Elaboration of an Ontology to Address Women's Presence on Computer Courses in Brazil

Vanessa Lazarin de Souza
vanessa.2011@alunos.utfpr.edu.br
https://orcid.org/0000-0002-7564-2628
Universidade Tecnologica do Parana, Brazil

Rita Cristina Galarraga Berardi
ritaberardi@utfpr.edu.br
https://orcid.org/0000-0002-0281-8952
Universidade Tecnologica do Parana, Brazil

Recibido: 21 de agosto del 2023 / Aceptado: 4 de octubre del 2023
doi: https://doi.org/10.26439/interfases2023.n018.6604

ABSTRACT. Women's presence in STEM (Science, Technology, Engineering and Math) has been growing in relevance, yet research on the topic is met with a lack of data for the construction of consistent analysis. The Equality for Leadership in Latin American STEM Network (ELLAS) aims to develop a Linked Open Data (LOD) platform to help fill this gap. This work is embedded in ELLAS and contributes to (1) a triplication (RDF) of the data in Inep's Higher Education Census, (2) the creation of a methodology for elaborating ontologies, to be used within the project, that enables an analysis of the presence and permanence of women in STEM areas and (3) its instantiation in the context of Higher Education in the field of computing in Brazil.

KEYWORDS: ontologies / gender / computer / STEM / Linked open data

ELABORACIÓN DE UNA ONTOLOGÍA PARA ABORDAR LA PRESENCIA DE LAS MUJERES EN LOS CURSOS DE INFORMÁTICA EN BRASIL

RESUMEN. La presencia de las mujeres en STEM (Ciencia, Tecnología, Ingeniería y Matemáticas) ha ido ganando relevancia, sin embargo, la investigación sobre el tema choca con la falta de datos para la construcción de análisis consistentes. La Red ELLAS (Equality for Leadership in Latin American STEM - Igualdad para el Liderazgo en STEM en América Latina) tiene como objetivo desarrollar una plataforma Linked Open Data (LOD) que ayude a llenar este vacío. El presente trabajo se ubica dentro de ELLAS y

colabora con (1) la triplicación (RDF) de datos del Censo de Educación Superior del Inep, (2) la creación de una metodología de elaboración de ontologías, para ser utilizada dentro del proyecto y que posibilite el análisis sobre la presencia y permanencia de mujeres en áreas STEM, y (3) su instanciación en el contexto de la Educación Superior en el área de informática en Brasil.

PALABRAS CLAVE: ontologías / género / informática / STEM / datos abiertos enlazados

DESENVOLVIMENTO DE UMA ONTOLOGIA PARA ABORDAR A PRESENÇA DE MULHERES NOS CURSOS DE INFORMÁTICA NO BRASIL

RESUMO. A presença de mulheres em STEM (Ciência, Tecnologia, Engenharia e Matemática) tem ganhado relevância, entretanto, a pesquisa sobre o tema enfrenta a falta de dados para a construção de análises consistentes. A Rede ELLAS (Equality for Leadership in Latin American STEM - Igualdade para o Liderança em STEM na América Latina) tem como objetivo desenvolver uma plataforma Linked Open Data (LOD) que ajude a preencher essa lacuna. O presente trabalho está inserido na ELLAS e contribui com (1) a triplificação (RDF) de dados do Censo da Educação Superior do Inep, (2) a criação de uma metodologia para a elaboração de ontologias, a ser utilizada no projeto, possibilitando a análise da presença e permanência de mulheres nas áreas STEM, e (3) sua instanciação no contexto do Ensino Superior na área de informática no Brasil.

PALAVRAS-CHAVE: ontologias / gênero / informática / STEM / dados abertos ligados

1. INTRODUCTION

High quality data is necessary to create reliable, accessible and easy-to-use information. To this end, open and connected data is required – the kind that can be utilized, reutilized and freely distributed. People, applications and processes must be able to access this data and use it, and it must be possible to connect it to other data sources in order to enable the creation of new data, information and ultimately of knowledge (Isotani y Bittencourt, 2015). Public interest – that is, governmental and civilian - relies on the potential of this form of data structure to contribute to the understanding of intricate and complex social, political and economic issues.

Reliable data, as described above, is the kind of data we need to broaden the debate about the underrepresentation of women in STEM and build an objective panorama of the subject. The Equality for Leadership in Latin America STEM Network (ELLAS) was created to help fill this gap, which contains the Latin American Open Data for Gender Equality Policies Focusing on Leadership in STEM Project. The project aims to contribute to the generation of cross-country comparable data to assess policies and interventions in order to reduce the gender gap in STEM, focusing on leadership and factors related to career development. A bigger panorama of the project can be seen on Berardi et al. (2023).

