The Potential of Open Data to Automatically Create Learning Resources for Smart Learning Environments [†]

Adolfo Ruiz-Calleja *, Miguel L. Bote-Lorenzo, Guillermo Vega-Gorgojo, Sergio Serrano-Iglesias, Juan I. Asensio-Pérez, Yannis Dimitriadis and Eduardo Gómez-Sánchez

GSIC-EMIC Group, ETSI Telecomunicaciones, Universidad de Valladolid. Campus Miguel Delibes, Paseo de Belén, 15, 47011 Valladolid, Spain; adolfo@gsic.uva.es (A.R.-C.); migbot@tel.uva.es (M.L.B.-L.); guiveg@tel.uva.es (G.V.-G.); sergio@gsic.uva.es (S.S.-I.); juaase@tel.uva.es (J.I.A.-P.); yannis@tel.uva.es (Y.D.); edugom@tel.uva.es (E.G.-S.)

* Correspondence: adolfo@tel.uva.es

† Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

check for updates

Abstract: Smart Education requires bridging formal and informal learning experience. However, how to create contextualized learning resources that support this bridging remains a problem. In this paper, we propose to exploit the open data available in the Web to automatically create contextualized learning resources. Our preliminary results are promising, as our system creates thousands of learning resources related to formal education concepts and physical locations in the student's local municipality. As part of our future work, we will explore how to integrate these resources into a Smart Learning Environment.

---

1. Introduction

Smart Learning Environments (SLEs) have arose as new Technology-Enhanced Learning (TEL) environments able to adapt students' learning experience and to provide them with personalized support considering their individual needs and context [1]. One of the main promises of SLEs is the support of ubiquitous and adaptive learning to bridge formal and informal settings [2] by exploiting mobile technologies. This bridging is an issue due to the different characteristics of formal and informal learning: while formal learning happens in a controlled environment and counts on a learning design, informal learning typically happens in the student's daily life in a way that it is difficult to foresee. Hence, a question emerges: how can we create resources that can a student can use in her daily life for her to further learn on the topics covered in her formal education?

As an example, we can consider a student of secondary school who is taking a course of History of Art and learning about Medieval architecture. It would be very interesting that, when she passes by a Gothic Cathedral, a mobile application suggests her to look at several details of the temple and to reflect on the characteristics of the Gothic art. Thus, she takes advantage of a very-interesting learning opportunity during her daily life.

One approach to support this scenario would imply developing a system capable of knowing the context of the learner (e.g., current location or topics being addressed in formal education) so as to trigger learning activities (e.g., via mobile devices) offering opportunities for informal learning. A derived challenge for such a system is creating the resources to be used in those eventual informal learning

activities. For instance, who (and how) compiles the details of the Gothic Cathedral that are worth looking at? A possible option is to automatically build this set of learning resources out of a database [3,4]. Nonetheless, the issue still persists as it would be required to create a dataset of multiple domains, related to the topics covered in the classroom and the student's contexts. Our key idea to overcome this problem is to exploit the open data available on the Web [5] in order to automatically create such dataset. In this Web of Open Data we can find semantically described and geolocalized entities, which we can potentially exploit to create learning resources.

This paper further unfolds this idea. More specifically, Section 2 presents the current state of the art and how our approach goes beyond it. Then, Section 3 describes our current working prototype, which lets us reflect in Section 4 on what are the most important difficulties of this approach, as well as our plans to overcome them.

2. Current State of the Art

The TEL community widely applied social technologies to crowd-source the creation of learning resources for communities of teachers (e.g., [6]). However, the creation and annotation of learning resources is a costly and time-consuming task that, on many occasions, teachers cannot afford. Moreover, the complexity of the used ontological models makes the resources hard to describe [7].

For these reasons, the research community explored the alternative solution of automatically creating learning resources out of semantic datasets. The main idea of this approach is to define a set of templates that the system applies to the factual knowledge (e.g., vocabulary definitions, relationship among concepts or isolated information about certain details) contained in an ontology, thus creating a vast amount of questions. Many of these systems (e.g., [8]) automatically generate Multiple Choice Questions (MCQs), which are later on used for (self-)evaluation of factual knowledge. Even if these MCQs are automatically created, the production of the ontology remains a problem. Moreover, the educational significance of many MCQs is put into question because of two reasons: they are decontextualized, and they cannot assess higher-level thinking, as they only assess factual knowledge ([3]).

