10. DATA INTEROPERABILITY AND THE INFORMATION TRANSDUCTION ENCAPSULATED IN DATA ACCESS

Ricardo César Gonçalves Sant'Ana[139]

10.1 INTRODUCTION

Reducing the distance and barriers between data and informational needs is the challenge herein. And when the data access scenario is considered, this issue also includes the location, interpretation, and use not only by individuals, but mainly by machines.

This demand implies turning these data into readable content, in situations and moments distinct from those in which They were created, generating the need to adopt standards and, consequently, principles and guidelines that, When shared, allow for the reuse of data collection.

Environments with high demand for data access, such as corporative environments or even environments related to public management, are composed of hundreds or even millions of systems, developed either internally or externally, setting a complex and diversified scenario. Such systems require integration of their data so that their effective use is enabled. (Reeve, 2013; Sant'ana, 2009; Dyché; Levy, 2006). However, much of the focus on data management is applied to the processes of collection, storage, and availability in the systems to the detriment of data flow among the different structures (Reeve, 2013). This space among systems tends to increase its complexity exponentially with the increase in data sources considered in the environment, and it is usually served by data interfaces (resources developed to enable data flow among systems).

Thus, the search for data availability in the right place, in the right format, and adherent to the informational needs intended to be met gains increasing relevance, which places data integration as a central condition to the success of data access (Kelleher; Tierney, 2018). This data integration implies, in its turn, in processes of transforming these data in an increasingly automated way so that They can be treated as a single set, generating the need of informational transduction (Sant'ana, 2019), which in its turn, demand definitions that derive from both technical and context knowledge (Reeve, 2013; Shkedi, 2019). In the dimension of knowledge about the context, lines of confrontation to integration challenges emerge, such as those that consider, for instance, the use of ontologies; this issue is beyond the scope of this text, though.

Considering the scenario of multiple systems, for multiple instances, such as in the case of sharing data among academia, industries, funding agencies and academic publishers, it is possible to predict the high level of com-

139 Associate Professor in Management Information Systems, Associate Professor at the São Paulo State University - UNESP, ricardo.santana@unesp.br

plexity that is presented. Seeking the development of shared means of expanding the reuse of data collections for situations such like this, representatives of these bodies met in a workshop held in Leiden, Holland, in 2014, called 'Joint Designing a Data Fairport' to "design and endorse, by mutual agreement, a concise and measurable set of principles named FAIR Data Principles." (Wilkinson *et al.*, 2016), denomination resulting from the concepts: Findable, Accessible, Interoperable and Reusable. A differential proposed by FAIR Principles guidelines is in the fact that they "emphasize specifically the improvement of machines capacity to automatically find and use data, besides supporting their reuse by individuals"; it is also noteworthy the intention that the guidelines are applicable to "algorithms, tools and workflows that led to these data" (Wilkinson *et al.*, 2016).

FAIR Principles, after adjustments and improvements, are presented with four principles, each one of them with their criteria: four related to Findable concept, two related to Accessible, one related to Reusable and three related to Interoperable. The latter, which is the focus of this chapter, are the following: (Wilkinson *et al.*, 2016):

I1. Metadata with formal, accessible, shared and widely applicable language for the knowledge representation;

I2. Metadata with vocabulary following FAIR principles;

I3. Metadata includes references qualified to other metadata.

These are generic principles, but they point to guidelines that can help increase the interoperability potential of datasets. Aspects such as the need for formalism, that is, meeting pre-established standards, can help to minimize divergent definitions of the same content. The 'accessible' concept, still provided in I1 principle, can also support the need of these standards and rules to be shared among those involved, and that they are also the target of dissemination strategies for their understanding, elements that lead to the concept, also part of I1 principle, 'shared'. All these factors could not be relevant if the feasibility of such definitions were not considered, which leads to the 'applicable' concept that completes I1 principle.

