# Towards Integration of Web Data into a Coherent Educational Data Graph

Davide Taibi
Institute for Educational Technologies
National Research Council of Italy
Via Ugo La Malfa, 153
90146 Palermo, Italy
davide.taibi@itd.cnr.it

Besnik Fetahu, Stefan Dietze
L3S Research Center
Appelstr. 9a
30167 Hannover, Germany
{fetahu, dietze}@l3s.de

## ABSTRACT
Personalisation, adaptation and recommendation are central aims of Technology Enhanced Learning (TEL) environments. In this context, information retrieval and clustering techniques are more and more often applied to filter and deliver learning resources according to user preferences and requirements. However, the suitability and scope of possible recommendations is fundamentally dependent on the available data, such as metadata about learning resources as well as users. However, quantity and quality of both is still limited. On the other hand, throughout the last years, the Linked Data (LD) movement has succeeded to provide a vast body of well-interlinked and publicly accessible Web data. This in particular includes Linked Data of explicit or implicit educational nature. In this paper, we propose a large-scale educational dataset which has been generated by exploiting Linked Data methods together with clustering and interlinking techniques to extract import and interlink a wide range of educationally relevant data. We also introduce a set of reusable techniques which were developed to realise scalable integration and alignment of Web data in educational settings.

## Categories and Subject Descriptors
H.3.4 [**Semantic Web**], C.0 [**Systems Application Architecture**], I.2.4 [**Ontologies**]

**General Terms** Design, Experimentation, Standardization

## Keywords
Recommender System, Linked Data, Semantic Web, TEL.

## 1. INTRODUCTION
While personalisation, adaptation and recommendation are central features of Web-based educational environments, recommender systems apply information retrieval techniques to filter and deliver learning resources according to user preferences and requirements. Widely used approaches deploy collaborative filtering or content-based filtering techniques to identify and deliver suitable educational resources to learners. Hence, the suitability and scope of possible recommendations is fundamentally dependent on the quality and quantity of available data, data about *learners*, and metadata about *learning resources*.

However, particularly with respect to the landscape of standards and approaches currently exploited to share and reuse educational resources, and in particular Open Educational Resources (OER), the metadata used to describe these types of resources is highly fragmented. A range of technologies are exploited by a wide educational resource repository providers to support interoperability. In addition, the widespread availability of content on the Web, in particular the Social Web, has led to a growing importance of informal learning on the Web, which exploits a wide range of not explicitly educational content for learning and knowledge acquisition. To this end, although a vast amount of educational content and data is shared on the Web in an open way, the integration process is still costly [4].

In the past years, TEL research has already widely attempted to exploit Semantic Web technologies in order to solve interoperability issues. However, while the Linked Data (LD) [1] approach has widely established itself as the de-facto standard for sharing data on the (Semantic) Web and has produced a wide range of highly relevant datasets, it is still not widely adopted by the TEL community. Linked Data is based on a set of well-established principles and (W3C) standards, e.g. RDF, SPARQL[1], aiming at facilitating Web-scale data interoperability. Despite the fact that the LD approach has produced an ever growing amount of data sets, schemas and tools available on the Web, its take-up in the area of TEL is still very limited [2]. Thus, the potential contribution of LD for learning analytics scenarios is two-fold:

1.) Educational data (such as OER resources metadata) as well as relevant but not explicitly educational data (such as academic publications, domain knowledge).

2.) LD techniques provide technical solutions to substantially alleviate interoperability issues and to improve quality, quantity and accessibility of TEL data.

In this paper we propose a *scalable approach* which takes advantage of both contributions mentioned above by harvesting educationally relevant data – Linked Data as well as non-LD Web data – and integrating and exposing them as a unified and well inter-connected educational graph. To this end, we provide (i) a well-integrated *educational data graph* which establishes links between previously disparate datasets and (ii) a set of *scalable techniques* which facilitate our work.

