


ELLAS Architecture and Process: Collecting and Curating Data on Women’s Presence in STEM


Rita Cristina Galarraga Berardi  [Federal University of Technology - Paraná | ritaberardi@utfpr.edu.br]


Pedro Henrique Stolarski Auceli  [Federal University of Technology - Paraná | pedroauceli@alunos.utfpr.edu.br]

Cristiano Maciel  [Federal University of Mato Grosso | California State Polytechnic University, Pomona | cmaciel@ufmt.br]

Rodgers Fritoli  [Federal University of Technology - Paraná | rfritoli@alunos.utfpr.edu.br]

Guillermo Davila  [Universidad de Lima | gdavila@ulima.edu.pe]

Indira R. Guzman  [California State Polytechnic University, Pomona | irguzman@cpp.edu]

Luana Mendes  [Foundation for Support and Development of the Federal University of Mato Grosso | luana-bm@hotmail.com]

Received: 15 November 2023 • Accepted: 21 May 2024 • Published: 30 May 2024

Abstract: The underrepresentation of women in STEM fields needs to be highlighted through data to assist decision-makers and public policy creators in addressing the issue effectively. However, the lack of structured, organized data published openly in this domain is still a reality. To address this problem, a Latin American research network called ELLAS was created. The project’s goal is to develop a platform with Semantic Web-based technologies to structure and concentrate data from Brazil, Peru, and Bolivia, initially. This paper presents the processes defined for the collection and curation of both unstructured and structured data, sourced from scientific articles, social networks, and existing open data. We explore the architecture design in a way that facilitates understanding of the details of the processes and the actors involved for each data source. We present the preliminary results from the application of these processes, and the strategies for future work, which include the data extraction and curation, and the ontology and knowledge graph development. We also present some of the undergoing work, such as the survey development and application as well as showing what still hasn’t been done, such as the platform development.

Keywords: Knowledge Graphs, Latin America, Female Leadership, STEM, Open Data

1 Introduction

Reflecting on the state of science, technology, and innovation in Latin America and the Caribbean, it’s evident that the economy of this region is not well-equipped to face the challenges of the knowledge society [Guzman *et al.*, 2020]. Part of the problem is the insufficient number of students engaging in Science, Technology, Engineering, and Mathematics (STEM), with women being notably underrepresented in countries such as Brazil [Kahn and Ginther, 2017] [Berardi *et al.*, 2022], Peru [OECD, 2022] [Garcia-Holgado *et al.*, 2020], or Bolivia [Egana-delSol *et al.*, 2022]. Another contributing factor is the even smaller proportion of women taking on leadership roles in industry or academia. In this context, higher education institutions have a critical role in supporting women and promoting an institutional environment that seeks gender equality and professional growth. This topic is, therefore, related to one of the United Nations’ Sustainable Development Goals (SDGs). Goal 5 is framed as “Achieve gender equality and empower all women and girls¹.”

Various programs and initiatives have been established to increase women’s representation in STEM. In the Brazilian context, one can mention the “Meninas Digitais” Program (PMD), endorsed by the Brazilian Computer Society (SBC). Through numerous partner projects in leading educational

institutions across Brazil and operating within the pillars of teaching, research, and/or extension, they disseminate computing knowledge to students and teachers nationwide [Maciel *et al.*, 2018]. In Peru, the National Council for Science, Technology and Technological Innovation – CONCYTEC has created the Committee for Women in Science, Technology and Innovation², responsible for the promotion of initiatives together with local and international partners. Some other initiatives have focused on the development of role models for enhancing the permanence of Women in STEM [Vidal *et al.*, 2020] or the strengthening of female students’ STEM identity, global perspective, and desire to pursue engineering for peaceful purposes by exposing them to the SDGs and engineering challenges [Tull *et al.*, 2018].

Yet, for initiatives to achieve greater effectiveness and for the formulation of public policies geared in this direction, it is crucially necessary to comprehend the factors influencing the absence of gender equality in STEM areas. These factors should be disseminated with reliability and visibility. Frequently, organizations, whether public or private, in larger or smaller contexts, possess a legitimate motivation to implement affirmative actions to attract, retain, and nurture women aspiring to pursue careers in STEM. However, the absence of data illustrating potential avenues for intervention hinders or diminishes the adoption of such actions.

Furthermore, despite the presence of data collection in cer-

¹<https://sdgs.un.org/goals/goal5>

²<https://mujercti.concytec.gob.pe/>

tain initiatives and projects, this data is seldom made available in an open and well-organized format. Typically, researchers collect and analyze data, which is often published in scientific articles. While this contributes significantly to advancing the field, there is no assurance that the data will be easily accessible for comprehension, public use, and reuse – a growing demand [Nunes *et al.*, 2020]. Consequently, STEM-related data tends to be unstructured and inadequately documented, particularly within the Latin American context. Another challenge stems from the diverse contexts and levels of analysis involved in each research endeavor, spanning projects, universities, local communities, and countries. Each context may attribute distinct meanings to the collected data, impeding its comprehension and utilization by third parties, particularly policymakers unaccustomed to extracting information from scientific articles.

In pursuit of solutions to issues like these, the international research network ELLAS - Equality in Leadership for Latin American STEM was born. It executes the project "Latin American Open Data for gender equality policies focusing on Leadership in STEM³", proposed and accepted in a call by the International Development Research Centre - IDRC. The purpose is to create an open data platform containing data related to the presence of women in STEM, useful for the formulation of research and public policies in this field [Maciel *et al.*, 2023]. The project's objective is to systematically gather and openly present data related to women engaged in courses, programs, or professions classified under STEM categories. This encompasses aspects like leadership roles, factors influencing career trajectories, and existing initiatives and public policies. Additionally, the proposal outlines the incorporation of data through ontologies to construct knowledge graphs on the subject. This structural approach ensures a more uniform and comparable representation of data across Latin American countries, akin to platforms exemplified by Hyvönen [2020]. By centralizing and organizing this information, the formulation of new policies and initiatives can be more streamlined, grounded in the identified issues highlighted by the data. The project proposal focuses on the participation of three neighboring South American countries: Brazil, Bolivia, and Peru, which have considerably different levels of development and whose groups have already been discussing these challenges [Guzman *et al.*, 2020]. The project is composed of teachers, researchers, and students with various backgrounds, including Computing, Economics, Psychology, Pedagogy, Mathematics, Knowledge Management, Social Sciences, and Administration, all with interests in research on women in STEM.

