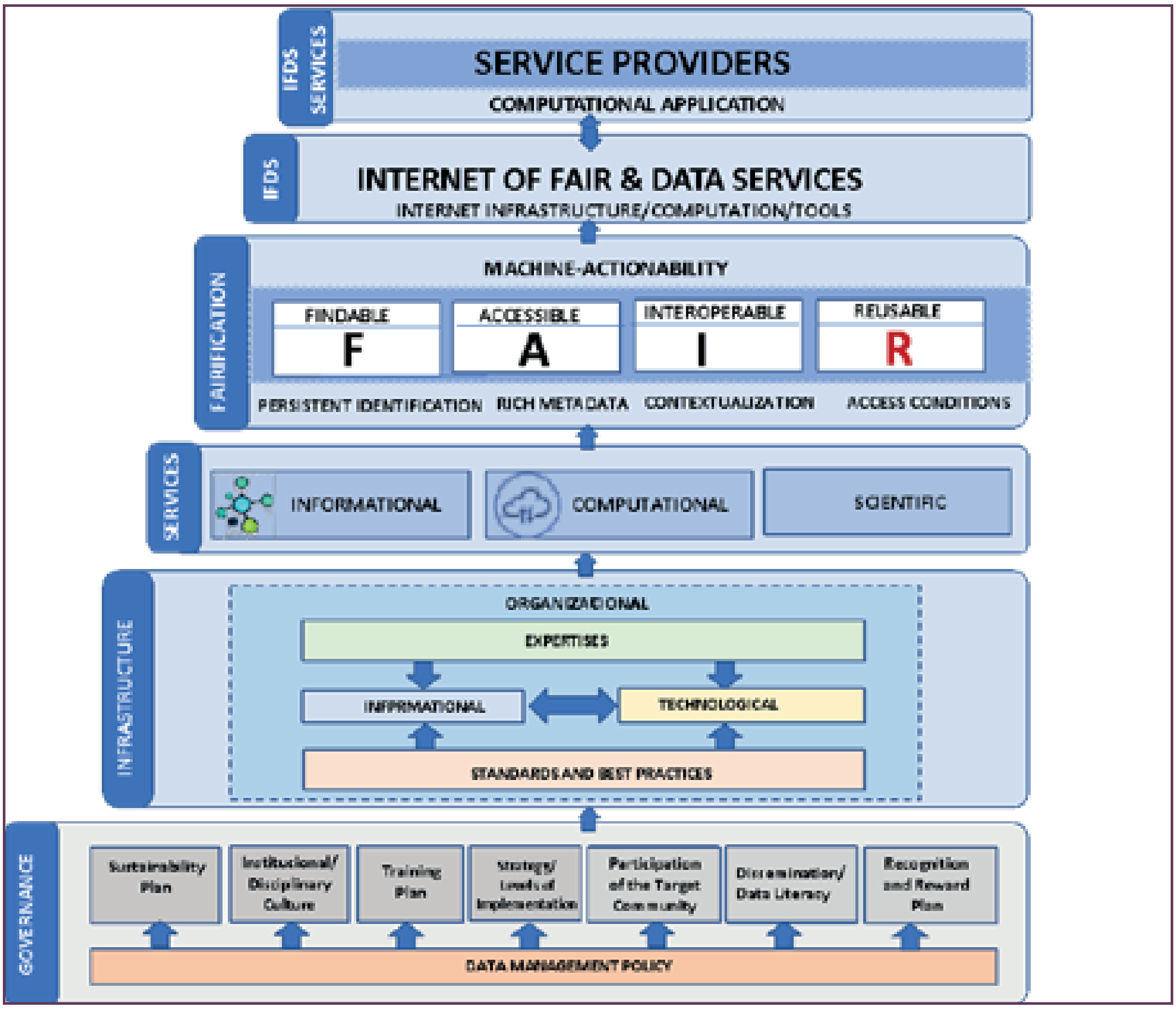
- **The Interoperability** principle is the most difficult to be implemented (Hong *et al.*, 2020), as it is conditioned to a high level of standardization in all its articulation. In general terms, when two or more digital resources are related to the same topic or entity, it must be possible for the machine to merge the information from each one of the resources in a unified and richer view of this topic or entity; similarly, when a digital entity is capable of being processed by an online service, a machine must be capable of automatically detecting this conformity and facilitate the interaction between data and this tool. It requires that the meaning (semantic) of each participant resource – whether data and/or services – is clear (Jacobsen *et al.*, 2020). In this sense, the Interoperability principle assumes that data and metadata are conceptualized, expressed and structured through widely accepted, published, trackable and accessible standards (that is, also FAIR). To achieve this objective, the implementation of this principle comprises a rigorous application of technical and semantic standards at all levels – scientific, computational and informational – such as in terms of variables, protocols, file formats, ontologies, and workflows.

