The Canadian Journal of Information and Library Science
La Revue canadienne des sciences de l'information et de bibliothéconomie

Ontologies and Research Data
A Theoretical and Methodological Overview

Caliel Cardoso de Oliveira and Thiago Henrique Bragato Barros

Volume 47, Number 2, 2024

Bobcatsss 2024 Special Issue
Numéro spécial Bobcatsss 2024

URI: https://id.erudit.org/iderudit/1115991ar
DOI: https://doi.org/10.5206/cjils-rcsib.v47i2.17714

See table of contents

Publisher(s)

Canadian Association for Information Science - Association canadienne des sciences de l'information

ISSN

1195-096X (print)
1920-7239 (digital)

Explore this journal

Cite this article

Cardoso de Oliveira, C. & Bragato Barros, T. (2024). Ontologies and Research Data: A Theoretical and Methodological Overview. *The Canadian Journal of Information and Library Science / La Revue canadienne des sciences de l'information et de bibliothéconomie*, *47*(2), 30–38. https://doi.org/10.5206/cjils-rcsib.v47i2.17714

Article abstract

This study examines the relationship between ontologies and open research data within the framework of the Semantic Web. The objective is to investigate the interconnectivity of both subjects through scholarly works that offer enhancements to the Web ecosystem and scientific research processes. The paper reviews critical theoretical frameworks related to the Semantic Web and the significance of metadata within this model. It also delves into the function that ontologies can fulfill in the Semantic Web landscape. The paper provides a historical overview of the emergence of ontologies and definitions in both Computer Science and Information Science. It chronicles how theorists in Information Science have progressively embraced the concept of ontologies since the late 20th century and assesses the current scholarly consensus on the subject. The study also addresses the importance of open research data in modern science by doing a systematic literature review; this study sources relevant publications from the Web of Science and Scopus databases, with a temporal focus from 2000 to 2023. The findings offer a comprehensive analysis of existing literature that bridges the two domains above, aiding in the theoretical and methodological systematization of the subject matter. The discussion section elaborates on the findings, offering insights into the evolutionary trajectory of the subject matter. Emphasis is placed on the utility of ontologies as tools for the sustainable and effective utilization of research data, accentuating the value of such data as a basis for future scholarly work. In conclusion, we advocate for information science to take a leading role in initiatives that leverage ontological frameworks to manage specialized knowledge in research data sets effectively, ensuring that such data remains an asset for advancing scientific understanding.

The Canadian journal of information and library science - La Revue canadienne des sciences de l'information et de bibliothéconomie (CJILS-RCSIB)
Vol. 47, No. 2 (2024) - Bobcatsss 2024 conference special issue
DOI: 10.5206/cjils-rcsib.v47i2.17714
Conference paper

Ontologies and Research Data: A Theoretical and Methodological Overview

Caliel Cardoso de Oliveira [1] and Thiago Henrique Bragato Barros [2]
[1]ORCALAB, Porto Alegre, Brazil
[2]UFRGS/UFSC, Porto Alegre, Brazil

This study examines the relationship between ontologies and open research data within the framework of the Semantic Web. The objective is to investigate the interconnectivity of both subjects through scholarly works that offer enhancements to the Web ecosystem and scientific research processes. The paper reviews critical theoretical frameworks related to the Semantic Web and the significance of metadata within this model. It also delves into the function that ontologies can fulfill in the Semantic Web landscape. The paper provides a historical overview of the emergence of ontologies and definitions in both Computer Science and Information Science. It chronicles how theorists in Information Science have progressively embraced the concept of ontologies since the late 20th century and assesses the current scholarly consensus on the subject. The study also addresses the importance of open research data in modern science by doing a systematic literature review; this study sources relevant publications from the Web of Science and Scopus databases, with a temporal focus from 2000 to 2023. The findings offer a comprehensive analysis of existing literature that bridges the two domains above, aiding in the theoretical and methodological systematization of the subject matter. The discussion section elaborates on the findings, offering insights into the evolutionary trajectory of the subject matter. Emphasis is placed on the utility of ontologies as tools for the sustainable and effective utilization of research data, accentuating the value of such data as a basis for future scholarly work. In conclusion, we advocate for information science to take a leading role in initiatives that leverage ontological frameworks to manage specialized knowledge in research data sets effectively, ensuring that such data remains an asset for advancing scientific understanding

Keywords: ontology, open research data, semantic web

Introduction

Ontologies have been adopted by Information Science as a research theme since the 1990s in the context of Knowledge Organization, in which they are understood as artifacts that translate the terminology of specialized knowledge and of the Semantic Web. Before their adoption by Information Science, within the context of Computer Science, ontologies could be understood as software artifacts that offer a conceptualization of the world, modelling a given part of the reality based on axiomatic constraints that stipulate what piece of knowledge about that domain would be considered information science as being true (Vickery, 1997). With these constraints, a system operating based on that represented knowledge could discern some basic statements about the domain it is dealing with, including inferring new knowledge from what was stipulated by the axiomatic constraints.