The project spans a three-year execution period (2022-2024). The primary emphasis during the first and second years involves data acquisition, platform design, and supplementary activities such as workshops and seminars to stimulate discussions about the concepts that should be incorporated into the platform [Maciel *et al.*, 2023]. In the third year, the platform will be accessible via a website, enabling direct interaction with the data on the site or through end-point connections to facilitate the development of external applications utilizing the constructed knowledge graph.

Thus, the objective of this article is to present the strategies adopted for the process of data collection and curation from various sources, whether structured or not, until making them open in a knowledge graph and available on the proposed platform. We also present the architecture in order to support the explanation of the involved processes and their focused user groups. This article is an extended and revised version of the paper by Berardi *et al.* [2023], which presented more broadly the strategies adopted for the development of the open data platform in STEM. From a methodological standpoint, the research follows a descriptive exploratory approach [Gil, 2008].

2 The Importance of Data

The utilization of statistical data is essential, and in Brazil, certain data sources are consistently and openly accessible. One example comes from the SBC⁴ (Brazilian Computing Society) which annually publishes a report with data from Brazilian university computing courses, with the data collected from the "Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira" (INEP). INEP's data encompasses various courses, and this compilation by the SBC is a report focused on computing. Despite its usefulness, it is released in an unstructured format (PDF), which makes analysis difficult.

In the context of Latin America, it's worth noting data collection practices in Peru and Bolivia, where the circumstances vary. In Peru, the Ministry of Education (MINEDU) issues reports in an unstructured format, specifically in PDF, with intervals between versions extending beyond a year (refer to Torres Manrique and Miñan Sánchez [2021] for example). The data originates from the Higher Education Information Collection System (SIRIES), possessing current information that is not openly accessible and requires specific requests for data provision to interested parties. The SIRIES data source includes fields such as the number of enrollments by public and/or private university, gender, area of knowledge, year, among others [MINEDU and DIGESU, 2023]. Moreover, MINEDU provides data about some initiatives, which is the case of "Niñas Talento Digital" and "200 Chicas STEM del bicentenario". However, the most common way to find initiatives is through social media, such as Facebook and Instagram, which is the preferred method to spread news and information about the initiatives, which is the case for: "AgileGirls-Peru", "Clubes de Ciencia Perú", "mujeresit", "Teens in AI Peru" and others.

In Bolivia, statistical data related to the Bolivian university system are provided by the Bolivian University Executive Committee⁵. This is the central body that coordinates and programs the goals and functions of the Bolivian University System, representing it through planning, organizing, execution, and management actions, recognized in the Organic Statute and the General Regulations of the Bolivian University System. Although statistical data are made available, they are not always accessible on their portal at the time of writing this article. An example of this is the work developed

³<https://ellas.ufmt.br/>

⁴<https://www.sbc.org.br/>

⁵<https://ceub.edu.bo/>

by Branisa *et al.* [2021], in which the authors had to collect data through interviews and focus groups to explore the issue of the low number of women in Information Technology in Bolivia. Although the statistical data is concentrated by a single committee, data about initiatives is scattered throughout the web. Most of them utilize social media to organize and spread their news and information, the most popular ones are Facebook and Instagram, some example of initiatives that utilize social media are: “Tremenda Bolivia”, “Derechoteca” and “Asociación Femenina de Ingeniería Bolivia”. In rare cases, some initiatives create their own websites, which is the case of: “EnerGEA STEM” and “Tecnonautas”.

In a narrower context, it is worth noting studies such as that conducted by Nunes *et al.* [2020], which underscore the necessity for open data concerning fundamental aspects to comprehend the actual circumstances of computing graduates. The authors, in their pursuit of insights into the employment landscape and the experiences of graduates in the Amazon region, found themselves obliged to devise their own questionnaires and independently analyze the data—operating in an essentially unconnected manner, without integration into a more structured initiative.

Another illustrative case demonstrating this need is exemplified by the work of de Mello *et al.* [2021], which independently collected data on the activities of projects in the southern region, associated with the PMD. The value of acquiring such data in a structured, organized, and openly accessible format extends beyond the confines of a single region.

Similarly, Gindri *et al.* [2021] and Pereira *et al.* [2022] encountered the necessity to gather data from social networks, the list of PMD projects, and/or websites, underscoring the ongoing need for more streamlined and open approaches to data collection.

These situations illustrate the challenge of collecting, processing, and structurally storing data on women in STEM, especially in an open format. Although initiatives exist, it is essential that this be a structured and organized movement to avoid creating information silos that do not contribute to a broader understanding of the issue and cannot support decision-making in this field. However, this is a complex task that requires theoretical, methodological, and technological efforts.

3 Concepts involved in the platform construction

Given the project's application context, the platform is under development with a foundation based on Semantic Web technologies. This choice is driven by several factors: (i) enabling the integration of data from diverse sources and countries in a standardized manner to accommodate cultural variations in understanding data within the domain, (ii) facilitating the publication of data on the web for potential future integrations with other data in the Linked Open Data Cloud (LOD), and (iii) ensuring effective support for data provision in multiple languages (English, Portuguese, and Spanish).

Therefore, the key concepts discussed in this section include (i) open data, emphasizing the availability of data in an open format along with comprehensive documentation and

provenance, adhering to licensing considerations for both publication and consumption of data, and (ii) knowledge graphs formed through ontology engineering techniques, aiming to seamlessly integrate the gathered data.

3.1 Open data

According to the Open Definition, open data is “data that can be freely accessed, used, modified, and shared by anyone for any purpose.” There are also 3 fundamental standards for data to be considered open data: (i) they must be possible to obtain via the Internet in a convenient and modifiable form, (ii) they must be distributed under terms that permit free reuse, and (iii) they must be accessible to anyone without discrimination [Isotani and Bittencourt, 2015]. In Brazil, Law No. 12,527 of November 18, 2011, deals with the opening of public data for access and use by any individual, which brings more transparency to data of interest to society [Brazil., 2011]. In Peru, Legislative Decree No. 1412 of September 13, 2018 approves the Digital Government Law, where open data is also understood as a form of transparency without violating personal rights. Bolivia does not yet have established laws on open access to information, despite being present in other documents such as Supreme Decree No. 27329, of January 31, 2004, which states that access to information must be guaranteed to the population.

According to Tim Berners-Lee, the creator of the web and the Linked Open Data (LOD) initiative, a five-star scheme was suggested that qualifies how open data is, and more than that, how connectable it is [Berners-Lee, 2006]. Connected data encompasses a set of techniques for structuring and integrating data on the web, which allows consumption by both humans and machines through a specific language, bringing greater scalability to the use of the web of data [Isotani and Bittencourt, 2015]. According to LOD, the more connected data is, the higher its quality and the more information and knowledge can be retrieved. When data is available on the web in some format, even if unstructured, it is classified with one star. When the data is available on the web in a structured form, even if using a proprietary format (like Microsoft Excel), it is classified as two stars. When data is available on the web, in a structured and open format (for example, CSV), this data is considered three stars. When data is available on the web in the form of triples (knowledge graph structure), even without links to other graphs, it is considered four stars. Finally, when data is available on the web in the form of triples and connected with other knowledge graphs, it is considered five stars. At this highest stage, the data is considered connected open data (or Linked Open Data).