- **Reusability** principle is a consequence of previous principles and reinforces important points advocated by them, such as the detailed description of data characteristics for human beings and computers, including the provenance, according to standards relevant to domain-specific communities. This allows a machine to decide: if a digital resource should be used – that is, if it is relevant for the task discussed; if a digital resource can be used and in what conditions – that is, if the resource meets the reuse conditions; and to whom to give credit in case the resource is reused. The degree of reusability points out the need for licenses appropriate to the conditions of use required.

15.4 DESCRIPTION OF THE PROPOSED MODEL

Research data management has many facets, but none of them can by itself fully explain the intrinsic complexity of its processes. Starting from this point, the model seeks to represent different aspects of research data management without losing sight of the interrelated nature of the dynamics of the activities that unfold in a data-intensive scientific environment whose objective is making data FAIR. As a convenient abstraction of reality that one wants to understand, a model is a cultural creation, a "mindfact", destined to represent a reality, or some of its aspects, to make them qualitative and quantitative describable and, sometimes, observable (Sayão, 2001). From this point on, it was decided to divide the model into six representational layers: 1) governance, where guiding principles of data management services project are discussed; 2) technical infrastructures which also include the necessary categories of expertise; 3) informational, computational and scientific services; 4) the results of the execution of these services manifested by data FAIRfication; 5) which, in turn, is consolidated in a global and shared environment, conceptualized as Internet of FAIR Data & Services (Sales *et al.*, 2020), where 6) Service Providers, through computational applications, offer diverse services. Figure 1 presents a general view of components grouped in layers and their interrelationships, which are discussed below.

Figure 1 – Model of implementation for Internet of FAIR Data & Services



Source: authors.

15.4.1 Research Data Governance : planning, policy, institutionalization, and sustainability

The organization configuration in which data management is performed can vary in relation to several aspects, such as the intensity of support to the management and level of investment applied. Some institutions such as scientific data reference centers and government statistical agencies may be entirely dedicated to data management, having it as their main purpose; in other settings, data management is part of a broader activity that connects to other research activities, such as, in the case of universities (National Research Council, 2015), whose data management activity is a result of their teaching, research and extension functions. However, even in the academic context, there are many ways to plan and execute data management tasks that vary according to objective references such as investment levels, available technical systems, volume, and type of data and how data management is integrated into its workflows and processes; and with more subjective perceptions such as disciplinary culture and academic prestige. In the present model, these parameters are equated by a more administrative level, understood by the term “data governance”. At a more conceptual level, data governance outlines the principles, policies, and strategies that are commonly adopted in an environment that needs a coherent data management; it also outlines the actions, functions, and roles that are required to implement these policies and strategies. Within a research institution, the principles, operationalized by management, govern the entire data lifecycle – from conceptualization to archiving and possible disposal. The data governance process treats data not only in its spatial aspect, but also along its temporal dimension (Solomonides, 2019), this requirement implies an increase in the degree of complexity and scope of governance commitments.

This structuring framework is necessary since digital research data can only be managed and preserved properly over time through a sustained institutional commitment (Mayermik, 2012, p. 1). To some extent, the consolidation of data management services reflects the level of organizational acceptance built into them and the degree of planning the various actions required: current sustainable budget, appropriate data policy, organic connection with target communities, compliance with ethical and legal codes, alignment with institutional strategic objectives and a development strategy that considers the possible paths for each institution. It is also necessary to consider the inevitability of the fact that the technological structures to access, interpret and preserve digital information are continually evolving; anticipating these problems and developing strategies to mitigate them is an activity relevant to governance commitments (National Research Council, 2015). Advanced data services that can appropriately support the entire life cycle of these information assets according to the interests of the various stakeholders involved can be developed based on these pillars. Considering these issues, we propose the following approaches as part of the model:

Institution Data Management Policy **–** Establishes the institution foundations, guidelines, and commitments regarding the management, use, ownership, compliance with ethical and legal ethical codes, adherence to funding agencies policies, national science, technology and innovation policies, to international guidelines and practices and, finally, but of critical importance, to culture, practices, and idiosyncrasies of communities and disciplinary domains: a comprehensive research data management policy must also identify the responsibilities of each of the actors – library, laboratories, information technology, management etc. – since data management involves different sectors of the institution (Mushi, 2020) and the project is considered as part of the institution research activities. It is necessary to emphasize that the process of developing an institutional data management policy requires extensive consultation with all stakeholders and the approval of relevant scientific communities and organizations (Wilson *et al.*, 2011). Policy guidelines must permeate the entire management cycle. "Policies can be an important motivating factor for FAIR data and other research objects (software, workflow, models, protocols etc.). Therefore, it is essential that "bottom-up" Community-based efforts are combined with policies with a "top-down" approach, complete Hong *et al*. (2020).

