International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

A Context-Independent Ontological Linked Data Alignment Approach to Instance Matching

Armando Barbosa, Federal University of Alagoas, Brazil
Ig I. Bittencourt, Federal University of Alagoas, Brazil
Sean W. Siqueira, Federal University of the State of Rio de Janeiro, Brazil
https://orcid.org/0000-0002-0864-2396
Diego Dermeval, Federal University of Alagoas, Brazil*
Nicholas J. T. Cruz, Federal University of Alagoas, Brazil

ABSTRACT

Linking data by finding matching instances in different datasets requires considering many characteristics, such as structural heterogeneity, implicit knowledge, and uniform resource identifier-oriented (URI-oriented) identification. The authors propose a context-independent approach to align linked data through an alignment process based on the ontological model's components and considering data's multidimensionality. The researchers experimented with the proposed approach against two methods for aligning linked data in two datasets and evaluated precision, recall, and f-measure metrics. The authors also conducted a case study in a real scenario considering a Brazilian publication dataset on computers and education. This study's results indicate that the proposed approach overcomes the other methods (regarding the precision, recall, and f-measure metrics), requiring less work when changing the dataset domain. This work's main contributions include enabling real datasets to be semi-automatically linked and presenting an approach capable of calculating resource similarity.

KEYWORDS

Domain-Independent, Instance Matching, Linked Data, Linked Open Data, Ontology Alignment, Schema-Independent, Semantic Web

INTRODUCTION

Publishing or maintaining Linked data on the Web goes beyond making datasets available through resource description framework (RDF) serializations, which is the innovations and applications cornerstone of semantic web and information systems (Avila-Garzon, 2020). Then, newly published data must be linked to other existing datasets. However, creating links between datasets requires careful analysis by an expert, which, despite being an effective approach, is not scalable, given that the amount of data published is constantly increasing. Consequently, the manual publishing process is unviable. Therefore, to efficiently build the Web of data, there must be solutions capable of linking data automatically or semi-automatically.

DOI: 10.4018/IJSWIS.295977
*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Automatically linking data is a problem recognized by many communities. In Databases, the problem is known by record linkage (Gu et al., 2003; Karr et al., 2019), which aims to identify and link resources that are judged to represent the same real-world entity. Additionally, it is possible to find other terms for this problem, such as the entity resolution problem (Menestrina et al., 2005; Ebraheem et al., 2017; Wu et al., 2020), deduplication (Sarawagi and Bhamidipaty, 2002; Xu et al., 2017; Yang et al., 2019), and Instance matching.

Instance matching is the term that the Linked data community uses to refer to the problem. In this community, the main goal is to find matching instances in different datasets (Abubakar et al., 2018). However, instance matching has additional characteristics (Castano et al., 2011; Mountantonakis & Tzitzikas, 2019; Azmy et al., 2019), such as (i) structural heterogeneity, which refers to variation in the structure of the instances; (ii) implicit knowledge, which refers to the characteristics and constraints exhibited by the domain; and (iii) URI-oriented identification, which refers to reusing URIs to identify new information about existing instances. Thus, there is a need for specific solutions for the correct execution of the instance matching process.

To identify and link resources on the Web, the community has been developing a growing number of solutions. The Ontology Alignment Evaluation Initiative (OAEI) conducts an annual evaluation consisting of aligning two predefined datasets and comparing the alignment generated by the solution with the reference alignment. However, according to Homoceanu et al. (2014), the solutions are not ready to automatically align data despite the good results. Most works are used only on conventional OAEI datasets with small ontologies (Ferranti et al., 2021), and there is a small number of real-world ontology matching application approaches (Otero-Cerdera et al., 2015; Ferranti et al., 2021). Also, no technique stands out from the others in all aspects (Xue & Tang, 2017).

This study proposes a context-independent approach for the alignment of Linked data through an alignment process that considers aspects of the ontological model's data and characteristics. Data properties and relationships drive the alignment of resources/instances. For this purpose, a cascade alignment approach is proposed. Moreover, the proposed approach addresses the alignment between real datasets, which enables reliable alignment of datasets distributed on the Web. This work provides the following contributions: i) development of a context-independent process for the alignment of Linked data; ii) enabling the execution of the alignment directly in the data storage; and iii) presenting a real-world case study dealing with heterogeneity and data quality issues.

Then, this research targets the following problem:

General Problem: How to determine that two instances refer to the same real-world entity? Currently, strategies based on similarity, learning, rules, and context (Castano et al., 2011; Abubakar, 2018) have been used to solve the problem. However, these instance-matching tools are not ready to reliably align real-world data automatically (Otero-Cerdera et al., 2015; Ferranti et al., 2021; Homoceanu et al., 2014).

Thus, the following specific research questions arise:

RQ1: How can the effectiveness of instance-matching tools be improved?
RQ2: How effective is the solution in a real-world scenario of instance-matching?

The first study is experimental. Its general objective is to evaluate the effectiveness of instance-matching tools (Risk Minimization based Ontology Mapping - RiMOM, AgreementMakerLight – AML, and the proposed approach), using real-world data (based on a specific dataset from the OAEI benchmark). The experiments' purpose is to show that the proposed approach – despite not containing computations specifically developed for the datasets – can effectively create instance matches. The experiment evaluates the instance matching tools considering precision, recall and f-measure, and used an OAEI dataset (DOREMUS task[1]) in which both RiMOM and AML were previously tested (IM- OAEI 2016).

Then, the authors performed a real-world case study using the proposed approach, matching information about publications and researchers' curriculum, publishing linked data and answering competence questions raised during requirements gathering.

This article is organized as follows. The first section (Related Work) compares this research with related works. The next section (Alignment Process) presents the alignment process proposed in this paper, describing the process's steps and implementation. The next section (Experiment) depicts the experimental design conducted in this research. Then, the authors present the case study conducted in a real-world scenario using the approach proposed in this paper. The final section presents the concluding remarks of this paper.

RELATED WORK

Zamazal (2020) presents a survey of ontology benchmarks for semantic web ontology tools. More specifically, Abubakar et al. (2018) present a literature review on instance-based ontology matching. They described a general architecture, where the ontologies to be compared are loaded, and a mechanism, for instance-matching is performed, then there is a similarity calculation that supports creating relationship matching, and the mappings are recorded. Abubakar et al. (2018) also presented the best performed OAEI instance matching track participants from 2010 to 2016, showing Risk Minimization based Ontology Mapping (RiMOM) consistency and AgreementMakerLight (AML) surging in 2016.

Table 1. Related work comparison summary

	AML	CoSum-P	LIMES	LINDA	MinoanER	MINTE	PARIS	RiMOM	Silk	WebPie	Proposed
Focus on instance (I) or entity (E) matching	E	I	I	I	I	E	E	E	E	E	E
Number of datasets processed by execution	2	2	2	2+	1?	2	2+	2	2	1?	2+
Ontology-based (B) or directed by ontology (D)	D	D	D	D	D	D	D	D	D	D	B
Specific algorithms for the datasets or exemplars (Y=Yes;N=No)	Y	N	Y	N	N	N	N	Y	N	N	N
Tested on an OAEI benchmark (IM or KG)	IM	N	N	KG	N	N	KG	IM	N	N	IM
Real-world dataset (Y=Yes;N=No)	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
Real-world case study (Y=Yes;N=No)	N	N	N	N	N	N	N	N	N	N	Y

Mountantonakis & Tzitzikas (2019) present a survey on large-scale semantic integration of linked data. The authors show eight scalable tools performing instance matching: Silk (Volz et al., 2009), LIMES (Ngomo and Auer, 2011), PARIS (Suchanek et al., 2011), WebPie (Urbani et al., 2012), LINDA (Böhm et al., 2012), MinoanER (Efthymiou et al., 2016), CoSum-P (Zhu et al., 2016) and MINTE (Collarana et al., 2017).

Table 1 shows a summary of related work. In this table, the focus on instance matching is considered when the work only describes dealing with the instances, not considering possible matchings between classes and instances (here called entity matching). The number of datasets is related to the input that the work mentioned. Although the works adopt different approaches, all of them consider the ontology or graph structure (directed by ontology), different from the work presented in this paper, based on the ontology structure.

Although the tools can find matches between instances, they are still deficient in terms of some criteria, such as their use of specific computations for the benchmark, in addition to their minimal utilization of ontologies, which are used only for metadata generation, intending to choose between the matching approaches available. Unlike the tools mentioned, the proposed tool uses ontologies to guide the process of instance matching. Additionally, this approach allows the user to define how the alignment should be performed.

Another difference between the proposed tool and previously developed tools is the cascade alignment, which uses instances of concepts related to the concept whose instances will be aligned. Cascade alignment goes beyond instances – it explores the existing relationships. From this, it is possible to find new matches. Additionally, the proposed tool allows matches between instances to be stored directly in the triple store database in which they are stored.

As evaluation collections, Mountantonakis & Tzitzikas (2019) presents the OAEI benchmark and Daskalaki et al. (2016) benchmarks. Both works agree that the most important challenge for judging the performance of instance matching techniques and tools is the OAEI. Initially, the OAEI evaluated only ontology alignment tools, beginning with evaluating solutions for aligning data in 2009. From 2009 to 2017, the OAEI track focused on instance matching was called instance matching (IM) track, and after that, a track for instance and schema matching related to knowledge graphs was introduced, called Knowledge graph track.

