Ontology-based teacher-context data integration

Nader N. Nashed[1], Christine Lahoud[2] and Marie-Hélène Abel[3]

***Abstract*— The divergent web-based data sources and the large amount of intentionally or automatically collected data offer a great opportunity to refine the existing knowledge in various fields, including education. In particular, the wide range of technologies provides a diversity of data such as resources for education, e-learning systems, teacher and learner analytic, educational institutions data, and the surrounding environment data. Educational data integration got the researchers' attention due to the continuous emergence of new technologies in addition to the data heterogeneity from different sources. This integration process has two main challenges: data discovery by finding relevant datasets and exploitation by measuring the effectiveness of these datasets. Teachers coexist in two main contexts: their living environment and the work/educational environment. Therefore, the multiple teacher's contexts cannot be represented through single-type datasets. In this paper, we introduce a general approach for data integration of teacher-related educational datasets to provide a rich and linked data for the educational domain. This approach provides the method of mapping the integrated datasets into the teacher-context ontology (TCO); which represents the teacher's coexistence in multiple contexts.**

I. INTRODUCTION

In the current modern era, the number of available data sources increases significantly. These data are generated by humans and machines equally. It can be used to improve knowledge in various application fields. In particular, the extensive usage of web-based educational systems and other productivity-enhancing technologies by teachers and learners are transforming the educational field into multiple complicated data-intensive domains. Accordingly, these systems generate huge amount of stored data from multiple data sources with different formats and different contextual levels [1]. Data integration is one of the persistent challenges that has been addressed in different disciplines, specifically, knowledge management, information retrieval, and machine learning. The ontology is another source of knowledge representation for the educational data. An ontology defines types, properties, and relationships of different entities in a common domain context. Teacher-context ontology (TCO) introduces a comprehensive description of the coexistence of one teacher in multiple contexts [2].

[1]Nader N. Nashed is with Université Française d'Égypte, Cairo, Egypt and HEUDIASYC, Université de Technologie de Compiègne, Sorbonne universités, Compiègne, France nader.nashat@ufe.edu.eg, nader.nashed@utc.fr

[2]Christine Lahoud is with Université Française d'Égypte, Cairo, Egypt christine.lahoud@ufe.edu.eg

[3]Marie-Hélène Abel is with HEUDIASYC, Université de Technologie de Compiègne, Sorbonne universités, Compiègne, France marie-helene.abel@utc.fr

This research proposes an automatic ontology-based data integration methodology for multiple diverse heterogeneous data sources to represent the teacher context. The next section explains a statement of our problem. The following section provides an overview of data integration trends and their application in the education domain. The fourth section introduces our approach to solve the data integration and the Linked Data problems. The fifth section discusses the proposed solution in this paper and its impact. The final section concludes this paper and our perspectives in this matter.

II. PROBLEM STATEMENT

The various educational data are categorized into academic, nonacademic, fidelity of implementation, and perception data according to the type of education [3]. Academic category includes institution hierarchical structure, resources, programs, and syllabuses while nonacademic category represents the context and environmental information that describes the cultural, social, financial backgrounds of this environment. However, the data corresponding to the fidelity of implementation category is obtained through evaluation of the educational process performance. Perception data is deduced through the computational process of the collected analytics. In that instance, complex computational analysis is needed to deduce relevant knowledge from these raw data. The availability of these different data sources unlocks a broad range of opportunities for new knowledge revelation. Nevertheless, it introduces a new data integration challenge that can be approached through the analysis of the available data sources. In fact, researchers deduced that educational context cannot be understood by the analysis of a single-perspective data [4]. Another challenge arises by the exponential growth of educational data and the automatic process of transforming these data into useful insights [5].

Two main challenges are linked with the research in the data integration field: data discovery and data exploitation [6]. Data discovery is concerned with identifying the relevant data sources for the intended application, while data exploitation is the insights extraction from collected information. Traditionally, research efforts in this context are obstructed by the data heterogeneity issue. This issue arises because the multiple heterogenous sources stem from independent activities and different structures. In order to get a unified single data source, these different sources undergo a management process that composes of scheme

mapping, entity resolution and data fusion. The scheme mapping solves the heterogeneity issue at the structural level while entity resolution solves it by assembling the different descriptions of the same entity at value level. Working also on the value level, the data fusion process combines these different descriptions into a single representation.