I2 principle aims at the issue of vocabulary used in metadata defined for datasets, recursively pointing to the Other FAIR principles. This issue aims at the issue of vocabularies used in metadata defined for datasets, recursively pointing to Other FAIR principles. It is a quite broad issue and requires remembering that vocabularies cannot always meet the specific needs, which can lead to publications of extensions of existing vocabularies or even the creation of new vocabularies (FORCE11, c2021), which is still a complicating factor.

I3 principle indicates the need of qualified references among metadata (Wilkinson *et al.*, 2016), which points to the need that machine resources can perform operations directly on data collected, which in its turn requires that metadata "must be syntactically parsable and semantically accessible by machine" (FORCE11, c2021). FORCE11 also points out that the syntax and 'semantic' of data models and formats used for metadata must be "easy to be identified and used, analyzed or translated by machines", and this is one of the guiding elements for the argumentation presented in this text

In this same line of analysis, there is a flexibilization provided in the proposal of FAIR Principles through the possibility of definitions emerging in a bottom-up movement,

> if a provider can prove that an alternative data model/format is unequivocally parsable as one of the FAIR formats adopted by the community, there is no particular reason such a format cannot be considered FAIR. Some types of data may simply not be 'capturable' in one of the existing formats and, in this case, maybe only part of the data elements can be analyzed (FORCE11, c2021)

Such flexibilization would increase the potential for adherence to the specificities inherent to the large number of situations and contexts in which data originate, while at the same time carrying with it the complexity from which the motivation for the proposal for FAIR principles originated. This effect is even foreseen by FORCE11 itself When it proposes that: "the ideal situation is restricted FAIR data release to the minimal possible of formats and standards adopted by the community", considering that it would be necessary to offer solutions to the new demands: "FAIRports will offer more and more guidance and assistance in these cases" (FORCE11, c2021).

Even considering that it is not a pre-requirement for determining the data adherence to FAIR Principles (Wilkinson *et al.*, 2016), machine access must be sought with the greatest autonomy possible – access and interpretation to such a point that it is possible to transform data collected in a new dataset more adherent to each need (Reeve, 2013).There is no way not to consider minimum levels of machine processing as a condition without which the data access process would not be able to cope with the data to which we are submitted, or as provided by FORCE11 (c2021), when it states that "providing machine-readable data as the main substrate for knowledge discovery [...] that works smoothly and sustainably is one of eScience greatest challenges".

But Where do such various informational elements necessary so that data can be properly collected and transformed for use come from? Much of it comes from their own fragmented essence, resulting from the necessary structure, native or obtained after treatment, but Always a requirement so that the algorithms can establish in a univocal and detailed way, step by step, what the machine must do in the processing of contents.