Table 2 shows the tools that participated in the OAEI IM and KG tracks from 2009 to 2020, where RiMOM and LogMap have the most appearances.

Complementary, Table 3 shows a brief overview of the results from OAEI tracks related to instance matching from 2016 to 2020, considering the f-measure. Three tools are highlighted: AML, LogMap and RiMOM. Until the 2016 edition, RiMOM presented the best results in general in the IM track. From 2016 to 2020, AML is the only tool that participated in all editions. Although AML results for 2017 to 2020 are not the best ones, it has good results compared to the others.

Despite all tools presented by Mountantonakis & Tzitzikas (2019) and the ones participating in OAEI editions, AML and RiMOM were selected to compare with the work presented in this paper due to their prominence and consistency in OAEI, specifically in the section regarding IM or, more recently, the Knowledge Graph track.

RiMOM (Li et al., 2009; Zhang et al., 2016) is a tool for aligning Linked data. It implements a considerable number of approaches for alignment, the choice of which is made based on the metadata extracted from the ontology. Furthermore, RiMOM-2016 uses an inverted index to index the objects and generate candidate pairs for possible alignment. Pairs are generated when two resources share at least one predicate and one object. RiMOM-2016 uses the ontologies only to align the properties but also as input for metadata generation. RiMOM was one of the main instance matching tools until 2016.

AML (Faria et al., 2016) is an ontology alignment tool initially based on lexical similarities and techniques, emphasizing the use of external sources as a background. AML relies on three alignment algorithms for instance matching: HybridStringMatcher, ValueStringMatcher, and Value2LexiconMatcher. The first algorithm uses several approaches (comparisons between sentences

Table 2. Tools that participated in OAEI Instance Matching and KG Tracks from 2009 to 2020

Tools	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
AFlood	✓											
AGM											✓	
ALOD2Vec												✓
AM or AML			✓					✓	✓	✓	✓	✓
ASMOV	✓	✓										
ATBox												✓
Baseline, baselineAltLabel, baselineLabel										✓	✓	✓
CODI		✓	✓									
DESKMatcher												✓
DOME										✓	✓	
DSSim	✓											
EXONA							✓					
FBEM	✓											
FCAMap-KG											✓	
HMatch	✓											
Holontology										✓		
I-Match									✓			
InsMT						✓	✓					
Legato									✓			
Lily					✓		✓					
LN2R		✓										
LogMap and variants				✓	✓	✓	✓		✓	✓	✓	✓
NjuLink									✓			
ObjectCoref		✓										
POMAP++											✓	
RiMOM	✓	✓			✓	✓	✓	✓				
SBUEI				✓								
semsim				✓								
SERIMI			✓									
SLINT+					✓							
STRIM							✓					
Wiktionary												
Zhishi.links			✓								✓	✓

and between words) to generate the similarity, and this hybrid approach also utilizes WordNet. The second algorithm uses value mapping to calculate the similarity, penalizing pairs in which annotations or data properties are not the same. Finally, the third algorithm unites the other two approaches. Although AML has different alignment algorithms in the tool, they all work only at the data level. Consequently, the characteristics of the properties are disregarded throughout the matching process. AML has presented one of the top results in the OAEI instance matching track and Knowledge Graph Track since 2016. While RiMOM was one of the main instance matching tools until 2016, AML has been one of the top tools since 2016 in the Knowledge Graph Track, which deals with instance matching.

Table 3. A brief overview of the OAEI tracks related to instance matching from 2016 to 2020

System	Instance Matching Track					Knowledge Graph Track		
	2016			2017		2018	2019	2020
	9HT	4HT	FPT	HT	FPT			
AGM	-	-	-	-	-	-	0.25	-
ALOD2Vec	-	-	-	-	-	-	-	0.87
AML	0.946	0.848	0.886	0.613	0.582	0.23	0.71	0.85
ATBox	-	-	-	-	-	-	-	0.84
Baseline	-	-	-	-	-	0.69	-	-
baselineAltLabel	-	-	-	-	-	-	0.84	0.84
baselineLabel	-	-	-	-	-	-	0.81	0.81
DESKMatcher	-	-	-	-	-	-	-	0.82
DOME	-	-	-	-	-	0.61	0.70	-
FCAMap-KG	-	-	-	-	-	-	0.84	-
Holontology	-	-	-	-	-	0.00	-	-
I-Match	-	-	-	0.129	0.101	-	-	-
Legato	-	-	-	0.930	0.990	-	-	-
LogMap	-	-	-	0.556	0.210	0.14	0.00	0.54
LogMapBio	-	-	-	-	-	0.00	0.00	0.00
LogMapIM	-	-	-	-	-	-	-	0.54
LogMapKG	-	-	-	-	-	-	0.54	0.54
LogMapLt	-	-	-	-	-	-	0.67	0.67
NjuLink	-	-	-	0.955	0.946	-	-	-
RiMOM	0.813	0.746	0.707	-	-	-	-	-
POMAP++	-	-	-	-	-	-	0.00	-
Wiktionary	-	-	-	-	-	-	0.79	0.87

Alignment Process

The process consists of four main steps: selecting datasets, identifying concepts, listing resources, and aligning data. Each step of the process will be described in the following subsections.

Step 1: Selecting Datasets

The step involving selecting datasets aims to determine which datasets will be aligned. As the scope of the process is Linked data, ontologies/vocabularies can support the data modeling in publishing processes. Then, the dataset is structured in triples and uses concepts modeled on ontologies/vocabularies.

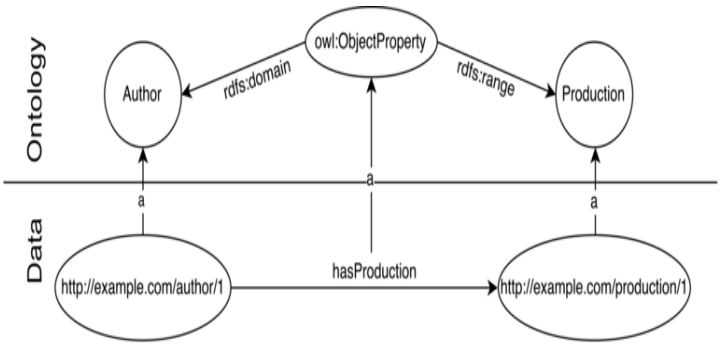
Step 2: Identifying Concepts

After choosing the datasets, Step 2 consists of choosing the main and related concepts. Then, two SPARQL queries were developed. The first query explores the ontology, especially the *rdfs:domain* and *rdfs:range relationships* of the object properties (see Code 1), whereas the second query explores the data and relationships established by the instances. In the query (Code 1), line 4 retrieves all the ontology or vocabulary concepts. In line 5, a restriction is applied, where the concepts must be in the domain or range of a relationship. Consequently, an instance of this concept will be the subject or object of a triple (see Figure 1).

Code 1. SPARQL query to identify a concept

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
#retrieve the concepts presented in the ontology/vocabulary,
#whether they are subjects or objects of a triple
select distinct ?Concept count(*) as ?count where {
#retrieve the concepts
[]a ?Concept.
#the concept must be a domain or range of a relation (RDF triple)
?Concept (^rdfs:domain|^rdfs:range) ?o.
}
#group by concept
Group by ?Concept
#defining descendent ordination
Order by desc(?count)
```

Figure 1. The relation between ontology and data



The query presented in Code 2 comprises two parts because the concept can model instances that are the subject or object of a relationship. In the first part, the selected concept represents the subject of the triple. It is possible to retrieve the concepts that model the related instances (objects)

Code 2. SPARQL query to retrieve related concepts