A possible solution to avoid producing such factual-knowledge datasets is to exploit the open datasets available on the Web of Data [5]. These datasets follow the Linked Data principles [9] as a common methodology to publish data that allows to interlink datasets from third parties. Although Linked Data has already been exploited for educational purposes, the TEL community has not deeply explored the automatic creation of learning resources out of them (see [10]). One interesting pioneer study is [11], where DBpedia (https://wiki.dbpedia.org) is used to populate local datasets that are later on used for programming exercises. We can also find several research proposals where Linked Data has been used for the automatic creation of MCQs related to many different domains. These questions are motivated to develop quiz games (e.g., [12]) or for (self-)assessment (e.g., [4]). However, as far as we know, no research publication reported their use in learning settings out of the classroom. Moreover, they all use factual knowledge from a single dataset. Thus, they do not fully exploit the potential of the Web of Data, as relevant information about the same entity may be published in different interlinked datasets. Finally, the questions generated are not suitable for SLEs, (where learning needs to be adapted to the learning space and context of the learner), as these questions are not related to any educational context, nor to any physical location. Note that these aspects could be stated by further exploiting the Web of Data, as several datasets contain the geolocalization of many physical entities (e.g., DBpedia, Wikidata (https://www.wikidata.org), LinkedGeoData (http://linkedgeodata.org/About)).

All in all, we consider that current state of the art does not fully take advantage of the data already available on the Web to create educational resources for SLEs. In our opinion, a much better support would be offered if:

- We created learning resources out of several integrated datasets available in the Web. Thus, we would be able to obtain a more complete collection of entities from different sources of the Web of Data.
- We automatically contextualized these learning resources taking into account their topic and the physical locations where they may be relevant. Thus, we would enable an SLE to offer learning resources to students according to their learning interests and their physical context.
- We do not only consider resources that assess factual knowledge. Thus, we would also promote higher-level thinking.

3. Technical Approach

We aim to automatically create contextualized learning resources related to physical locations and the student's learning interests out of the data available in the Web. This problem can be divided into two: the creation of a domain knowledge base out of the data from the Web; and the creation of a set of learning resources -and their corresponding metadata- out of such domain knowledge base. Figure 1 shows the system architecture.

As depicted in Figure 1, the architecture of the system includes two main components: a *Web of Data crawler* ([5] chap. 6) and a *Learning resource generator*. The *Web of Data crawler* collects data from different sources of the Web of Data and integrates it to create a *Domain Knowledge Base*; while the *Learning resource generator* applies a set of templates in order to create learning resources (e.g., a resource that invites students to look at some details of a Cathedral) and their metadata (e.g., relationship between the resource and the topics covered in the classroom or the geolocation where the resource could be relevant) out of the *Domain Knowledge Base*. Next we provide more details about the functionality of these two components.
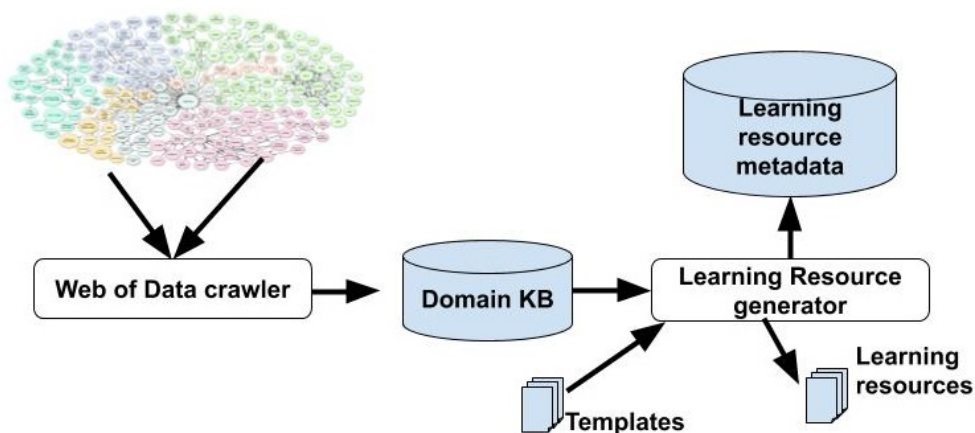

Figure 1. System architecture.