There is a set of best practices for the publication and connection of data; these are essential for the data to be automatically used by software agents [Isotani and Bittencourt, 2015]. The ELLAS network platform adheres to these best practices and aims to be classified as five stars, according to Berners-Lee's scheme.

3.2 Knowledge Graphs

Knowledge graphs are data structures that adhere to the RDF/OWL format (Resource Description Framework/Web

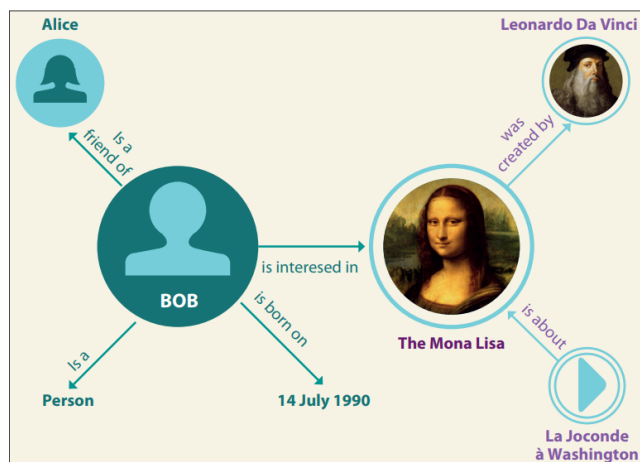


Figure 1. Example of an RDF graph [Isotani and Bittencourt, 2015]

Ontology Language), representing information in triples to express relationships between data. In Figure 1, for instance, the connection between "The Mona Lisa" and "Leonardo da Vinci" is explicitly articulated through the relation "was created by," forming what is referred to as a triple. Unlike mere storage structures like databases, knowledge graphs are designed with consideration for the meaning that data holds within a specific domain, encapsulating the semantics of the data.

A knowledge graph seamlessly combines the definition of data through ontological models, outlining the data schema, with the instances of these definitions – the actual data. The incorporation of ontological modeling brings the advantage of applying logic to infer new relationships between data based on explicit relationships. This is made possible through inference engines capable of discovering and adding new triples to the knowledge graph [Fensel *et al.*, 2020].

Ontology is originally a field of study in philosophy that encompasses the nature of being, existence, and reality itself. In computing, the definition of ontology is the formal representation of the concepts within a domain [Noy *et al.*, 2001]. Ontologies define a particular domain through classes, properties, and restrictions. The classes serve to describe the set of objects within the domain, properties are the relations between classes and data, and restrictions are responsible for the logical part. Graphically, these elements can be identified because a knowledge graph is a directed graph, composed of vertices and edges, where the vertices represent classes and the edges represent the relations.

3.2.1 Ontology Engineering

The development of ontologies is a widely discussed subject where various authors propose certain sets of steps; some include automated processes, others semi-automated, and there are also authors who argue that the creation of an ontology is an essentially human process involving ontologists (individuals with technical knowledge of computational languages) and domain experts (individuals with domain knowledge, but not necessarily with technical know-how in modeling ontologies) [Noy *et al.*, 2001]. Following this last approach, it can be pointed out that the creation of ontologies is an iterative process based on two steps: (i) creation of the ontology (ii)

review and refinement. The following steps adhere to the "Ontology 101" development process by Noy *et al.* [2001].

The first step in developing an ontology is domain analysis, through which the necessary terms for constructing the ontology's vocabulary are identified. There are several methods for domain analysis, but two stand out: literature analysis, which uses technical texts to extract terms, and the involvement of domain experts, where the experts define the terms [Smiraglia, 2015]. The analysis does not need to be carried out exclusively by just one method. In the case of Hippolyte *et al.* [2018], both literature analysis and the involvement of domain experts were the methods used for analysis. This mix of methods provides flexibility in domain analysis, and the choices made by the ontologist should be based on the resources available.

In the process of modeling, it becomes imperative to establish the classes and properties that will constitute the ontology. This commences with the compilation of a list comprising the most pivotal terms within the domain, effectively representing the core concepts in that specific domain. Within this modeling endeavor, the success of the ontology hinges on the collaborative efforts of two key roles: the ontologist and the domain expert. The ontologist bears the responsibility of transforming these identified terms into classes and properties that govern the relationships between classes. Meanwhile, the domain expert plays a crucial role in generating the list of terms and scrutinizing the defined classes and properties. Additionally, the domain expert is tasked with determining the key questions within the domain that the model should be proficient in answering, often referred to as competency questions.

With all classes and properties created, data instantiation is necessary, in other words, the addition of data to the ontology. The result of this step is the knowledge graph, which contains the ontology and the instantiated data. For processing the data present in the knowledge graph, it is necessary to store it in a tool designed to work with data structured in the form of triples, a Triplestore. These are responsible not only for storing but also for enabling access to these data, and for this purpose, the SPARQL (SPARQL Protocol and RDF Query Language) query language is used, a programming language that uses the concept of triples to search for data.

In other words, to make data open, it is necessary to create an entire work method aimed at publishing high-quality open data in a format suitable for consumption by both software agents and humans [Rodrigues and Maciel, 2022]. Considering the application of these concepts, the architecture of the platform was developed, described below.

4 The architecture of the platform

Figure 2 illustrates the data architecture designed by the ELLAS research network, divided into three main groups, considering different processes and users (from bottom up): Data Sources, Processing Layer, and Application Layer.

