Integrating User Profiles from Academic and Professional Web Data Sources

André Alencar, Edemberg Rocha, Damires Souza
Informatics Academic Unit
Federal Institute of Paraíba (IFPB)
João Pessoa, Brazil
andrealencar@outlook.com, edemberg@gmail.com, damires@ifpb.edu.br

***Abstract*—Semantic Web concepts and technologies have been considered as a way for enhancing integration of datasets on the Web. Based on that, this work applies a semantic Extract-Transform-Load approach to integrate user profiles from academic and professional web data sources. Considering a given application, which needs those integrated user profiles, we take into account the application data requirements in order to refine the proposed approach and to easy the data integration tasks. In this paper, we present the principles underlying our approach and some obtained results in the light of a scenario including some real web academic and professional data sources.**

***Keywords*— professional social network; academic social network; data integration; user profile; semantic web**

I. Introduction

With all the time spent using social networks, more and more information about their users have been generated. A social network is usually a place for sharing content in different forms, e.g., feelings such as personal likes on Facebook, or even establishing professional or academic networks as usually done on LinkedIn. In the light of academic and professional networks, some are widely used today, namely: LinkedIn, one of the largest and most complete networks on professionals; Research Gate, which has the proposal to manage data on researchers and, also, organizes a system of notes for them; and Academia, whose focus is the publication of research results conducted by its users, enabling them to monitor who are reading these publications. In addition to these social networks, the Lattes platform is currently characterized in Brazil as the major scientific web data source on researchers.

What is more interesting about these user networks or platforms relies in a twofold fact: (i) users themselves are responsible for keeping all the information up to date in the way they are important to be employed; and (ii) users usually are members of some of these platforms/networks at the same time. As a consequence, the former points out that the information provided is generally updated and consistent. The latter is a means to provide integrated user profiles in order to achieve a comprehensive and more complete view on users.

This introduces the concept of academic and professional user profile as a collection of settings and information associated with a given user, which are acquired from academic and professional platforms/networks (hereafter called as professional web data sources). Information collected from integrated user profiles may be used by commercial and non-commercial applications (e.g., recommender systems) in order to increase the quality of personalized usage. To this end, user data integration issues have to be dealt with.

In order to have an integrated view of user profiles which belong to web data sources, we usually have to deal with some steps such as data Extraction, Transformation and Loading (ETL). This implies in an ETL process on the web [1]. To this end, Semantic Web concepts and technologies may be used as a way to assist these steps [2].

With these ideas in mind and considering the need for diverse applications to be able to benefit from integrated profiles of researchers or professionals, this work employs an ETL approach. The developed approach is based on semantic web standards and assists the necessary steps to integrate user profiles from professional and academic web data sources. To this end, we analyze user profiles of some relevant professional and academic web data sources with respect to the data requirements of a given application. The principle of our work is to produce integrated user profiles that meet the application data requirements. Although the development and evaluation of the approach have been carried out with respect to a given application, the approach can be applied or extended to other scenarios and applications as well.

Our contributions are summarized as follows:

1) We employ a semantic-based approach on how to extract, transform, and load user profiles from academic and professional web data sources;

2) We propose a User Vocabulary in order to assist the data integration and data conversion steps.

3) We present results in terms of a developed tool and of some accomplished experiments regarding the effectiveness of our work.

The remainder of this paper is organized as follows: Section 2 introduces some concepts, a motivating scenario and the research problem; Section 3 formalizes our applied approach; and Section 4 presents some evaluation results. Related works are discussed in Section 5. Finally, Section 6 draws our conclusions and points out some future work.

II. Concepts, Scenario and Research Problem

In this section, we provide some concepts regarding Data Integration. Then, we describe our motivating scenario and define the research problem.

A. Data Integration

Data Integration has commonly been defined as the problem of combining data residing at different heterogeneous sources, and providing the user with a unified view of these data [3]. Semantic Web standards such as linked data principles have been considered as a way to provide a lightweight data integration method [4]. Linked data principles are based on technologies such as HTTP (Hypertext Transfer Protocol), URI (Uniform Resource Identifier) and RDF (Resource Description Framework) [2]. With their usage, it is possible to integrate data sources by describing them in terms of a semantic vocabulary, and also by establishing links among the entities belonging to those sources [4].

By using the RDF model, data or resources are published on the Web in the form of triples (composed by a subject, a predicate and an object), where each resource is individually identified by means of a URI. RDF links allow client applications to navigate between data sources and to discover additional data, providing unified views of data.

In the light of Data Integration solutions, the Extract-Transform-Load (ETL) strategy has been used. It refers to a process that extracts data from different data sources, transforms them to fit standardization and usage needs, and loads them into a target repository [1]. Extraction involves acquiring data from appropriate and chosen data sources. Transformation usually includes the cleansing, normalization and conversion of data to comply with the target schema. Load implies in the persistency of the integrated data into a data repository.