This work is embedded within the project and focuses on structuring data from the Higher Education context for IT courses in Brazil. It maps data sources on women's presence in the field and uses it to create and instantiate an ontology model, which is also to be reutilized in the project and to be connected with other data in similar contexts in Peru and Bolivia.

1.1 Objectives

This work aims to elaborate a methodology for the creation of ontologies that allows us to answer questions about the presence and abidance of women in STEM areas, to then instantiate it in the context of Higher Education computer courses in Brazil and to restructure the data from the Higher Education Census of 2021, collected by Inep[1], to be employed with said ontology.

To test and validate the instantiated ontology, Competency Questions (CQs) were elaborated. Noy y McGuinness (2001) names competency questions as a way to determine the scope of an ontology, show what type of questions the ontology can answer and at which level of granularity. CQs do not aim to be exhaustive, instead they aim to provide a glimpse of the ontology's scope within its domain.

1 The National Institute of Educational Studies and Research Anísio Teixeira (Inep) is a Brazilian governmental institution that gathers data on all formal education levels and is related to the National Education Ministry.

2. METHODS

To create a new methodology for ontology elaboration, two previously established ones were referenced: NeOn (Suárez-Figueroa et al., 2012) and SABio 2.0 (de Almeida Falbo, 2014). The latter was chosen for its emphasis on collaboration and documentation and the former for its perspective on the reutilization of non-ontological resources. The result is the ELLAS Methodology, which establishes 6 stages to model ontologies by, considering the context of the ELLAS network. Its phases and their use on ELLASCompBra that addresses INEP data about Brazil's High Education computer courses instantiation are as follows:

Scope definition

Phase goal: To determine the domain and the purpose of the ontology, as well as the team roles.

ELLASCompBRA: The domain of this research is the presence of women in STEM areas in Brazil – specifically on the academic spectrum. On this instantiation, no roles were assigned. The CQs were designed to delineate the potential of the ELLASCompBRA ontology to answer inquiries about the domain, taking different factors into account: geographical, funding (public or private), institution category (federal institutions, state universities, etc), courses and their categories (bachelor, technologist, etc). The Competency Questions and their results are shown in section 3.

Resource selection

Phase goal: Evaluation and selection of resources related to the domain, ontological and non-ontological, to be reutilized within the ontology.

ELLASCompBRA: The semantic representation proposed on this work was built using a non-ontological resource, first thought to belong to the Computation Brazilian Society (SBC) and later traced to the Inep Census. This Census is Brazil's most complete research instrument about Brazilian higher education. It provides trustworthy statistics since 1997 and, since 2009, publishes data in CSV format (Comma Separated Values).

Resource restructuring

Phase goal: Cleaning, reorganizing and structuring of selected resources.

ELLASCompBRA: The first filter utilized was to only consider courses from the General Area of Computation and Information and Communication Technology (ICT) in the Inep file. The original data was divided into two CSV files, one for the courses and one for the institutions. A unique file was put together using vertical search to connect courses and HEIs[2], while columns relevant to the domain were preserved. The decision to translate

2 High Education Institutions.

the column names to English came from the consensus on its use in ontology modeling. A single new column, CourseID, was created as a unique identifier for each entry and the final file has 30 columns and close to 40 thousand lines.

Ontology conceptualization

Phase goal: Elaboration of an ontology, with its classes and properties, and a Data Dictionary to specify the terms used.

ELLASCompBRA: The ELLASCompBRA ontology, seen on Figure 1, was modeled using Protégé[3] . The final ontology has 9 classes (blue circles), 18 object properties, which establish relationships between classes (arrows with blue names) – and 15 data properties, which connect classes to literals (arrows with green names). Their descriptions can be found on the ELLAS Wiki[4].

Ontology instantiation

Phase goal: Instantiate and transform the reference ontology into an operational one.

ELLASCompBRA: The triples created in the triplication process were stored in Ontotext GraphDB[5], a graph database with RDF and SPARQL[6] support that allows us to build knowledge graphs and query them. The process consisted of merging the Inep data with the reference ontology, thus creating an operational ontology, and using it to answer the CQs formulated. The triplication of the data – that is, the transformation of the relational dataset (Comma Separated Value file) into triples – was done using the OntoRefine[7] tool on the GraphDB platform. As a result, we created a RDF (Resource Description Framework) file – where each record is a triple (<subject>, <predicate>, <object>). An example of a triple is <HEI> <has_course_id> <Course_ID>. This step and the ontology conceptualization were both conducted iteratively until the data was adjusted to the reference ontology and able to answer the proposed QCs. During the process of transforming the relational data into triples, the relationships between the columns in the CSV file are established.