The *Web of Data crawler* follows the best practices suggested by Heath and Bizer ([5] chap. 6) (see Figure 2, which is particularized for our current prototype as later on explained). It includes a set of scripts to collect data from datasets that follow the Linked Data principles (5-star dataset, according to the well-known ranking by Tim Berners Lee [9]), others to parse data from other open datasets published on the Web (3-star datasets), and others to integrate the data collected. More specifically, we considered the following five scripts:

- **Extractor**. This script collects entities from an open data source that includes a SPARQL endpoint and relate them to the ontology used in the *Domain Knowledge Base*.
- **Descriptor**. This script collects the description of the entities extracted from that same data source and relates it to the ontology used in the *Domain Knowledge Base*. The descriptions obtained should include the `owl:sameAs` relationships stated in the data source for each element.

- **Enricher**. This script further describes each entity by extracting descriptions from other data sources and relates the data collected to the ontology used in the *Domain Knowledge Base*. This is done by exploiting the `owl:sameAs` relationships obtained by the *Descriptor*.
- **Parser**. This script collects data from other datasets that are available on the Web of Data but are not offered through a SPARQL endpoint, nor provide explicit relationships to the previous datasets (these datasets are typically offered in Open Data portals as downloadable files). The script relates the data provided by these datasets to the ontology used in the *Domain Knowledge Base*.
- **Integrator**. This script integrates all the data obtained by the previous scripts. As the entities are described using the same ontology, the integration focuses on resolving the entities of not-linked datasets.
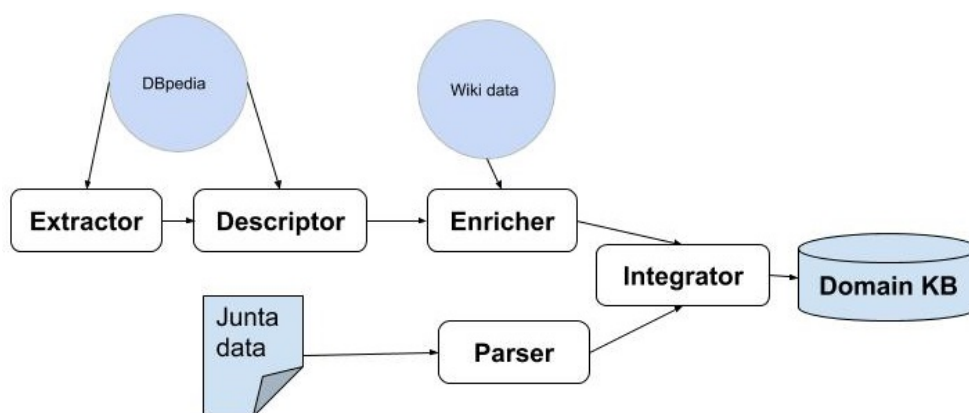

Figure 2. Current prototype of the Web of Data crawler.

Regarding the *Learning resource generator*, its technical functionality is very similar to other proposals that create MCQs or learning resources from a close dataset (e.g., [3]) or from the Web of Data (e.g., [12]). It simply applies a set of templates that select entities from the *Domain Knowledge Base* according to a set of rules (e.g., belonging to certain class or being described by certain parameters) and use the entity descriptions to create learning resources and their corresponding metadata.

We developed an initial version of the *Web of Data crawler* and the *Learning resource generator*. This initial version focuses on creating a *Domain Knowledge Base* that includes descriptions of historical buildings in Castilla y Leon (Spain), which will be later on exploited to create learning resources about History of Art. They are developed using Javascript. As depicted in Figure 2, the current version collects descriptions of these buildings from DBpedia, Wikidata and the Open Data Portal of Castilla y Leon (https://datosabiertos.jcyl.es/web/jcyl/set/es/cultura-ocio/monumentos/1284325843131). The resulting *Domain Knowledge Base* contains descriptions of 2600 buildings from Castilla y Leon (see [13] for more details). Later on, we applied several templates and we obtained several thousands of learning resources with their corresponding metadata.