10.2 THE FRAGMENTED NATURE OF DATA AND FAIR PRINCIPLES

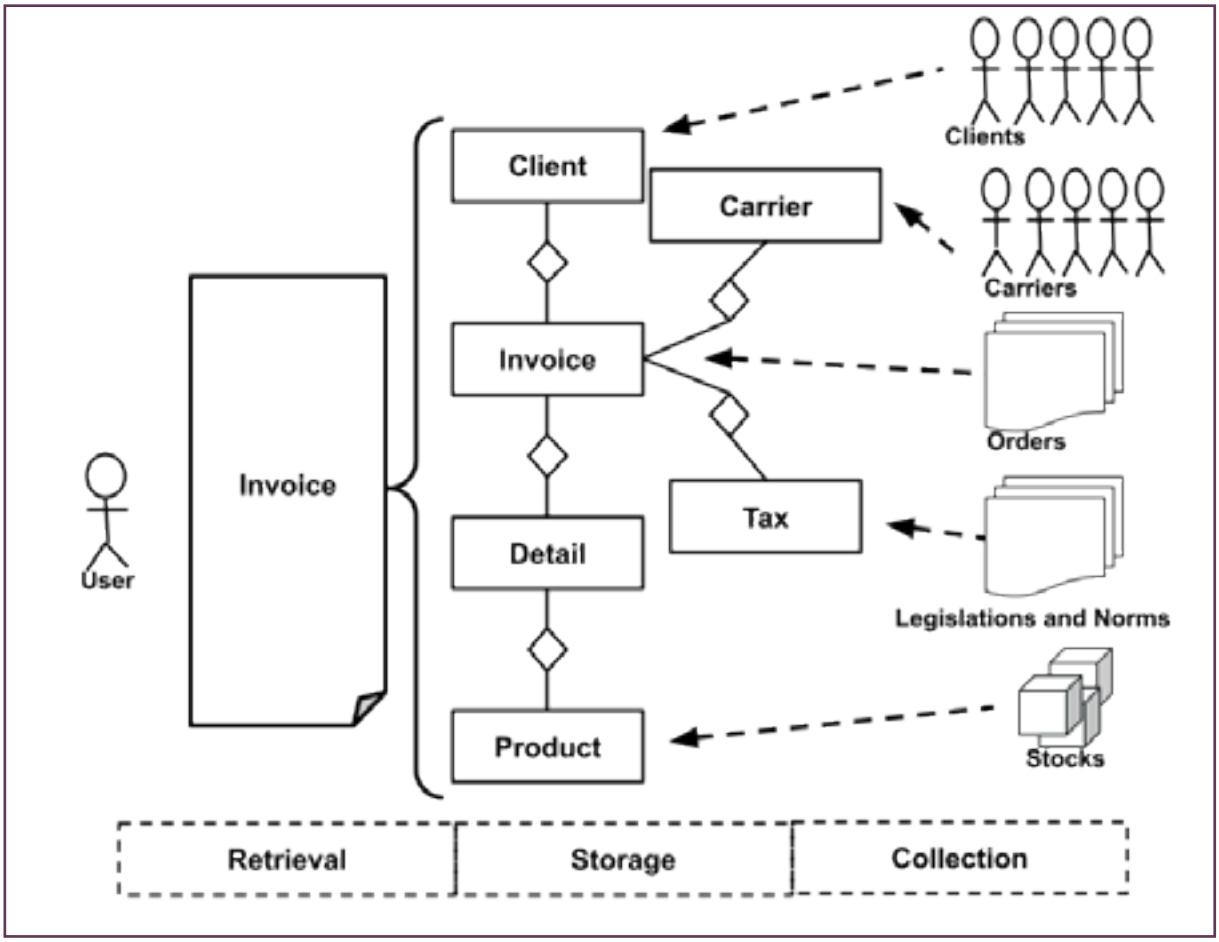
By their nature, machine resources are only capable of executing absolutely detailed and precise instructions, hence the need to establish, exactly and formally, what should be done with each informational particle of the contents to be treated. This process algorithmizing leads to a necessary and inevitable fragmented nature of data, so that new informational layers can be established with metadata that support syntactic and semantic elements to each of the fragments, thus composing a minimal structure of meaning – the triad: entity, attribute and value <e,a,v> (Santos; Sant'ana, 2015) – which, in turn, provides the viability of algorithmizing of machine treatment of contents.

This fragmentation generates, in the phase of collection (Sant'ana, 2016), the need to allocate specific values, related to each one of the transactions and registered facts, in specific 'attributes', which in turn, will be linked to 'entities' related to each one of the information identified as relevant. This link emerges from logical mapping of these entities, respecting principles such as those related to data normalization, avoiding redundancies and bringing coherence to datasets.

Thus, if we consider, for example, a document referring to a sales transaction, an invoice (Figure 1), we will have all the data trajectory, from the collection phase (Sant'ana, 2016), with the identification of information about the client, order details, products, and quantities involved, carrier responsible for delivery, classifications and tax

calculation, among other information. This information is then recorded, persisted, in the respective semantic structures (entities) with its respective label (attributes), and related to each other in such a way that it is possible for its visualization as a document.

Figure 1 – Fragmented data structure



Source: Designed by author.