This inferential capacity would assist in the functionalities of automated systems, facilitating the acquisition of new knowledge and speeding up operations that could be delegated to automated applications capable of interacting with ontological artifacts. The application of Ontology and the construction of ontological artifacts in Computer Science had their importance ratified with the proposal of the Semantic Web model, which advocated the importance of building web environments enriched with useful content, allowing automated applications to reach conclusions about a given domain of reality and optimize information retrieval on the Web (Berners-Lee, Hendler, & Lassila, 2001).

The combination of constant computational advances and the theoretical-methodological framework of Information Science opened the way for developing new approaches using ontologies. The results of this union were innovations in the creation processes of Knowledge Organization Systems (KOS) and the organization of specialized knowledge domains in a structured and accessible manner. In general terms, the collaboration between Information Science and

Correspondence concerning this article should be addressed to Caliel Cardoso de Oliveira: caliel.co@gmail.com

Computing encouraged academics to develop ontological artifacts that combine traditional knowledge organization practices with the dynamics of emerging digital environments. These environments are characterized by formal structuring and logical languages, allowing for a precise representation of knowledge (Almeida, 2020).

In both Computing and Information Sciences, the essence of constructing ontological artifacts is the structured and formal representation of knowledge, making it comprehensible to machines. In Knowledge Representation aimed at AIs, the focus is on the formal representation of specific knowledge necessary for the operations of automated systems. In Information Science, specialized knowledge, built on the consensus of experts, is the basis for the development of ontologies, which, like other KOS, are crucial for the discipline of Knowledge Organization.

The main application of Ontology, both in Computing and Information Science, focuses on the representation of finalized knowledge and not on research data, which is the basis for specialized knowledge. Given the growing relevance of research data, this research aims to systematize the relationship between ontologies and research data repositories, addressing both theoretical and methodological aspects. The adopted methodology is the bibliographic review, covering a decade of literature in the Scopus and Web of Science databases, as detailed in this work's Analysis and Results section. With this study, we seek to highlight the intersections between Applied Ontology and Open Research Data in Information Science to identify new contributions that can advance the field of Knowledge Organization.

Open Linked Data and Ontologies

Although there is some proximity in how ontological artifacts are defined in Computing and Information Science, their understanding in both areas remains distinct. Initially, however, it is appropriate to contextualize the framework of the Semantic Web and its implications for the development of ontologies. Proposed in 2001 by a group of prominent Computer Science theorists, the Semantic Web model proposed a virtual scenario where each node of the network of web pages and digital resources was semantically enriched, bringing not just metadata but also content loaded with meaning and readable by automated systems (Berners-Lee; Hendler; Lassila, 2001). Although they have slightly different definitions when addressed in Computer or Information Sciences, metadata can generally be understood as "structured information that describes, [...], or otherwise makes it easier to retrieve, use, or manage an informational resource" (NISO, 2004, p.4). In this sense, metadata is traditionally used in web content tagging to facilitate interoperability between different systems, containing information that makes using those resources easier for human agents or automated applications (Mayernik, 2020), such as data on provenance, authorship and uniform identification.

Thus, the Semantic Web model proposed, beyond the use of metadata for aggregating value in web pages, integration of an additional layer of meaning in these pages to represent their content via discrete and stable web resources, formally defined in a software artifact readable by automated systems. This artifact would be an ontology containing controlled vocabulary and formal definitions for the page content and the systematization of the web of relationships, which are part of the elements that make up that content. With the integration of ontologies, the expectation was that the Semantic Web environment would facilitate information retrieval by making use of inferences that systems would be capable of operating once they could interpret the web page content, which would be represented in a formal and uniform language (Berners-Lee; Hendler; Lassila, 2001). In this sense, it is understood that the importance of ontologies in the context of the Semantic Web lies in their role as world-restricting artifacts and formal content representation, allowing automated applications to obtain more precise results in information retrieval processes, integrating into their search operations the networks of relationships established between different web resource nodes and defined in ontological artifacts (Silva, Martins, & Siqueira, 2018).

The Open Linked Data model is related to the Semantic Web framework. A semantically enriched digital environment depends on an infrastructure capable of encompassing the different elements that allow the representation of meaning in formal language. As explained earlier, the representation of meaning depends on restricting the content on a web page to different uniform and stable resources, which can be integrated into the networks of relationships provided by ontologies and consulted, considering their role in these networks. Tim Berners-Lee proposed some web architecture principles that should be considered during the process of structuring a web page; these four principles, known as the Principles of Linked Data, according to Heath and Bizer (2011), consist of:

1. the use of Uniform Resource Identifiers (URIs) to identify web resources;

2. the use of these URIs under the Hyper Text Transfer Protocol Secure (HTTPS), which allows the search for these identifiers;

3. the integration of helpful information during the search for a URI through the use of metadata standards such as Resource Description Framework (RDF) and query protocols like the Protocol and RDF Query Language (SPARQL);

4. Links to other URIs should be included so the user can find more content of their interest based on the relationship resources formalized in an ontological artifact.