```
# This part is not part of the code
#
# http://example.com/concept/uri is an example of URI
# https://dbpedia.org/ontology/Company in case testing with dbpedia data
# https://dbpedia.org/sparql DBpedia endpoint
#
#########################
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
# retrieving subject's or object's type from a RDF triple of the ontology
select distinct ?type where {
values ?Concept{<http://example.com/concept/uri>}
# retrieving object's type
?instance rdf:type ?Concept; ?p ?o.
?p rdf:type owl:ObjectProperty.
?o rdf:type ?type.
union
# retrieving subject's type
?s ?p ?instance.
?p rdf:type owl:ObjectProperty.
?o rdf:type ?Concept.
?s rdf:type ?type.
```

using the instances' relationships. In the second part, the inverse occurs: the concept represents the triple's object, and the concepts that represent the subjects are retrieved.

As a result of Code 2, a list containing the concepts related to the (main) concept chosen is provided. At this point, users must choose which related concepts they want to use to improve the alignment of the chosen concept. This decision will influence both the time that the process will take to conclude and the number of resources aligned at the end of the process because, for each related concept, there will be a new execution of steps (iii) and (iv). This loop is necessary because some alignments will only be possible through the relationship between these concepts.

Step 3: Listing Resources

The step of listing resources can be understood as retrieval of the resources about the concepts. It is important to highlight that the listing/retrieval of resources from the knowledge database can be executed more than once during the process, which generates a set of resources for each concept chosen. Additionally, this step is responsible for generating candidate pairs, in which the resources of Dataset D1 are compared with the resources of Dataset D2.

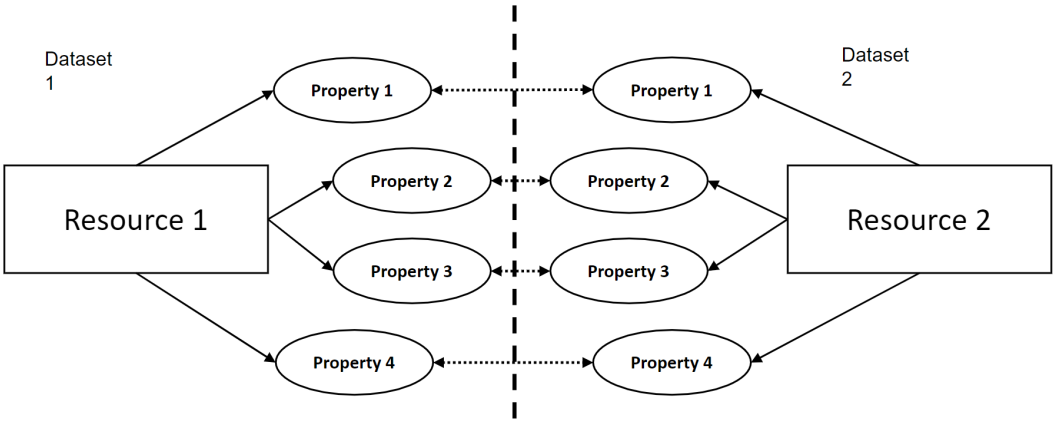
Step 4: Aligning Data

The data alignment step is divided into two activities, (i) simple alignment and (ii) cascade alignment, detailed in the following sections.

Simple Alignment

It is necessary to perform procedures to align the resources, including data processing, resource comparison, and similarity analysis. The first procedure – data processing – refers to transformations in the properties of the resources. These transformations are necessary to assist the similarity algorithms to analyze the similarity between the resources better. In the comparison procedure, each of the properties is analyzed. If a property does not pertain to one of the resources, it is exempted from the comparison. Figure 2 shows a comparison between the properties of each resource.

Figure 2. Comparison between resources



Two equations are used to define the similarity between the instances. Equation (1) defines the set of properties considered during the comparison between the resources. The set is obtained from the difference between the largest set of properties and the set that should be disregarded. Therefore,

$$Pf = Max(Pr1, Pr2) - Pd \quad (1)$$

where:

- $Pr1$ – Resource 1 set of properties;
- $Pr2$ – Resource 2 set of properties;
- Pd – Set of properties that must be not considered;
- $Max(Pr1, Pr2)$ – Retrieve a set with the maximum number of properties;

Equation (2) concerns the similarity function between resources – this equation can be understood as the mean of the similarities between two resources. This approach was chosen not to favor any of the partial similarities. However, there may be other more appropriate functions for calculating the similarity between resources.

$$SR = \frac{1}{\left|P_f\right|}\sum_{i=1}^{P_f} S\left(V\left(R_1, P_f\left[i\right]\right); V\left(R_2, P_f\left[i\right]\right)\right) \tag{2}$$

where:

- S – Similarity function;
- V(R,P) – Property value of P in a resource R;
- *R*1 – Resource 1;
- *R*2 – Resource 2.

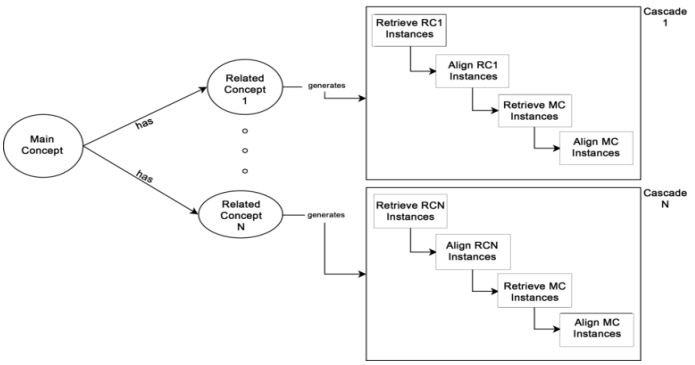
Cascade Alignment

Cascade alignment is named so because of the linkage between the necessary activities: (i) retrieving instances that pertain to the related concept, (ii) aligning instances that pertain to the related concept, (iii) retrieving instances that pertain to the main concept, and (iv) aligning instances that pertain to the main concept. The name is also intended to refer to the cascade development model (Royce, 1987), the first software development model.

The cascade development model and the cascade alignment approach share some similarities. The shared similarities include how the activities are executed, which is sequential. Additionally, each activity can only begin when the previous activity is completed. Another shared characteristic is the fact that the whole project is planned before execution.

Unlike a project that uses the cascade model, in which the entire project must be completed in the final stage, it is assumed that matching between instances will only be considered to be concluded when all the related concepts have been used in the matching process. It is worth highlighting that a "cascade" is generated for each related concept selected. Figure 3 shows the relationship between the concepts and the cascade.

Figure 3. The relation between a related concept and cascade

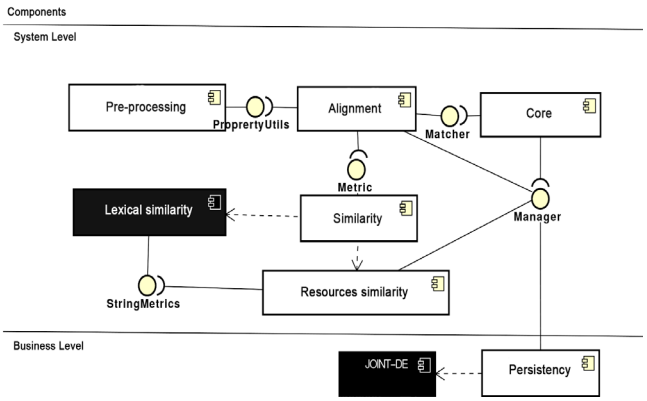


Let us imagine that someone wants to discover which authors are present in two datasets simultaneously. For this, he/she wishes to use the papers registered in both databases. Thus, the main concept and the related concept are the Author and Publication, respectively, Retrieving Instances from a Related Concept, Aligning Instances of the Related Concept, Retrieving Resources of the Main Concept, and Aligning Retrieved Resources.

Implementation of the Process

In the process implementation process definition, the authors relied on existing literature reviews (Feitosa et al., 2018; Barbosa et al., 2021). Some components were developed to perform the proposed process: similarity analysis, data persistence, alignment generation, logic between the steps, and pre-processing (see Figure 4). Some components have been reused from other work (in black); an example is the lexical similarity, which contains several algorithms for detecting similarity between texts. Other components are newly developed (white), such as those responsible for detecting resource similarity, alignment, etc.

Figure 4. Components used in the process implementation



Although various solutions contain algorithms for the calculation of similarity and alignment between resources, the development of an approach that contemplates problems faced when working with real databases (e.g., accentuation, absence of properties, and formatting) was chosen (Castano et al., 2011; Ferrara et al., 2008).