- *Institutional/disciplinary Culture-the* implantation of research data management services platform must be preceded by an analysis of requirements that considers the institutional, community and disciplinary context and culture and its unique characteristics. This process is expected to help define a more effective portfolio of data management services to support the research practices of the institution and its communities (Mushi, 2020; Coates, 2014; Reed, 2015). It is also important to recognize that some disciplines require different types of technical solutions to obtain the same benefits from FAIR data (Hong *et al.*, 2020).

- *Sustainability Plan* **–** One of the great challenges of a data management infrastructure implementation program is to ensure that each phase of the Project is sustainable as a continuous service over time (Wilson *et al.*, 2011). Once research data management is recognized as necessary for research activities, its costs must be estimated and its funding sources - especially perennial ones - identified. In this way, a project to implement research data management services needs to be associated with a sustainability plan that outlines a possible commitment to the present and the future. Creating and committing to a long-term strategy for services can more clearly reveal the resources needed for continuity of services and the infrastructure needed to do so. This, therefore, may include a succession plan (Mushi, 2020).

- *Dissemination/Literacy in data* **–** For the implementation of a FAIR research environment, it is necessary that the communities involved develop a shared understanding of what is limited by FAIR concept and Principles. In general, researchers and other stakeholders have a low level of perception about the importance of data management practices and the management and sharing requirements of funding agencies and data deposit commitments established with scientific editors, in addition to the ethical and legal issues involved in publishing the data. For example, in relation to the FAIR concept, Hong *et al.* (2020) observe that the researcher does not know what FAIR data is and often thinks it is the same as open data. This indicates that planning and dissemination and awareness actions are needed to elicit these issues. A dissemination program in this direction should include the development of didactic material (booklets and guides), courses, events, workshops, among others.

- *Knowledge/participation of the target Community* **–** As creators and users of research data, the engagement of researchers is crucial in the development of data management services. The provision of any service needs to be based on a close understanding of standards and flows of research that is developed in the institution, its motivations, characteristics, and priorities. Therefore, the precise definition of service requirements needs to be established with the commitment and contribution of the researchers' community; without these considerations, the characteristics of services may not be in accordance with researchers' goals. The community must be accompanied by changes in interest in data, and its participation in the development and choice of sharable standards for practices and for FAIR infrastructures must be recognized and institutionalized. The proximity, interaction, and alignment of communities with national and international organizations that deal directly with FAIR data management such as GO FAIR, RDA, CODATA, DCC and others, should be encouraged.

- *Training plan* **–** To offer complete services in data management, libraries need to have technologically qualified staff or greatly increase technological training for existing staff (Tenopir *et al.*, 2012). Human sustainability is critical to ensure the continuity and consistency of service offerings over time. However, few formal programs in informational studies include data management in their curricula; thus, research data managers are normally trained in service in the specific disciplines where they work (Borgman, 2007, p. 155).

- *Strategy/levels of implementation* **–** The development and implantation of data management infrastructure, in addition to many resources, require time to reach its full maturity and mirror the demands of the scientific communities, which implies the need to establish levels of implementation of infrastructure and services. Research libraries, for example, have, often, proactively sought to meet data management needs for their user communities. This often happens without additional financial support for the development and availability of data services. Therefore, libraries have to start on a simpler scale, building a base on which to develop more sophisticated services (Erway *et al.*, 2016, p. 5), starting with basic services that only require resources from the library itself, until they reach more complex services that require a high level of institutional commitment and more financial, technological and human resources (Kouper *et al.*, 2017).

- *Reward and recognition* **–** Research data management consumes time, resources and requires great dedication from the researcher; however, this effort is rarely noticed by the academic reward system, except When linked to publications in scientific journals. Therefore, to encourage this new task for researchers and highlight its importance, it must be properly recognized and that it is considered in the evaluation, promotion and hiring criteria.

15.4.2 Infrastructures of Research Data

Infrastructure is a broad and multidimensional notion. It can have a technical, legal, organizational connotation and, often, it is essential to also consider social, cultural and political aspects. Indeed, it is so in the science domain: research infrastructure projects are simultaneously a technological issue, a matter of identifying research needs in specific disciplinary areas, and a political issue. This more general perspective applies to institutional research data management infrastructures that need to provide technologies and tools, processes, policies, resources, and training for the various and diverse stages of data management.

Thus, just as institutions must provide basic infrastructure for research – such as laboratories, instrumentation, high-performance computing, networks, reagents and much more – they must also take steps to properly manage data. This assumes a broad spectrum of managerial, technological and informational activities that include information professionals trained to support researchers in the planning and management of their data, in the access of secure to secure storage devices and backups during project development and availability of access and long-term preservation platforms, necessary after the end of the research (Strasser, 2015); it is also essential to have a body of norms, standards and good practices that allow, mainly, a dialogue at different levels of systems and services, both local and global, which can be translated by interoperability.