As touched upon elsewhere in this study, ontology artifacts, much like metadata, also have distinct, albeit close, definitions in both Computer and Information Sciences. In Computer Science, the view shaped by Knowledge Representation prevails, with the understanding being that ontologies restrict the universe being modelled in the digital environment, serving as a comprehensive list of all entities that exist in that universe. For example, Sowa (2001) states that an ontology is a formal catalogue of types of things that are supposed to exist in a given domain from the perspective of someone using a specific language to refer to this domain. Correspondingly, Maurício de Almeida (2020) defines ontologies in Computing as formal representational artifacts composed of a taxonomy of entities and relationships. Ding and Foo (2002) refer to ontologies as domain application reference models, which aim to improve the consistency and reusability of information and interoperability between systems and knowledge sharing. The possibilities of restricting a given universe/domain and, therefore, defining what would be actual knowledge for the system were the main themes in the development of ontologies within Computer Science, starting from the 1980s. The combination of a conceptual model representing the given universe and a formal logical language, as well as the integration of ontologies, would result in automated systems capable of considering the semantics of the domain models when performing inference operations (Sowa, 2001; Almeida, 2020). When built rigorously, based on the knowledge of experts in the domain being represented and with attention to translating that knowledge into a logical language, ontologies serve as knowledge restrictors, increasing the precision of computational systems that operate by drawing information from that domain. However, despite the existing potential of ontologies in Computer Science, especially considering the proposals of the Semantic Web model, Coneglian and Santarém Segundo (2022) point out that ontologies as a research subject had lost space, already by the 2010s to other AI construction proposals, such as the use of natural language in developing chatbots or virtual assistants.

In Information Science, ontologies are usually associated with other Knowledge Organization Systems (KOS), such as taxonomies and thesauri. Vickery (1997) describes ontologies as a method of organization that uses semantic categories to represent the significant concepts of a given domain, including definitions of each concept and explaining the relationships between the concepts. Almeida (2020) further states that the application of ontologies is related to two subareas of Information Science: Theory of Classification and Information Retrieval. In the first case, Information scientists have always used Applied Ontology for classification activities, cataloguing, and the general activities that comprise Knowledge Organization. As for Information Retrieval, the author points out that technological advances that led to the migration of recorded knowledge from physical media to digital media led to the emergence of ontologies as a solution to the issue of translating controlled vocabulary into a formal language, understandable by machines and usable in the virtual medium (Almeida, 2020). Thus, in the context of Information Science, ontologies are characterized as content representation artifacts, representing concepts in a specific domain and aiming to reduce ambiguity and facilitate information retrieval (Santos, Corrêa, & Lapa, 2013; Santos, 2014).

Open Research Data

The importance of primary data produced during scientific efforts began to stand out with the advancement of the Internet and the increasing possibilities for connectivity and dialogue among researchers worldwide relating to the context of e-science (Sayão & Sales, 2016). Whereas, in the past, certain aspects of data (such as its production cycle, iterations, and lineages) were relegated to the background of the research that originated them, this perspective began to change with e-science. It culminated in the movement for open scientific data (or "open research data"), which advocates that specific data should be made publicly available for free, without copyright, patent, or other control mechanisms (Rodrigues et al., 2010; Sayão & Sales, 2016).

Different factors support initiatives of this kind. First, one must consider the universal, far-reaching nature of Science: under the assumption that scientific knowledge belongs to Humanity as a whole, research data are of interest to the general population, which has the right to access them easily, especially considering that, in many cases, it is a publicly-funded governmental budget that maintains scientific research (Murray-Rust, 2008). Additionally, the issue of Big Data must be considered, i.e., the enormous volumes of data that are constantly produced thanks to technological advances on the Web: without adequate treatment, preservation, and availability, large volumes of scientific data with intrinsic value can get lost amid the immensity of content that emerges at all times, wasting any potential use these data packages may have (Sayão & Sales, 2016).

This intrinsic value of research data for the scientific community is the main argument in favour of their openness: data reuse increases exponentially, both in scientific and non-academic activities (Downs, 2021), for two reasons. The first is due to the function of data packages as tools for verification and validation of results, which is facilitated when primary data are available for the repetition of experiments (ALPSP; STM, 2006; Murray-Rust, 2008). The second reason is that, with the normalization of reuse, the flow of scientific production transforms so that the availability of data can lead to new efforts to expand the original research, to the study of phenomena that escaped the scope of the original research if conducted by scientists from another area, and even to the development of products and resources that use these data,

which could not exist if there was no access to them (Downs, 2021). In this way, Open Data practices positively affect research activities as a whole, as the value of research data is linked to its potential for reuse and reinterpretation in areas other than where it was generated, combined with the context of e-science and the increasing interconnectivity between different academic communities, the openness of research data allows the advancement of scientific knowledge and the exchange between different communities beyond geographical barriers (Sayão & Sales, 2016).

The means for sharing data packages, whether scientific or not, are manifest through creating and managing data repositories, which require legal and intellectual protection directives regarding the ownership of the data deposited there (ALPSP; STM, 2006). The principles that govern the dynamics of availability are the so-called FAIR Principles (Abadal, 2021), an acronym for Findable, a principle that stipulates that research data and the metadata that describe them must be possible to locate by search tools after their publication; Interoperable, which stipulates that both data and metadata should be described following guidelines of the academic community, favouring their exchange and reuse; Reusable, which stipulates that data and metadata must have easily discernible provenance and usage conditions, aiming at the reuse by other researchers.