The pre-processing component's function is to perform treatments on the texts that will be applied to the similarity function. Some examples of treatments are the treatment of accents and punctuation.

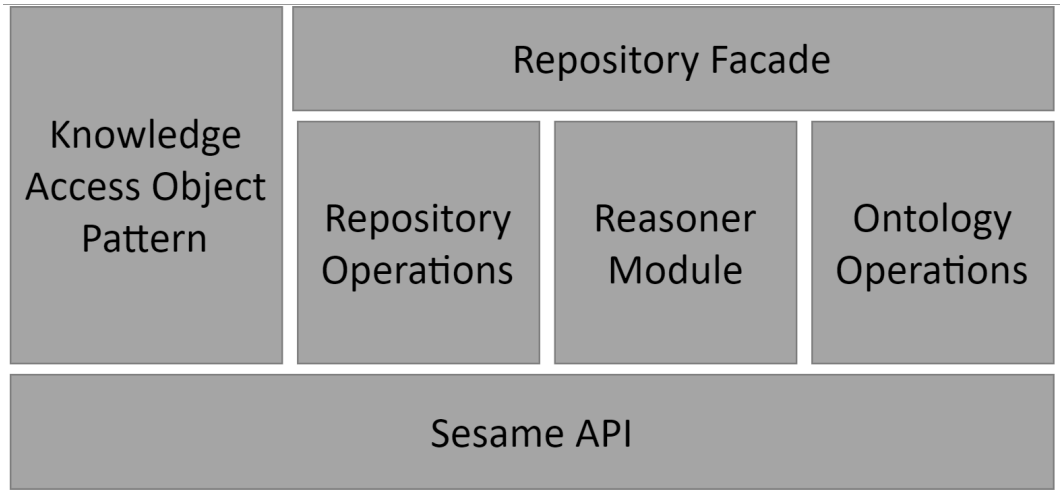
The similarity component is divided into two subcomponents: lexical similarity and resource similarity. The first uses metrics that analyze the similarity between words and texts, which include the Levenshtein (1966), cosine (Singhal, 2001), and Jaro-Winkler (Winkler, 1990) techniques, amongst others. The second component, which refers to the resource similarity, uses the first component, and its function is to generate the similarity between resources. An approach based on the semantic subgraph technique (Wang and Xu, 2009) calculates the resources' similarity. In practice, a semantic subgraph refers to the triples related to any resource and agrees with the ontological modeling. Figure 1 represents a subgraph that relates an author and the author's publications. In addition to object properties, a subgraph also has data properties associated with both the main resource (author) and related concepts (Production).

The alignment component is responsible for determining – in accordance with the values obtained in the similarity step – whether the analyzed resources are related to the same real-world entity. This component utilizes acceptance thresholds, which are determined beforehand, to determine whether alignment must be performed. For this reason, the alignment process is not an automated task because it requires the values to be adjusted. There are various methods for determining the threshold value, ranging from executing multiple times and analyzing the best cost/benefit between precision and recall to using techniques that update the threshold value dynamically.

The core component is responsible for concentrating and coordinating the settings during execution. Currently, the core component has three modalities for matching between instances: simple alignment, which is executed in all modalities and can also be executed independently; cascade alignment, which is performed when a concept related to the concept whose instances are to be aligned is chosen; and multi-cascade alignment, which occurs when more than one related concept is chosen.

The persistence component is responsible for materializing the matches found by the alignment component. For this, JOINT is used (see Figure 5). According to Holanda et al. (2013), JOINT is a framework to facilitate the development of ontology-based applications. The features presented by the JOINT tool allow operations (Virtuoso, OWLim, etc.) to be performed directly on the triple server. Additionally, this tool also supports SPARQL queries' execution, which, through a translation system, transforms the triples into Java language objects.

Figure 5. JOINT tool architecture



The Experiment

This study presents a context-independent data alignment process consolidated into a tool for performing instance matching. In this experiment, the evaluation of instance matching tools uses two datasets from OAEI 2016 Instance Matching (IM) track and their reference alignments. OAEI 2016 IM track was the only edition combining RiMOM and AML tools, while DOREMUS tasks were chosen because they contain real-world data from two major French cultural institutions. The two datasets used in the experiment were: nine heterogeneities (9-heterogeneities) and false positives trap (falsepositives-trap), as the first considers different types of heterogeneities and the second focuses on false positives matching.

Figure 6. Experiment execution process

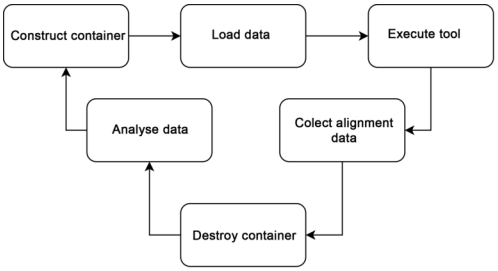


Table 4. Summary of precision, recall, and f-measure data per scenario.

Scenario	Tool	Precision	Recall	f-measure
C1	Proposal	1	0.875	0.933
	AML	0.966	0.875	0.918
	RiMOM	0.813	0.813	0.813
C2	AML	0.921	0.854	0.886
	Proposal	0.906	0.707	0.794
	RiMOM	0.707	0.707	0.707

The experiment looks to evaluate if the proposed approach improves instance-matching tools' effectiveness, then it is related to RQ1 (How can the effectiveness of instance-matching tools be improved?).

Formally, the objective of this experiment can be defined as follows: to analyze instance matching tools to compare them in terms of their effectiveness from the point of view of generating matches between instances – in the context of data alignment between datasets – to use the best approach in a real-world case study.

The instance matching tools (AML, RiMOM-2016, and the proposed tool) were compared in terms of effectiveness (Precision, Recall and F-measure).

- Precision (P): In instance matching, this variable indicates the number of relevant matches concerning all the matches generated by the tools.
- Recall (R): In instance matching, this variable indicates the number of relevant matches concerning the set of all possible matches (mirror).
- F-measure (F): Harmonic mean between precision and recall. The purpose of this variable is to transform (P) and (R) metrics into one.

Considering the different experiment classifications (Montgomery, 2017), the present experiment was classified as complete factorial with blocking. The blocking was chosen to suppress the effects of the datasets on the response variables. For each scenario, all the tools were executed, thus ensuring the completeness of the experiment.

Figure 6 shows the execution steps of each alignment:

- Construct container with the settings reset;
- Load data;
- Execute tool inside the container;
- Collect alignment data;
- Destroy container;
- Analyze data.

The experiment has two possible scenarios to evaluate the instance matching tools, with one execution per tool, thus totaling six executions. Scenario C1 was applied on the 9-heterogeneities dataset, and scenario C2 was applied on the falsepositives-trap dataset.

The following instruments will be used to conduct the instance-matching experiment:

- Intellij IDEA 2016.3 for development of the code and execution of the proposed tool;
- Virtuoso RDF Store - 07.20.3217;
- OpenJDK 64-Bit Server VM (build 25.111-b14, mixed mode).

The entire experiment was conducted in a container, making it possible to isolate the applications and the executions' effects. Then, the applications were executed in identical environments without generating side effects on each other.

After the execution, the resulting alignments were analyzed based on precision, recall, and f-measure. Figures 7 and 8 show the results corresponding to each of the tools in each of the scenarios. Table 4 summarizes the data obtained for each scenario.

As presented in Table 4, the proposed tool had a precision of 1 for scenario 1 and 0.906 for scenario 2. Given the proposed tool's results in the two alignment scenarios, with a recall of 0.875

Figure 7. Results of the tools for scenario 1

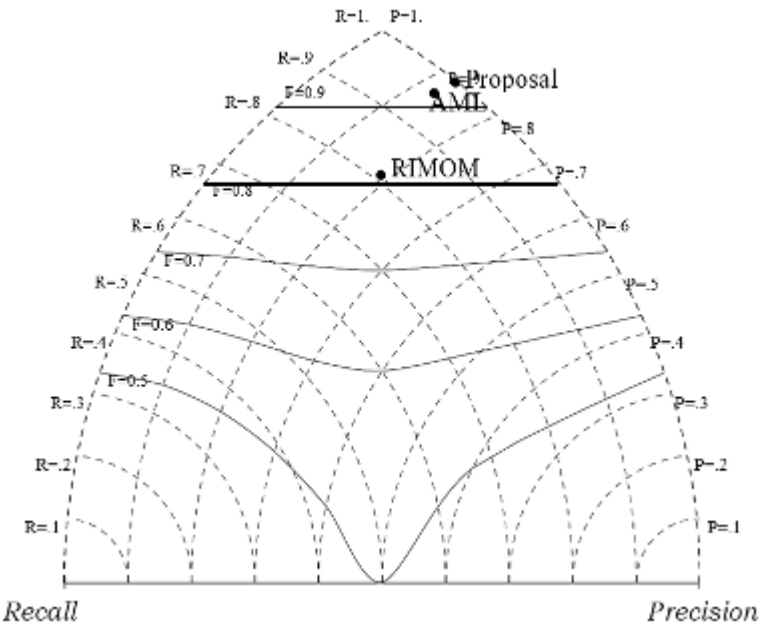
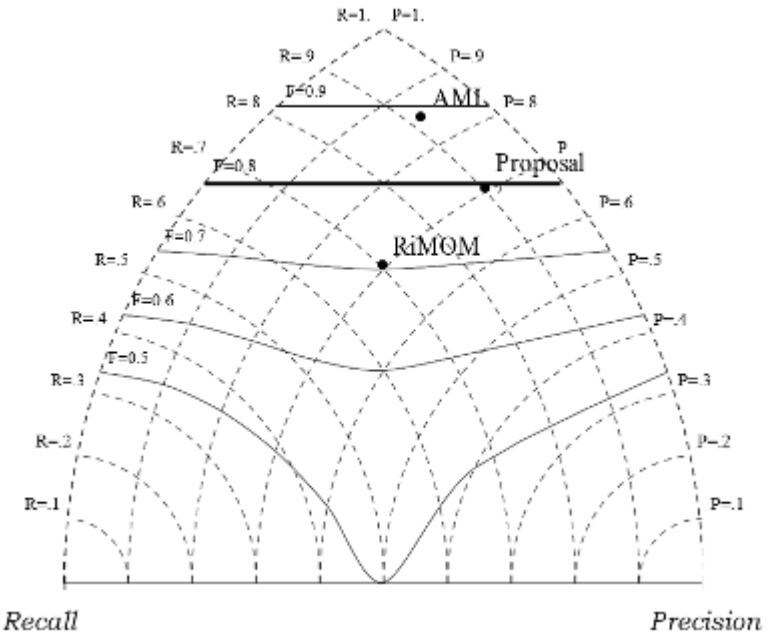


Figure 8. Results of the tools for scenario 2