When we compare traditional academic publishing with data publishing, we verify that the underlying infrastructures of academic publishing create an epistemological bridge between disciplines, having as aggregation point the research libraries that select, collect, organize and make publications of all kinds and all fields. Due to their nature, social institutions work to stabilize particular practices and forms of knowledge. In a certain sense, the institutions are social infrastructures in themselves. Therefore, the technical infrastructure is intertwined with the social infrastructures of the institutions, many times mediated by standards, protocols, documents, and devices that link the social and technical aspects of the infrastructures (Leonardi, 2010). However, there is no infrastructure of this magnitude for data. Some few areas have consolidated mechanisms to release data; others are in the stages of development of standards and practices to add their data and become the most widely accessible. "The lack of infrastructure for data amplifies discontinuity in academic publication" (Borgman, 2007, p. 155).

The infrastructure frameworks used for data management are diverse and fragmented in terms of flows, complexity, application and topology, and organized differently across various disciplines and in different countries (Graaf; Waaijers, 2011). However, the infrastructures increasingly shape standards and data management practices.

Therefore, knowledge about the origin, disciplinary domain, degree of processing, collection system, workflows etc. seem to be of essential importance in the conception of infrastructures for data management (Sayão; Sales, 2020).

In this model proposal, we consider five necessary types of infrastructure: standardization, technological, informational and organizational.

Standards and Best Practices-standards are consensual ways of codifying knowledge that circulates transversally through communities to ensure uniformity and similarity in our products and processes through time and space. They reflect more current knowledge about professional practices and increase interoperability, consistency, preservation, reusability, security, and protection of digital collections. Therefore, ensuring that in a scientific ecosystem, in which infrastructures are globally dispersed, its products are aligned with FAIR Principles, as well as they have a satisfactory degree of quality and excellence and are appropriate to researchers' needs, requires a body of standards and practices widely adopted and shared. Considering this fact, it is proposed that a consensual body of standards and best practices establishes infrastructures that must underlie the data management processes. This is because it is expected that data collections are suitable to be used for a wide variety of purposes – and not only for the purpose for which they were initially collected. To do so, they need to be added to other collections in other systems, shared, accessed, analyzed and archived using a wide spectrum of technologies. This condition makes a body of standards and common practices an essential infrastructure for the management and curation of research data. As the principle and practices of research data management develop, they begin to acquire knowledge as a distinct field of knowledge and to draw the attention of organizations interested in their improvement, such as DCC, Codata, GOFAIR, DataOne, DataCite, among many others. In this regard, standards and procedures commonly adopted for data management are taking part in many disciplines and sectors and are being redefined in other disciplines. As a result, practices improved to ensure digital data quality and durability have been continuously established. (National Research Council, 2015).

- *Technological Infrastructure* – Comprises a broad set of activities, equipment, processes, and expertise that can enable operational technological requirements necessary to data management cyberinfrastructures, such as logical, physical and virtual data organization; devices for high-performance processing, grid computing and storage of local or cloud data collections; local networks, communication, external connections, internet, web services; acquisition/development of scientific codes, workflow software; equipment for data analysis and view; physical, logical and network security strategy.

- *Informational Infrastructure* **–** Comprises persistent representation and identification schemes; descriptive, technical, administrative, preservation and disciplinary metadata; apart from taxonomy, ontologies, classification schemes; databases; it also includes repositories, digital libraries and reliable platforms for long-term archiving.

- *Personnel Infrastructure* **–** The many research institutions develop the most diverse approaches to data management. This assumes support teams made of different professionals (Pinfield; Cox; Smith, 2014). Roles as data stewards and data scientists are emerging in the world of contemporary science and joining the more traditional of researchers, lab technicians, research assistants and analysts; on the other hand, within the scope of specialized libraries and repositories, new stakeholders as librarians and data archivist and curators make the connection between libraries and laboratories and support the management of disciplinary idiosyncrasies of data life cycles (Ball, 2012). However, an essential requirement – especially when it comes to services associated with curation – is the need to know the disciplines and domains in which data are collected, processes and used. Without some familiarity with the problem to be addressed, he disciplinary culture, the goals to be pursued, as well as the methods used, nomenclature and practices of Fields in which digital assets are used, curators will not be able to make the most correct decisions to manage these assets for current and future use (National Research Council, 2015).

- *Organizational Infrastructure* – The infrastructure framework assumes, like governance, an anchoring based on some organizational structure aimed at research, as a university, research institution, or even a company whose projects depend on data management. These organizations offer technologies and tools, processes, policies resources and training to several and diversified stages of data management.

These infrastructure aspects – which enable interrelation of knowledge and practices that are underlying to equipment, installations, methodologies and mainly people – provide several services, tools, and processes that continuously put research objectives in line with FAIR principles. These limits are not always clear, for instance, the repositories are points of aggregation of technologies, standards, informational resources and expertise around archiving of research objects and constitute an essential link to reach Internet of Fair Data & Services, bring various stages of data management life cycle together in their research environments.