## 2. RELATED WORK
While vast amounts of *Open Educational Resources* (*OER*) became freely available online their availability is a common objective for universities, libraries, archives and other knowledge-intensive institutions and provides an important resource for

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
*WWW 2013 Companion,* May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] www.w3.org/TR/rdf-sparql-query/.

educational recommender systems. This raises a number of issues, particularly with respect to Web-scale *metadata interoperability* or legal as well as *licensing aspects*. Several competing standards and educational metadata schemata have been proposed over time, including IEEE LTSC LOM[2] (*Learning Object Metadata*), one of the widest adopted, IMS[3], Ariadne, ISO/IEC MLR - ISO 19788[4] Metadata for Learning Resources (MLR) and Dublin Core[5]. The adoption of a sole metadata schema is usually not sufficient to efficiently characterize learning resources [5]. As a solution to this problem, a number of taxonomies, vocabularies, policies, and guidelines (called *application profiles*) are defined [4]. Some popular examples are: UK LOM Core[6], DC-Ed[7] and ADL SCORM (see also [11]). Repositories also exploit diverse interface mechanisms such as OAI-PMH[8] or SQI[9]. Due to the diversity of exploited standards, existing *OER repositories offer very heterogeneous datasets*, differing with respect to schema, exploited vocabularies, and interface mechanisms.

Regarding the presence of *educational information in the linked data landscape*, two types of linked datasets need to be considered: (1) datasets directly related to educational material and institutions, including information from open educational repositories and data produced by universities; (2) datasets that can be used in teaching and learning scenarios, while not being directly published for this purpose. This second category includes for example datasets in the cultural heritage domain[10] as well as by individual museums and libraries[11,12]. It also includes information related to research in particular domains, and the related publications[13], as well as general purpose information for example from Wikipedia (see DBpedia.org).

The Open University in the UK was the first education organization to create a linked data platform to expose information from across its departments, and that would usually sit in many different systems, behind many different interfaces (see http://data.open.ac.uk) which includes around 5 Million triples about 3,000 audio-video resources, 700 courses, 300 qualifications, 100 Buildings, 13,000 people [14]. Many other institutions have since then announced similar platforms, including in the UK the University of Southampton (http://data.southampton.ac.uk) and the University of Oxford (http://data.ox.ac.uk). Outside the UK, several other universities and education institutions are joining the Web of Data, by publishing information of value to students, teachers and researchers with LD[14,15,16]. In addition, educational resources metadata has been exposed by the mEducator project [12]. A more thorough overview of educational data and Linked Data is offered by the Linked Education[17] platform and in [4]. The approach proposed in this paper to enhance educational data is strictly related to the explicit semantic analysis (ESA) [7], in which Wikipedia is used as a knowledge base. In our approach DBpedia, the semantic representation of Wikipedia is used. A detailed description is presented in Section 5.

# 3. SCALABLE LINKING OF EDUCATIONAL DATA & RESOURCES

As shown above, there is an abundance of educationally relevant data and knowledge available on the Web, where the main obstacle towards Web-scale integration is the lack of interoperability, integration and, fundamentally, *links* between different datasets. Additionally, given the diversity of available resources, a *classification* of the available datasets, indicating their *main purpose*, *nature* and *educational relevance* is required.

As a first step a representative list of datasets was selected, which reflect the criteria described above, i.e., which are diverse with respect to their data representation (schema, vocabularies) as well as their content (purpose, domain). To this end, both explicitly educational datasets (such as *OpenLearn* or the *mEducator Educational Resources*) as well as implicitly educationally relevant datasets (such as *BBC Programmes* or the *ACM Library Metadata*) were selected (a detailed description of the datasets is reported in section 3.2). The imported data lacked any logical or semantic integration which would allow a cross-dataset exploration. to provide a coherent, well-aligned dataset (or, an *educational graph*), our method comprises the following activities:

1) *Schema-level integration and dataset categorisation*: schema-level mappings are defined by means of an upper level RDF schema which aligns disparate schemas used by different datasets. In addition, to enable an initial classification of different resource types, the upper schema introduces a vocabulary of educational resource types as vocabulary to describe the nature of each dataset.
2) *Instance-level integration – scalable enrichment*: out-of-the-box named entity recognition (NER) and disambiguation techniques are used to detect *entities* (e.g., *people*, *subjects*, *locations*) in semi-structured resource descriptions and enrich these with references to structured entities in jointly used vocabularies, such as DBpedia.
3) *Clustering and correlation*: the enrichments from step (2) are exploited to identify previously disconnected educational resources which are related, by, e.g., addressing similar subjects.