From another point of view, some authors consider Data Integration as a process composed by three steps [5]: schema matching, entity resolution and data fusion. The first one is used to circumvent the problem of sources with different schemes. Considering schemas of distinct sources, one can map, for example, between the "Author" entity, in one source, with the entity "Writer", in another data source. Entity Resolution has the objective of identifying if two instances represent the same entity of the real world. Considering two instances of entities in distinct sources, one can, for example, indicate that the author "Neving, C." is the same as the author "Claire, Neving" in another data source. In addition, Data Fusion refers to the process of synthesizing raw data from several sources to generate more meaningful or complete information. Integrated data should be of greater value than a single source data.

In fact, a data integration process may use a combination of steps to cope with its challenges. In this sense, we argue that the use of semantic technologies should be included in a data integration process in order to easy the needed steps. Moreover, we propose a semantic ETL process to accomplish data integration, where the transformation step is indeed composed by schema matching, entity resolution and data fusion. By using semantic domain vocabularies (i.e., ontologies), the linked data principles, and a semantic ETL process, an RDF dataset with integrated user profiles will be produced and then stored in an RDF store. This dataset may be published on the web by means of a data API (Application Programming Interface).

B. Motivating Scenario

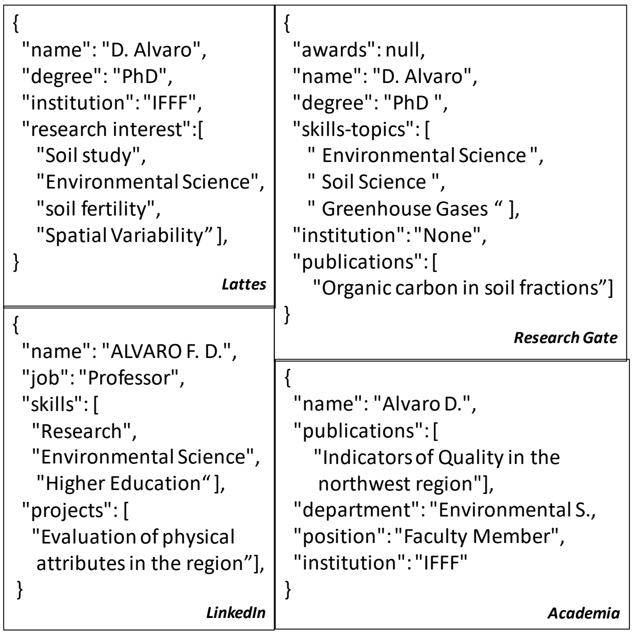
Extracting and integrating academic and professional user profiles becomes relevant to some specific applications. In our institution, for instance, we have a business problem which regards a recommendation system. This application will be responsible for suggesting reviewers to submitted productions (e.g., articles, projects). This is a real demand in a system that supports the reviewing process of a scientific journal.

Usually, in many systems, the reviewers themselves define keywords or themes associated with their areas of research and expertise. In other situations, manually, editors look for publications and related projects of researchers that indicate some affinity with the subject of a production to be evaluated. To help matters, user profiles obtained from the Lattes platform and from some professional and academic social networks may be used in order to provide some indicators. In this work, we consider four web data sources to be used, namely: the Lattes platform, and the Academia, LinkedIn and Research Gate networks.

With these ideas in mind, consider that the application at hand (i.e., the described recommendation system) is, hereafter, called as **A**. Particularly, A has the following general requirements: (i) Get Researcher Profile; (ii) Analyze Researcher Profiles versus Production to be evaluated; and (iii) Suggest the best Researchers (Reviewers) for that Production Evaluation. **A** would indeed benefit from integrated views of researchers. Thus, **A** would need to combine information regarding a given user (i.e., a researcher) from multiple professional web data sources to build his/her integrated profile.

Also, in order to analyze what user data should be acquired, consider that **A** has the following data requirements (defined as a set of entities and properties): Researcher, ProjectTitle, PublicationTitle, Institution, name, expertise and others. As an illustration, according to these data requirements, we have extracted some information regarding a researcher, called “Alvaro D.”, from the four mentioned web data sources. Under this scenario, we assume that data are acquired in JSON format. Also, property names, which belong to Lattes, have been normalized (translated to English) in order to easy understanding and data integration. These data are depicted in Figure 1.

In order to have an integrated data view of the researcher “Alvaro” from the professional web data sources, it is necessary to deal with ETL steps. We argue that if we use semantic standards and references we can easy the whole process and provide ways to resolve conflicts and matching. Each phase of the ETL process has specific technical issues to be addressed. To facilitate this process, identifying the relevant data to crawl, creating domain-specific feature extractors, and building a domain vocabulary to align the data are important steps to be done.



```
{
  "name": "D. Alvaro",
  "degree": "PhD",
  "institution": "IFFF",
  "research interest":[
    "Soil study",
    "Environmental Science",
    "soil fertility",
    "Spatial Variability” ],
}
```

Lattes

```
{
  "name": "ALVARO F. D.",
  "job": "Professor",
  "skills": [
    "Research",
    "Environmental Science",
     "Higher Education“ ],
  "projects": [
    "Evaluation of physical
     attributes in the region” ],
}
```

LinkedIn

```
{
  "awards": null,
  "name": "D. Alvaro",
  "degree": "PhD ",
  "skills-topics": [
    " Environmental Science ",
    " Soil Science ",
    " Greenhouse Gases “ ],
  "institution": "None",
  "publications": [
    "Organic carbon in soil fractions”]
}
```

Research Gate

```
{
  "name": "Alvaro D.",
  "publications": [
    "Indicators of Quality in the
     northwest region"],
  "department": "Environmental S.,
  "position": "Faculty Member",
  "institution": "IFFF"
}
```

Academia

Fig. 1. Illustration of some obtained profiles for a researcher.