15.5 SERVICES FOR DATA FAIRFICATION

To begin with, it is necessary to clarify that we are dealing here with services offered by the various data management platforms to provide research objects with degrees of alignment with FAIR Principles. These services are distinct in nature from the services offered by IFDS to humans and computing agents, for of researchers and other stakeholders. Therefore, services for FAIRfication can be classified as informational, computational and scientific:

- *INFORMATIONAL SERVICES* **–** They comprise the services offered by information professionals within organizations such as scientific libraries and information centers: persistent identification of research objects and researchers; development of representation structures such as metadata schemas, taxonomy, and ontologies; cataloging and indexing of research objects; data release; disclosure; researchers literacy; development of data collections; support for the elaboration of data management plans; long-term archiving/preservation; linking/contextualization.

- *COMPUTATIONAL SERVICES* – It comprises availability of software tools and computing resources to support the processing analysis and visualization of research data; recommend how data can best be structured and stored, and work, if necessary, with researchers in structuring databases and text marking (Wilson *et al.*, 2011); these services may also include specific training for the research team in the resources offered and, in more advanced situations, offer high-performance processing and grid computing.

- *SCIENTIFIC SERVICES* - They comprise services that are limited to the scientific environment, such as laboratories, and performed by researchers or data stewards specialists with disciplinary knowledge. These are services related to preparing data for wider uses and may include activities such as evaluation, cleaning, normalization, file organization, appointment and, when necessary, anonymization, and other strategies for preserving privacy, disciplinary indexing; code documentation, workflow and processing, data aggregation. Even considering that these services are carried out by researchers themselves, they need considerable computational support.

The services that support FAIRfication processes towards an Internet of Fair Data & Services, have as focal point some essential concepts for the materialization of their assumptions and reuse. They are: machine-actionability; metadata; and access conditions.

- *FAIR IS ABOUT ACTIONALITY BY MACHINE* **-** "Recognition that computers must be able to access data released autonomously, without the help of human operators, is central for FAIR Principles", categorically state Mons *et al.* (2017, p. 51); therefore, "FAIR principles place a privileged emphasis on improving the potential of machines to find and use data, in addition to supporting their reuse by human beings" confirm Wilkinson *et al.* (2016, p. 1). This is clear when one observes that much of the data life cycle, such as indexing, retrieval via API, processing and reliable analysis of sensory data, are computer-assisted and executed procedures, highlighting the concept of "machine-actionable". In general, this concept assumes a continuum of possible states in which a digital object provides increasingly detailed information to a computational data explorer of autonomous action. The "computational stakeholders", as Wilkinson *et al.* (2016), called them, such as application programs and computational agents, are explorers who act on our behalf – human beings - , performing an increasingly relevant role in data retrieval and analysis. In this constant transitioning context, it is necessary, therefore, to consider that human beings are not the only critical interlocutors in the data ecosystem. FAIR principles are also, and primarily, for machines. Considering the primary limitation of human beings to operate at the scope, scale, and speed required by the level of complexity of contemporary research, especially in the scope of eScience, it is evident the need for machines to be able to act autonomously and appropriately when faced with the broad spectrum of types, formats, protocols, and access mechanisms encountered in exploring the global data ecosystem. "One of the great challenges of intensive data science is, therefore, to improve the discovery of knowledge through the assistance of human beings and their computational agents" (Wilkinson *et al.*, 2016, p. 3). This interlocution is of great importance in retrieval, access, integration and for "the types of deep and broad integrative analyzes that constitute most contemporary eScience" (Wilkinson *et al.*, 2016, p. 3).

These configurations and conditions of current Science have a profound impact on the processes of modern data management platforms, and the full or partial adoption of FAIR principles as part of the backbone of these management-technical systems is an important step towards machine actionability, as it enables them to optimize the use of data resources through appropriate technical implementation choices. For example, the digital resource can be used as an agent or subtract in analyses based on machine learning or artificial intelligence.

Finally, it should be noted that not all data can or should comply with the condition of being automatically processed. There are numerous circumstances that making data machine actionable reduces its usefulness – for example, when adequate tools capable of efficiently processing certain formats are lacking s (Mons *et al.*, 2017).

- *FAIR IS ABOUT METADATA* – Making an essential bridge between machine actionability and metadata, Wilkinson *et al.* (2016) clarify that a resource that lies on a continuum of possible machine-actionable state, provides increasingly detailed information to a computational explorer, and it is applied in two main contexts: first, it refers to contextual metadata that involves the digital object, that is, recognizing the digital object; second, when it refers to the content of the digital object strictly speaking (how to process / integrate it?). In this matter, this information – depending on the quantity, structuring and quality – allows an agent who is faced with a digital object not previously found to identify the kind of object in relation to structure and intention; to identify its utility in the context considered; to determine if it can be used according to its license, consent, level of sensibility or limits of use; and take appropriate actions, similar to what a human would do. Therefore, assisting machines to find and explore data through technology applications and standards at the level of data platforms becomes the main priority of a good data management and highlights the importance of the concept of metadata. The metadata standards have an important role in scientific communication flow, whose emphasis extends the methodological and transparency requirements of the scientific report for data management domain. As such, FAIR Principles emphasize the importance of metadata and its standards in data management, focusing on the concept of "metadata" across its 15 guiding principles. "FAIR Principle's key message is that metadata and metadata standards should be articulated and made publicly available to the greatest extent possible" (Boeckhout; Zielhuis; Bredenoord, 2018, p. 932).