During the following sections, we elaborate the implementation stages taken to realise the above steps and demonstrate the gradual improvement of our dataset. The preliminary selection of datasets used in our experiments took into consideration the heterogeneity of the data with the aim of creating an integrated dataset which combines a wide variety of educational as well as educationally related resources. Included educational datasets are:

- *LinkedUniversities:* this dataset consists of more than 14,000 video lectures of 27 different academic institutions [6], such as the *Open University (UK)* or the *Khan Academy*.

[2] http://ltsc.ieee.org/wg12/par1484-12-1.html
[3] http://www.imsglobal.org/metadata/
[4] http://www.iso.org/iso/
[5] http://dublincore.org/documents/dces/
[6] http://zope.cetis.ac.uk/profiles/uklomcore/
[7] http://www.dublincore.org/documents/education-namespace/
[8] Open Archives Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html
[9] Simple Query Interface: http://www.cen-ltso.net/main.aspx?put=859
[10] http://www.europeana.eu/
[11] http://collection.britishmuseum.org/
[12] http://data.bnf.fr/
[13] http://www.ncbi.nlm.nih.gov/pubmed/ and http://thedatahub.org/dataset/bio2rdf-pubmed
[14] http://data.uni-muenster.de
[15] http://lodum.de
[16] http://openbiblio.net/2011/09/08/ntnu/
[17] http://linkededucation.org

- *mEducator Linked Educational Resources:* metadata about educational material in the medical domain and has been produced in the framework of the EU funded project mEducator [12].

In addition, educationally relevant resources are represented by relevant multimedia artefacts, TV broadcasts, academic publications or knowledge items. The datasets chosen for this category are the following:

- *Europeana:* this dataset provides metadata of resources related to European culture such as texts, books, film, museum artefacts [9].
- *DBLP Bibliography Database:* collected information related to academic publication in the computer science sector.[18]
- *ACM Library:* scientific papers published by the Association for Computing Machinery (ACM). The dataset contains paper metadata such as: authors, abstract.[19]
- *BBC Programmes:* metadata describing BBC broadcasts from all BBC channels (TV & radio) spanning several decades and domains. Given the BBC's strong involvement in distance education and reputation for world-leading educational programmes and documentaries, these artefacts constitute relevant resources also from an educational perspective. [10].
- *DBpedia*[20] & *Freebase*[21]: these datasets were used to provide additional structured knowledge about resources. Note that this data has not been imported but added in a more selective and elaborate "enrichment" process (see Section 5).

Data was imported into an OpenRDF/OWLIM store which provides a publicly accessible SPARQL endpoint[22] in itself. The repository created comprises about 97 million of triples and 21.6 GB of educational resources and related data.

# 4. DATASET CATALOGING AND MAPPING: THE LINKED EDUCATION SCHEMA

This first step of integration facilitates schema-level integration by means of an upper-level RDF schema which aims at: (a) *describing the general notions* in our dataset and their properties, to provide the basis for (b) *classifying the nature of the resources/datasets*, e.g. to distinct between dedicated educational resources (such as OER) and related resources (such as academic publications) and to (c) *enable mappings* with the diverse schemas of the different imported datasets. Thus, the schema will provide a general and reusable vocabulary to describe educational Web datasets, resource types and their relations and exploits established RDF schemas and vocabularies such as the *Vocabulary of Interlinked Datasets (VoID)*[23].