C. Problem Definition

Based on the described scenario and on the presented concepts, we define our research problem as follows:

Given an application A, which has data requirements D, how can we extract, transform and load user profiles UP_i, which belong to web professional data sources, into an RDF integrated dataset UI, which is complete w.r.t. UP_i according to D?

This means that we have to analyze available user (researcher) profiles by means of the mentioned professional web data sources with the intention of extracting, transforming and integrating the data into an RDF dataset. The generated integrated view (UI) must meet the application data requirements and be complete w.r.t. the original user profiles.

III. Proposed Applied Approach

The main idea underlying our approach is to bring the domain semantics into the ETL process aiming to facilitate data integration. The activity of converting different professional profiles w.r.t. a determined user U produces an integrated view of U defined in terms of a given domain vocabulary (ontology). In this section, we present some definitions regarding our approach. Then, we present the User Vocabulary and the strategies which compose the proposed semantic ETL process.

A. Some Definitions

A user model is an abstract model defining type and meaning of information stored about users [6]. Usually, it contains information such as user preferences, goals, context, behavior, or even background. A user profile is a data instance of a user model and contains the data of a determined user [6]. In our work, a user profile regards data belonging to specific academic or professional web data sources. Our scenario may be defined as follows.

Let S = {s1, s2... s_n} be the set of professional web data sources and U = {u1, u2... u_m} be the set of users, where a source s $\in$ S is characterized by information regarding professional and/or academic data of users u $\in$ U. Also, Let D = {d1, d2... d_k} be the set of data requirements which belong to application A. A needs integrated views of U from S in accordance with D.

In this scenario, we define a User Profile (UP) w.r.t. professional web data sources as the following:

Definition 1.User Profile (UP). A User Profile UP_i is a triple <u, s, t> that denotes a profile instantiation obtained for user u on web source s at time slice t, where t represents the time when the user profile was acquired.

Since a user u may have different and/or complementary profiles in diverse s, our goal is to define and to validate a user integrated profile which represents relevant professional and academic information regarding u. We define a User Integrated (UI) profile as follows.

Definition 2. User Integrated Profile (UI). A user integrated profile UI is composed by the union of concepts and properties which belong to UP_i.

UI meets D and is semantically formalized by a domain vocabulary.

On the Semantic Web, vocabularies define the concepts and relationships (also referred to as “terms”) used to describe and represent an area of concern. Particularly, we need to use semantic vocabularies regarding users and their professional data, thus we are able to bridge the conceptual differences or similarities among the web data sources.

Reuse of appropriate vocabularies is becoming easier nowadays, since most of them are available on the web. Nevertheless, vocabularies regarding User information may be heterogeneous with respect to focus and coverage. A comprehensive view on professional and academic user profiles is required for choosing and using vocabularies suitable for D. We evaluated existing available vocabularies which could be potentially useful for reuse, such as FOAF, DBPEDIA, SWPO and VIVO [7]. However, none of them has fulfilled all or the major elements of D. Thereby, we found out the need of building a specific user vocabulary for D.

The User Vocabulary should be comprehensive, include concrete facts from existing professional web data sources, and also extensible to cover further possible concepts. The goal was to implement the vocabulary in the form of an ontology. Also, we will make it publicly available for reuse, as well as employ it for user profile integration and, consequently, for application A. We define the User Vocabulary (UV) as follows.

Definition 3.User Vocabulary (UV). User Vocabulary UV is a domain specific ontology for aggregating user profiles from s $\in$ S as well as for providing semantic terms to data conversion from UP_i into UI in accordance with D.

B. User Vocabulary

Vocabularies provide the semantic glue enabling data to become meaningful data. To capture the variety of available professional and academic user profiles, in such a way that they meet D, a User Vocabulary (UV) has been designed and developed.

For the creation of UV, the following phases were performed: (i) Analysis of the information present in the profiles of academic and professional web data sources, especially Lattes, Academia, Research Gate, and LinkedIn; (ii) Selection of information from researchers, including their competences and expertise, in accordance with D; (iii) Identification of the metadata (concepts and properties) to be used; (iv) Verification of the possibility of reuse of metadata from open and recommended vocabularies; (v) Ontology construction; (vi) Ontology Instantiation and Validation for Tests. The steps were fulfilled and a version of the ontology was generated.

Following the best practices of producing linked open datasets [8], we reused terms from some open vocabularies in accordance with D. Table 1 shows the vocabularies used to compose UV, including itself, and some of their reused or defined terms.