for scenario C1 and 0.707 for scenario C2, despite the good performance, the proposed tool produced values equal to at least one of the other tools in each of the scenarios.

According to the results presented in Table 4, the proposed tool ranked first in scenario 1, with an f-measure of 0.933, and second in scenario 2, with 0.794.

The researchers used Fisher's test (Fisher, 1922) to compare the pairs of metrics. The results obtained in the experiment were used as input according to the configuration listed in Table 4. As a result, the statistical test yielded a p-value of 0.8333, which indicates that the tools have similar effectiveness in terms of the metrics.

The experiment conducted sought to evaluate the effectiveness of the Linked data alignment tools in terms of precision, recall and f-measure. These variables were evaluated separately in each of the two scenarios, C1 and C2. In scenario C1, the data exhibited nine types of heterogeneity (e.g., multilinguality, difference in the catalogues, phonetic difference, and different degrees of description). In scenario C2, the data exhibited similar instance sets, with only one match possible, which indicates that the other instances are false positives. Given the statistical test results, the tools have similar effectiveness for both of the scenarios analyzed. However, from the analyses performed on the metrics, it was possible to verify that the proposed tool stood out in one of the scenarios, C1. Nevertheless, it should be emphasized that the proposed approach does not use specific implementations for the analyzed datasets, which enables it to be easily used in other contexts.

Although the experiment was designed to minimize possible threats that would compromise its conclusions, some threats should be mentioned. One possible internal threat to the experiment's validity is the selection of the experimental units because the OAEI benchmark provided the datasets used in the experiment, and no other datasets and benchmarks were used.

The experimental units were executed in only one configuration setting and only one version of the tools. It is possible that the number of tools and scenarios is not sufficient for observing significant differences in the effectiveness between the approaches used for instance matching. Additionally, one must consider that the response time was not considered in the experiment.

Table 5. Questions suggested by the community

ID	Questions
Q01	How many Informatics in Education (IE)'s researchers are in the community?
Q02	Who are the IE researchers in Brazil?
Q03	Where are the IE researchers in Brazil? (State)
Q04	Where are the IE researchers working in Brazil? (University)
Q05	Which IE researchers in Brazil are doctors?
Q06	How many IE researchers in Brazil are doctors?
Q07	Which IE researchers in Brazil own a registered trademark?
Q08	Where did the IE researchers in Brazil get their doctorate degree?
Q09	Where did the IE researchers in Brazil get their post-doc?
Q10	How many papers does the author "z" have in the event "x" (SBIE/WIE) in the IE field of study?
Q11	How many papers were published in the event "x" (SBIE/WIE) in the IE field of study?
Q12	How many authors published papers in the event "x" (SBIE/WIE) in the IE field of study?
Q13	How many papers were published in journal "y" (RBIE) in the IE field of study?
Q14	List of doctors that published on RBIE and their e-mails
Q15	List of authors in the IE community in Brazil, with their area of research and e-mail
Q16	List of papers published in RBIE – General.
Q17	How many IE community researchers receive PQ/DT scholarships, and at which level?
Q18	Which are the main fields of study of the IE community?
Q19	Which concepts do the IE researchers in Brazil explore?
Q20	Which are the most researched IE themes in Brazil?
Q21	Which IE researchers in Brazil cooperate with each other?
Q22	Which Brazilian institutions in which the IE researchers work cooperate with each other?
Q23	Which are the related papers published in SBIE, WIE and RBIE?
Q24	How do the concepts exploited in IE papers evolve over time?
Q25	Tendency map for IE researchers in Brazil set in a timeline
Q26	List of productive scholarship IE researchers in Brazil
Q27	In which institutions the IE researchers in Brazil work?
Q28	In which institutions the IE researchers in Brazil work?
Q29	Which authors from the IE community in Brazil published in conference "X"?
Q30	How many IE researchers in Brazil are enrolled in post-graduate computation programs?
Q31	Who are the top experts in digital resources and learning objects in Brazil?

Figure 9. Dac ontology taxonomy

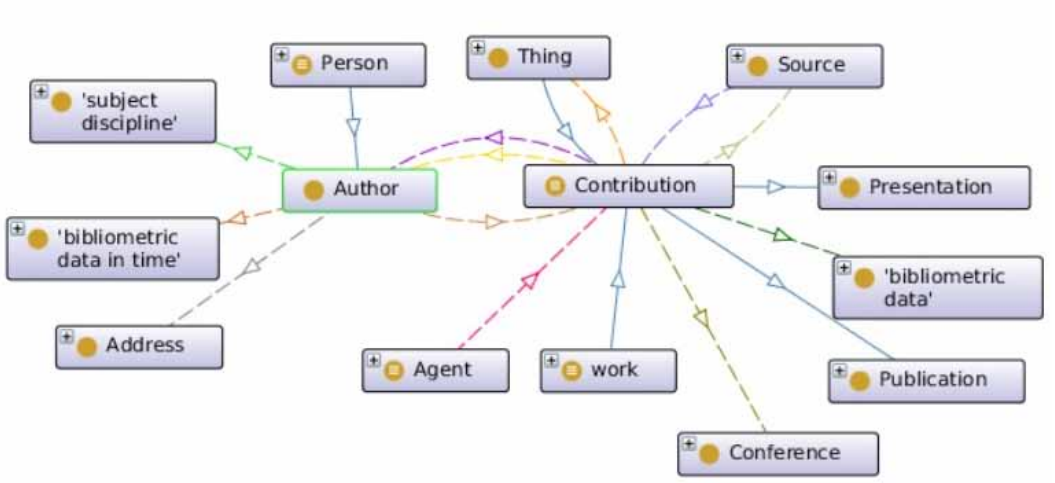
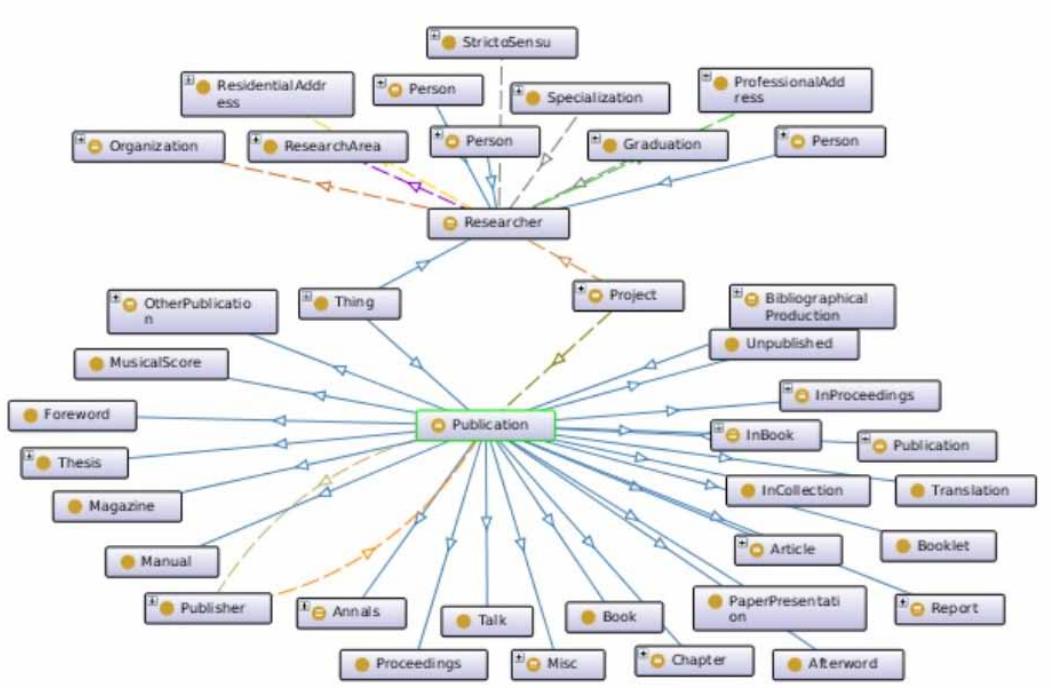


Figure 10. Lattes ontology taxonomy



Owing to the small amount of data per dataset, the number of instances per dataset may not be sufficient to observe significant differences in the associated metrics.

CASE STUDY: BRAZILIAN COMPUTERS AND EDUCATION COMMUNITY DATA

Members of the Special Committee on Informatics in Education (IE) were asked to suggest questions of interest (possible competence questions for the study, which is part of the requirements gathering) to the Brazilian IE community. As a result, a set containing more than 30 questions was produced. Table 5 lists the questions proposed by the members.

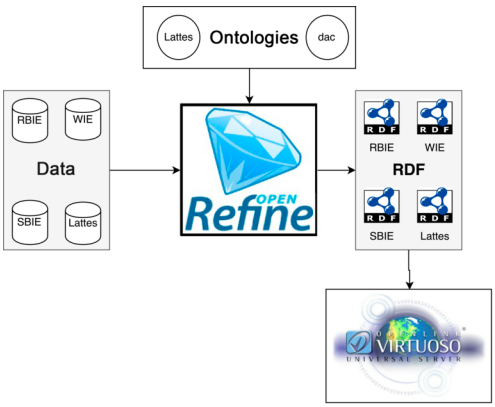
To answer some of the questions asked, it is necessary to cross-reference different sources of information from the researchers and their publications, namely the Brazilian Journal of Informatics in Education (Revista Brasileira de Informática na Educação - RBIE)[2], Workshop on Informatics in School (Workshop de Informática na Escola - WIE)[3], Brazilian Symposium on Informatics in Education (Simpósio Brasileiro de Informática na Educação – SBIE)[4], and the researchers' curriculum, which in Brazil is available at a platform called Lattes[5]. The authors assume there may be a dependence of distinct elements of our datasets (i.e., instance equivalence, class equivalence, and class-instance equivalence) in the same dataset and considering different datasets. It is worth highlighting that the datasets were made available as XML files and it is assumed the datasets are complete, corresponding to the whole population of the corresponding publications and curriculum.

The solution proposed for the data models assumed that the data conform to the schemas (i.e., data from different schemata need to be mapped to a common schema before performing instance matching). Then, the XML files had to be converted to RDF, and dac[6] and Lattes ontologies were used to model the data. The first ontology is intended to model the publishing domain (see Figure 9), whereas the second was constructed to model the lattes domain (see Figure 10).