The FAIR principles align with the four Principles of Linked Data presented earlier in this paper. Similar to the principles of Linked Data and establishing its dialogue with the theme of Open Data and the Semantic Web, there is another crucial element when discussing the sharing of open research data. It is the 5-star implementation scheme for Open Data[1], a sharing model proposed by Tim Berners-Lee, among others. The scheme classifies Open Data into five levels, according to the format in which they are shared, considering the degree of accessibility and connectivity that the format possesses. The five levels are as follows, according to the 5 Star Open Data (2012) page:

1. One star: data are available on the Web in any format under an open license. It is the simplest way to publish data and allows users to access them, but the data are "trapped" in the document, being inaccessible to applications that try to access them;

2. Two stars: data are made available in a structured manner, such as in an Excel spreadsheet format. Here, the data remain attached to the document but are presented in a structured manner, allowing reading by applications that can access them;

3. Three stars: data are available in a structured and non-proprietary format, such as in Comma-separated Values (CSV). This level is more complex than the previous ones, as it requires converters to translate the data from the proprietary format to the new format, but in exchange, does not leave access to the data conditioned to the use of specific software;

4. Four stars: usage of URIs for resource identification. In this format, the data become nodes in the large scheme of the Web, but this requires time and resources to represent the data and apply standards to them appropriately. The most commonly used standard for use at this level are RDF triples;

5. Five stars: data are connected with data of other Web users, contextualizing them in the broader network. In this case, some risks exist, such as the lack of maintenance in access links. However, at this level, the linking with other data and the description through standardized metadata add value to the available data.

Thus, the question of Open Data and the theme of Open Science relate to the more significant problem of open and linked data that populate the Semantic Web model.

Dialogues between Ontologies and Research Data

As we pointed out previously, there is a broad field for dialogue between the development and application of ontologies and the treatment of research data for their availability and potential reuse. Differences in formats, different specialties within the same research team, and immense volumes of data are risk factors, but damage to the progress of research and the proper treatment of resulting data can be greatly reduced through the use of ontological artifacts (Gonçalves, 2020). This section of the study presents some examples of initiatives that specifically aim at the application of ontologies in efforts in the representation of data in general.

In the field of Archaeology, Niccolucci (2020) highlights the use of the AO-Cat ontology, derived from the CIDOC Conceptual Reference Model, as a tool for facilitating Interoperability in the ARIADNE web environment, developed by the European Union for the integration of cultural heritage collections. In this specific example, the great diversity of formats in which a package of archaeological research data can manifest is singled out as a key point for inserting an ontological artifact capable of restricting the type of data being worked with. The use of AO-Cat was so successful that the author states that, among the 4 FAIR principles, the principle of Interoperability is the best developed in the field of Archaeology, with emphasis on the permanent dialogue that exists between the various specialists in the area for the construction of a comprehensive and efficient ontological model (Niccolucci, 2020).

An experiment in integrating scientific studies through datasets with a wide range of formats and granularity was

[1] https://5stardata.info/en/

conducted by Gonçalves (2020), achieving positive results in constructing a method for semantic data integration. The author points out that the participation of experts in the construction of ontologies (or "ontologists") in the development of the model contributed to the proper conceptual mapping of the data sets being worked with. Gonçalves also mentions the importance of including domain experts in data integration efforts to verify and point out potentialities in existing domain ontological models; the author also highlights the participation of data scientists, capable of offering clarifications on issues of competencies that exceeded the model's inference capabilities in the face of the represented data. Gonçalves' work used the Hadatac ecosystem and the Human-Aware Science Ontology, or HAScO.

Introduced in 2018, HAScO is a top-level ontology that proposes to identify acquisition and knowledge production processes in scientific studies and encode metadata capable of expressing the data produced during these studies. Starting from the understanding that "science is organized knowledge," HAScO uses top-level ontologies to represent research data and related elements in general while using domain ontologies for more precise modelling of specific scientific fields (Pinheiro et al., 2018). Recognizing that the value of research data lies in facilitating the reproducibility of scientific studies and promoting advances, HAScO proposes to integrate these data from heterogeneous sources with the specific knowledge of the domains, under the perspective of opening and extending the model according to the new needs of the scientific community (Pinheiro et al., 2018).

The examples presented in this section demonstrate that the approximation between the themes of ontologies and open research data has been occurring constantly, with significant projects and efforts in developing systems and methods capable of uniting the two themes. Even though using ontologies offers new challenges to scientists who are little proximate to Knowledge Organization, it greatly eases the efforts to use research data to its full potential. As explained, it is from this use that the value of the data increases exponentially.

Analysis and Results

Based on a search in journal databases, this study made an effort to establish the state of the art regarding the relationship between ontologies and research data. Searches were conducted in the *Scopus* and *Web of Science* databases, which index scientific works from various knowledge areas and formats. For practicality and to observe the evolution of the theme, a time filter was applied, covering works from 2000 to 2023, although no results before 2010 were retrieved in either case. In both databases, a filter was also applied to recover only open-access results, thus facilitating analysis.

In *Web of Science*, the search was initially conducted using expressions like *data reuse*, *ontology**, *dataset*, *research data*, and *semantic Web* in the topic field, along with the previously mentioned filters. This search yielded a low number of works, just 15 results, leading to the removal of the expression semantic Web and the retention of the others, resulting in 34 results.