We largely reused SWPO ontology for research aspects description. We reused the DBPEDIA and VIVO ontologies for modeling some academic concepts. Regarding the FOAF vocabulary, we only reused the term "Person" since its context is usually related to people in general, but not with research facets. Since some data requirements are very particular to Brazil's context, mainly because of the Lattes platform, some concepts could not be reused. Thus, they had to be created as own terms of UV, as the ones shown in Table 1. Besides the concepts depicted in Table 1, object and data properties were also reused, when possible.

TABLE I. REUSED VOCABULARIES AND UV ADDED TERMS

Vocabulary	Short Description	Reused Terms
FOAF	Contains people-related terms	Person
SWPO	It has been created to serve as a conceptual backbone for community portals	Researcher, Topic, Publication, Article, TechnicalReport, Inproceedings, PublicationContainer, Journal, Proceedings, PublishingCompany
DBPEDIA	It provides the classes and properties used in the DBpedia dataset	EducationalInstitution, Occupation, Project, ResearchProject
VIVO	An ontology of academic and research domain, developed in the VIVO project	Department, AcademicDepartment
UV	Proposed vocabulary	User, EducationalLevel, GeneralKnowledgeArea, KnowledgeArea, Keyword, Language, DevelopmentProject, ExtensionProject

Thus, UV comprises some primary concepts such as *User*, *Researcher*, *Project*, *Publication*, *KnowledgeArea* and *EducationalInstitution*. Properties and relationships are defined by means of data and object properties depending on the type of data field being represented. UV allows a reuse of information while keeps the operational information focused on D. Figure 2 shows a high-level view of UV with the main concepts, according to the ontograf notation [9].

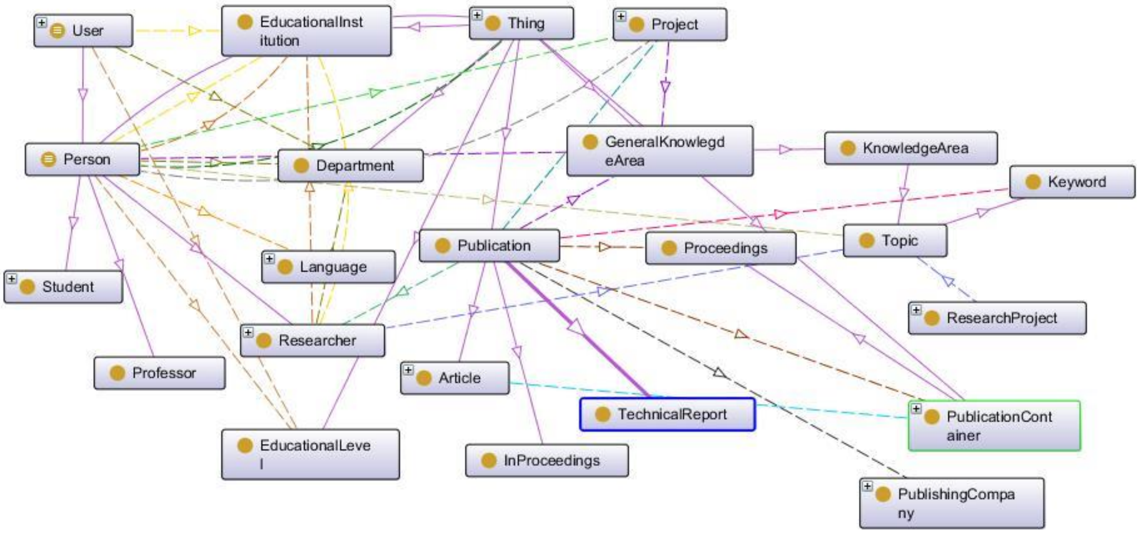


Fig.2. The User Vocabulary Main Concepts

C. *Semantic ETL Process*

The proposed semantic ETL process consists of the three ETL major phases as depicted in Figure 3. The use of semantic technologies is introduced in the Transformation phase as a means to enhance data integration. Phases are discussed in detail in the following.

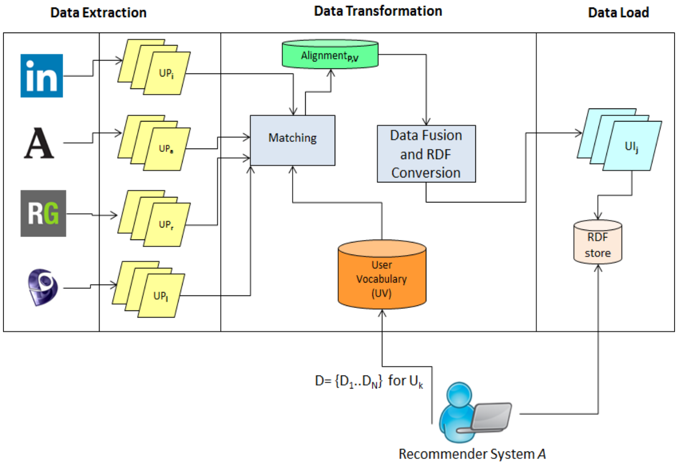


Fig. 3. Semantic ETL Process

Data Extraction

In the data extraction step, instance data and their properties (metadata) are extracted from user profiles belonging to the chosen professional web data sources. To this end, a web crawler was written to obtain user profiles in JSON format. In addition, metadata are identified according to the names of properties in collections of name/value pairs.