With respect to schema mappings, it emerged that well-known schemas such as Dublin Core (DC) or FOAF[24] are already widely used, reducing the work load on the required manual mappings. But at the same time, dataset-specific concepts, types and properties are used, differing at the semantic and the syntactic level. Thus, our mappings did not aim at an exhaustive alignment of all schemas, but primarily at an integration of properties where a semantic congruency of properties is given. This includes in particular higher-level properties such as *title* or *description*, which carry essential information about the content of the described resource. Mappings were simply defined by introducing upper level properties (such as *led:title*[25]) from which diverse variations (such as *dc:title*) were derived as sub-properties. Inserting our upper-level schema into our dataset, inference mechanisms automatically are able to consider the aligned schemas of integrated datasets.

# 5. SCALABLE ENRICHMENT OF EDUCATIONAL DATA

While previous steps involved the graph alignment at the schema-level, *data enrichment* and *named entity recognition (NER)* aim at the instance-level alignment. This is required due to (i) the diversity of used taxonomies across imported datasets and (ii) the widespread use of unstructured text (for instance, as part of resource descriptions).

## 5.1 Overview

Even though schema-level alignment provides a first step towards cross-dataset queries, discovery of resources across distinct datasets is still challenging. Hence, our enrichment phase aims at adding a common descriptive layer to the datasets which addresses (a) *identification of (common) named entities* from unstructured descriptions, (b) *disambiguation* and (c) *expansion of the limited resource descriptions with additional background knowledge*. Our approach takes advantage of established datasets such as DBpedia and Freebase by exploiting the DBpedia Spotlight API[26]. Spotlight has been chosen as it combines a number of required features such as term recognition, NER and disambiguation functionalities which enable the interlinking of semi-structured data with structured entities within the DBpedia graph with sufficient precision/recall as has been shown in previous work in the educational area [4]. Initially, our current implementation enriches resource titles and descriptions, as these are the most frequently used properties. This approach feeds each unstructured title and description from each resource into DBpedia Spotlight and associates the retrieved entity descriptions with the corresponding resource in our RDF graph. While this constitutes the ideal, most exhaustive but computationally expensive way of deploying DBpedia Spotlight, it is deemed *exhaustive approach* in the following, as opposed to the *scalable approach* described in the next section.

For instance, poorly structured descriptions, such as the title of one of our imported resources "*Linear Equations in Standard Form*" is automatically linked with the DBpedia concepts "*Linear_equation*" and " *Standard*" which, in turn, are linked to related knowledge within DBpedia. Enrichments allow further reasoning on related concepts and also enable users to query for resources by using well-defined concepts as opposed to ambiguous free text. It is also important to highlight that DBpedia Spotlight disambiguation features enable the correct recognition and association of terms, acronyms and similar texts. For instance, regarding the use of acronyms, the words "*dns*", "*gmt*" have been

[18] http://dblp.l3s.de/
[19] http://acm.rkbexplorer.com/
[20] http://dbpedia.org
[21] http://www.freebase.com/
[22] http://linkedup.l3s.uni-hannover.de:8880/openrdf-sesame/repositories/linked-learning-selection?query
[23] http://vocab.deri.ie/void/
[24] http://xmlns.com/foaf/spec/
[25] Note, *led* represents the namespace of our schema (http://data.linkededucation.org/ns/linked-education.rdf)
[26] http://spotlight.dbpedia.org/

successfully associated respectively with the concepts: *Domain_Name_System*, *Greenwich_Mean_Time*. Concerning synonyms, for instance, the term "*e^x*" has been enriched with the concept *Exponential_function* and the word "*globe*" has been enriched with the concept *Earth*. Finally, it is relevant to report that the same word can be associated to different concepts depending on the context it is used, few examples from our repository are the word "*apple*" enriched with the concepts *Apple_III*, *Apple_Inc*.

## 5.2 Scalable Approach

The enrichment process is a very crucial step when considering Linked Data scenarios, providing useful information for querying, clustering and interlinking of different datasets. However the process described above is only computable with small datasets and text corpora. This bottleneck, is caused, for instance, by, the computational complexity of NER and disambiguation tasks, the large amount of remote HTTP requests required when interacting with Web APIs such as DBpedia Spotlight and the comparably high response times of RDF storage and query mechanisms.