Another source of knowledge management for educational field is the ontology. TCO represents the main concepts of the multiple contexts that affect a teacher's career in combination with the main concepts of an educational process [2]. The living environment is one of the multiple contexts where the teacher coexists, and it forms the teacher's cultural background. While the working environment represents the context where the teacher interacts with his learners and the educational institution's structure. TCO concepts include living environment of teacher and learner, and educational institution environment representation. Contrastingly, TCO comprises the educational institution's structure from the teacher point-of-view through the classrooms, courses, lessons, and resources concepts. The educational institution's structure with the resource concept cover the academic education data category while the living one completes the nonacademic category.

This research proposes a solution for the stated challenges by finding and integrating the appropriate data sources that represent the different educational contexts from a teacher's point-of-view. The primary objective is presenting an automated approach for data mapping of multiple data sources in the educational context into the ontology representation of TCO.

III. RELATED WORK

In the pursuit of finding an optimal solution for the integration problem, we offer an overview of the related work and research that may propose a solution for the integration problem for data sources broadly and educational resources particularly.

The data integration problem arises in many research fields including life science. One of the remarkable efforts towards life science data integration is Bio2RDF project [7]. This project integrates 30 biomedical databases and datasets into RDF extended dataset with 10 billion triples. In order to enrich the cancer research field, the approach by [8] integrates 23 cancer-related datasets from five different categories. The research integrates different representations of datasets (e.g., JavaScript Object Notation (JSON), Tab Separated Value (TSV), and Comma Separated Value (CSV)) with the disease ontology (DO) [9].

The educational environment is categorized into full face-to-face classroom education, computer-based education, and mixture from both. Similarly, current educational information systems take many approaches such as Learning management system (LMS), Massive open online course (MOOC), and intelligent tutoring system (ITS). These different trajectories generate varieties of data that can aid the cause of solving many educational problems [10].

There are different sources of educational data collection such as the interaction of teachers and learner with the educational system (e.g., discussion board messages, search processes, user inputs in various modules, navigation behavior, etc.), organizational data (e.g., educational institution data, teacher data, etc.), demographic data (e.g., gender, age, education, etc.), environmental data (e.g., location, area type, cultural background, etc.), and affectivity (e.g., emotion, motivation) [11]. Therefore, the different educational environments tend to store a huge amount of data from various sources into different formats.

The data integration and linking trends and challenges in the education field was discussed as early as 2009 [12]. The authors categorize the educational datasets into two types: educational resources and institutions datasets, and teaching scenarios datasets. This review builds an important overview of the available educational datasets and its usage and nature. At learner level, there are remarkable efforts by FOAF to unify the learner's modeling and by Contextualized Attention Metadata [13] to introduce a comprehensive modeling of the learner activities and interactions. At the educational content level, EEE Learning Object Metadata (LOM) is the most significant initiative for the standardization and modeling of the learning objects. Another approach by [14], targets the aggregation of different educational content models through a data mapping algorithm. Recently, educational institutions start to expose their data following the data integration and linking approaches [15].

Most educational data mining and learning analytics researchers use self-built datasets in order to find a solution for a certain educational problem. This task is time-consuming and difficult to handle [11]. Another review [16] highlights the importance of learner modeling through the integration of recorded data from different learning systems or tools. The research that was conducted in [17], illustrates an approach of collecting and management of learners and content data from different sources (e.g., services and applications). The recent research by [18] integrates the learners activities data from Connectivist Massive Open Online Course (cMOOC) with the Social Networks Analysis (SNA) data to measure the impact of the published posts of learners on various social networks on their progress in the enrolled courses.

The MOOCLink research in [19], [20] introduces a new approach of educational courses data aggregation to compare the syllabi of courses of particular subject. Interestingly, the approach by [21] achieves a linkage between Coursera and Udacity MOOC courses to identify the similarity in the

taught ESCO (European Skills, Competences, Qualifications and Occupations) [22] skills that are provided by the same subject courses on both platforms.