- *FAIR IS ABOUT ACCESS UNDER WELL-DEFINED CONDITIONS* - "FAIR is not the same as open", assertively state Jacobsen *et al*. (2020). The "A" in the context of FAIR is understood "Accessible under well-defined conditions", which makes it different from open without restrictions. Mons *et al.* (2017, p. 51) point out that may be legitimate reasons to shield data and services generated with public funds from indiscriminate access. These types of data include: sensitive personal data, data on geolocations of endangered species, on patentable processes, national security, among others. Furthermore, several sectors, such as industrial and medical, for legal, ethical, contractual or competitiveness reasons need appropriate security for their data and require additional authorization and authentication measures, both for human explorers and for computational agents; in practice, the Internet of Fair Data & Services cannot function without these mechanisms (Jacobsen *et al.*, 2020). Although maintaining primary connections with Open Science, FAIR Principles explicitly and deliberately do not address ethical and moral questions about the degree of openness of data, their availability is entirely at the discretion of the data custodian. FAIR Principles only address the need to describe a process – automatic or manual – for accessing discovered data; a requirement to describe extensively and openly the context in which these data were generated.

The principles do not require FAIR data to be "open" or "free"; however, they do require clarity and transparency about the conditions that govern their access and reuse; they also require that data have an accessible and clear license, preferably machine-readable. "Transparent but controlled access to data and services, rather than the generic and ambiguous concept of "open", allows the participation of a wide range of sectors – public and private - [...] around the world", concluded Mons *et al.* (2017, p. 52).

15.6 "FAIRFICATION" TOWARDS IFDS

The fundamental idea of implementing an Internet of Fair Data & Services is not made real by itself. To this end, it is necessary a multidimensional data management process that can effectively add value, over time, to research objects; the level of adherence of research products to FAIR Principles is linked to the scope and depth of management to which they are subjected. This assumes the need for a multi-layered framework – scientific, technological, informational and governance, as presented in the previous sections, which address the numerous ethical, methodological and organizational problems that interpose between the flows of sharing, integrity, reproducibility, provision of research accounts, as well as the new needs and opportunities for large-scale analysis and reanalysis (Wilkinson *et al.*, 2016, p. 1).

To clarify the meanings embedded in the acronym FAIR, Mons *et al.* (2017) offer a FAIRfication scale – here understood as the level of depth and coverage of management that make digital research adherent to FAIR Principles. During this process, at the lowest level of this scale are objects with no potential for reuse, which correspond to unreleased data, or released in unstable environments such as a web page. These objects do not have **machine-resolvable persistent identifiers** that lead to both data elements and corresponding metadata; these, in turn, are not machine-readable. The minimum path towards FAIRfication is to assign a persistent identifier to a dataset.

However, without a set of **machine-readable metadata,** it will be difficult to find the resource, unless its identifier is known in advance. This indicates that the identifier is necessary but insufficient, and that we need to go further. The next step is assigning metadata, which can have two origins: "intrinsic metadata", which is signaled at the time data is captured, usually by automated processes carried out by the instruments or workflow that generated the data, for example, file format, time stamp and location; metadata marked by the researchers who created/collected the data, information professionals and the stakeholders who reused it in the form of, for instance, annotations, which provide provenance and contextualization to the data and increase its degree of FAIRfication. Therefore, the addition of rich metadata – and also FAIR – is an essential step in this journey. Thus, "the persistent identification and aggregation of metadata already has a profound effect on the reuse potential of research objects, since they can be identified and retrieved" (Mons *et al.,* 2017).

However, even if data is technically FAIR, access may be restricted for clear and fair reasons such as contracts, protection of endangered species, legal and ethical issues; that said, we understand that the maximum standard of FAIRfication should happen When data elements themselves are available under well-defined conditions, for open reuse by any other interested party.

Going even further on the FAIRfication scale, Mons *et al.* (2017) propose that when data are linked to other FAIR research objects we will have reached " FAIR Data Internet"; since an increasing number of applications and services can link and process FAIR data, it can be said that the "Internet of Fair Data & Services" will have been achieved, meaning a "global and shared environment focused on data-driven research and innovation" (Sales *et al.*, 2020, p. 3), where all researchers can access, store, analyze and reuse data for research, innovation and educational purposes. Based on the contours of this territory, an ecology of data activated by associated services is established, which, for the different user segments, translates into a continuum of benefits triggered by computational applications

Just like the current internet, which does not have a centralized governance and is based on a minimal but rigorous set of standards and protocols that support an immense variety of implementations, the concept of "Internet of Data Fair & Services" assumes maximum freedom of development for all interested parties. In this sense, the scalable and transparent routing of DATA, TOOLS, and COMPUTATION – which processes (executes) the tools – is the central feature of a desired Internet of Fair Data & Services, where all types of service providers, public and private, can begin prototyping FAIR data and service applications FAIR (Go Fair, [20--?a]).