Data Transformation

In this phase the extracted metadata as well as their corresponding instances data are converted to RDF triples. At first, matching of extracted metadata against UV terms is done. Entity resolution is accomplished as well. Then data fusion is performed while data are converted to RDF. These steps are explained in the following:

Matching

Identified profiles metadata are used to match against UV terms. Thus, at first, our approach maps s properties to terms in UV. The gathered user profiles are aligned to UV by means of a matcher [10]. This alignment is produced by a linguistic matcher [11]. Each correspondence is defined with a confidence measure (between 0 and 1). Accomplishing tests, we have defined a threshold of 0.8 to identify correspondences to be set as equivalences.

The output of the matching process between s properties and UV properties is called an alignment A_{pv}. A_{pv} contains a set of equivalence correspondences indicating which properties correspond to each other. This alignment is saved to be used later.

Entity Resolution

Since we proceed with the data integration process as a whole to one given user, we still do not cover the entity resolution step. It is indeed being accomplished in a manual way.

Data Fusion and Conversion

In order to carry out this step, we need to resolve some data conflicts. Particularly, we resolve data conflicts based on deciding strategies which choose a preferred value among the existing values for a given property. To this end, we take into account user preferences as a support in decisions about which data are worth using. According to user preferences, a specification of properties priority is generated. We define a Priority specification (Pr) as follows.

Definition 4.Priority Specification (Pr). Given n instances of a property p from s ∊ S, a priority specification Pr over n is an ordered collection {n1... n_a} where n_a+1 is preferred over n_a w.r.t. Pr, denoted as $n_a+1 >> n_a$.

Therefore, data fusion and conversion are accomplished according to the set of correspondences defined at matching level and to a priority specification based on user preferences. Pr indicates the order to be used when choosing a given value for p. An algorithm regarding the data fusion and conversion step is shown and explained in Section 4.1.

Data Load

The generated UI is persisted in an RDF store and made available on the web as linked data. In general, this means that it is available for querying (via SPARQL queries), analytics, or to be used in data-driven applications such as A. Particularly, UI represents enriched and more complete versions of the user profiles, what may be used to enhance recommendations in A. The more information in an UI, the more likely it is to provide better recommendations on that user.

IV. Implementation and Results

In this section, we describe some implementation issues, by means of a high-level main algorithm and a developed tool. Also, we present some experimental results.

A. *The Algorithm*

The principle of our approach is to integrate user profiles by using a semantic ETL process. For the sake of clarity, a high level view of the main algorithm is presented in Figure 4. It regards the data fusion and conversion phase.

In order to provide data fusion and conversion to RDF triples, the algorithm (Figure 4) performs the following tasks.

At first, it instantiates a graph which will receive the user integrated data. Then, it identifies the user at hand (U) (line 02) and retrieves his/her name in order to generate his/her URI (line 03). Then, it acquires all of the user profiles and his/her respective JSON objects and puts them into a collection to be iterated with (line 04). For each user object, it verifies its properties.

```
---------------------------------------
Algorithm_DataFusionandConversion()
---------------------------------------
INPUT: U, UPi, Pr, UV, Apv
   //User U, User profiles, Priority
     Specification, User Vocabulary and
     Alignment
OUTPUT: UI
     //User Integrated data in the form of
      RDF triples
Begin
01: new graph G
02: resName = retrieveName(U)
03: researcherURI =
combine(UV.namespace, resName)
04: jsonObjects = retrieveJsonDocs(UPi)
05: for each object ∊jsonObjects do
06:for eachproperty∊object do
07:   if (property != null) and
         (property != '')  then
08:     useProp = verifiesPriorityandUse
                  (Pr,property)
09:      If useProp = true then
10:        If property is simple then
11:          propname =
      getName(property)
12:          predicate =
getCorrespondence(propname)
13:          objValue =
getValue(property.value)
14:          G = addTriple (s=researcherURI,
                 p=predicate, o=objValue);
15:        Else
16:          propname =
getName(property)
17:           predicate =
getCorrespondence(propname)
18:          objValueList =
getValues(property)
19:          G = addTriple (s=researcherURI,
                 p=predicate,
o=objValueList)
20:        Endif;
21:      Endif;
22:Endif;
23: UI = Save(G.rdf)
24: Return (UI);
EndDataFusionandConversion;
```

Fig. 4. The data fusion and conversion Algorithm

For each property, it verifies if it is not null or even empty and decides if its value will be used according to the priority specification (line 08). If so, it gets the name of that property and its value (or values, if it regards a list of objects). Then, it generates an RDF triple, where the subject is the user at hand, the predicate is the obtained property (from the alignment), and the object may be a single one or even a list. This process is accomplished for each property belonging to that object; the same happens for each object belonging to the user profiles. After these tasks, the graph is saved as an RDF dataset (line 23), and a user integrated profile (UI) is returned (line 24).