As an example, part of the description of the resource "Monasterio de San Juan de Duero" is reproduced next. Note that many of these parameters include data from two or three sources of the Web of Data. For example, DBpedia (http://es.dbpedia.org/resource/Monasterio_de_San_Juan_de_Duero) states that the architectural style is "Romanesque", while the open data published by the Junta of Castilla y Leon states that it is "Romanesque" and "Mudejar". Note that these differences between data sources can very well be exploited by the *Learning resource generator* to create learning resources. Indeed, a template could filter the religious buildings described with one style in a data source (considering it the "predominant style") and by this same one and others in another data source (considering that some elements of the

building belong to such style). Then, this template may state how to create a learning resource -and its corresponding metadata (e.g., geolocation, age group or related learning interest)- that asks students to find out the elements related to the non-predominant architectural style (see [13] for more details and other examples). Applying this template, it is possible to obtain the learning resource depicted in Figure 3.

```
{"id": "Monasterio_de_San_Juan_de_Duero",
"label": {
      "dbpedia": ["Monasterio_de_San_Juan_de_Duero"],
          "JCYL": ["Monasterio_de_San_Juan_de_Duero"]},
"abstract": {
          "dbpedia": ["San_Juan_de_Duero,_conocido..."],
          "JCYL": ["Es_uno_de..."]},
"image": {
          "dbpedia":[".../Soria_-_San_Juan_de_Duero09.jpg"]},
"style": {
          "dbpedia": ["Estilo_romanico"],
          "JCYL": [["Mudejar","Romanico"]]},
"coordinates": {
          "latitude": {
                  "wikidata": "41.768398888",
                  "JCYL": "41.768582"},
          "longitude": {
                  "wikidata": "-2.454345",
                  "JCYL": "-2.454413"}},
"municipality": {
          "JCYL": ["Soria"]}
"type": {
      "dbpedia": [["Monasterio"]],
          "JCYL": [["Monasterio_/_Convento"]]},
"historical_period": {
      "JCYL": [["Edad_Media"]]},
"source": [
          "http://es.dbpedia.org/.../Monasterio_de_San_Juan_de_Duero",
          "http://www.wikidata.org/entity/Q3813454",
          "https://opendata.jcyl.es/.../monumentos.json#25178"]}
```

The main style of the Monastery of San Juan de Duero (which is in front of you) is Romanesque. However, it also includes Mudejar elements. **Could you detect Mudejar elements in this temple?**

Figure 3. Example of a learning resource related to the "Monasterio de San Juan de Duero".

An important aspect is that these scripts are not ad-hoc developed for a specific domain or for some specific datasets. The scripts use open standards and vocabularies heavily used in the Web of Data. For this reason, these same scripts can be used to create learning resources for other domains (e.g., botanic or literature) collecting data from other datasets (e.g., Spanish National Library (http://datos.bne.es/inicio.html) or the Spanish forest indicators (https://www.miteco.gob.es/es/biodiversidad/temas/inventarios-nacionales/)). For the scripts to be adapted to these other domains and datasets, it would only be required to state how the ontology of the *Domain Knowledge Base* is related to the ontology of these datasets; if not-linked datasets need to be integrated, then it would also be required to define how this integration should be done.

4. Conclusions

In this paper, we argue that the automatic creation of learning resources for Smart Learning Enviornments (SLEs) is underexplored. We also consider that the data available on the Web offers a very interesting opportunity for this creation of learning resources because there are descriptions of a vast amount of entities, many of them geolocalized and explicitly related to certain topics. These descriptions can well be exploited to automatically create learning resources associated to the topics covered in the student's formal education and related to entities in their daily life. Thus, we expect these resource to be offered to students by a context-aware recommender system that bridges between formal and informal learning processes.