Another assumption is that the OpenRefine[7] tool can transform the data to RDF, with the extension to support RDF. This tool was selected owing to its ease in creating the transformation templates. After the data transformation, the ontologies and the data were persisted in Virtuoso[8]. Figure 11 illustrates the data conversion process.

With the data conversion process, 1.1 million triples were generated, which were distributed as follows: 96%, or 1,094,307, triples pertaining to Lattes; 1.61%, or 18,363, triples pertaining to

Figure 11. The conversion process to RDF



the SBIE; 1.21%, or 14,601, triples pertaining to the WIE; and 1.1%, or 12,503, triples pertaining to the RBIE.

This research assumed that (i) authors may publish one or more papers in any of the publication datasets; (ii) each paper must have at least one authors; (iii) each paper may have one or more authors; (iv) a paper's author (in the RBIE, SBIE and WIE datasets) may not be registered in the Lattes dataset, thus there may be no corresponding Lattes entity for some authors; and (v) although there may be different authors with similar names, it is possible to find the corresponding entity in the Lattes dataset as it also contains the publications for each author.

Some concerns are related to the data quality in the datasets, i.e., there may be different name formatting (e.g., first name + middle name initials + surname in one instance and first name + last name in another instance); and untrue information (e.g., false e-mail). It is expected that the matching algorithms can support dealing with these issues, but the Lattes dataset is more reliable and should help to solve the inconsistencies.

Execution of the Process

This section describes how each of the steps of the matching process was performed

Selecting Datasets

This step refers to the selection of datasets used as input for the instance matching process. It should be noted that two or more datasets can be selected. Thus, the datasets RBIE, SBIE, WIE, and Lattes were selected.

Identifying Concepts

This step involves selecting the concepts (main and related) that are used in the process. Currently, there is only one restriction regarding the selection of the concepts, in that it is possible to select only one main concept.

Main Concept: To identify the main concept by the user, the query presented in Code 1 was executed. Table 6 presents the results obtained from the execution of the query.

Table 6. Ontology concepts and quantity of instances

Concepts	quantity
http://www.ic.ufal.br/dac/Contribution	868752
http://www.ic.ufal.br/dac/Author	155680
http://www.ic.ufal.br/lattes/DoctoralDegree	2387
http://www.ic.ufal.br/lattes/Graduation	2195
http://www.w3.org/2002/07/owl#Class	1680
http://www.ic.ufal.br/lattes/Course	1186
http://www.w3.org/1999/02/22-rdf-syntax-ns#List	204
http://www.w3.org/2002/07/owl#Restriction	161
http://www.w3.org/2002/07/owl#ObjectProperty	128
http://www.w3.org/2000/01/rdf-schema#Class	60
http://www.w3.org/2002/07/owl#Ontology	24
http://www.w3.org/1999/02/22-rdf-syntax-ns#Property	23
https://purl.org/dc/terms/AgentClass	3

The Author concept, which represents the second-highest number of instances in the data, was selected as the main concept. The concept was chosen not because of the number of instances but for strategic issues as the purpose of this case study was to cross-reference information about researchers.

Related Concept: The query presented in Code 2 was executed to select the related concept that would be used. The Contribution concept was selected and used during the cascade alignment as the related concept returned from the query. Note that more than one related concept can be selected.

Listing Resources

The resource list is generated automatically based on the previously selected concepts. From the list of resources, the candidate pairs are assembled. It is worth noting that in the study in question, a dataset can contain more than one instance for the same real-world entity (e.g., more than one URI for the same researcher). Thus, candidate pairs were generated within the same dataset, which characterized the internal alignment.

Aligning Data

The alignment step is responsible for determining the matches between the instances. In this process, there are two alignment approaches, simple and cascading. In the simple approach, the resources are compared directly, utilizing the properties and their characteristics. In the cascading approach, the resources are compared based on the related resources.

Main Concept: As in other instance matching approaches (Zhang et al.., 2016), functions that analyze the similarity between two resources were also used. The function presented in Code 2 was used to determine the similarity between the pairs. This similarity function generates values between 0 and 1, with 0 indicating totally distinct and 1 denoting equal. In addition to the function used, thresholds were defined. It means that similarity values greater than the threshold were considered to be matching.

Initially, the threshold value was defined arbitrarily and later adjusted with the help of tests. So, the same dataset was aligned several times using a threshold value for each execution. Finally, the threshold value was set to 0.88.

Related Concept: Cascade alignment consists of aligning instances of the main resource based on related resources. This step of the process is performed for each of the related concepts selected in the concept identification step, and it consists of three activities:

- Aligning related resources: simple alignment is performed between the instances that pertain to the related concept;
- Retrieving instances of the main concept: based on the alignment between instances of the related resources, the instances that pertain to the main concept are retrieved;
- Aligning instances of the main concept: based on the retrieved instances, new candidate pairs are generated and become input for the simple alignment.

Results

The results presented in this section are separated into two parts. The first part consists of the alignments generated, whereas the second part addresses the answers to the community's questions.

Alignments

After alignment by the tool, a survey was conducted (a crowdsource with Brazilian researchers in Computers in Education), from which it was possible to generate information – such as the number of resources that are repeated in the databases, the total number of resources aligned with the Lattes profile, and the precision, recall, and f-measure (Goutte and Gaussier, 2005) – to analyze the reliability

Table 7. Alignment results concerning Lattes data. * with repetitions. ** with no repetitions

Dataset	RBIE	SBIE	WIE
Initials	1118	1687	1952
Finals	806	1032	1098
Lattes Profiles*	92.03% (1029)	71.54% (1207)	75.20% (1468)
Lattes Profiles**	89.06% (717)	70.16% (792)	67.48% (741)
P | R | F	0.97 | 1 | 0.98	0.94 | 1 | 0.97	0.84 | 1 | 0.91

Code 3. Query to retrieve the number of researchers per Brazilian State (UF)