Data Sources: This group represents the data in original formats (without any manipulation), which will be collected for insertion into the knowledge graph. The origin of the data is classified into two categories: primary and secondary data

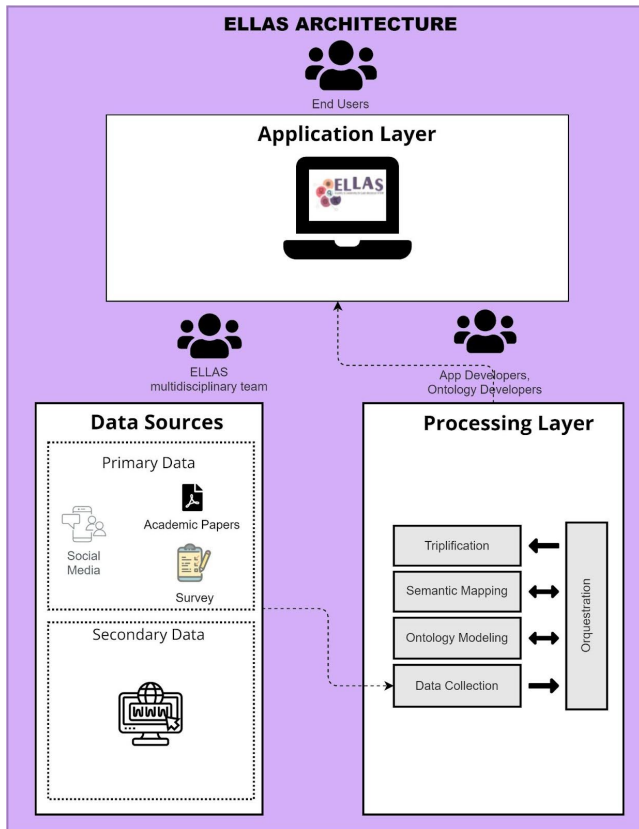


Figure 2. General Platform Architecture

sources. In this project, primary data refers to data collected from bibliographic sources, social networks, and through a survey created specifically for this purpose; secondary data are collected from sources of open structured data available on websites like UNESCO and INEP. The users involved in this group are the internal teams of the ELLAS research network, strategically formed, detailed in Section 5.1. The Triplestore will be the database responsible for storing and retrieving data through SPARQL queries. The data sources are used in the data collection process described in the next group.

Processing Layer: This category encompasses tasks associated with processing and transforming data, initiating with the collection of both structured and unstructured data, and advancing through the creation of ontologies and instantiation of data. Pipelines, which automate these procedures and incorporate the data into the knowledge graph through triplification actions, are implemented in this layer. The orchestration of these pipelines is conducted utilizing the Pentaho tool. Users within this group comprise platform developers affiliated with the ELLAS network team, encompassing ontology developers and application developers.

Application Layer: This category involves the query interface primarily crafted for policymakers, decision-makers, STEM researchers, gender researchers, and other pertinent profiles. Moreover, within this layer, querying through an API (SPARQL endpoint) will be accessible for web applications that may be developed by external parties.

While the architecture depicts actions related to the data collection process within a singular processing box, in practice, a distinct pipeline is executed based on the type of data source (primary or secondary data). Further details on these

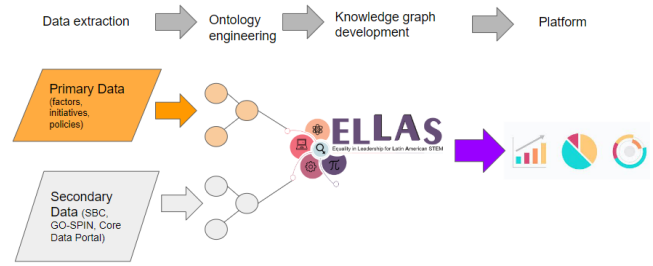


Figure 3. Overview of the platform construction (author's own)

processes are elaborated in Section 5.1.

5 Strategies for the data extraction and curation process

This section is divided into 5 subsections that provide more details on the strategies for the data extraction and curation process, from data collection to ontology creation, knowledge graph development, platform development, and data maintenance. Figure 3 provides a macro-level illustration of the pipeline, with details explained in the following sections.

5.1 Data collection

Data collection is an essential step in creating the knowledge graph, and each type of data source (primary or secondary) has its own data collection methods tailored to their characteristics and complexities. See below.

5.1.1 Primary data

Acknowledging that the majority of data concerning women in STEM remains predominantly within the academic domain and is often presented in an unstructured format (PDF) across various publication sources, it became imperative to devise a methodology. This methodology aims to facilitate the comprehensive collection of information about women in STEM, with a particular emphasis on prevailing policies, initiatives, and factors influencing their presence in these fields. Moreover, the methodology is designed to mitigate biases in interpreting the significance of data extracted from bibliographic articles. To achieve this, strategic teams characterized by diversity in gender, age, cultural background (linked to the country), and field of expertise were established to conduct the collection process.

The team responsible for collecting and curating primary data consists of researchers who work on the topic of gender equality in STEM in the different countries involved in the project. They are responsible for obtaining data on three topics related to the presence of women in STEM fields: factors that impact the choice and retention in the field, initiatives developed in the form of projects, and existing policies aimed at mitigating the low representation of women in these areas, this division can be seen in the data extraction layer shown in Figure 4. According to the ontology engineering methodologies discussed in Section 3.2.1, team members are referred to as Domain Experts because they have the ability to search for data and identify its relevance in the subject.

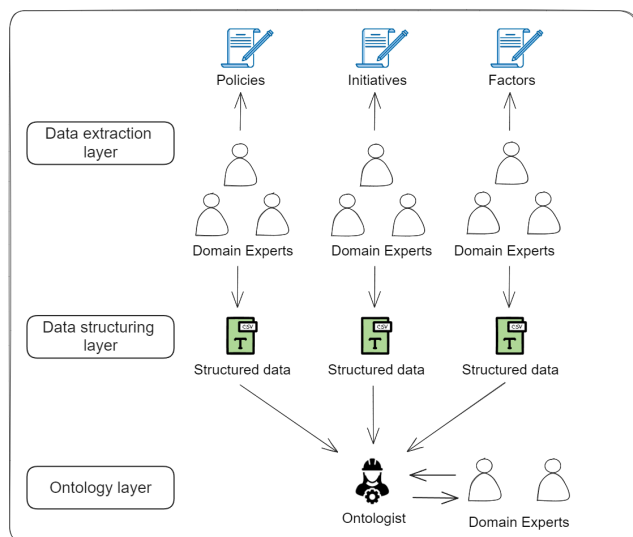


Figure 4. Methodology for collecting primary data

With the division of teams, each one was responsible for conducting a systematic literature review (SLR) [Kitchenham, 2004], in which the team not only analyzed scientific articles but also grey literature Kitchenham [2004]. The SLR defines organized and systematic steps for reading articles with well-defined criteria for sources and inclusion of articles in the corpus. The goal is that through SLRs, data described in the articles are extracted and structured in spreadsheets, which are used in subsequent processing by the Processing Layer of the architecture to model the integration ontology.

The curation process during and after data collection is characterized by internal review within the teams, where each member reviews the data collected by their team members, and any inconsistencies in interpretation are discussed internally. There must be consensus on the interpretation of the data being collected, and the interpretations used should be recorded. This methodological process reflects the collaborative and multicultural nature of the ELLAS network platform. In this way, the chance of introducing bias into the collected data is reduced.