The Educational Resource Discovery Index for Data Science (ERuDIte) [23] describes over 11,000 training resources (e.g., courses, video tutorials, conference talks, etc.) using Schema.org[1] metadata. Then, the resources are tagged with concepts from Data Science Education Ontology [24]. The teaching analytics are not fully discovered by researchers but the research by [25] introduces a novel approach to link the teaching analytics data with the learning analytics data. This approach provides the teacher with the necessary knowledge to understand and improve their teaching outcomes.

We can conclude that none of the reviewed research proposes an automatic integration methodology for multiple diverse data sources. Moreover, the accumulated research does not direct enough effort towards the educational data integration with different category of data sources. Therefore, a novel methodology is required to handle the data integration in an automated manner and based on the semantic representation of an ontology.

IV. PROPOSED APPROACH AND METHODOLOGY

A. Data Sources Layer

The coexistence of a teacher in multiple contexts forces the data extraction from different sources and their integration into one form of data format [4] as shown in Fig.1. A person can be either a teacher or a learner and is represented by the integration of user-provided information, organizational information, and user-recorded activities. These analytics are collected during the real-time usage of teacher to an educational system such as a recommender system [26]. The multiple contexts of a teacher are illustrated through environment-related and institutions datasets that are publicly published by governmental and research entities. This data integration approach highlights the necessity of resources to aid teachers and enhance their performance. Accordingly, resources' data are collected from multiple datasets that target different types of resources in the educational field.

Table I summaries the selected datasets and categorizes them into environment, institution, and resource. In this work, we selected the French territories as our case study that can be extended afterwards to other regions. Thus, "French employment, salaries, population per town" dataset is selected for the best representation of France [27]. This dataset is provided by the INSEE (L'Institut national de la statistique et des études économiques) and it gathers geographical (e.g., coordinates, cities, regions, departments, etc.),

[1]https://schema.org/

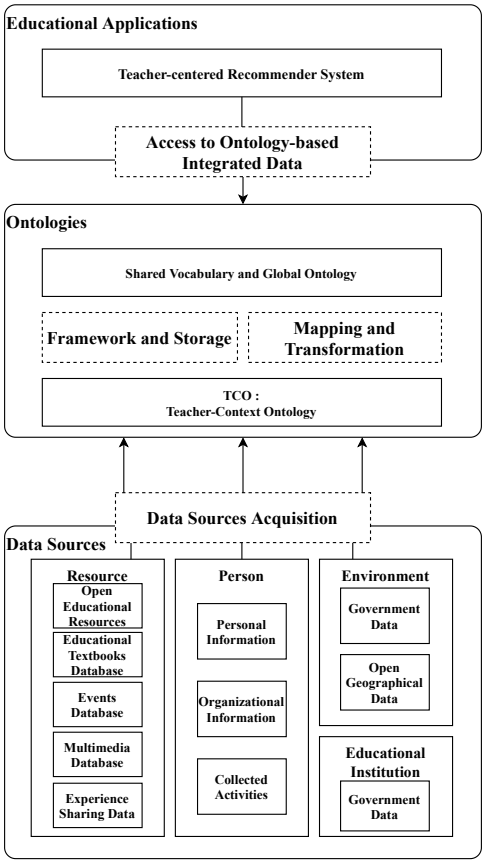

Fig. 1. An overview of the proposed approach's layers.

demographical (e.g., births, deaths, population density, etc.), and economical data (e.g., salary, firms, etc.). It is divided into four tables: *base_etablissement_par_tranche_effectif*, *name_geographic_information*, *net_salary_per_town_per_category*, and *population*. The *name_geographic_information* and *population* tables are the most important for the living environment description by providing the geographic information of all areas with demographic information about the living conditions. In order to represent the working environment, *base_etablissement_par_tranche_effectif* and *net_salary_per_town_per_category* tables provide an overview of the working conditions according to the region in terms of job category, salary, and firm size.