In *Scopus*, the search was conducted using the same expressions, but the low return on searches led to the breakdown of the compound expressions and the inclusion of the boolean operator AND. Thus, expressions such as *data*, AND *reuse*, AND *ontolog**, AND *dataset*, AND *research*, AND *data* were searched. This search resulted in 36 results.

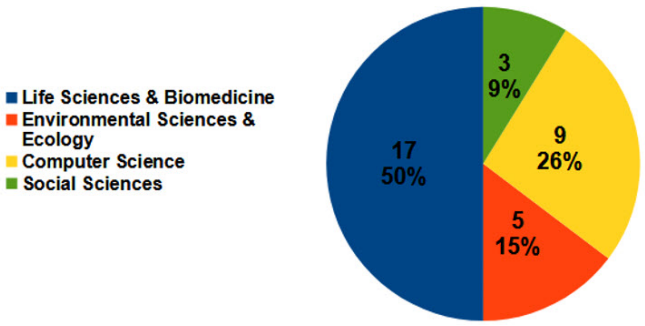
After conducting the searches and quick analysis to discard duplicates, 53 works remained out of the 70 results recovered. Despite having filtered only for open-access results, some results, especially from the *Web of Science* search, turned out to be closed-access works. Excluding these cases, the corpus for analysis consisted of 46 works.

The initial analysis of our corpus involved reading the abstracts of the works and assessing their fit with the research objectives. Works that addressed initiatives where there were points of contact between ontologies and research data and works that, within the context of the Semantic Web or Web of Data, presupposed an interface between the two themes were considered relevant to this research. Works that only addressed one point addressed in this study were excluded from the second analysis stage. Thus, the textual corpus that advanced to the second stage of the work consisted of 34 works. These works were systematically analyzed, and the observations from these analyses are presented in the following section.

It is worth noting that a preponderance of works has been produced in the broad area of health sciences, such as pharmacology and biomedicine, followed by some works in natural sciences. A smaller number of works are divided among the other areas of knowledge, with a notable low incidence of works from the Social Sciences and from Exact Sciences, disregarding approaches of the two within works in the field of Health Sciences. The following graph illustrates the proportion of works by field of knowledge:

Figure 1

Proportion of works analyzed in this study according to their field of knowledge



From a temporal perspective, it was noted that works from the first six years covered by our search presented similar research problems, indicating a lack of effective metadata to promote reuse and interoperability. Works from the start of the 2010s, such as those by Van Assem et al. (2010), Webb and Ma'ayan (2011), and Mena-Gárces et al. (2011), presented attempts to translate raw tabulated or semi-structured data into machine-readable language. In particular, the work of Mena-Gárces et al. (2011) describes an attempt to translate metadata that existed in an ecological markup language into OWL. This language, the Ecological Metadata Language (EML), though widely used in its field, was based on textual descriptions that were difficult for machines to read, and its conversion into OWL would facilitate its reuse and sharing from the use of automated systems. In the Health field, especially Biomedicine, the issue of Big Data was already looming over researchers, considering the volume of data produced during the professional and academic practice of the area. For example, the work of Deus et al. (2012) presents concerns about the standardization of research data in interoperable formats. The authors point out that, at the time of the work, there was a large number of ontologies for the semantic description of data from -omic sciences; however, these ontological artifacts had little coordination among themselves, leading to overlaps and integration problems, creating both redundancies and divergences in the definition of concepts. Tilahun et al. (2014) point out that even though a vast amount of data was available in repositories, many of these datasets did not have formal structuring, which, as previously exposed, is one of the pillars of interoperability. The authors attempted to structure the knowledge in the datasets by applying controlled vocabularies and ontologies, aiming at developing a system that allowed interconnected visualization of biomedical data for end users. Another initiative to build a structured metadata model was made by Musen et al. (2015), who described how the Center for Expanded Data Annotation and Retrieval (CEDAR) helps authors select the best metadata model for their datasets from a bank of preexisting metadata and structured knowledge described in ontologies. Approaching the end of the 2010s, the retrieved works presented more specific dilemmas beyond annotating existing datasets or converting simple metadata into ontological languages. The work of Salguero et al. (2019) deals with creating an ontology capable of enriching data from different sensors present in domestic smart devices. The authors revisit existing ontologies in the area, which gave rise to sets of metadata for the description of data from experiments with smart devices and build an ontology capable of representing temporal information about these events, facilitating the representation and annotation of the data obtained. A similar theme is addressed in the work of Woznowski, Tonkin, and Flach (2018), which describes the functioning of a specific ontology to describe knowledge derived from data obtained by vital signs monitoring sensors in caring and medical facilities. The work exposes the relationship between the construction and evaluation stages of an ontology and the creation of metadata based on it, with its authors estimating that a metadata model for sensor description will be included in future iterations of the project. A work focused on the formal representation of legislative content is presented by Pandit et al. (2018), who introduced the GDPRtEXT, an attempt at ontological modelling of the General Data Protection Regulation (GDPR), a legal instrument aimed at data privacy in the European Union, implemented in 2018. In this project, automated search for the information presented in the GDPR is facilitated by the formal structuring of the knowledge contained therein, as well as by the use of previously-existing metadata for the description of legislative knowledge within the European Union. That demonstrates how structuring knowledge facilitates the reuse of less rigid data and content, typical of the Humanities areas, but whose formal definitions are fundamental for compliance with current legislation. Works in the Health field during this period stand out for similar complaints, such as an excess of repositories and datasets that exist without the support of adequate metadata and ontologies for their standardization and automated management, as well as for the proposed solutions. The work of Legaz-García et al. (2016) points out the risks of heterogeneity regarding the format and the very nature of the data produced in biomedicine, whose storage hardly takes into account the possibilities of automated retrieval of the knowledge contained in the datasets. The authors proposed the creation of a metadata repository suitable for the Semantic Web, developed from the construction of an ontology that systematized the content dispersed in different documents and databases; the description of the approach used by the authors would allow the reproduction of the procedures, facilitating the structuring of knowledge in related Health fields and the application of this structured knowledge to clinical activities. The difficulties of searching produced by heterogeneity in formats are also the object of the work of Rivault, Dameron, and Le Meur (2019), who points out that the lack of interoperability in the management of medical-administrative data can negatively influence decision-making in hospital environments. The lack of data structuring prevents their exploration through SPARQL language, which is suitable for querying in knowledge bases structured in RDF, making it impossible to retrieve relevant information in these data sets. The authors then present queryMed, a library of pre-defined SPARQL queries developed in R format suitable for exploring statistical data. Through queryMed, queries can be carried out without prior knowledge of SPARQL, facilitating users' decision-making and data exploration without specific language training. The functioning of queryMed occurs through integrating biomedical ontologies and using plugins available in specialized databases, which allow external consultation and exploration of the datasets stored there.