The team assigned to the collection and structuring of data on initiatives is responsible for identifying projects and programs aimed at enhancing the representation of women in STEM. Their task involves documenting and analyzing both successful and less successful initiatives.

Concurrently, the team in charge of gathering and structuring data on factors endeavors to map the various elements influencing the career development of women in STEM. It is crucial to document factors that exert both positive and negative influences on the career choices of female STEM students, as noted by Ribeiro [2020].

Simultaneously, the team dedicated to structuring data on existing policies is assigned the task of mapping policies and interventions designed to narrow the gender gap.

All primary data is initially recorded in spreadsheets for minimal structuring, forming the data structuring layer as depicted in Figure 4. This includes definitions extracted from articles, and provenance information from the articles is meticulously collected and documented. At the conclusion of the literature review, three CSV format spreadsheets are

generated, each corresponding to one of the three topics.

Subsequently, these spreadsheets undergo an additional curation process led by the ontologist team. This phase involves a thorough review of data characteristics, addressing potential redundancies or contradictions that may have been introduced. Curation discussions are integrated with both the ontologist team and multicultural teams, and after reaching a consensus, knowledge graphs are generated from the spreadsheets, as illustrated in the ontology layer of Figure 4.

In addition to scientific articles, the data collection process on social networks followed a similar dynamic to the SLR, where information on social networks was manually analyzed from multidisciplinary projects. These data are included in the same spreadsheets generated by the SLR, following the same structure of information collected in scientific articles, with a few changes, such as the inclusion of the URL of the consulted social network.

In this scenario, the team overseeing initiative data determined the necessity for 36 columns for data description, and a total of 250 data rows were incorporated. The public policy team specified 28 columns and identified 29 rows of data. Meanwhile, the factors team outlined 58 columns and discovered 170 data rows. The amalgamation of all this data, excluding provenance data, resulted in a total of 5,129 triples.

Beyond acquiring data on factors, initiatives, and public policies through systematic literature review (SLR) and social networks, additional data is being collected in the domain of STEM Leadership. This involves a survey devised by a team within the ELLAS network. The survey team formulated a questionnaire grounded in literature and adhering to best practices for survey development. This survey was then analyzed by an Ethics Committee, which accepted the proposed survey after some changes. The survey project and protocol were submitted for consideration by specific ethics committees in Brazil, Bolivia and Peru, with Brazil as the coordinating center and receiving Certificate of Presentation for Ethical Appreciation No. 66296922.6.0000.5690.

The purpose of applying this survey is to collect new data not present in the analyzed literature and thus enhance data on the scenario of women's leadership in STEM in Latin America. It is worth noting that its application is being carried out on a large scale by companies contracted by the project through a Support Foundation, with a focus on application in universities and the job market for women in STEM in Brazil, Peru, and Bolivia, which was concluded in all three countries. Even though it was concluded, the data still hasn't been made available for the ELLAS network team by the contractors, as of the time of writing. Despite the project's focus on women, the survey will also collect data from other genders to support comparative analyses.

The primary data collection processes are quite manual for the reasons already explained. To ensure data quality, the strategy of involving domain experts as much as possible was adopted to cover the understanding of the subject. Since there was no known modeling of this domain that could assist in extracting these data with quality, this approach was chosen. In the ELLAS team, several people with the mentioned diversities were involved in carrying out this activity.

As of the time of writing, the first round of systematic literature reviews was concluded. Which resulted in 3 spread-

sheets with data from public policies, initiatives and factors that influence the presence of women in STEM. A new round of SLRs is planned for 2024 which will update the spreadsheets with the data from papers published in 2023. They will also be updated with new data originated from the review of grey literature. So far 24 new sources of data were found and are being analyzed. The estimated time to finish both of these updates is the end of 2024.

5.1.2 Secondary data

In addition to the data obtained from literature, there is potential information that may not explicitly address the involvement of women in STEM but can indirectly contribute, such as gender-related statistics in higher education. This type of data, known as secondary data, is being gathered through open data portals like INEP/SBC and GO-SPIN (an observatory featuring science and technology data supported by UNESCO).

To obtain secondary data for each platform, another team of master's and undergraduate students from the ELLAS Network team was assigned. Various data sources will be analyzed, and their data will be inserted into the ELLAS Network platform. Unlike the manual characteristic of the primary data collection process, the process for secondary data is through automated pipelines. This is possible because it involves data that already has some structure on the web and does not depend on human interpretation, such as reading bibliographic texts. To facilitate understanding, the process of analyzing and inserting data from two data sources is exemplified below.

The first one was the data source provided by INEP⁶, which contains data about the higher education courses offered in Brazil. In this data source, it is possible to check the number of male and female students entering and graduating from a certain course offered at a university. This data source is extensive due to the number of universities in Brazil and because it contains data for an entire year. It is necessary to make a "cut" in the amount of data since the data source contains all courses from all universities. However, the platform's scope is data related to STEM courses, so unrelated courses should be removed. A pipeline is also being developed with Pentaho and Python for the constant update of the knowledge graph whenever new data is made available. Some issues have already been detected for the update since a new data source is released every year, and changes in file standards may occur.

The second data source is supplied by UNESCO and includes information on initiatives across different countries. In this particular instance, two specific data sources were chosen: the GO-SPIN⁷ data source and the Core Data Portal⁸. Similar to the data source provided by INEP, the UNESCO data source serves as a portal for data collection with filters. Defining filters is necessary to gather data within the ELLAS scope. Only data concerning initiatives that seek to encourage women's participation in STEM fields and are located in Brazil, Bolivia, or Peru will be incorporated. Presently,

these data sources have limited information within the Latin American context. An automated update pipeline is also in progress to apply filters and automatically collect data for future updates.

Since these data sources have data from various years the ELLAS team had to decide the starting year for the data collection. For INEP's data the starting year is 2009, UNESCO's Core Data Portal is 2005 and GO-SPIN is 2022.

Once gathered and organized, the ontology can be developed, and the data instantiated, thereby creating the knowledge graph.

As of the time of writing, the data from INEP, Core Data Portal and GO-SPIN have been gathered and inserted into the knowledge graph. In 2024 more data sources are being analyzed and are going to be integrated into the knowledge graph. So far 8 new sources are being analyzed, 4 from Peru and 4 from Bolivia, the estimated time for the conclusion of the analysis and integration is the end of 2024.

5.2 Ontology creation

For the creation of the ontology, methodological steps of the aforementioned Ontology Engineering are considered. The chosen method for domain analysis was a combination of literature analysis and involvement of domain experts. The systematic review process conducted on primary data corresponds to the literature analysis step, where various terms related to each topic were identified. Domain experts were responsible for selecting and including the terms found in the literature analysis. It is expected that with the expertise of the experts, the selected terms will be more coherent with reality.