```
PREFIX dac:<http://www.ic.ufal.br/dac/>
PREFIX lattes:<http://www.ic.ufal.br/lattes/>
# retrieving researchers' quantity by Brazilian state (UF)
SELECT ?uf count(distinct ?o) as ?total
# data source
FROM <http://www.ic.ufal.br/dac/wie/>
FROM <http://www.ic.ufal.br/dac/author/wie/lattes/alignments/>
FROM <http://www.ic.ufal.br/dac/author/wie/wie/alignments/>
FROM <http://www.ic.ufal.br/dac/lattes/>
WHERE {
# retrieving researchers
?s a dac:Author; owl:sameAs{,5} ?o.
# subject must be in wie dataset
filter regex(?s,"http://www.ic.ufal.br/dac/author/wie/(\\d)+$","i")
# object must be in lattes dataset
filter regex(?o,"http://www.ic.ufal.br/dac/author/lattes/(.)+$","i")
# there might be no triple/relation in alignment database pointing to the subject
filter not exists {
graph<http://www.ic.ufal.br/dac/author/wie/wie/alignments/>{?k owl:sameAs ?s}
# retrieving researchers' professional address
?o lattes:hasProfessionalAddress ?address.
?address lattes:uf ?uf
# ensuring owl:SameAs type relation
filter exists {?elem owl:sameAs ?g}
# grouping by Brazilian state (UF)
group by ?uf
# descendent ordenation
order by desc(?total)
```

of the alignments (see Table 7). It is worth noting that the reference alignment was generated manually with the help of domain experts.

Responses

The instance matching process identifies instances that refer to the same entity and allows complementary information to be integrated. Thus, it was necessary to consult more than one database simultaneously to answer the community's questions. Owing to the number of questions asked, only a few of them will be presented below.

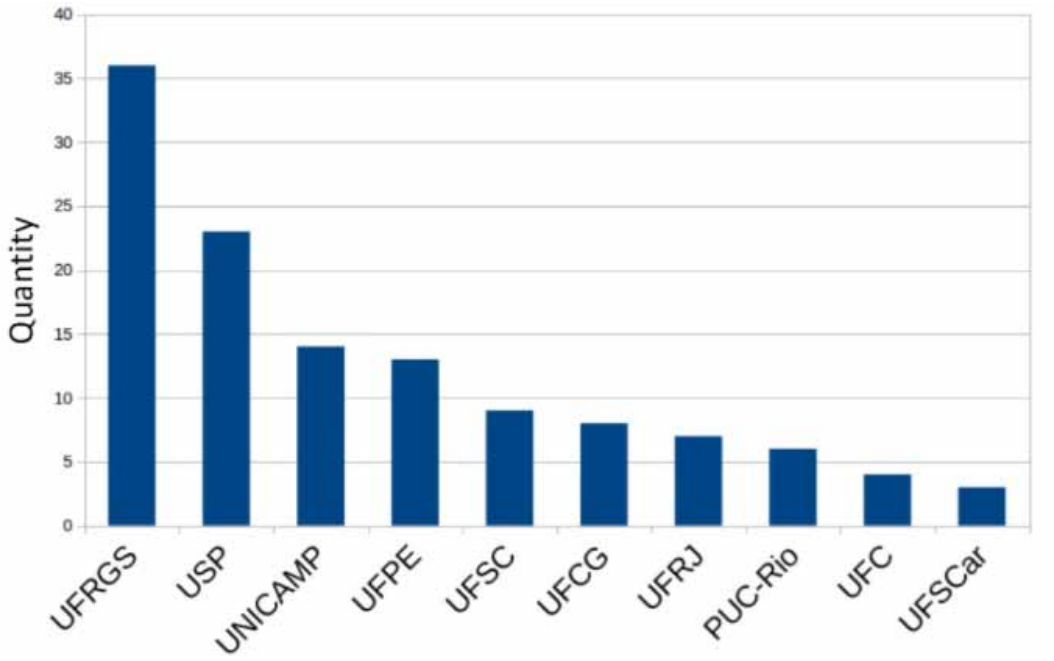
Another factor that should be highlighted is the problem with the data that were provided by the authors (such as multiple authors for the same paper and different name formatting). Additionally, untrue information was provided – e.g., 77 authors used the same e-mail (URL: author@email.com) to avoid spam.

Q1 – Where in Brazil (states) are IE researchers located?

Through this question, it is possible to know where the researchers who published in the RBIE, SBIE, or WIE work. This information is obtained through the professional address found in the Lattes curriculum profile of each researcher. Thus, to answer this question, it is necessary to identify these researchers' profiles in the Lattes curriculum.

Query presented in Code 3 retrieves the professional address of the researchers who published in the WIE (for other publication venues, it is easily adapted). In line 9, the *owl:sameAs* transitivity property is used to retrieve all of the matching profiles. For the query not to loop, a query of up to five elements composing the transitivity was established. The five elements were chosen manually, and with this value, it was possible to attain all the possible properties through transitivity.

Figure 12. The number of researchers per university where they got their Ph.D.



Q8 – Where did the IE researchers in Brazil conduct their doctoral studies?

Through this question, it is possible to know where the researchers who published in the RBIE, SBIE, and WIE concluded their doctoral studies. Similar to the professional address information, this information can also be obtained by cross-referencing between these publishing databases and Lattes. Query presented in Code 4 retrieves the institution where the researchers who published in the WIE completed their doctoral studies. Figure 12 shows the concentration of doctorates completed per university.

Code 4. Query to retrieve the number of researchers per University they got their Ph.D.