In order to alleviate these issues, we provide an alternative enrichment process which has been altered towards higher scalability and applicability to large-scale datasets, while at the same time, ensuring minimal impact on precision/recall. First, we detect *most probable enrichment candidate terms* by identifying all terms belonging to the part of speech (POS) tag *noun phrase* {NN, NNS, NNP, NNPS}. This has been identified as viable measure since previous analysis of DBpedia concepts has shown that over 92% of the existing concepts belong to this category. While this step already singles out terms from their context, i.e. co-occurring terms and introduces ambiguity, our evaluation (Section 7) has shown only insignificant variations in precision results.

In addition, to further minimize the amount of required HTTP requests, the simplifications of the previous step allows *tokenizing the text corpora* for the individual resources of a dataset. Tokens with different numbers of consecutive terms belonging to specific part of speech tags are considered. The created tokens from all resources are stored in a map data-structure, thus avoiding any duplicates, ensuring that we enrich only once similar tokens.

Applying the previous step of tokenization, the extracted tokens are taken out of the resource context what reduces the precision and recall of NER techniques. To ensure a still reasonable return of enrichments from DBpedia Spotlight, request parameters (*confidence*, *support*, and *context*) were reduced to zero, which has been experimentally identified as useful measure. In contrast, the exhaustive approach uses a value of 0.6 for the *confidence* parameter, allowing higher precision but lower recall. Valid enrichments are considered only those that match the number of terms involved in the token, and for tokens which are subsets of other tokens, only the supersets are kept.

In this way the text corpora, or any other means used for describing a resource can be reduced by removing terms that do not belong to certain POS tags, while the tokenizing process massively reduces the amount of computationally expensive enrichment requests, since for tokens appearing multiple times only one enrichment is needed. This has a very strong impact on fairly homogeneous datasets whereas for highly heterogeneous datasets with small numbers of frequently used terms and phrases the advantage is less visible (see Section 7).

# 6. CORRELATION & CLUSTERING

The previous section described the use of enrichments to add additional knowledge using a unified reference vocabulary (DBpedia) to the resources. The enrichment procedure has not the only benefit of providing a common base for queries across datasets, but it is also used to detect correlations between inherently related resources [13], that is resources which target similar topics. Following a similar approach used to classify documents, a *Resources-Enrichments-Matrix* has been created. The generic element of the matrix contains the frequency of enrichments in a resource. This matrix has been used as the starting point for the elaboration of three correlations methods based on different similarity measures.

The first method is a naïve method in which resources have been correlated if they share at least one enrichment. This method does not take into consideration the normalization of the number of enrichments. The latter is suggested by the wide variation of enrichments per resource (see Section 7.1) caused by the varied nature of resource descriptions in particular datasets, ranging from comprehensive texts to brief one-liners, what generates highly diverse amounts of enrichments per resource and hence, differing probabilities for correlations. For this reason, two further methods have been developed with the aim of taking into consideration the normalisation through the calculation of an adapted version of the *tf-idf (Term Frequency–Inverse Document Frequency)* usually adopted in linguistic analysis. In our context the *ef-irf (Enrichment Frequency-Inverse Resource Frequency)* index has been defined as follows. We consider a set of data with $n$ enrichments and $m$ resources the generic element of the matrix:

$$efirf_{\mathrm{i,j}} = ef_{ij} \times irf_{ij}$$

with

$$ef_{ij} = \frac{n_{ij}}{r_j}$$

where $n_{ij}$ indicates the number of occurrences of the enrichment $i$ in the resource $j$, and $r_j$ is the number of enrichments for the resource $j$. Furthermore, we define

$$irf_{ij} = \frac{\log(N)}{\log(r_i)}$$

where $N$ is the total number of resources and $r_i$ is the number of resources containing the enrichment $i$.