On the other hand, the educational institution data are essential in a teacher's context. These data are covered through three government datasets for each educational level. In this approach, the collected data represent the schools, high schools, and universities from educational priority networks in France (REP) [37]. Other institutions are not included due to the lack of complete datasets for this matter and the unnecessity of special treatment for these institutions. Three selected datasets are Écoles Education Prioritaire [28], Collèges Education Prioritaire [29], and Académies Education Prioritaire [30]. Then, the datasets are fused into one

TABLE I
SELECTED DATASETS SUMMARY

Dataset name	Targeted data	Format	No. of records
French employment, salaries, population per town [27]	Environment	CSV	34142
Écoles Education Prioritaire [28]	Environment + Institution	CSV + JSON	417147
Collèges Education Prioritaire [29]	Environment + Institution	CSV + JSON	47552
Académies Education Prioritaire [30]	Environment + Institution	CSV + JSON	1658
Coursera All Courses [31]	Resources	CSV	5157
edX All Courses [32]	Resources	CSV	3082
Udemy Courses [33]	Resources	CSV	42375
IT-Software Courses Udemy [34]	Resources	CSV	22750
Goodreads-books [35]	Resources	CSV	10352
Wikibooks Dataset [36]	Resources	CSV	185692

dataset using a conflict resolution function f defined on one dataset D and conflicting on another dataset S as in (1) [38]. The used CONCAT function f returns the concatenated records with annotation of the data source's name.

$$f : D \times ... \times D \rightarrow S \tag{1}$$

The resource data are the last part of the educational data integration. Six datasets are selected to comprise two type of resources: online MOOC courses (four datasets), and textbooks (two datasets). The MOOC courses datasets are collected through the application programming interfaces (APIs) of three different MOOC platforms: Coursera[2], edX[3], and Udemy[4]. The entity resolution problem arises because of the existence of different records from the different platforms referring to the same course. Therefore, a similarity-based technique is required to combine similar records/courses. For example, given two courses c_1 and c_2 from two different datasets with the following common properties

```
(course_id, course_title, course_author,
course_length)
```

the two courses are declared a match if

$$\begin{aligned} & w_1.f_1(\texttt{course\_title}) \\ & + w_2.f_2(\texttt{course\_author}) \\ & + w_3.f_3(\texttt{course\_length}) \geq \tau \end{aligned} \tag{2}$$

where f_i is the cosine similarity function of two characters' vectors in (3) [39], w_i is the weight for each property p_i, and a τ is the threshold.

$$Cosine(\vec{v}, \vec{w}) = \frac{\vec{v} \bullet \vec{w}}{\mid \vec{v} \mid\mid \vec{w} \mid} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{3}$$

The two textbooks datasets are collected using APIs of Goodreads[5] network and Wikibooks[6] open-content textbooks. Similarly, the entity resolution problem is solved using the similarity-based technique using the following properties

```
(book_title, book_author, book_language)
```

and then is followed by the data fusion concatenation.

After removing the records' conflicts, the data are integrated into one database that is mapped onto the TCO representation using D2RQ mapping language[7].

B. Ontologies and Mapping Layer

Teacher context ontology (TCO) was designed to facilitate the representation of the multiple contexts in an educational process from a teacher's point-of-view. TCO is assumed to be used as a part of an educational resources recommender system [2]. Fig.2 shows a partial representation of selected concepts from TCO that are concerned with the integrated datasets. This representation of the environment concept features the difference between the living and working environments. The living environment represents the nature of the geographical location of teachers' and learners' place of residence. On the contrary, the working environment portrays working conditions, financial situation, and educational level of the educational institution. During the coexistence in the working environment, a person interacts with resources with different types.

In order to link the integrated datasets to TCO, there are two fundamental steps in this implementation: the relational database mapping into virtual RDF graphs, and SPARQL endpoint's setup for relational data access. The INSEE dataset's tables are mapped to "TCO:Environment" concept in the graph while the government datasets are mapped to "TCO:WorkingEnvironment" and "TCO:EducationalInstitution" concepts. The other integrated datasets are mapped directly to "TCO:Resource" concept.