```
PREFIX dac:<http://www.ic.ufal.br/dac/>
PREFIX lattes:<http://www.ic.ufal.br/lattes/>
# retrieving the number of researchers by university
SELECT ?nameInstitution count(distinct ?g) as ?Count
FROM <http://www.ic.ufal.br/dac/wie/>
FROM <http://www.ic.ufal.br/dac/author/wie/lattes/alignments/>
FROM <http://www.ic.ufal.br/dac/author/wie/wie/alignments/>
FROM <http://www.ic.ufal.br/dac/lattes/>
WHERE {
# retrieving researchers' profiles
?s a dac:Author; owl:sameAs* ?g.
# retrieving researcher's name
?g foaf:name ?oname.
# subject must be in wie database
filter regex(?s,"http://www.ic.ufal.br/dac/author/wie/(\\d)+$","i")
# object must be in lattes database
filter regex(?g,"http://www.ic.ufal.br/dac/author/lattes/(.)+$","i")
# there might be no triple/relation in the alignment database pointing to the subject
filter not exists {
graph<http://www.ic.ufal.br/dac/author/wie/wie/alignments/>{?k owl:sameAs ?s}
# retrieving researcher's academic degree
?g lattes:hasAcademicDegree ?t.
?t a lattes:DoctoralDegree.
# retrieving instittution in which researcher received his degree
?t lattes:hasInstitution ?institution.
# retrieving institution's name
?institution foaf:name ?nameInstitution
# descendent ordenation
ORDER BY DESC(?Count)
```

CONCLUSION

This study presented a semiautomatic approach for aligning real-world datasets and scenarios. This proposed approach is necessary due to the need for solutions capable of reliably aligning data with the domain's least possible knowledge. Additionally, the solution lets the alignment be executed directly within the triple storage such that there is no need to generate files to align.

A case study was performed to evaluate the proposed approach in a real scenario. The proposed approach aligns RBIE, SBIE, WIE, and Lattes datasets. The use of the alignment solution enabled various questions to be answered. Additionally, it was possible to note problems related to the information provided by authors submitting their work. This research also conducted an experiment to evaluate the proposed approach and compared it to effectiveness with other tools using the precision, recall, and f-measure metrics. In the experiment, these metrics were evaluated in two alignment scenarios, in which the proposed approach obtained good results compatible with the top instance matching tools. Despite not having the best values in either evaluation, the proposed approach stands out due to the absence of specific implementations for the dataset or benchmark (for instance, it does not use the reference alignment for improving its processing), requiring less work when a context change is necessary. This paper presents the following contributions:

1. An alignment process for Linked data based on a general approach, independent of specific algorithms for the datasets and exemplars, with good f-measure results.
2. The execution of the alignment directly within the triple storage.
3. A real-world case study that deals with heterogeneous datasets and data quality issues.

Future studies can be performed to analyze the tool's effectiveness with datasets with different characteristics (e.g., domain, number of triples, and quality). Moreover, other studies will be conducted with the following objectives:

- Automating the whole alignment process – one possible method to do this would be to choose the related concepts automatically.
- Optimizing the performance – given that the related concepts can be aligned in parallel, possible approaches include parallelism and distribution.
- Improving the quality of the calculation of similarity between resources – one possible approach would be the composition of similarity functions and the identification of the most significant characteristics for identifying similarity.

REFERENCES

Abubakar, M., Hamdan, H., Mustapha, N., & Aris, T. N. M. (2018, February). Instance-based ontology matching: a literature review. In *International Conference on Soft Computing and Data Mining* (pp. 455-469). Springer.

Avila-Garzon, C. (2020). Applications, Methodologies, and Technologies for Linked Open Data: A Systematic Literature Review. *International Journal on Semantic Web and Information Systems*, *16*(3), 53–69. doi:10.4018/IJSWIS.2020070104

Azmy, M., Shi, P., Lin, J., & Ilyas, I. F. (2019). Matching entities across different knowledge graphs with graph embeddings. arXiv preprint arXiv:1903.06607.

Barbosa, A., Bittencourt, I. I., Siqueira, S. W. M., de Amorim Silva, R., & Calado, I. (2021). The use of software tools in linked data publication and consumption: A systematic literature review. *Research Anthology on Digital Transformation, Organizational Change, and the Impact of Remote Work*, 1868-1888.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic web. *Scientific American*, *284*(5), 34–43. doi:10.1038/scientificamerican0501-34 PMID:11681174

Böhm, C., De Melo, G., Naumann, F., & Weikum, G. (2012, October). LINDA: distributed web-of-data-scale entity matching. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2104-2108). ACM.

Castano, S., Ferrara, A., Montanelli, S., & Varese, G. (2011). Ontology and Instance matching. In *Knowledge-driven multimedia information extraction and ontology evolution* (pp. 167–195). Springer. doi:10.1007/978-3-642-20795-2_7

Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G., ... Lambrix, P. (2015, October). *Results of the ontology alignment evaluation initiative 2015*. Academic Press.

Collarana, D., Galkin, M., Traverso-Ribón, I., Vidal, M. E., Lange, C., & Auer, S. (2017, June). MINTE: semantically integrating RDF graphs. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-11). Academic Press.

Daskalaki, E., Flouris, G., Fundulaki, I., & Saveta, T. (2016). Instance matching benchmarks in the era of Linked Data. *Journal of Web Semantics*, *39*, 1–14. doi:10.1016/j.websem.2016.06.002

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., & Tang, N. (2017). *DeepER—Deep Entity Resolution*. arXiv preprint arXiv:1710.00597.

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., & Tang, N. (2017). DeepER—Deep Entity Resolution. arXiv preprint arXiv:1710.00597.

Efthymiou, V., Stefanidis, K., & Christophides, V. (2016, March). Minoan ER: progressive entity resolution in the web of data. *19th International Conference on Extending Database Technology, EDBT 2016*.

Faria, D., Pesquita, C., Balasubramani, B. S., Martins, C., Cardoso, J., Curado, H., . . . Cruz, I. F. (2016, January). OAEI 2016 results of AML. In *11th International Workshop on Ontology Matching co-located with the 15th International Semantic web Conference, CEUR Workshop Proceedings* (*Vol. 1766*). Academic Press.

Feitosa, D., Dermeval, D., Ávila, T., Bittencourt, I. I., Lóscio, B. F., & Isotani, S. (2018). A systematic review on the use of best practices for publishing linked data. *Online Information Review*, *42*(1), 107–123. doi:10.1108/OIR-11-2016-0322

Ferranti, N., Soares, S. S. R. F., & de Souza, J. F. (2021). Metaheuristics-based ontology meta-matching approaches. *Expert Systems with Applications*, *173*, 114578. doi:10.1016/j.eswa.2021.114578

Ferrara, A., Lorusso, D., Montanelli, S., & Varese, G. (2008, October). Towards a benchmark for Instance matching. In *The 7th International Semantic web Conference* (p. 37). Academic Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222*(594-604), 309-368.

Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European conference on information retrieval* (pp. 345-359). Springer.

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, *3*, 83.

Holanda, O., Isotani, S., Bittencourt, I. I., Elias, E., & Tenório, T. (2013). JOINT: Java ontology integrated toolkit. *Expert Systems with Applications*, *40*(16), 6469–6477. doi:10.1016/j.eswa.2013.05.040

Homoceanu, S., Kalo, J. C., & Balke, W. T. (2014, June). Putting Instance matching to the Test: Is Instance matching Ready for Reliable Data Linking? In *International Symposium on Methodologies for Intelligent Systems* (pp. 274-284). Springer. doi:10.1007/978-3-319-08326-1_28

Karr, A. F., Taylor, M. T., West, S. L., Setoguchi, S., Kou, T. D., Gerhard, T., & Horton, D. B. (2019). Comparing record linkage software programs and algorithms using real-world data. *PLoS One*, *14*(9), e0221459. doi:10.1371/journal.pone.0221459 PMID:31550255

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics, Doklady*, *10*(8), 707–710.

Li, J., Tang, J., Li, Y., & Luo, Q. (2008). Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, *21*(8), 1218–1232.

Menestrina, D., Benjelloun, O., & Garcia-Molina, H. (2005). Generic entity resolution with data confidences. Academic Press.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Mountantonakis, M., & Tzitzikas, Y. (2019). Large-scale semantic integration of linked data: A survey. *ACM Computing Surveys*, *52*(5), 1–40. doi:10.1145/3345551

Ngomo, A.-C. N., & Auer, S. (2011). Limes-a time-efficient approach for large-scale link discovery on the web of data. *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI'11)*, 2312–2317.

Nguyen, K., Ichise, R., & Le, B. (2012, December). Interlinking Linked data sources using a domain-independent system. In *Joint International Semantic Technology Conference* (pp. 113-128). Springer.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, *42*(2), 949–971. doi:10.1016/j.eswa.2014.08.032

Royce, W. W. (1987, March). Managing the development of large software systems: concepts and techniques. In *Proceedings of the 9th international conference on Software Engineering* (pp. 328-338). Academic Press.

Sarawagi, S., & Bhamidipaty, A. (2002, July). Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 269-278). doi:10.1145/775047.775087

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35–43.

Suchanek, Abiteboul, & Senellart. (2011). Paris: Probabilistic alignment of relations, instances, and schema. *Proc. VLDB Endow., 5*(3), 157–168.

Urbani, J., Kotoulas, S., Maassen, J., Van Harmelen, F., & Bal, H. (2012). WebPIE: A web-scale parallel inference engine using MapReduce. *Journal of Web Semantics*, *10*, 59–75. doi:10.1016/j.websem.2011.05.004

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk—A link discovery framework for the web of data. *Proceedings of the World Wide Web Workshop on Linked Data on the Web*.

Wang, P., & Xu, B. (2009). Lily: Ontology alignment results for oaei. In *Int. Semantic web Conf* (pp. 1-4). Academic Press.

Winkler, W. E. (1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Academic Press.

Wu, R., Chaba, S., Sawlani, S., Chu, X., & Thirumuruganathan, S. (2020, June). Zeroer: Entity resolution using zero labeled examples. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (pp. 1149-1164). doi:10.1145/3318464.3389743

Xu, L., Pavlo, A., Sengupta, S., & Ganger, G. R. (2017, May). Online deduplication for databases. In *Proceedings of the 2017 ACM International Conference on Management of Data* (pp. 1355-1368). doi:10.1145/3035918.3035938

Xue, X., & Tang, Z. (2017). An evolutionary algorithm based ontology matching system. *Journal of Information Hiding and Multimedia Signal Processing*, *8*(14), 551–556.

Yang, C., Hoang, D. H., Mikolov, T., & Han, J. (2019, May). Place deduplication with embeddings. In *The World Wide Web Conference* (pp. 3420-3426). doi:10.1145/3308558.3313456

Zamazal, O. (2020). A Survey of Ontology Benchmarks for Semantic Web Ontology Tools. *International Journal on Semantic Web and Information Systems*, *16*(1), 47–68. doi:10.4018/IJSWIS.2020010103

Zhang, Y., Jin, H., Pan, L., & Li, J. Z. (2016, October). RiMOM results for OAEI 2016. In OM@ ISWC (pp. 210-216). Academic Press.

Zhu, L., Ghasemi-Gol, M., Szekely, P., Galstyan, A., & Knoblock, C. A. (2016, October). Unsupervised entity resolution on multi-type graphs. In *International semantic web conference* (pp. 649-667). Springer.

ENDNOTES

1 http://islab.di.unimi.it/content/im_oaei/2016/
2 http://www.br-ie.org/pub/index.php/rbie
3 http://www.br-ie.org/pub/index.php/wie
4 http://www.br-ie.org/pub/index.php/sbie
5 http://lattes.cnpq.br
6 https://github.com/josmarios/dac/blob/master/Ontologies/dacV2.1.owl
7 http://openrefine.org
8 https://virtuoso.openlinksw.com

Armando Barbosa received his MS in Informatics from the Federal University of Alagoas in 2017. In 2014 he won first place in an innovative ideas competition with a real estate recommendation system based on linked open data. Currently, he works as Data Engineer in Getrak, an IoT company focused on car tracking.

Ig Ibert Bittencourt is an Associate Professor at Federal University of Alagoas (Brazil) and Co-Director of the Center of Excellence for Social Technologies. He received his Ph.D. in Computer Science in 2009 from Federal University of Campina Grande (Brazil) and his Post-Doctoral degree in Computer Science in 2013 from University of Campinas (UNICAMP, Brazil). Ig Ibert was capable of finishing his PhD in two years and five months, he was one of the most productive graduate students in the history of Computer Science Program (since 1973) and he was awarded with the PhD Thesis Distinction. During the PhD, he proposed a theoretical and computational model to build Semantic Web-based Educational Systems (the main paper has more than one hundred citations). His research career has been devoted to Artificial Intelligence in Education (AIED), working on the design, development and experimentation of educational technologies. Ig Bittencourt was also a visiting researcher University of São Paulo (Brazil), Japan Advanced Institute of Science and Technology - JAIST (Japan), Mannhein University (Germany) and Beijing Normal University – BNU(China). He was the president of the Special Committee of Computers and Education from Brazilian Computer Society (leading around 2500 researchers), W3C Advisory Committee Representative and Brazilian Computer Society Representative of IFIP TC on Education (TC 3). Prof. Ig Bittencourt co-founded an awarded company called MeuTutor (now eyeduc) and he stand out from his peers by creating one of the most innovative companies in the field of educational technology in Brazil (and Latin America). As a result, MeuTutor won three important awards, including Innovation Hall Award at the RioInfo – the biggest event in Software Industry in Brazil. His new company, called eNeuron – Cognitive Computing, developed an algorithm to automatically correct essays in Portuguese (that can be adapted to other 20 languages). In his career, Ig Ibert won more than 30 awards (including national and international awards) and he was the first Latin-American to be awarded with the IEEE TCLE Early Career Researcher Award (2019). He believes in innovative social entrepreneurship as a model for promoting a sustainable economic and social development to mankind.

Sean W. M. Siqueira is professor at the Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil, where he teaches courses in Information Systems, Databases, Web Science and Social & Semantic Web. He holds a M.Sc. (1999) and a Ph.D. (2005) in Computer Science, both from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. His research interests include knowledge representation and management; collaborative systems; social web; semantic web; ontologies; information integration; semantic models; user models; and learning & education. He has experience in the Computer Science area, with focus on Web Science, Information Systems and Technology Enhanced Learning. He has participated in some international research projects and has written more than 130 papers for conferences, journals, and book chapters. He is member of the special committee on Computers in Education (CEIE) and of the special committee on Information Systems (CESI), both from the Brazilian Computer Society.

Diego Dermeval is an Adjunct Professor at the Federal University of Alagoas. He received his Ph.D. from the Federal University of Campina Grande (UFCG - Brazil) with a sandwich period at the Department of Computer Science at the University of Saskatchewan (U of S - Canada). He has been dedicated to research in the field of Artificial Intelligence in Education (AIED), working on the design, development, and experimentation of intelligent educational technologies. He was a visiting researcher at the Institute of Mathematical and Computer Sciences (ICMC) of the University of São Paulo (USP-São Carlos, 2017 and 2018) and was selected to participate in the British Council Researcher Links Workshop: Higher Education for All: International Workshop on Social, Semantic, Adaptive and Gamification techniques and technologies for Distance Learning in Maceió, Brazil (2017). Diego is the author/co-author of over 50 papers published in journals or conferences in the areas of Artificial Intelligence, AIED, Software Engineering, and Human-Computer Interaction, and has published in reputable venues in these areas. He is an associate editor of the Brazilian Computers and Education Journal. He is a co-director of the Center for Excellence in Social Technologies (NEES - IC / UFAL). He is a reviewer of journals (e.g., IEEE Transactions on Learning Technologies, Frontiers in Artificial Intelligence, Smart Learning Environments, British Journal of Educational Technology) and conferences (i.e., AIED, ITS, and ICALT) in the field of computers and education. Diego Dermeval is currently a consultant to the Ministry of Education of Brazil in designing Evidence-Based Educational Technology Policies. Within the scope of innovation, he is co-founder of the spin-off 2KnowBetter, an educational technology company whose goal is to develop computational solutions that provide augmented intelligence for teachers.