For the modeling and construction of the ontology, a team of experts is responsible for analyzing the spreadsheets generated from the collection of primary and secondary data and creating an ontology for each of them. For each spreadsheet, there is a process of ontology creation, validation, and modeling in collaboration with the teams that collected the data (domain experts). If a problem is identified, the ontology will need to be modified, and the review process will be repeated. This cycle continues until the results are satisfactory.

After the creation of the ontologies, a search for vocabularies to describe the classes and properties present in the ontologies will be conducted, as it is a good practice to reuse domain terms. Finally, the integration of all ontologies with their respective data will generate the knowledge graph about women in STEM. To validate whether the ontologies are correctly representing the data, competence questions defined by domain experts were used. These questions are relevant and are what the platform is expected to help answer, a full list of all the competence questions can be found here⁹. Examples of some defined competence questions are:

- Which/How many initiatives are carried out <in a given country>?
- What types of gender policies/processes/practices have been implemented in Bolivia, Brazil and Peru since <a given year>

⁶<https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos>

⁷<https://gospin.unesco.org/frontend/home/index.php>

⁸<https://core.unesco.org/en/home>

⁹<https://docs.google.com/document/d/1ncDpha6rnUnb3x-P-HILMDpuOmBR9qgITjFXx9W5TB4/edit?usp=sharing>

- What are the contextual factors that impact Positively/Negatively <on a given impact> in <a given country>?

Since the ontology creation process is cyclical, it will be adjusted as needed based on new data sources, whether they are primary or secondary.

The terms related to the ontology for the three topics presented in the primary data have already been defined and added to the ontology. However, this is not the final version, as the spreadsheets are still undergoing data quality curation, which may require changes to the ontology.

Providing an overview of the initial version of the ontology generated from the structuring of the data, the following structure is in place:

- Ontology entities related to Public Policies: 3 classes, 5 subclasses, 2 object properties, and 10 data properties.
- Ontology entities related to factors influencing women in STEM: 2 classes, 1 subclass, 1 object property, and 6 data properties.
- Ontology entities related to Initiatives: 5 classes, 16 subclasses, 4 object properties, and 19 data properties.

As a means of ontology validation, the literature suggests using the answers to competency questions. To that end, it has been tested that with this initial version of the ontology, it is already possible to answer the competency questions presented earlier. It's worth noting that the entities related to survey data have not yet been created because the survey is still ongoing.

5.3 Knowledge Graph Creation

The construction of the graph takes place as the ontological model is populated with both primary and secondary data extracted from spreadsheets. Consequently, an RDF-format file containing all the triples is generated, and this file is then inserted into a Triplestore for data consumption. Notably, this mapping is dynamic, meaning it is employed to generate new graphs as the data undergoes updates.

The mapping process adheres to a CSV-format spreadsheet, where the strategy assumes that the column names serve as variables for instantiating the data within those columns. This mapping involves creating triples using the columns from the spreadsheet and the ontology terms. Figure 5 provides an example of a triple generated in the OntoRefine program. In this instance, the "COUNTRY" column serves as the subject of the triple, the predicate is the property "A," and the object is the class "Country" as defined in the ELLAS ontology, forming the information that the column containing countries are instantiated as the concept country of the ontology *ellas*.

For example, if the multidisciplinary teams state that a policy should have its name, country of application, and authors, then the classes Policy, PolicyName, PolicyCountry and PolicyCreator are created in the model. Then, using the OntoRefine software, a mapping is created indicating that the instance of the Policy class is found in the "policy name" column of the CSV, for example, and so on, an example can be seen in Figure 5.

The mapping is completed once all the columns in the spreadsheet and all the ontology terms have been used in a triple. It's worth noting that a column or ontology term can be used multiple times to follow the logic created in the ontology. In the end, the knowledge graph for that spreadsheet is expected as a result of the entire process, which should be repeated for all available spreadsheets. This is a part of the triplification process in the Processing Layer of this paper's architecture.

In order to validate the created graph the SPARQL language was used to answer the competency questions presented in Section 5.2. Some of these were selected for the creation of charts and tables which are presented in the sequence. The chosen questions were:

- Which/How many initiatives are carried out <in a given country>? Where the chosen countries were Brazil, Peru and Bolivia, the resulting chart can be seen in **Figure 6**.
- What types of gender policies/processes/practices have been implemented in Bolivia, Brazil and Peru since <a given year>? Where the chosen year was 2015, the resulting table can be seen in **Table 1**.
- What are the contextual factors that impact Positively/Negatively <on a given impact> in <a given country>? Where the chosen impact was "Increase female representation", the impact was positive and the countries selected were all from Latin America, the resulting table can be seen in **Table 2**.

From this point onwards, the data is structured, open, and made available in the knowledge graph.

5.4 The Platform

With a back-end based on ontologies, the platform will provide integrated data from projects, initiatives, and actions carried out in some Latin American countries, as well as statistical data from open sources such as the INEP data source and research such as the project's own survey, and data from well known portals such as UNESCO's GO-SPIN and Core Data Portal. This integration will facilitate gender research, as having centralized data makes the process of searching for information much easier.

One of the project team's considerations pertained to the potential complexity introduced by the use of ontologies, as it might necessitate prior subject knowledge, potentially impeding platform usage. In response, the design and functionality of the website will prioritize a user-centered approach, emphasizing usability, communicability, and accessibility. The goal is to explore and implement techniques for human-computer interaction and human-data interaction, thereby eliminating the necessity for users to possess specific knowledge of SPARQL. For this purpose, researchers in the field of Human-Computer Interaction are part of the project team for the platform.

In order to facilitate the utilization of the platform's data as input for visualization applications by other systems, the SPARQL endpoint for the Triplestore will be provided. This functionality enables connections with other knowledge graphs.