Please note, our ef-irf measure exploits the existence of a unified vocabulary (in the form of enrichments which provide links to a common reference vocabulary), and hence, aims at generating more precise results compared to traditional tf-idf measures applied to highly heterogeneous data such as ours. Our matrix of ef-irf values has been used as a starting point to calculate the Cosine and Jaccard similarity indices for resources. The cosine similarity index measures the similarity between any two vectors representing resources (in the enrichments space) evaluating the cosine of the angle between the two vectors. The output of these elaborations is a symmetric square matrix in which the generic element represents the similarity value between two resources.

The generated matrix has been used to create a weighted graph, in which Linked Education resources are the nodes and the edges are weighted taking into consideration the similarity value between the resources. The graph has been created considering only the values of similarity that overcome a predefined threshold.

Different views of the graph have been generated according to different thresholds of the similarity indices. In order to aggregate resources with similar features, the graph has been clustered using the "*Edge Betweenness Clustering"* approach proposed by Girvan and Newman [7]. A comparison between the three approaches is summarized in Table 1.

**Table 1: Nodes, edges and clusters**

| | Cosine | | | Jaccard | | | Naive | | |
|---|---|---|---|---|---|---|---|---|---|
| Threshold | 0.3 | 0.5 | 0.7 | 0.3 | 0.5 | 0.7 | 3 | 5 | 7 |
| # Edges | 1530 | 357 | 78 | 285 | 108 | 7 | 1141 | 120 | 8 |
| # Nodes | 683 | 191 | 57 | 126 | 57 | 6 | 184 | 83 | 7 |
| # Clusters | 59 | 26 | 8 | 16 | 7 | 1 | 13 | 10 | 2 |

Different thresholds generate different networks and thus different cluster aggregations between the resources. An example of network is shown in Figure 1.

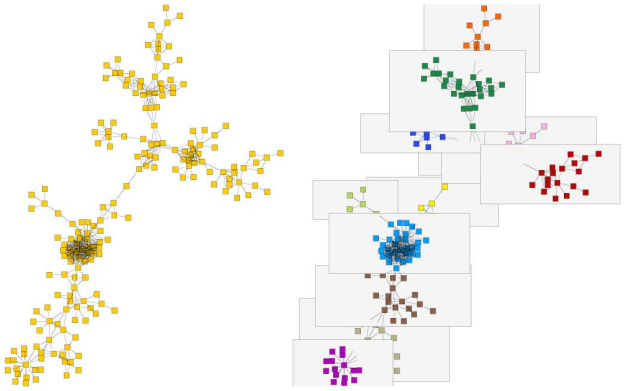

**Figure 1:Network of resources before and after clustering**

It is important to highlight that the aggregation of resources in the cluster is not only based on the DBpedia concepts they were enriched with, but is also based on the topology of the relationships that the resources have in the network. Consequently, resources can belong to the same cluster even if they do not have enrichments in common. The aggregation of resources in clusters enables more efficient search of similar resources based on concepts, thus providing an innovative approach for educational recommender systems.

# 7. EVALUATION

The evaluation procedure compares the benefits, and drawbacks of using the different approaches proposed for enrichment, and interlinking in terms of quantitative, qualitative, and performance gains, for a smaller subset from the original dataset with approximately *250* resources for each context. The original dataset contains around *5,953,623* distinct resources, with *Europeana, DBLP,* and *ACM* contexts having the largest number of resources. For the qualitative evaluation, we rely on relevance judgments for both of our enrichment approaches.

Enrichments from both the exhaustive and the scalable approach were evaluated independently to allow the comparison of recall and precision. In addition, the performance evaluation analyses the efficiency of the different enrichment approaches, in terms of considered data, and execution time.

## 7.1 Quantitative data assessment

To give a clear picture of the educational resources involved in our evaluation, we show the different contexts (original datasets) the resources belong to, context entity type associations, and number of distinct enrichments made for each of the contexts using the scalable approach. In Table 2, the number of educational resources per context is displayed. Note that here we refer to our evaluation subset of the Linked Education graph.