As an abstraction, the IFDS models itself in the shape of a three-blade propeller that corresponds to the fundamental elements – data, tools, and computation – that are "routed" to find each other at the right time and place and to be used and reused more efficiently. In this context, tools are mainly defined as software services that act on data, such as virtual machines packaged to travel through the IFDS doing distributed analysis of data or even a data repository and computing as the infrastructure that enables action. As in the hourglass model of the internet, the helix axis corresponds to the minimum set of standards and protocols, as the growth of the ISDF is based on the mantra of the GO FAIR network: "Only a necessary minimum set of protocols and standards to support a wide variety of implementation choices for data, tools and computing elements".

IFDS would run more smoothly if the underlying infrastructure operated on a strong, common, and globally interoperable network and on an engine that efficiently routed data to tools, tools to data, and both to the computation needed, as these three elements are increasingly not residing in large super storage systems and HPC facilities, but are distributed throughout the internet (Go Fair, [20--?a]; Go Fair, [20--?b]).

15.7 FINAL CONSIDERATIONS

Contemporary science, data-intensive by nature, requires data management whose scale goes beyond the most conventional measures, and needs to continually put these assets and other research objects ready for reuse – the ultimate goal of management – by human beings and, above all, by service providers, through computational applications, thus expanding their potential for reuse, repurpose and resignification for various segments, including those outside the world of research. The difficulties of humans operating at the scale and speed required by the complexity of data intensive sciences, especially science, reinforce the need for computational explorers to act autonomously and appropriately in the face of a global data ecosystem.

However, to reach this state of continuous supply, a chain of processes is necessary, ranging from the establishment of policies to a high degree of standardization that requires an infrastructural framework whose density depends on the level and depth of management. But what can be seen is that this effort, sometimes entropic and with diffuse objectives, needs organization and a horizon. The application of FAIR principles realigns these efforts and establishes clear objectives for the management of research objects synthesized in its four fundamental principles, dimensioned by its fifteen guiding principles.

In this complex ecology, the model sought to deconstruct the building blocks that make up a generic architecture to reach a level of FARIfication that allows the achievement of the desired IFDS by articulating the various conceptual modules– guidelines, policies, services, tools, infrastructures etc.,– in the form of pieces that can be adjusted according to the depth, scope, and philosophy of each institution or discipline, thus providing a possible scale to support the measurement of the maturity level of management services projects.

Even considering the general approach of the model, it is necessary to consider that in the implementation of FAIR practices and infrastructures, the specific context of the scientific communities and the possibilities of adoption must be observed. The importance of each principle may depend on the priorities and maturity of the community and the generation and use of certain research objects. This condition implies that different disciplines find technical solutions and need different infrastructural and organizational frameworks and management services to achieve the degree of FAIRification required by their communities. But it should be noted that although scientific imperatives are different between disciplines – which still present different types of organization and culture -, which makes them seek their solutions and follow particular strategies towards FAIR data, the difficulties, and challenges, as well as facilities, are generally shared, as there is a common core of interest. Furthermore, when the scope of FAIR principles is expanded to include other research objects, it is necessary to consider that many of these objects belong to a specific disciplinary domain, which reinforces the finding that FAIR guidelines and practices are also discipline-specific.

REFERENCES

BALL, A. **Review of data management lifecycle models**. Bath, UK: University of Bath, 2012.

BOECKHOUT, M.; ZIELHUIS, G. A.; BREDENOORD, A. L. The FAIR guiding principles for data stewardship: fair enough? **European Journal of Human Genetics**, v. 26, n. 7, p. 931-936, 2018. Available on: https://www.nature.com/articles/s41431-018-0160-0.pdf. Access on: 25 apr. 2024.

BORGMAN, C. **Scholarship in the Digital Age**: Information, Infrastructure, and the Internet. London: The MIT Press, 2007.

COATES, H. L. Building Data Services from the Ground Up: Strategies and Resources. **Journal of eScience Librarianship,** v. 3, n. 1, 2014. Available on : https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1063&context=jeslib. Access on:25 apr. 2024.

ERWAY, R. *et al*. **Building Blocks:** Laying the Foundation for a Research Data Management Program. Dublin: OCLC, 2016. Available on: https://files.eric.ed.gov/fulltext/ED589141.pdf. Access on: 25 apr. 2024.

GO FAIR. **The internet of FAIR Data & Service**. [20--?a]. Available on: https://www.go-fair.org/resources/internet-fair-data-services/. Access on: 25 apr. 2024.