B. Developed Tool

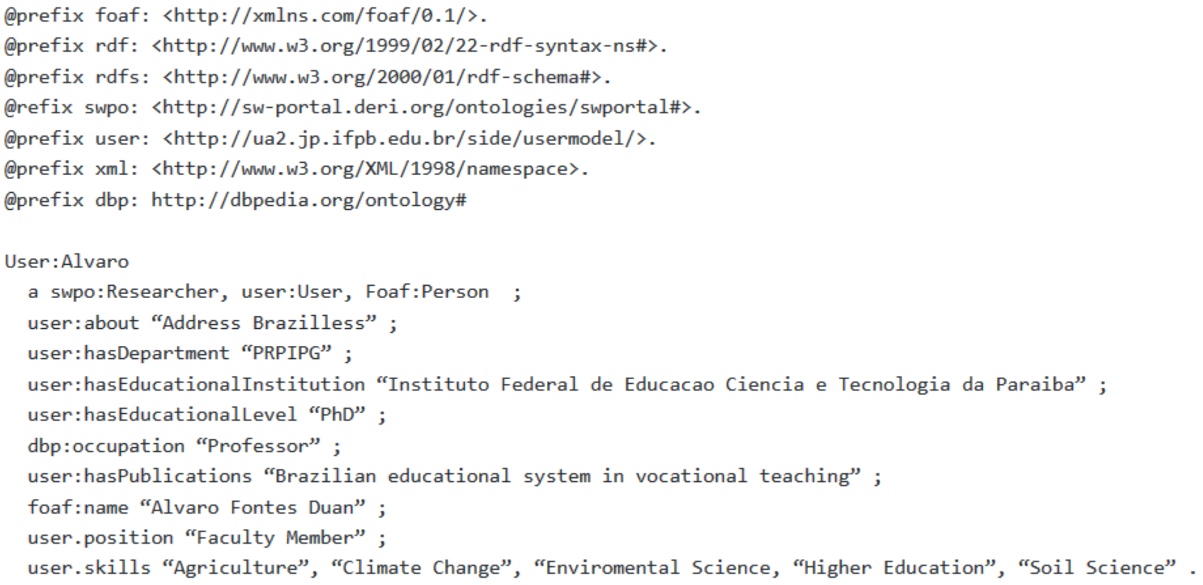
We have developed the approach and the presented algorithm in Python. To this end, we have used RDFLib, a Python library to work with RDF data [12]. In addition, to provide data extraction from the web data sources (LinkedIn, Academia and Research Gate), we have used different strategies and technologies, such as the LXML library [13] and the Ruby Linkedin-Scraper library [14]. In some cases, data were extracted in CSV files. We have developed an algorithm to convert them into JSON files.

As an illustration, consider the example presented in Figure 1. The tool is able to generate the user integrated profile as the one depicted in Figure 5. In this case, it applies the semantic ETL process in order to integrate profiles from user “Alvaro”, by considering three professional and academic networks (LinkedIn, Research Gate and Academia). It provides data fusion and conversion to RDF triples in accordance with the algorithm described in Section 4.A.

C. Experiments

We have conducted some experiments to verify the effectiveness of our approach. The goal was twofold: to check if there is any difference when using the proposed user vocabulary instead of other available ones, and (ii) to measure the data completeness in terms of using only one of the considered professional web data sources versus the integrated user profile w.r.t. D. The former aims to identify the degree of recall and precision regarding the use of domain vocabularies when integrating data sources. The latter intends to verify whether or not all the data necessary to meet D are available in each web data source or in the integrated profile. In this particular evaluation, we have used profiles belonging to nine users from three web professional web data sources: Academia, Research Gate and LinkedIn. Most of them have profiles in all of the mentioned web data sources.

Regarding the first goal, as domain vocabularies, we have used the SWPO, VIVO Ontologies, and the proposed UV. All of them belong to academic or professional user domains, with appropriate terms to be used.



```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@refix swpo: <http://sw-portal.deri.org/ontologies/swportal#>.
@prefix user: <http://ua2.jp.ifpb.edu.br/side/usermodel/>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix dbp: http://dbpedia.org/ontology#

User:Alvaro
  a swpo:Researcher, user:User, Foaf:Person  ;
  user:about “Address Brazilless” ;
  user:hasDepartment “PRPIPG” ;
  user:hasEducationalInstitution “Instituto Federal de Educacao Ciencia e Tecnologia da Paraiba” ;
  user:hasEducationalLevel “PhD” ;
  dbp:occupation “Professor” ;
  user:hasPublications “Brazilian educational system in vocational teaching” ;
  foaf:name “Alvaro Fontes Duan” ;
  user.position “Faculty Member” ;
  user.skills “Agriculture”, “Climate Change”, “Enviromental Science, “Higher Education”, “Soil Science” .
```