The declarative language that is provided by D2RQ Platform, is used to describe the link between relational models and ontologies [40]. This platform contains

[2]https://www.coursera.org/
[3]https://www.edx.org/
[4]https://www.udemy.com/
[5]https://www.goodreads.com/
[6]https://www.wikibooks.org/
[7]http://www.d2rq.org/

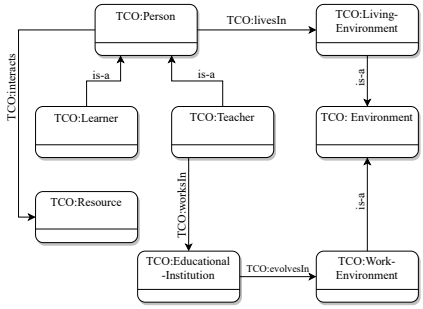

Fig. 2. Selected concepts representation from TCO.

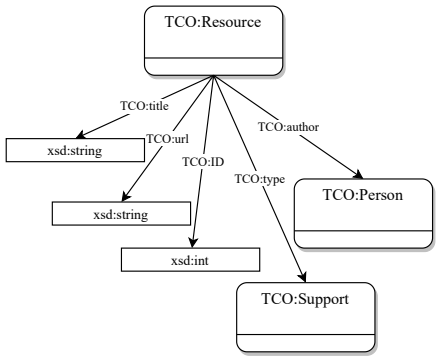

Fig. 3. Resource concept example from TCO.

an automatic map-generation tool that creates a default customizable mapping file through the $generate-mapping$ D2RQ command.

For example, the resources data are mapped according to the "TCO:Resource" representation shown in Fig.3. Through a semi-automated process, the generated mapping file is customized to match TCO concepts as shown in Fig. 4.
D2RQ Server is a built-in tool in D2RQ platform to publish a relational dataset on Semantic Web according to the previous mapping file. This tool uses the mapping file to connect the database with an RDF graph that can be used by SPARQL to extract data through the $d2r-server$ command. The integrated data are represented by a reasoning model that is more coherent and can be easily navigated through Semantic Web by any user.

V. DISCUSSION

In this research, three main types of data sources are handled to represent the teacher's context. The environmental Linked Data covers the whole French territories and provides summarized information about the population, area type, and financial conditions for all French provinces and cities. The environmental data include 34,142 out of 36,681 metropolises that are distributed over the French communes. These data are integrated with the 466,357 educational institutions from French prioritized education network. The learning resources non-redundant data are collected from 73,364 online courses over three major MOOC platforms and 196,044 textbooks with contrast in categories and languages.

Fig. 4. Content of customized mapping file for D2RQ

```
# Table Resource
map:Resource a d2rq:ClassMap;
      d2rq:dataStorage map:database;
      d2rq:uriPattern "Resource/@@Resource.id@@";
      d2rq:class tco:Resource;
      d2rq:classDefinitionLabel "Resource";
      .
map:Resource_label a d2rq:PropertyBridge;
      d2rq:belongsToClassMap map:Resource;
      d2rq:property rdfs:label;
      d2rq:pattern "Resource_@@Resource.id@@";
      .
map:Resource_id a d2rq:PropertyBridge;
      d2rq:belongsToClassMap map:Resource;
      d2rq:property tco:ID;
      d2rq:column "Resource.id";
      .
map:Resource_title a d2rq:PropertyBridge;
      d2rq:belongsToClassMap map:Resource;
      d2rq:property tco:title;
      d2rq:column "Resource.title";
      .
map:Resource_author a d2rq:PropertyBridge;
      d2rq:belongsToClassMap map:Resource;
      d2rq:property tco:author;
      d2rq:column "Resource.author";
      .
map:Resource_type a d2rq:PropertyBridge;
      d2rq:belongsToClassMap map:Resource;
      d2rq:property tco:resouce_type;
      d2rq:column "Resource.type";
      .
map:Resource_URL a d2rq:PropertyBridge;
      d2rq:belongsToClassMap map:Resource;
      d2rq:property tco:url;
      d2rq:column "Resource.url";
      .
```

The previous efforts concerning the educational Linked Data targets either educational resources [19]–[21], [23] or learning/teaching analytics [17], [18], [25], [26]. These research efforts lack the connection between the environmental data that represent the context of the person, the educational resources, and the educational analytics. This paper forms an initial step to find new ontology-based data with the previously mentioned connection. Also, educational applications, such as recommender systems, can be built on top of the single ontology-based data integration layer. Through this approach, the system will be able to access the global ontology-based reasoned data. Nevertheless, this approach needs more improvement to handle a multiple-ontology approach in terms of dynamicity and inclusivity.