We presented our first attempt to automatically create a collection of learning resources out of several datasets available on the Web. We explored the topic of historical buildings of Castilla y Leon, creating a local knowledge base that contains descriptions of 2600 buildings and allows us to obtain thousands of learning resources. This first attempt also lets us reflect on the most critical steps for this automatic

creation of learning resources. More specifically, we consider four very relevant aspects that will be tackled in our future work:

- Definition of an ontology. The definition of the ontology for the domain knowledge base is a critical step as it states the vocabulary for the description of the domain. This vocabulary should include the abstractions used by teachers and students; it should also cover the concepts that are relevant for a particular topic in a particular course, so it should take into account the course syllabus.
- Integration of datasets. The integration of datasets is a very well-known issue that is facilitated for those 5-star linked datasets. Unfortunately, not all the relevant datasets published on the Web are rated with 5 stars. For those not-linked open datasets, the identity resolution becomes a problem. For our example, we tried to overcome this problem by exploiting the entities' geolocalization (i.e., we understood that two entities are the same if, and only if, they are located in the same place); however, this approach seems not to be enough. Hence, we will explore other algorithms to overcome this problem.
- Definition of resource's templates. The definition of templates becomes very relevant to obtain resources out of the domain knowledge base. In our current prototype, these templates are defined by a technician. However, we will explore how to allow teachers to define these templates by manipulating a resource-creation application.
- Integration of resources in an SLE. Another relevant issue is how the resources created can be exploited in the context of an SLE. We foresee that a mobile application could be integrated into an SLE in order to offer relevant resources to the students according to their contexts. Some gamification techniques may also be useful to help the adoption of such mobile application.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hwang, G. Definition, framework and research issues of smart learning environments—A context-aware ubiquitous learning perspective. *Smart Learn. Environ.* **2014**, *1*, 1–14.
2. Gros, B. The design of smart educational environments. *Smart Learn. Environ.* **2016**, *3*, 15.
3. Alsubait, T.; Parsia, B.; Sattler, U. Ontology-Based Multiple Choice Question Generation. *KI-Künstliche Intell.* **2016**, *30*, 183–188.
4. Foulonneau, M. Generating Educational Assessment Items from Linked Open Data: The Case of DBpedia. T*he Semantic Web: ESWC 2011 Workshops*; García-Castro, R., Fensel, D., Antoniou, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 16–27.
5. Heath, T.; Bizer, C. *Linked Data: Evolving the Web into a Global Data Space*, 1st ed.; Synthesis Lectures on the Semantic Web: Theory and Technology; Morgan & Claypool, San Rafael, California (USA), 2011. Available online: http://linkeddatabook.com/editions/1.0/ (accessed on 7 November 2019).
6. Tarasowa, D.; Khalili, A.; Auer, S. Crowdlearn: Crowd-sourcing the creation of highly-structured e-learning content. *Int. J. Eng. Pedagog.* **2015**, *5*, 47–54.
7. Šimko, M.; Barla, M.; Bieliková, M. ALEF: A framework for adaptive web-based learning 2.0. In Proceedings of the IFIP International Conference on Key Competencies in the Knowledge Society, Brisbane, Australia, 20–23 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 367–378.
8. Leo, J.; Kurdi, G.; Matentzoglu, N.; Parsia, B.; Sattler, U.; Forge, S.; Donato, G.; Dowling, W. Ontology-Based Generation of Medical, Multi-term MCQs. *Int. J. Artif. Intell. Educ.* **2019**, *29*, 145–188.

9. Berners-Lee, T. Linked Data-Design Issues. 2006. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 7 November 2019).

10. Nahhas, S.; Bamasag, O.; Khemakhem, M.; Bajnaid, N. Added Values of Linked Data in Education: A Survey and Roadmap. *Computers* **2018**, *7*, 45–70.

11. Zavala, L.; Mendoza, B. On the Use of Semantic-Based AIG to Automatically Generate Programming Exercises. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education, Baltimore, MD, USA, 21–24 February 2018; ACM: Baltimore, MD, USA, 2018; pp. 14–19.

12. Vega-Gorgojo, G. Clover Quiz: A trivia game powered by DBpedia. *Semant. Web J.* **2019**, *10*, 779–793.

13. Ruiz-Calleja, A.; Bote-Lorenzo, M.L.; Vega-Gorgojo, G.; Serrano-Iglesias, S.; Asensio-Pérez, J.I.; Dimitriadis, Y.; Gómez-Sánchez, E. Exploiting the Web of Data to bridge formal and informal learning experiences. In Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 19), León, Spain, 16–18 October 2019; ACM: León, Spain, 2019, in press.