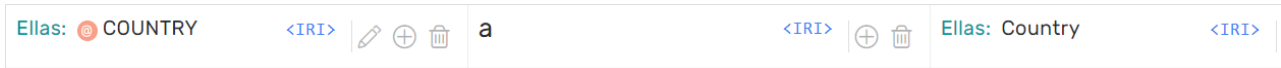


Figure 5. Example of triples in the mapping

Table 1. Types of gender policies since 2015.

| Types | Amount |
|--|--------|
| Assessing the gap of women’s participation in STEM | 4 |
| Zero tolerance for sexual harassment and violence at the workplace | 1 |
| Attracting women’s participation in STEM fields | 8 |
| Support Women’s Career Growth | 2 |
| Retention/empowerment during STEM training | 1 |
| Identifying Women’s Participation in STEM | 2 |
| Improving awareness about importance of participation of women in STEM | 1 |
| Providing financial support for female students in STEM | 1 |
| Mentoring programs for young women (meet women engineers, etc) | 1 |
| Gender quota (ensure women’s participation) | 1 |
| Mentoring programs for young women | 1 |

Table 2. Contextual factors that positively impact the female representation.

| Contextual Factor | Amount |
|--|--------|
| Respectable and comfortable environments | 1 |
| University administration support | 1 |
| Successful pathways through alumni | 1 |
| Mentoring and connecting with alumni | 1 |
| Mentoring | 1 |
| Role models | 4 |
| Projects and activities in school | 1 |

Number of initiatives for each country

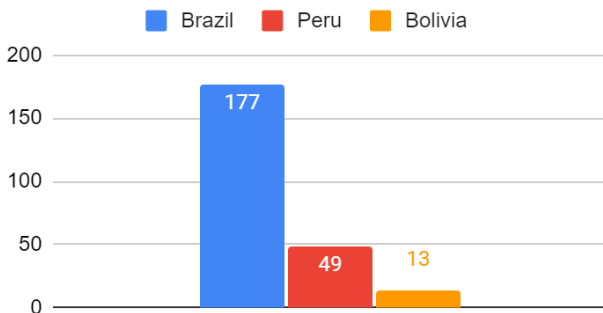


Figure 6. Amount of initiatives per country)

5.5 Data Maintenance

The project platform, in addition to providing access to information, will also offer the functionality to add new primary data when a new project, initiative, or policy, is created. This way, the platform will stay up-to-date and relevant for use in new gender-related STEM research. To achieve this goal, the semantic mapping created in the knowledge graph creation process, along with a script, will automate these functions to transform new primary data into a graph and then insert it into the project’s Triplestore. There will also be a quality control process for new data following techniques such as those developed by Bertucini *et al.* [2023]. The technique consists of defining quality dimensions to be checked every time new data is inserted into the knowledge graph, using

the SHACL¹⁰ (Shapes Constraint Language) language and Python.

For secondary data, especially open data sources that are frequently updated, a pipeline will be developed to automate the data updates. This pipeline will fetch the data sources as soon as they are updated, process the data to fit the project’s scope, transform them into RDF format using semantic mapping, and finally insert the graph into the Triplestore. One limitation of this process is that the data is made available by third parties, and if they decide to stop publishing their data there won’t be any new updates in the ELLAS platform. at least when it comes to secondary data.

6 Final considerations

In Latin America and the Caribbean, there is a notable underrepresentation of women in STEM fields, impacting technological development directly. Efforts to involve women in STEM are often undertaken through initiatives, policies, and actions. However, information about these endeavors lacks centralization, posing challenges for analysis and the development of more effective strategies. Furthermore, open data concerning women in leadership roles also suffers from decentralization and a lack of structured formats, hindering reusability.

To tackle these issues, the ELLAS research network and its project were established. The primary goal is to address

¹⁰<https://www.w3.org/TR/shacl/#references>

the decentralization and lack of structure in data related to factors, initiatives, and policies promoting gender equality in STEM. The project also aims to provide open statistical data on women in STEM fields in Brazil, Peru, and Bolivia. Although the project is still in progress, the overarching objective is to create a Semantic Web-based platform facilitating the distribution and connection of this data. The platform will feature a knowledge graph describing and consolidating data to ease research in the field.

The project has 3 phases divided into 3 years. The first phase was conducted in 2022 and it was the formal initiation of the project. The focus on this phase was to convene a wide array of stakeholders, which encompasses policymakers, academic institutions, local and governmental representatives, as well as civil society initiatives; in this phase a training was also conducted to instruct the domain experts on ontology and knowledge graphs. The second phase was conducted in 2023 and the focus was on the collection and structuring of the primary and secondary data, as well as the development of the ontology and knowledge graph. The third phase is currently being conducted in 2024 and has the focus of updating the data, the development of the platform and the dissemination of the platform to the stakeholders.

This article presents the platform's architecture concisely, focusing on the collection and curation of primary and secondary data from literature and existing sources. Carefully adopted strategies mitigate potential biases in the data collection process. Additionally, preliminary results on data collection are shared. Challenges such as the multicultural team composition, diverse expertise across knowledge areas, and heterogeneity in data vocabularies are addressed through ontology modeling.

Future work is required, especially for documenting the entire data collection and transformation process in greater detail, including information about data collection from social networks and via surveys. In this way, ELLAS will not only be offering a platform and its data but also the entire methodology and decisions made by the group of researchers to achieve its goal. It is believed that this will be an important step in the initiative to open and publish data related to this topic, which can be further enhanced even after the conclusion of ELLAS network activities

Declarations

Acknowledgements

The authors would like to thank the International Development Research Centre - IDRC for the opportunity and support for the project, the National Council for Scientific and Technological Development - CNPq for the scholarships granted, and the UNISELVA Foundation for the administrative and financial management of the project. They would like to give special thanks to the entire team involved in the project, working on different fronts for its success, and advocating for more women in STEM.

Funding

This work was carried out with the support of the International Development Research Centre - IDRC International Development Re-

search Centre - IDRC and also with several research grants in the contexts of the universities involved.

Availability of data and materials

The data and materials used in the study can be found in the project's website <https://ellas.ufmt.br/pt/inicio/> and also on the platform scheduled for launch in 2025