VI. CONCLUSION AND PERSPECTIVE

The qualitative and quantitative growth of data sources requires an efficient data integration mechanism. However, there are no rules for the relational databases that are shared across multiple applications. Ontology-based data integration is the solution for such problem. Therefore, data can be smoothly connected and shared between multiple applications and for multiple purposes. However, the educational data generally and teacher-centered data especially do not emerge as other domains.

In this paper, we demonstrate an approach of educational ontology-based data integration for three conceptual elements in a teacher context: environment, educational institutions, and learning resources. We choose datasets from governmental sources for the environment and institution

description to guarantee the accuracy level of these data. As for the resources data, we select research datasets which takes into consideration the selection standards according to the inclusion level and the usability degree. The integrated data are then published as Linked Data using D2RQ Platform that offers a mapping technique between ontology and relational data.

This research is the first step of multi-level data collection and integration. The other unmentioned TCO concepts will be covered through future research, in addition to teacher's sentiment data. Moreover, another dynamic data management approach will be investigated to balance the static-dynamic data diversity.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data science in massive open online courses," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 1, p. e1187, 2017.

[2] N. N. Nashed, C. Lahoud, and M.-H. Abel, "Tco : a teacher context ontology," in *2021 IEEE 24rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2021.

[3] H. Eshach, "Bridging in-school and out-of-school learning: Formal, non-formal, and informal education," *Journal of science education and technology*, vol. 16, no. 2, pp. 171–190, 2007.

[4] M. Varelas, *Identity construction and science education research: Learning, teaching, and being in multiple contexts*. Springer Science & Business Media, 2012, vol. 35.

[5] R. Baker, "Big data and education," *New York: Teachers College, Columbia University*, 2015.

[6] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan, "Data integration: After the teenage years," in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems*, 2017, pp. 101–106.

[7] M. Dumontier, A. Callahan, J. Cruz-Toledo, P. Ansell, V. Emonet, F. Belleau, and A. Droit, "Bio2rdf release 3: a larger connected network of linked data for the life sciences," in *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, vol. 1272. Citeseer, 2014, pp. 401–404.

[8] F. Jeanquartier, C. Jean-Quartier, T. Schreck, D. Cemernek, and A. Holzinger, "Integrating open data on cancer in support to tumor growth analysis," in *International Conference on Information Technology in Bio-and Medical Informatics*. Springer, 2016, pp. 49–66.

[9] T.-J. Wu, L. M. Schriml, Q.-R. Chen, M. Colbert, D. J. Crichton, R. Finney, Y. Hu, W. A. Kibbe, H. Kincaid, D. Meerzaman, *et al.*, "Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis," *Database*, vol. 2015, 2015.

[10] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[11] C. Romero, J. R. Romero, and S. Ventura, "A survey on pre-processing educational data," in *Educational data mining*. Springer, 2014, pp. 29–64.

[12] C. Keßler, M. d'Aquin, and S. Dietze, "Linked data for science and education." *Semantic Web*, vol. 4, no. 1, pp. 1–2, 2009.

[13] C. Roda, *Human attention in digital environments*. Cambridge University Press, 2011.

[14] K. Niemann, M. Wolpers, G. Stoitsis, G. Chinis, and N. Manouselis, "Aggregating social and usage datasets for learning analytics: Data-oriented challenges," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 2013, pp. 245–249.

[15] M. d'Aquin, A. Adamou, and S. Dietze, "Assessing the educational linked data landscape," in *Proceedings of the 5th annual ACM Web science conference*, 2013, pp. 43–46.