From this document, which is available in the retrieval phase, it is possible for the user to visualize the transaction data, therefore, the fact. In addition, it is this result from the composition of respective data from each entity (dataset) available as a document that will be shared with the others involved, such as the client himself, tax authorities, carrier, and the Other systems of the organization itself, such as the financial system, accounting system, stock, among others.

From this fragmentation, two points of reflection emerge in this text: the informational transduction and the encapsulated complexity.

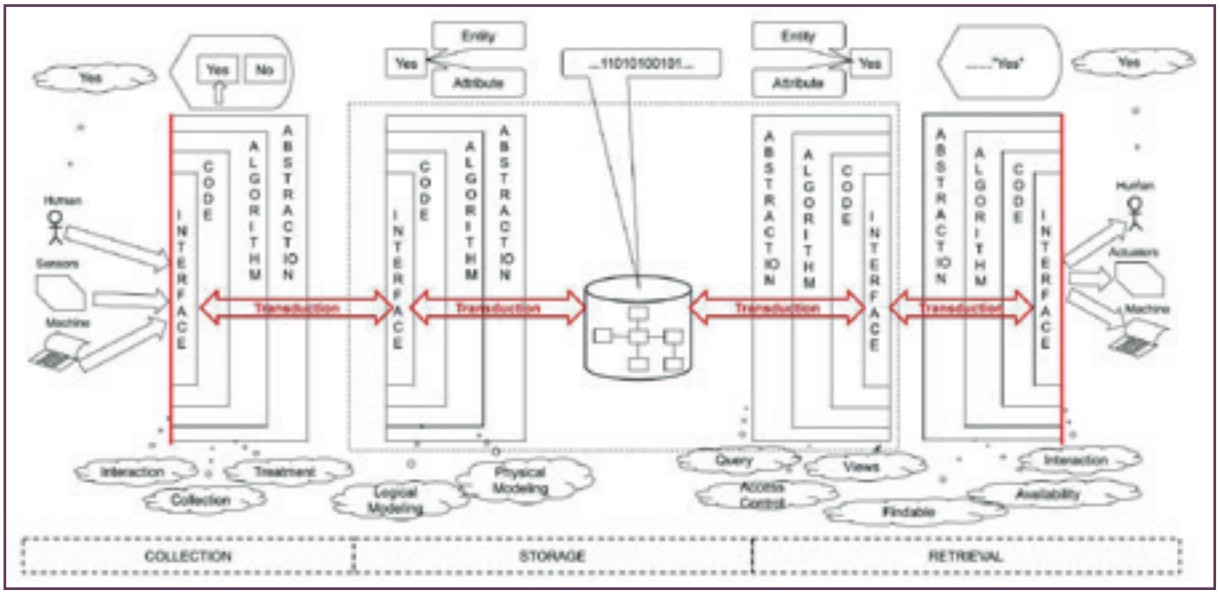
The contents, flowing between source and user, collected, stored and available for retrieval in different data life cycles (Sant'ana, 2016), undergo changes both in energy and format scope and even the content itself. These informational transductions allow for the user's informational needs to be met, meeting the demands through greater adherence to the use context, such as personalization, adequacy, adaptation, resulting in the lowest cost possible when accessing data.

Even though it is fully dominated by those who work in the abstraction layers closest to the machine-analysts, developers and administrators, whether in the programming (software) dimension or in the data dimension – most users do not have, and could not even have, the perception of this fragmented structure to which data are submitted so that they can be used, as such complexity would make the computational resources unfeasible.

10.2.1 Informational Transduction

Data path is long and complex, starting from its apprehension from the fact, going through the several transformations imposed by the interfaces and adjustments to the models of data structure, reaching its records in the digital supports so that, finally, They can be recomposed in a document format. (figure 1). Each one of these changes (Figure 2) implies conversions not only in format and content, but also in energy. Such changes, herein defined as Informational Transductions (Sant'ana, 2019) are necessary so that mediating mechanisms can treat the contents; however, They increase the complexity of the process, making its understanding more difficult, which would make it impossible to use the systems involved

Figure 2 – Informational Transduction in Data Access Source: Designed by the author.