GO FAIR. **GO FAIR Initiative**. [20--?b]. Available on: https://www.go-fair.org/go-fair-initiative/. Access on: 25 apr. 2024.

GRAAF, M. V. D.; WAAIJERS, L. **A surfboard for riding the wave**: Towards a four country action program on research data. Copenhagen: Knowledge Exchange, 2011.

HONG, N. C. *et al*. **Six recommendation to implementation of FAIR Practices**. Bruxelas: European Commission, 2020. Available on: https://ec.europa.eu/info/publications/six-recommendations-implementation-fair--practice_en. Access on: 25 apr. 2024.

JACOBSEN, A. *et al*. FAIR principles: Interpretations and implementation considerations. **Data Intelligence**, n. 2, p. 10–29, 2020. Available on: emhttp://www.inf.ufes.br/~gguizzardi/102-Annika_Jacobsen-1_GRFHSzW.pdf. Access on: 25 apr. 2024.

KOUPER, I. *et al.* Research Data Services Maturity in Academic Libraries. *In*: JOHNSTON, L. R. (eds.). **Curating Research Data**: Practical Strategies for Your Digital Repository. Chicago: Association of College and Research Libraries, 2017. p. 153-170. Available on: https://experts.illinois.edu/en/publications/research-data-services--maturity-in-academic-libraries. Access on: 25 apr. 2024.

LEONARDI, P. M. Digital materiality? How artifacts without matter, matter. **First Monday**, v. 15, n. 6-7, 2010. Available on: https://journals.uic.edu/ojs/index.php/fm/article/view/3036. Access on: 25 apr. 2024.

MAYERMIK, M. S. *et al*. The data conservancy instance: infrastructure and organizational services for research data curation. **D-Lib Magazine**, v. 18, n. 9-10, Sep./Oct., 2012. Available on: http://www.dlib.org/dlib/september12/mayernik/09mayernik.html. Access on: 25 apr. 2024.

MONS, B. *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, v. 37, n. 1, p. 49-56, 2017.

MUSHI, G. E., PIENAAR, H., VAN DEVENTER, M. Identifying and Implementing Relevant Research Data Management Services for the Library at the University of Dodoma, Tanzania. **Data Science Journal**, v. 19, n. 1, p. 1-9, 2020. Available on: https://datascience.codata.org/articles/10.5334/dsj-2020-001/. Access on: 25 apr. 2024.

NATIONAL RESEARCH COUNCIL. **Preparing the workforce for digital curation**. Washington, D.C.: The National Academies Press, 2015.

PINFIELD, S.; COX, A. M.; SMITH, J. Research data management and libraries: Relationships, activities, drivers and influences. **PLoS One**, v. 9, n. 12, p. e114734, 2014. Available on: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114734. Access on: 25 apr. 2024.

REED, R. B. Diving into data: Planning a research data management event. **Journal of Escience Librarianship**, v. 4, n. 1, 2015. Available on: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517608/. Access on: 25 apr. 2024.

SALES, L. *et al*. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. Available on: https://direct.mit.edu/dint/article/2/1-2/238/10004/GO-FAIR-Brazil-A-Challenge-for-Brazilian-Data. Access on: 25 apr. 2021.

SAYÃO, L. F. Modelos teóricos em ciência da informação-abstração e método científico. **Ciência da informação**, Brasília, v. 30, n. 1, p. 82-91, 2001. Available on: https://revista.ibict.br/ciinf/article/view/941. Access on: 25 apr. 2024.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa? **BIBLOS**, v. 34, n. 2, 2020. Available on: https://www.seer.furg.br/biblos/article/view/11875. Access on: 12 apr. 2024.

SOLOMONIDES, A. Research Data Governance, Roles, and Infrastructure. *In*: RICHESSON, R.; ANDREWS, J. (eds.). **Clinical Research Informatics**. Cham: Springer, 2019. p. 291-310.

STRASSER, C. **Research data management**. Baltimore: NISO, 2015. Available on: https://wiki.lib.sun.ac.za/images/2/24/PrimerRDM-2015-0727.pdf. Access on: 25 apr. 2024.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016. Available on: https://www.nature.com/articles/sdata201618.pdf. Access on: 25 apr. 2024.

WILSON, J. A. J. *et al*. An institutional approach to developing research data management infrastructure. **The International Journal of Digital Curation**, v. 6, n. 2, 2011. Available on: http://ijdc.net/index.php/ijdc/article/view/198. Access on: 25 apr. 2024.

How to cite this chapter: SAYÃO, Luís Fernando; SALES, Luana Farias. An implementation model for the internet of FAIR data & services. *In*: SALES, Luana Faria; VEIGA, Viviane Santos de Oliveira; VIDOTTI, Silvana Aparecida Borsetti Gregório; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **FAIR Principles Applied To Research Data Management**: Brazilian Experiences. Brasília, DF: Editora Ibict, 2024. cap. 15, p. 177-194. DOI: 10.22477/9788570131959.cap15.