[16] M. C. Desmarais and R. S. d Baker, "A review of recent advances in learner and skill modeling in intelligent learning environments," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1, pp. 9–38, 2012.

[17] J. L. Santos, K. Verbert, J. Klerkx, S. Charleer, E. Duval, and S. Ternier, "Tracking data in open learning environments," *Journal of Universal Computer Science*, vol. 21, no. 7, pp. 976–996, 2015.

[18] S. Joksimović, V. Kovanović, J. Jovanović, A. Zouaq, D. Gašević, and M. Hatala, "What do cmooc participants talk about in social media? a topic analysis of discourse in a cmooc," in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 2015, pp. 156–165.

[19] S. Kagemann and S. Bansal, "Mooclink: Building and utilizing linked data from massive open online courses," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, 2015, pp. 373–380.

[20] C. Dhekne and S. Bansal, "Mooclink: An aggregator for mooc offerings from various providers," *Journal of Engineering Education Transformations*, vol. 2018, no. Special Issue, 2018.

[21] M. Zotou, A. Papantoniou, K. Kremer, V. Peristeras, and E. Tambouris, "Implementing" rethinking education": Matching skills profiles with open courses through linked open data technologies," 2014.

[22] European Commission. "esco: European skills, competencies, qualifications and occupations," mar. 16, 2021. [Online]. Available: https://ec.europa.eu/esco

[23] J. L. Ambite, J. Gordon, L. Fierro, G. Burns, and J. Mathew, "Linking educational resources on data science," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9404–9409.

[24] National Center for Biomedical Ontology. "dseo: Data science education ontology," mar. 01, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/DSEO

[25] I. G. Ndukwe and B. K. Daniel, "Teaching analytics, value and tools for teacher data literacy: A systematic and tripartite approach," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–31, 2020.

[26] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[27] L. national de la statistique et des études économiques, "French employment, salaries, population per town," https://www.insee.fr/, updated: 2021-01-19.

[28] E. Nationale, "Écoles éducation prioritaire," https://www.data.gouv.fr/fr/datasets/ecoles-education-prioritaire/, updated: 2021-02-17.

[29] ——, "Collèges éducation prioritaire," https://www.data.gouv.fr/fr/datasets/colleges-education-prioritaire/, updated: 2018-06-26.

[30] ——, "Académies education prioritaire," https://www.data.gouv.fr/fr/datasets/academies-education-prioritaire/, updated: 2018-06-10.

[31] S. A. Patil, "Coursera all courses," https://github.com/santoshapatil/Web-scrapping-all-coursera-Courses/, updated: 2020-11-13.

[32] ——, "edx all courses," https://github.com/santoshapatil/Edx-All-Courses/, updated: 2020-11-22.

[33] S. Song, "40k udemy courses dataset 2020," https://datastudio.google.com/u/0/reporting/dc4b3168-e6c3-4da0-898e-65fce6f9b7f2/page/1xZU/, updated: 2020-10-04.

[34] J. Kothari, "It - software courses udemy," https://www.kaggle.com/jilkothari/it-software-courses-udemy-22k-courses/, updated: 2020-10-12.

[35] S. Ranjan, "Goodreads-books," https://www.kaggle.com/jealousleopard/goodreadsbooks/, updated: 2020-03-09.

[36] D. Dave, "Wikibooks dataset," https://www.kaggle.com/dhruvildave/wikibooks-dataset/, updated: 2021-02-18.

[37] A. Stéfanou, "Éducation prioritaire," *Éducation et Formations*, no. 95, pp. 87–106, 2017.

[38] J. Bleiholder and F. Naumann, "Declarative data fusion–syntax, semantics, and implementation," in *East European Conference on Advances in Databases and Information Systems*. Springer, 2005, pp. 58–73.

[39] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *International conference on intelligent data engineering and automated learning*. Springer, 2013, pp. 611–618.

[40] C. Bizer and R. Cyganiak, "D2r server-publishing relational databases on the semantic web," in *Poster at the 5th international semantic web conference*, vol. 175, 2006.