**Table 2: Resource Context Distribution**

| Context | #Resources | #Enrichments | #Entity Types |
|---|---|---|---|
| ACM | 249 | 200 | 239 |
| mEducator | 250 | 495 | 355 |
| BBC | 250 | **1364** | **769** |
| LinkedUniversities | 243 | 166 | 283 |
| DBLP | 250 | 295 | 161 |
| Europeana | 249 | 938 | 672 |
| **Total** | **1491** | **3458** | **937** |

As described previously in Section 5.2, we analyzed the textual content of our resources and educational data, considering only POS tags of *noun phrase* as possible enrichment candidates. From the measures the choice of tags which are the most likely to contain DBpedia concepts cover *63%* of the whole text contained in the original educational resources metadata. This result supports the viability of our scalability adjustments (see Section 5.2). Another important aspect is the association of resources with the different entity types found during the enrichment phase. This is crucial, since it facilitates the clustering, and interlinking of related contents with higher accuracy. Additionally, this gives important insights on the categorization of the resources based on their associated entity types. In total there were *938* distinct types associated to the resources and educational data considered, with some types having more than thousands of assignments

Other quantitative measures of interest are the number of disambiguated and enriched entities found at particular resources. This is directly related to the length of the text used in a resource, where in our case during the enrichment phase, for resources with longer text, we were able to identify up to *87* distinct entities, and more than *200* entity type associations.

## 7.2 Qualitative evaluation

In this section, we evaluate the results of the enrichment process, and the accuracy of detected clusters. For the enrichment process, each disambiguated entity is evaluated if it's relevant to the context of the resource to which it belongs. The relevance judgments are taken using crowd-sourcing, where in our case we used Crowdflower[27]. We evaluated *2000* enrichments for both enrichment approaches. In order to achieve fair relevance judgments, we limit the number of tasks (*200* tasks) that can be completed from a single user, thus we have a more representative set of relevance judgments. We are aware that more than one relevance judgment is needed for evaluating an enrichment in order to aggregate and have more trustful judgments, but since the evaluation is done in the same fashion for both approaches, we think this represents a fair comparison. The number of users involved for the first approach, is *32* with an average of *63* completed tasks, whereas for the second approach, there were *23* users with an average of *87* completed tasks. From the evaluation results, we get an accuracy of *82%* for the exhaustive approach,

[27] https://crowdflower.com/

while for the scalable approach as defined in Section 5.2 we get an accuracy of *77%*. We measure the values for the recall metric, for both approaches for the set of evaluated enrichments, where the first approach achieves a recall value of approximately 43*%* and the second approach achieves a recall value of *69%*. We take the exhaustive approach where the whole content of a resource and educational data is considered for enrichment as the baseline, and compare the second approach (scaled) against the baseline. Results show only minor deviations in precision, whereas with respect to recall, the scalable approach outperforms the exhaustive approach by 26%.

## 7.3 Performance evaluation

The gains in terms of performance are measured to assess the scalability of our approach proposed in Section 5.2. For this reason, we evaluate two aspects of the enrichment process. First we consider the reduction of terms to be analyzed during the enrichment phase, where by taking only the terms with POS tag, we reduce the amount of text by almost *40%*. Taking into account this factor, for a token containing a single term, we can reduce the number of tokens considered for enrichment for up to *86%*.

In summary, a significant gain is achieved by following the previous two steps, where the amount of text and set of tokens for enrichment is reduced drastically. Other performance attributes are the NER complexity task from DBpedia Spotlight, and the reduction of HTTP requests. As attempt to underline the overall performance gain, we measured, the time required to conduct the enrichment process with both approaches (exhaustive, scalable) on a randomly selected set of resources, coming from different datasets, with over *1700* resources. The whole process was run on a PC, using only a single CPU. While it took approximately *20* minutes to finish the scalable approach, the exhaustive enrichment process took approximately 3.5 hours.