The 2020s present a significant increase in the output of works that correlate ontologies and reuse scientific data: Out of the total documents comprising our analysis corpus, 16 works were carried out between 2020 and 2023. The Health field remains the source of most studies, and some problems identified throughout the 2010s are still present in more recent productions.

The work of Celebi et al. (2020) points out that there was low reproducibility potential on data produced in works in the field of Pharmacology, which hampers future experiments in the area, as these are often expensive and involve high-value products. The authors offer a solution by applying ontologies for structuring scientific workflows, creating a unified model for adapting existing scientific workflows to the FAIR principles based on semantic models of Pharmacology and related fields. The challenges and approaches in the biomedical field regarding data interoperability and standardization, as addressed by Pereira et al. (2023) and Queralt-Rosinach et al. (2022), highlight a significant issue in research data management. Pereira et al. emphasize the lack of interoperability among repositories in the biomedical field, largely due to the absence of standardized vocabularies. They suggest that building a structured model based on ontologies, which can map natural language terms into controlled vocabulary, could mitigate this problem and better align with the FAIR (Findable, Accessible, Interoperable, Reusable) principles. This approach would facilitate the translation of diverse terminologies into a unified format, enhancing data accessibility and utility across different repositories.

Queralt-Rosinach et al. (2022) also discuss the low level of interoperability in Healthcare, even among data producers such as university hospitals or national health systems, who theoretically operate within integrated networks. To address this, the authors propose using specialized ontologies and vocabularies to ensure data conforms to the FAIR principles right from the source. They exemplify this with data from the COVID-19 Pandemic in a university hospital setting, demonstrating how using data in its structured form enhances its value as a source of knowledge for future research.

Moreover, Al-Fayez et al. (2023) extend the discussion beyond the biomedical sciences by constructing a domain ontology that structures knowledge about global terrorism movements and acts. This ontology is built by integrating various databases on the subject and structuring the data sets derived from these databases. While their primary concern is less about aligning the data with the FAIR principles and more about formal structuring for representation in the ontological artifact, their efforts in modelling the domain-specific metadata contribute to interoperability and increase the potential for data reuse.

Collectively, these studies underscore the importance of ontologies in enhancing data interoperability and reuse, not only in biomedical sciences but also in other domains like global security. They highlight the crucial role of structured, standardized approaches in data management and the potential of ontologies to bridge gaps between disparate data sources, thereby facilitating more effective and wide-ranging research collaborations.

Conclusion

As we conclude this study, we believe our objective of conducting a comprehensive review connecting the themes of applied ontology and open research data, focusing on the representation and organization of knowledge from an Information Science perspective, has been achieved. Our analysis reveals that while professionals in other fields often employ core competencies of Information Science, the explicit presence of information scientists is not consistently recognized. We found that the theoretical-methodological framework of Information Science, combined with technical knowledge from Computer Science, is used to model specialized knowledge domains, even without a deep understanding of the principles of our field.

Throughout our review, covering fields from biomedical sciences to global terrorism studies and ecology, we observed the broad applicability of ontologies, which confirms the role of Information Science as a post-modern field whose theories and methods can influence all areas of specialized knowledge. We also noticed an iterative aspect in the analyzed works, where ontological artifacts often build on previous versions or adapt to the FAIR Principles, emphasizing the importance of making scientific knowledge available for reuse in new research. This observation suggests an under-explored potential, where a lack of perspective can limit the use of data and, consequently, its value as tools for scientific advancement. Thus, the Information Scientist can play a crucial role in adapting the scientific production of other fields to the FAIR Principles, significantly contributing to the efforts of Knowledge Organization and Representation.