Fig. 5. Generated UI for User “Alvaro”

We consider recall and precision as follows. Recall measures the ratio of correctly found properties or terms (true positives) over the total number of expected properties or terms (true positives and true negatives) [15]. In order to achieve the expected number of properties, we have produced gold standards regarding the integration of profiles for the nine users, considering the most appropriate terms to be used. These gold standards have been manually produced by participants of our group. On the other hand, precision measures the ratio of correctly found properties (true positives) over the total number of returned properties (true positives and false positives) [15]. This measure was applied only considering UV and the returned integrated profiles produced by our tool. Formulas are presented in the following.

$$\text{Precision}(\text{CorrectProp}, \text{ReturnedProp}) = \frac{\#\text{CorrectProp}}{\#\text{ReturnedProp}} \qquad (1)$$

$$\text{Recall}(\text{CorrectProp}, \text{ExpectedProp}) = \frac{\#\text{CorrectProp}}{\#\text{ExpectedProp}} \qquad (2)$$

Where

CorrectProp is the number of correct applied properties (terms);

ExpectedProp is the total number of all possible properties that could be used; and

ReturnedProp is the total number of all retrieved properties produced by the tool (i.e., correct or incorrect ones).

A summary of the results regarding the recall measure along with the usage of SWPO, VIVO and User Vocabulary for nine users is shown in Figure 6.

We are able to observe that the usage of a suitable domain vocabulary makes all the difference. In this work, we have defined a specific vocabulary by making reuse of recommended terms when possible. New terms which belong to mainly used academic and professional web data sources have been defined in the User Vocabulary. As a result, it has covered almost 100% of the required data.

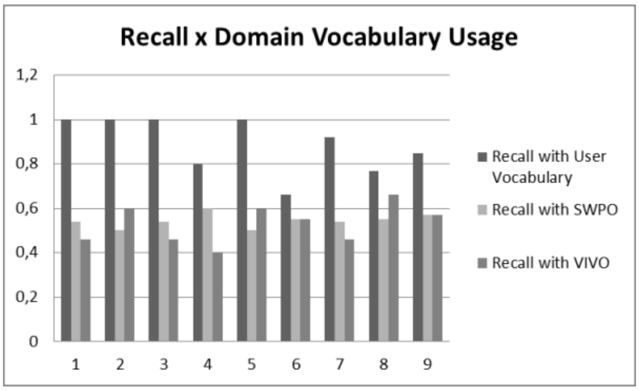


Fig. 6. Recall w.r.t. the choice of a domain vocabulary

We could verify that the precision of the generated integrated profiles, when using UV, was also high. On average, for the nine users and their various profiles, the precision was about 80%.

Regarding the second goal, we have obtained results regarding data completeness w.r.t. D (Figure 7). Particularly, the integrated profile obtained better results when comparing to the data completeness of each one of the considered web data sources.

Comparing user by user, we can see that some of the web data sources are indeed originally incomplete. This trend remains when the integrated profile is generated. The more data a web professional data source provides the more complete the integrated profile becomes. Thus we can conclude that the integrated profile can meet the application data requirements and increase usefulness of the data.

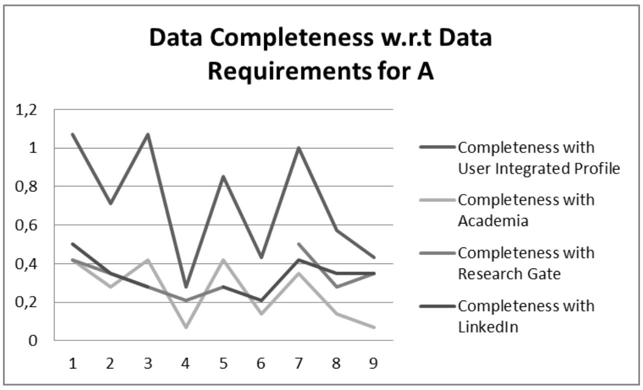


Fig. 7. Data Completeness w.r.t. D

V. Related Work

Some works for data integration on user profiles have been considered in literature. Examples are the works of Magalhães et al. [16], Wang, Zhang and Vassileva [17], Taheriyan et al. [18], Xiang [19], and Tao et al. [20]. Other ones have been proposed in the context of big data [1].

Magalhães et al. [16] presented a semantic approach to integrate user profiles from web data sources. The integrated profiles can be used to improve the quality of a recommender system. To this end, they have defined a heuristic that quantifies the importance of each data source for a given user. Wang, Zhang and Vassileva [17] proposed a user-centric approach for social data integration and recommendation, based on a new ontology of user social data. This approach applies some machine learning techniques to learn users' preferences and blend user friends.

Taheriyan et al. [13] proposed an approach which exploits the knowledge from domain ontologies and known semantic models of data sources to automatically learn a semantic model for a new data source. A semantic integration was performed on different data sources which belong to a medical domain in Xiang's work [14]. The data integration is used to enhance prediction of diseases. Tao et al. [15] proposed an approach for creating RDF-based user profiles on Twitter according to the frequency of the entities extracted from the user's tweets. In this case, user profiles are modeled using the FOAF vocabulary. The approach scores the interests based on simple term frequency technique.