Source: Designed by author.

This eventual barrier is overcome by hiding non-essential details from users, which originates the second point of reflection in this text: complexity encapsulation.

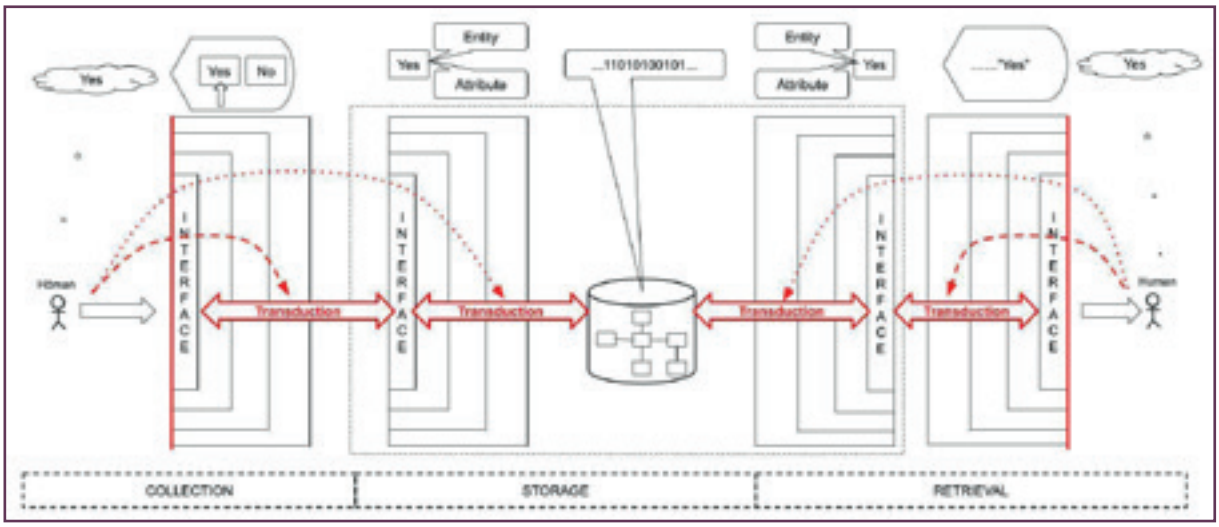
10.2.2 Complexity encapsulation

One of the main factors for a system to be viable in its use is the learning curve required from its users. A very illustrative example is the adhesion to the Internet, which in its first years of access to the public, presented a great complexity of use, requiring simple access to a certain content to use complex commands, true lines of code, with information on address and operation, for example, always with a high level of syntactic formalism. This model kept the big public at bay, and only 'beginners' in technology took the risk of using it. This barrier was broken with the proposal of the Web model, which concealed such developments necessary to the operations,

through graphic interfaces that allowed a simple user action, in the recent and so far, little used mouse, to be converted 'internally' in complex commands performed by the machine.

This same concealment process occurs in all spheres of technology use and, thus, this complexity encapsulation (Figure 3), allows the user to concentrate only on elements strictly necessary for his interaction with the systems. If we return to the example presented in Figure 1, we can infer that the responsible for data input in the system focuses on data content, without realizing the fragmented way in which they are converted in entities, and even less how they will be physically treated in digital supports. Information such as the hard disk area where the information will be recorded or how the device memory will treat these contents, or even how different programs will interoperate, such as the Invoicing System and the Database Management System (Figure 3).

Figure 3 – User inability on Informational Transduction processes



Source: Designed by the author.

This process has advanced so much that today we have the feasibility of direct customer interaction with the interface systems of companies, the e-commerce, which combines the ease of use and ubiquity of the Internet, linked to the evolution of interface systems, exclude the participation of those who until then were responsible for data selling and input in the systems.