References

- Berardi, R. C. G., Amador, B. O., Hoger, M. D. V., Turato, P. A., da Silva Santos, L. M., and Bim, S. A. (2022). The demand for stereotype-free computing courses for elementary school teachers. *Journal on Interactive Systems*, 13(1):410–418. DOI: <https://doi.org/10.5753/jis.2022.2854>.
- Berardi, R. C. G., Auceli, P. H. S., Maciel, C., Davila, G., Guzman, I. R., and Mendes, L. (2023). Ellas: Uma plataforma de dados abertos com foco em lideranças femininas em stem no contexto da américa latina. In *Anais do XVII Women in Information Technology*, pages 124–135. SBC. DOI: <https://doi.org/10.5753/wit.2023.230764>.
- Berners-Lee, T. (2006). Linked data. world wide web consortium (w3c). Available at: <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed on 29 May 2024.
- Bertucini, O. T., Berardi, R. C., Belizario, M. G., and Koziévitch, N. (2023). Garantindo a qualidade de dados na fusão de dados conectados: Um caso de uso de shacl em dados abertos de mobilidade e educação de curitiba. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 31–40. SBC. DOI: <https://doi.org/10.5753/erbd.2023.229429>.
- Branisa, B., Cabero, P., and Guzman, I. (2021). The main factors explaining it career choices of female students in bolivia. *AMCIS 2021 Proceedings*.
- Brazil. (2011). Law no. 12,527, of november 18, 2011. regulates access to information provided for in xxxiii of art. 5, ii of § 3 of art. 37 and § 2 of art. 216 of the federal constitution; amends law no. 8,112, of december 11, 1990; repeals law no. 11,111, of may 5, 2005, and provisions of law no. 8,159, of january 8, 1991; and provides other measures. Available at: <https://presrepublica.jusbrasil.com.br/legislacao/1029987/> Accessed on 29 May 2024.
- de Mello, A. V., Finger, A. F., Gindri, L., and Melo, A. M. (2021). Mapping the actions carried out by partner projects of the meninas digitais program in the southern region. In *In Proceedings of the XV Women in Information Technology*, pages 91–100. SBC. DOI: <https://doi.org/10.5753/wit.2021.15845>.
- Egana-delSol, P., Bustelo, M., Ripani, L., Soler, N., and Viollaz, M. (2022). Automation in latin america: are women at higher risk of losing their jobs? *Technological Forecasting and Social Change*, 175:121333. DOI: <https://doi.org/10.1016/j.techfore.2021.121333>.
- Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., Wahler, A., Fensel,

- D., et al. (2020). Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10. DOI: https://doi.org/10.1007/978-3-030-37439-6_1.
- Garcia-Holgado, A., Deco, C., Bredegal-Alpaca, N., Bender, C., and Villalba-Condori, K. O. (2020). Perception of the gender gap in computer engineering studies: a comparative study in peru and argentina. In *2020 IEEE Global Engineering Education Conference (EDUCON)*, pages 1252–1258. IEEE. DOI: <https://doi.org/10.1109/EDUCON45650.2020.9125224>.
- Gil, A. C. (2008). *Methods and techniques of social research. São Paulo: Atlas.*
- Gindri, L., Araújo-de Oliveira, P., Melo, A. M., Maciel, A., Vargas, K. D. A. R., Otokovieski, M. B., and dos Anjos, R. (2021). Mulheres na computação: de norte a sul-uma ação de extensão na pandemia na busca pela integração das diferentes regiões do brasil. In *Anais do XV Women in Information Technology*, pages 101–110. SBC. DOI: <https://doi.org/10.5753/wit.2021.15846>.
- Guzman, I., Berardi, R., Maciel, C., Cabero Tapia, P., Marin-Raventos, G., Rodriguez, N., and Rodriguez, M. (2020). Gender gap in it in latin america. *AMCIS 2020 Proceedings*.
- Hippolyte, J.-L., Rezgui, Y., Li, H., Jayan, B., and Howell, S. (2018). Ontology-driven development of web services to support district energy applications. *Automation in Construction*, 86:210–225. DOI: <https://doi.org/10.1016/j.autcon.2017.10.004>.
- Hyvönen, E. (2020). Linked open data infrastructure for digital humanities in finland. In *Digital Humanities in the Nordic Countries*, pages 254–259. CEUR.
- Isotani, S. and Bittencourt, I. I. (2015). *Connected Open Data: In Search of the Web of Knowledge*. Novatec Editora.
- Kahn, S. and Ginther, D. (2017). Women and stem. Technical report, National Bureau of Economic Research.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.
- Maciel, C., Bim, S. A., and da Silva Figueiredo, K. (2018). Digital girls program: disseminating computer science to girls in brazil. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, pages 29–32.
- Maciel, C., Guzman, I., Berardi, R., Caballero, B., Rodriguez, N., Frigo, L., Salgado, L., Jimenez, E., Bim, S., and Tapia, P. (2023). Open data platform to promote gender equality policies in stem. *Proceedings of the Western Decision Sciences Institute (WDSI)*.
- MINEDU, M. d. E. and DIGESU, D. G. d. E. S. U. (2023). Sistema de recolección de información para educación superior (siries). Available at: <https://www.gob.pe/minedu> Accessed on 29 May 2024.
- Noy, N. F., McGuinness, D. L., et al. (2001). *Ontology development 101: A guide to creating your first ontology*.
- Nunes, L. H. C., Reis, J. R., Paxiúba, C. M., Ponte, M. J., Nascimento, M. W., and Nascimento, R. P. (2020). Perfil dos egressos de computação do interior da amazônia no mercado de trabalho. In *Anais do XIV Women in Information Technology*, pages 254–258. SBC. DOI: <https://doi.org/10.5753/wit.2020.11305>.
- OECD (2022). Women in peru are under-represented among stem graduates, though less so than across the oecd: Share of graduates in stem subjects (% of women graduates), 2019 or last year available, in gender equality in peru: Towards a better sharing of paid and unpaid work, gender equality at work. OECD Publishing, Paris, <https://doi.org/10.1787/a5e150db-en>.
- Pereira, L. R. R., de Souza, K., dos Santos Nunes, E. P., Maciel, C., et al. (2022). Perfis em mídia social para meninas e mulheres com interesse na área stem e steam. In *Anais do XVI Women in Information Technology*, pages 227–232. SBC. DOI: <https://doi.org/10.5753/wit.2022.223162>.
- Ribeiro, K. d. S. F. M. (2020). Gênero, carreira e formação: O desenvolvimento da carreira das estudantes do ensino médio integrado em informática. Thesis (Doctorate in Education). Institute of Education, Federal University of Mato Grosso, Mato Grosso.
- Rodrigues, F. A. and Maciel, C. (2022). Um método para captura e compartilhamento de dados abertos educacionais via um processo etl. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 133–144. SBC. DOI: <https://doi.org/10.5753/wcge.2022.223023>.
- Smiraglia, R. (2015). *Domain analysis for knowledge organization: tools for ontology extraction*. Chandos Publishing.
- Torres Manrique, D. S., P. P. A. J. C. G. F. D. N. V. A. N. O. M. J. A. C. A. J. M. . and Miñan Sánchez, L. F. (2021). National survey of university higher education students 2019: main results.
- Tull, R., Jangha, S., Medina, Y., Bell, T., and Parker, R. (2018). Sharing peace engineering with us-based minority students, through the un's sustainable development goals, in peru. In *2018 World Engineering Education Forum-Global Engineering Deans Council (WEEF-GEDC)*, pages 1–6. IEEE. DOI: <https://doi.org/10.1109/WEEF-GEDC.2018.8629764>.
- Vidal, E., Castro, E., Montoya, S., and Payihuanca, K. (2020). Closing the gender gap in engineering: Students role model program. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1493–1496. IEEE. DOI: <https://doi.org/10.23919/MIPRO48935.2020.9245186>.