Bansal and Kagemann [1] presented an approach to integrate big data by using an ETL framework. They present a conceptual ETL framework which uses semantic technologies to provide data integration. They also show a case study related to various open online courses.

Some of the discussed related works have similarities with ours. Our approach uses a specific user ontology to provide a common vocabulary for the integrated data as other ones do. Nevertheless, some differences regard the proposed vocabulary which has been built in accordance with issues specifically related to the domain of academic and professional user profiles. Also, the presented approach includes some semantic enhancements to the ETL process and generates integrated RDF data as part of the transformation step.

VI. Conclusions and Further Work

The integration of available academic and professional user profiles into a meaningful user integrated profile that allows querying and use by applications is an important issue. Based on a developed user vocabulary, we have developed a user-centric approach for integration of those profiles. The proposed semantic ETL process focuses on providing semantics to the steps thereby facilitating richer data integration. Integrated user profiles have potential to be used and to facilitate the creation of data-driven applications such as the described recommender system.

Accomplished experiments show that our approach is promising. By using the user vocabulary, it is able to produce complete integrated profiles w.r.t. the original web data sources and to the data requirements of a given application.

For future work, we intend to deal with the entity resolution problem thus enabling its automatization. The approach will be generalized in such a way that it can be used in diverse data scenarios. In addition, the User Vocabulary will be exposed as a web service.

References

[1] S. Bansal and S. Kagemann, "Integrating Big Data: A Semantic Extract-Transform-Load Framework". Computer, 48(3), pp.42-50. 2015.

[2] T. Heath and C. Bizer, "Linked data". Milton Keynes, UK: Morgan & Claypool. 2011.

[3] A. Doan, A. Halevy, and Z. Ives, "Principles of data integration". Amsterdam: Elsevier/Morgan Kaufmann. 2012.

[4] T. Knap, J. Michelfeit, M. Necaský, "Linked Data Integration with Conflicts". CoRR, abs/1410. 7990. 2014.

[5] X. Dong, D. Srivastava, "Big Data Integration". In Proceedings of VLDB'2013. Vol: 6. Nº 11. Italy. 2013.

[6] T. Plumbaum, "User modeling in the social semantic web". Doctoral Thesis. Technische Universität Berlin. 2015. Available at https://depositonce.tu-berlin.de/handle/11303/5222. Last access on March, 2018.

[7] LOV (2018). Linked Open Vocabularies. Available at https://lov.okfn.org/dataset/lov/. Last access on April, 2018.

[8] World Wide Web Consortium (W3C) 2018. [online] Available at: https://www.w3.org/. Last access on April, 2018..

[9] OntoGraf - Protege Wiki. [online] Available at: https://protegewiki.stanford.edu/wiki/OntoGraf. Last access on April, 2018.

[10] A. Silva, L. Chaves, D. Souza, "A Domain-based Approach to Publish Data on the Web". In Proceedings of the iiWAS2013. 2-6 December, Vienna, Austria. ACM, New York, NY, USA. DOI: http://dx.doi.org/10.1145/2539150.2539233. 2013.

[11] J. David, J. Euzenat, F. Scharffe, and C. Trojahn dos Santos, "The alignment api 4.0". In Semantic web journal 2 (1): 3–10, 2011.

[12] Rdflib.readthedocs.io. (2018). rdflib 4.2.2 — rdflib 4.2.2 documentation. [online] Available at: https://rdflib.readthedocs.io/en/stable. Last acess on April, 2018.

[13] LXML (2018). The LXML library. Available at http://lxml.de/. Last access on April, 2018.

[14] Scrap (2018). The Ruby Linkedin-Scraper Library. Available at https://github.com/yatish27/linkedin-scraper. Last access on April, 2018.

[15] C. J. Rijsbergen, "Information Retrieval", 2nd Ed. Stoneham, MA: Butterworths, 1979.

[16] J. Magalhães, C. Souza, P. Silva, E. Costa, and J. Fechine, "Improving a recommender system through integration of user profiles: a semantic approach". In Proceedings of UMAP Workshops. 2012.

[17] Y. Wang, J. Zhang, and J. Vassileva, “A user-centric approach for social data integration and recommendation”. In Proceedings of the 3rd International Conference on Human-Centric Computing (HumanCom) pp. 1-8. 2010.

[18] M. Taheriyan, C. Knoblock, P. Szekely and J. Ambite, “Learning the semantics of structured data sources”. Web Semantics: Science, Services and Agents on the World Wide Web, 37-38, pp.152-169. 2016.

[19] J. Xiang, “Social data integration and analytics for health intelligence”. In Proceedings of the VLDB 2014 PhD Workshop. 2014.

[20] K. Tao, F. Abel, Q. Gao and G. Houben, “TUMS: Twitter-based User Modeling Service”. In Proceedings of the International Workshop on User Profile Data on the Social Semantic Web (UWeb), pp. 269-283. 2011.