A common challenge identified in our review is the lack of standardization and openness of scientific datasets. Although the importance of data as a substrate for knowledge production is being more recognized in fields such as Health Sciences, Information Science can lead initiatives to standardize metadata and structure knowledge in various fields. Applying the theoretical and methodological framework of Information Science to the treatment, custody, and sharing of data can maximize the effectiveness of these practices and benefit specialized knowledge areas.

We finish this study with the expectation that our review has provided a comprehensive overview of how the interactions between ontology development and the reuse of research data are evolving across various knowledge fields. With its interdisciplinary approach, Information Science is well-positioned to lead new initiatives that promote closer integration between these themes. The practices and conceptions of our field can

benefit research in other fields as much during the research stages, offering clarity for standardizing the products of these investigations by the time the research is completed, aiming to build dataset repositories and apply metadata standards that offer structured meaning to these data.

Thus, research originating from Information Science itself, projects that demand a focus on the organization of specialized knowledge, and platforms that benefit from information treatment and management are three key areas where Information scientists can play a prominent role. In this way, our field will continue to consolidate itself as an important discipline in the diverse and constantly evolving informational context, where the Internet, Science, and society intertwine in the contemporary world.

References

Abadal, E. (2021). Ciencia abierta: Un modelo con piezas por encajar. *Arbor*, *197*(799), Article 799. https://doi.org/10.3989/arbor.2021.799003

Al-Fayez, R. Q., Al-Tawil, M., Abu-Salih, B., & Eyadat, Z. (2023). GTDOnto: An Ontology for Organizing and Modeling Knowledge about Global Terrorism. *Big Data and Cognitive Computing*, *7*(1), 24. https://doi.org/10.3390/bdcc7010024

Almeida, M.B. de (2020). *Ontologia em Ciência da Informação: teoria e método*. (Vol.1). Editora CRV.

Association of Scientific, Technical and Medical Publishers & the Association of Learned and Professional Society Publishers (2006). *Databases, data sets, and data accessibility – views and practices of scholarly publishers*. STM-Assoc. https://www.stm-assoc.org/2006_06_01_STM_ALPSP_Data_Statement.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5). https://doi.org/10.1038/scientificamerican052001-yL7Vw7HIOZ4iSjlnEeVsJ

del Carmen Legaz-García, M., Miñarro-Giménez, J. A., Menárguez-Tortosa, M., & Fernández-Breis, J. T. (2016). Generation of open biomedical datasets through ontology-driven transformation and integration processes. *Journal of Biomedical Semantics*, *7*(1), 32. https://doi.org/10.1186/s13326-016-0075-z

Celebi, R., Rebelo Moreira, J., Hassan, A. A., Ayyar, S., Ridder, L., Kuhn, T., & Dumontier, M. (2020). Towards FAIR protocols and workflows: The OpenPREDICT use case. *PeerJ Computer Science*, *6*, e281. https://doi.org/10.7717/peerj-cs.281

Coneglian, C. S., & Segundo, J. E. S. (2022). Inteligência artificial e ferramentas da Web Semântica aplicadas a recuperação da informação: Um modelo conceitual com foco na linguagem natural. *Informação & Informação*, *27*(1), 625. https://doi.org/10.5433/1981-8920.2022v27n1p625

Deus, H. F., Prud'hommeaux, E., Miller, M., Zhao, J., Malone, J., Adamusiak, T., McCusker, J., Das, S., Rocca Serra, P., Fox, R., & Marshall, M. S. (2012). Translating standards into practice – One Semantic Web API for Gene Expression. *Journal of Biomedical Informatics*, *45*(4), 782–794. https://doi.org/10.1016/j.jbi.2012.03.002

Ding, Y., & Foo, S. (2002). Ontology research and development. Part 1—A review of ontology generation. *Journal of Information Science*, *28*(2), 123–136. https://doi.org/10.1177/016555150202800204

Downs, R. R. (2021). Improving Opportunities for New Value of Open Data: Assessing and Certifying Research Data Repositories. *Data Science Journal*, *20*, 1. https://doi.org/10.5334/dsj-2021-001

Gonçalves, J.E.A. (2020). *Método ágil de integração semântica de dados científicos baseado em ontologias.* 2020. [Doctor's Thesis, Universidade Federal de Minas Gerais]. https://repositorio.ufmg.br/handle/1843/34013

Heath, T.; Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web Theory and Technology*, *11*(2). https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Mayernik, M. (2020). Metadata. In B. Hjørland & C. Gnoli (Eds.). *Encyclopedia of Knowledge Organization. International Society for Knowledge Organisation*. https://www.isko.org/cyclo/metadata

Mena-Garcés, E., García-Barriocanal, E., Sicilia, M.-A., & Sánchez-Alonso, S. (2011). Moving from dataset metadata to semantics in ecological research: A case in translating EML to OWL. *Procedia Computer Science*, *4*, 1622–1630. https://doi.org/10.1016/j.procs.2011.04.175

Murray-Rust, P. (2008). Open Data in Science. *Nature Precedings*. https://doi.org/10.1038/npre.2008.1526.1

Musen, M. A., Bean, C. A., Cheung, K.-H., Dumontier, M., Durante, K. A., Gevaert, O., Gonzalez-Beltran, A., Khatri, P., Kleinstein, S. H., O'Connor, M. J., Pouliot, Y., Rocca-Serra, P., Sansone, S.-A., Wiser, J. A., & and the CEDAR team. (2015). The center for expanded data annotation and retrieval. *Journal of the American Medical Informatics Association*, *22*(6), 1148–1152. https://doi.org/10.1093/jamia/ocv048