Ontology evaluation

Phase goal: Evaluate the potential of the ontology built to answer the specified CQs.

ELLASCompBRA: The CQs were transformed into SPARQL queries to evaluate the

3 Open-source software for knowledge management and ontology editor. https://protege.stanford.edu/

4 https://shorturl.at/boGZ8

5 https://www.ontotext.com/products/graphdb/

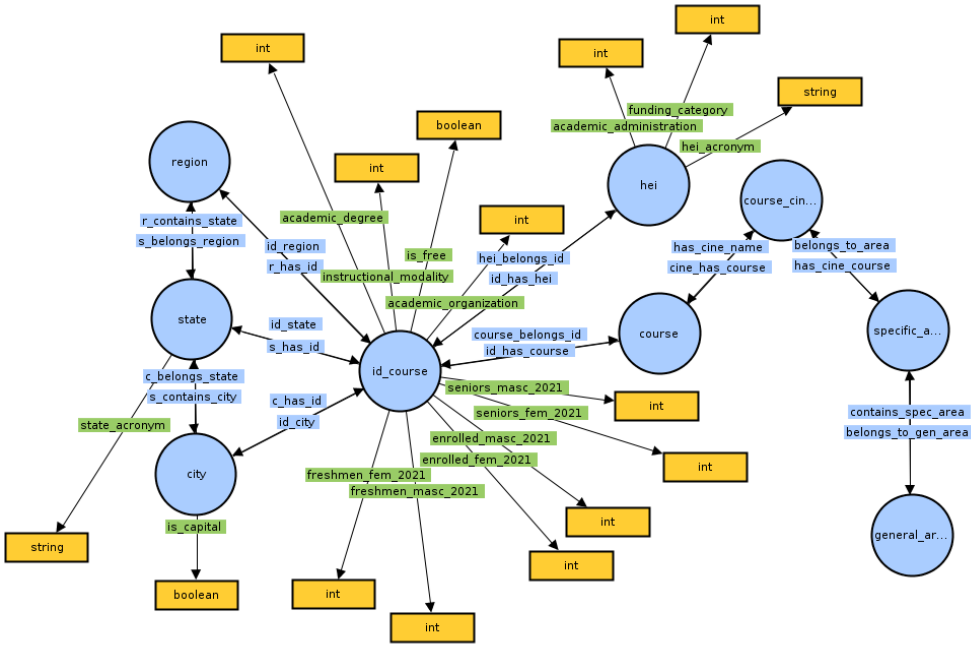
6 SPARQL Protocol and RDF Query Language (SPARQL) is a language for querying data on RDF.

7 OntoRefine is a tool attachment on the GraphDB platform based on Open Refine that is used for transforming relational datasets to RDF format.

effectiveness of the operational ontology. Figure 2 shows the query for CQ 3 and its result. The operational ontology built as per ELLAS methodology was able to answer all the competency questions established and met the research objectives of answering questions about the presence of women on higher education IT courses in Brazil. Mathematical formulas were applied to the CSV file to verify the numbers related to each CQ and validate the values yielded by the SPARQL queries.

Figure 1

First version of the ELLAS-CompBRA ontology



3. RESULTS

The first result of this work is the ELLAS Methodology, which was instantiated with the Inep data resulting in the ELLASCompBRA - our second result. The results of this instantiation were materialized with the queries created in SPARQL[8] to answer the CQs (Figure 2) and both the RDF graph and the OWL ontology file can be accessed on the ELLAS Wiki. All competency questions were executed in GraphDB using the CSV that resulted from the triplication of the Inep Census. The results are the following:

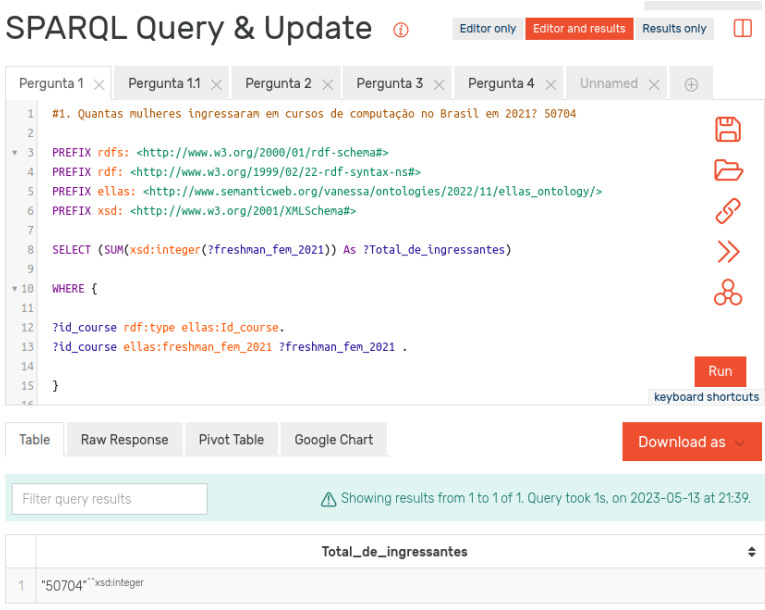
1. How many women started IT undergraduate courses in Brazil in 2021? And how many have finished? Result: 50740 started, 9780 finished.

8 https://www.w3.org/TR/rdf-sparql-query/

2. How many women started IT undergraduate courses in Curitiba, Brazil, in 2021? Result: 1391.
3. How many women finished IT bachelor's degrees in public universities in Brazil in 2021? Result: 727.
4. Among students who finished an Information System's degree in capital cities in Brazil in 2021, what is the percentage of women? Result: 14,2%.
5. How many women were enrolled in Federal Institutes of Education, Science and Technology on IT courses in Brazil's Northeast in 2021? Result: 1662.

Figure 2

SPARQL query used to answer CQ 3, g the number of women that started IT undergraduate courses in Brazil in 2021



The triplication of the Inep data - a process that translates data to a format we can query and matches the operational ontology - is also an important result of the research. A Data Dictionary was also created that establishes the entities' meanings and classification. It specifies how the columns from the original Inep CSVs were mapped into the ELLAS-CompBRA ontology, along with their description.

4. DISCUSSION

The instantiation of the ELLAS methodology yielded results that demonstrate its ability to contribute to fill the data gap on the presence of women in STEM areas. Using the

ELLAS methodology and the restructured data gathered from the Inep Census, we built an operational ontology that traced the numbers on women in IT courses in Brazil, the ELLAS-CompBRA ontology. The answers we obtained show why this work is so important. If we look at QC 3, we see that less than 2 out of 10 people who graduated in Information Systems in capital cities in Brazil were women – something most people in the field already suspected, and we now have data to make this perception more objective.

Having concrete numbers is essential in order to establish how serious the gender gap is and to start building policies to address it. This work allows us to build a panorama about gender on higher education IT courses in Brazil according to geographical location, institution, category, branch of IT and course – a granularity lacking in SBC reports and other research.

The potential of this work is further enhanced when put in the context of ELLAS, where the structured data will be connected to research from other sources – such as public and private policies related to gender in STEM. For this reason, it is important for all data to be structured in triples, in the same format. This will make data easy to access and work with through the ELLAS Platform in the future.

There is still plenty to be done and we can count on other research within ELLAS to further this work – some already use the methodology for ontology elaboration following ELLAS Methodology. In future works, we foresee a remapping of the fields in data sources to their literal meaning instead of a number pointing to a meaning specified on the Data Dictionary. For example, the field <academic_degree> contains numbers one through four, whose meaning is established on the Data Dictionary instead of the actual values of "Bachelor" or "Technologist".

It's important to mention that the current ontology was built for a very specific context – IT courses in Brazil – and extrapolating it is one of the most important points of development. Another central point of improvement is the definition of a strategy to deal with time linearity, a task that will require substantial updates on the ontology.

Other points of improvement are the inclusion of remote courses and the improvement of naming conventions for IRIs and IDs, as well as aligning these with terms used on already established ontologies. That will allow us to connect to other LOD data available in the Semantic Web and enable more complex analysis and usage by people and applications.

REFERENCES

Berardi, R. C. G., Auceli, P. H. S., Maciel, C., Davila, G., Guzman, I. R., & 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* (pp. 124-135). Porto Alegre: SBC. https://doi.org/10.5753/wit.2023.230764

De Almeida Falbo, R. (2014). *SABiO: Systematic Approach for Building Ontologies*. Onto. Com/odise@Fois, 1301. https://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf

Isotani, S., & Bittencourt, I. I. (2015). *Dados Abertos Conectados: em Busca da Web do Conhecimento*. Novatec.

Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: a guide to creating your first ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05; Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn methodology for ontology engineering. En *Ontology Engineering in a Networked World* (pp. 9-34). Springer Berlin Heidelberg.