# 8. CONCLUSION

In this work we have described our efforts in creating a Linked Education dataset, by exploiting several methods to enrich, disambiguate and interlink large-scale educational Web data into a coherent educational data graph. As shown by the evaluation results, our enrichment procedures provided reasonable precision results, for both, the exhaustive and scalable approaches. Hence, in particular the scalable approach described here introduces a number of improvements which significantly increase performance in order to offer a scalable approach for integration and alignment of disparate (educational) datasets. Even though recall values (0.429; 0.687) still leave room for improvement, our processing also provided a means to detect correlated educational resources and data out of completely previously disparate and highly heterogeneous datasets. Future work will deal with the integration of our dataset into recommender systems which would allow the retrieval of educational data and resources according to specific learning contexts, learner histories and preferences.

# 9. ACKNOWLEDGMENTS

This work is funded in part by the LinkedUp project (Grant Agreement: 317620) under the FP7 programme of the European Commission.

# 10. REFERENCES

[1] Bizer, C., Heath, T., Berners-Lee, T. 2009. Linked data - The Story So Far. *Special Issue on Linked data. International Journal on Semantic Web and Information Systems*.5(3):1-22

[2] Dietze, S., 2012. Linked Data as facilitator for TEL recommender systems in research & practice. *In Proceedings of 2nd Workshop on Recommender Systems for Technology Enhanced Learning, at 7th European Conference on Technology-Enhanced Learning*, Saarbrücken.

[3] Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. and Taibi, D. 2012. Linked Education: interlinking educational Resources and the Web of Data. *In Proceedings of the 27th ACM Symposium On Applied Computing, Special Track on Semantic Web and Applications*, Riva del Garda (Trento), Italy.

[4] Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H., Giordano, D., Marenzi, I., Pereira Nunes, B. 2013. Interlinking educational Resources and the Web of Data – a Survey of Challenges and Approaches. Accepted for publication in *Emerald Program: electronic Library and Information Systems*. 47, 1.

[5] Duval, E., Hodgins, W., Sutton, S. and Weibel, S. 2002. Metadata Principles and Practicalities. *D-Lib Magazine*. 8, 4.

[6] Fernandez, M., d'Aquin, M., and Motta, E. 2011. Linking Data Across Universities: An Integrated Video Lectures Dataset. *In Proceeding of the 10th International Semantic Web Conference*, Bonn, Germany.

[7] Gabrilovich E., Markovitch S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *In Proceedings of The Twentieth International Joint Conference for Artificial Intelligence,* Hyderabad, India, 2007.

[8] Girvan M. and Newman. M. E. J. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. 99, 12, 7821-7826.

[9] Haslhofer, B., Isaac. A. 2011. data.europeana.eu - The Europeana Linked Open Data Pilot. *In Proceeding of the International Conference on Dublin Core and Metadata Applications*.

[10] Kobilarov, G., Scott T., Raimond Y., Oliver S., Sizemore C., Smethurst M., Bizer C., Lee R. 2009. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Conections. *In Proceedings of the 6th European Semantic Web Conference*.

[11] Koutsomitropoulos, D.A., Alexopoulos, A.D., Solomou, G.D. and Papatheodorou, T.S. 2010. The Use of Metadata for Educational Resources in Digital Repositories: Practices and Perspectives. *D-Lib Magazine*.

[12] Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P., Bratsas, C. and Woodham, L. 2011. Connecting Medical Educational Resources to the Linked Data Cloud: the mEducator RDF Schema, Store and API, *in Linked Learning 2011, Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age*, CEUR-WS, Vol. 717.

[13] Nunes, B. P., Kawase, R., Dietze, S., Taibi, D., Casanova, M.A., Nejdl, W. 2012. Can entities be friends?. *In Proc. of Web of Linked Entities (WOLE2012), Workshop at The 11th International Semantic Web Conference*, Boston, US.

[14] Zablith, F., d'Aquin, M., Brown, S. and Green-Hughes L. 2011. Consuming Linked Data Within a Large Educational Organization. *In Proceedings of the 2nd International Workshop on Consuming Linked Data* at International Semantic Web Conference*.* Bonn, Germany.