National Information Standards Society (2017). *Understanding Metadata: What is metadata, and what is it for?: A primer*. https://www.niso.org/publications/understanding-metadata-2017

Niccolucci, F. (2020). From Digital Archaeology to Data-Centric Archaeological Research. *Magazén*, *1*, JournalArticle_3443. https://doi.org/10.30687/mag//2020/01/002

Pandit, H. J., Fatema, K., O'Sullivan, D., & Lewis, D. (2018). GDPRtEXT - GDPR as a Linked Data Resource. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, & M. Alam (Eds.), *The Semantic Web* (Vol. 10843, pp. 481–495). Springer International Publishing. https://doi.org/10.1007/978-3-319-

93417-4_31

Pereira, A., Almeida, J. R., Lopes, R. P., & Oliveira, J. L. (2023). Querying semantic catalogues of biomedical databases. *Journal of Biomedical Informatics*, *137*, 104272. https://doi.org/10.1016/j.jbi.2022.104272

Pinheiro, P., Bax, M. P., Santos, H., Rashid, S., Liang, Z., Liu, Y., Mccusker, J., Mcguinness, D., & Ne'eman, Y. (2018). *Annotating diverse scientific data with hasco*. https://repositorio.ufmg.br/handle/1843/51523

Queralt-Rosinach, N., Kaliyaperumal, R., Bernabé, C. H., Long, Q., Joosten, S. A., Van Der Wijk, H. J., Flikkenschild, E. L. A., Burger, K., Jacobsen, A., Mons, B., Roos, M., BEAT-COVID Group, & COVID-19 LUMC Group. (2022). Applying the FAIR principles to data in a hospital: Challenges and opportunities in a pandemic. *Journal of Biomedical Semantics*, *13*(1), 12. https://doi.org/10.1186/s13326-022-00263-7

Rivault, Y., Dameron, O., & Le Meur, N. (2019). queryMed: Semantic Web functions for linking pharmacological and medical knowledge to data. *Bioinformatics*, *35*(17), 3203–3205. https://doi.org/10.1093/bioinformatics/btz034

Rodrigues, E., Saraiva, R., Ribeiro, C., Fernandes, E.M. (2010). *Os repositórios de dados científicos: estado da arte.* Grupo de trabalho conjunto da Universidade do Minho e da Universidade do Porto. https://hdl.handle.net/10216/23806

Salguero, A. G., Delatorre, P., Medina, J., Espinilla, M., & Tomeu, A. J. (2019). Ontology-Based Framework for the Automatic Recognition of Activities of Daily Living Using Class Expression Learning Techniques. *Scientific Programming*, *2019*, 1–19. https://doi.org/10.1155/2019/2917294

Santos, M.T. dos. (2014). *Estudo do processo de apropriação da ontologia pela Ciência da Informação no Brasil*. [Master's Thesis, Universidade Federal de Pernambuco]. https://repositorio.ufpe.br/handle/123456789/12945

Santos, M.T. Dos; Corrêa, R. F.; Lapa, R.C (2013). Estudos sobre a apropriação da Ontologia pela Ciência da Informação. In: *Anais do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB),* XIV. Universidade Federal Fluminense. XIV. http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/2333

Sayão, L. F., & Sales, L. F. (2016). Algumas considerações sobre os repositórios digitais de dados de pesquisa. *Informação & Informação*, *21*(2), 90. https://doi.org/10.5433/1981-8920.2016v21n2p90

Silva, M. F., Martins, D. L., & Siqueira, J. (2019). Web semântica em repositórios: Ontologia para representação de bibliotecas digitais. *Ciência da Informação em Revista*, *6*(1), 99-113. https://doi.org/10.28998/cirev.2019v6n1f

Sowa, J.F. (2001). Building, Sharing, and Merging Ontologies. *jfowa.com*. https://www.jfsowa.com/ontology/ontoshar.htm

Tilahun, B., Kauppinen, T., Keßler, C., & Fritz, F. (2014). Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation. *JMIR Medical Informatics*, *2*(2), e31. https://doi.org/10.2196/medinform.3531

Van Assem, M., Rijgersberg, H., Wigham, M., & Top, J. (2010). Converting and Annotating Quantitative Data Tables. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, & B. Glimm (Eds.), *The Semantic Web – ISWC 2010* (Vol. 6496, pp. 16–31). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17746-0_2

Vickery, B. C. (1997). Ontologies. *Journal of Information Science*, *23*(4), 277–286. https://doi.org/10.1177/016555159702300402

Webb, R. L., & Ma'ayan, A. (2011). Sig2BioPAX: Java tool for converting flat files to BioPAX Level 3 format. *Source Code for Biology and Medicine*, *6*(1), 5. https://doi.org/10.1186/1751-0473-6-5

Woznowski, P. R., Tonkin, E. L., & Flach, P. A. (2018). Activities of Daily Living Ontology for Ubiquitous Systems: Development and Evaluation. *Sensors*, *18*(7), 2361. https://doi.org/10.3390/s18072361