Complexity encapsulation leads to a perception of content, treated by systems, totally based on data visualization, under the form of documents or reports, always aimed at searching for adherence to informational need, concealing, therefore, the structures of entities used and the linking ways that allow relationships among data entities. This user's inability to informational transductions keeps him away from eventual actions and definitions necessary so that the interoperability is viable.

10.3 REFLECTIONS ON ANALYSIS AND FUTURE DEVELOPMENTS

As part of FAIR principles, the increase in the interoperability, so relevant for full data access, is deeply affected by the inability of individuals involved in the process of several physical and logic transductions among the data collection, storage, and retrieval phases.

It is reinforced here the difficult perception, for those involved, of the structurally fragmented nature of data and the need to group them, which in turn, bring their features to layers of abstraction that incorporate semantics to these data and allow, finally, their interpretation.

Logical mappings, in their subsequent layers, incorporate data on data (metadata) that need, among other purposes, to allow the interpretation by humans, and increasingly, by machine treatments, which leads to the need for sharing physical, logical and semantic-complex standards.

Data collected in their respective environments receive treatment and are prepared to be stored, predicting, in most cases, their use in the context established at the moment of collection. However, their use tends to be increasingly disseminated, and it is necessary that these semantic layers, aggregated to the data, can be used by unforeseen or inexistent contexts in the moment of collection and storage. On the other hand, factors such as those related to possible limitations to these data also require that these openings, for unforeseen uses, are explicit, not only to the holders of the resources involved in the data life cycle, but also to users and eventual referenced by these data.

Information Science can, and must, participate in the process of identifying factors involved in informational transductions' encapsulation necessary to the process of data, and collaborate, not only in the search for improvement and increase in the potential for integrating informational content in data, but also, and mainly, contribute for the dissemination, in the Society, of the potential of data use, which once aggregated and meeting interoperability requirements, can represent a higher value than those represented by sets when individually considered.

REFERENCES

DYCHÉ, Jill; LEVY, Evan. **Customer Data Integration**: reaching a single version of the truth. Hoboken, New Jersey: John Wiley & Sons, 2006.

FORCE11. **Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing:** version b1.0. c2021. Available from: https://www.force11.org/fairprinciples. Access on: 07 oct. 2024..

KELLEHER, John D.; TIERNEY, Brendan. **Data Science**. Cambridge: The MIT Press, 2018.

REEVE, April. **Managing Data in Motion**: Data Integration best practice techniques and technologies. Waltham, EUA: Elsevier, 2013.

SANT'ANA, Ricardo César Gonçalves. **Tecnologia e gestão pública municipal**. São Paulo: Cultura Acadêmica, 2009. (Coleção PROPG Digital - UNESP). Available from: http://hdl.handle.net/11449/109104. Access on: 07 oct. 2024.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116–142, dec. 2016. Available from: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940. Access on: 29 dec. 2016.

SANT'ANA, Ricardo César Gonçalves. Transdução Informacional: impactos do controle sobre os dados. *In*: MARTÍNEZ-ÁVILA, D; SOUZA, E. A.; GONZALEZ, M. E. Q. (orgs). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução? Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2019. p.117-128.

SANTOS, Plácida L. V. Amorim da Costa; SANT'ANA, Ricardo César Gonçalves. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, p. 199-209, 2015. Available from: https://revista.ibict.br/ciinf/article/view/1382. Access on: 07 oct. 2024.

SHKEDI, Asher. **Introduction to Data Analysis in Qualitative Research**: practical and theoretical methodologies with use of a software tool. [*S. l.: s. n.*], 2019.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, [*s. l.*], v. 3, n. 1, p. 1-9, 2016. Available from: https://doi.org/10.1038/sdata.2016.18. Access on: 15 aug. 2020.

How to cite this chapter: SANT´ANA, Ricardo César Golçalves. Data interoperability and the information transduction encapsulated in data access. *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. 10, p. 125-132. DOI: 10.22477/9788570131959.cap10.