Expert Systems With Applications 213 (2023) 118892

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

e-LION: Data integration semantic model to enhance predictive analytics in e-Learning

Manuel Paneque [a,b,*], María del Mar Roldán-García [a,b], José García-Nieto [a,b]

[a] Khaos Research, ITIS Software, Universidad de Málaga, Arquitecto Francisco Peñalosa 18, 29071, Málaga, Spain
[b] Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Málaga, Spain

ARTICLE INFO

Keywords:
E-learning
Ontology
Open data
Data analysis
Knowledge graph

ABSTRACT

In the last years, Learning Management systems (LMSs) are acquiring great importance in online education, since they offer flexible integration platforms for organising a vast amount of learning resources, as well as for establishing effective communication channels between teachers and learners, at any direction. These online platforms are then attracting an increasing number of users that continuously access, download/upload resources and interact each other during their teaching/learning processes, which is even accelerating by the breakout of COVID-19. In this context, academic institutions are generating large volumes of learning-related data that can be analysed for supporting teachers in lesson, course or faculty degree planning, as well as administrations in university strategic level. However, managing such amount of data, usually coming from multiple heterogeneous sources and with attributes sometimes reflecting semantic inconsistencies, constitutes an emerging challenge, so they require common definition and integration schemes to easily fuse them, with the aim of efficiently feeding machine learning models. In this regard, semantic web technologies arise as a useful framework for the semantic integration of multi-source e-learning data, allowing the consolidation, linkage and advanced querying in a systematic way. With this motivation, the e-LION (e-Learning Integration ONtology) semantic model is proposed for the first time in this work to operate as data consolidation approach of different e-learning knowledge-bases, hence leading to enrich on-top analysis. For demonstration purposes, the proposed ontological model is populated with real-world private and public data sources from different LMSs referring university courses of the Software Engineering degree of the University of Malaga (Spain) and the Open University Learning. In this regard, a set of four case studies are worked for validation, which comprise advance semantic querying of data for feeding predictive modelling and time-series forecasting of students' interactions according to their final grades, as well as the generation of SWRL reasoning rules for student's behaviour classification. The results are promising and lead to the possible use of e-LION as ontological mediator scheme for the integration of new future semantic models in the domain of e-learning.

1. Introduction

Since the last decade, the progress in the access to new technologies experimented by most of the society is also reflected in online education at any level, which is indeed accelerated by the breakout of COVID-19, hence leading academic institutions to revise their educational strategies. In this context, a plethora of e-Learning tools and resources are appearing to facilitate a similar methodology to the traditional system, that connects teachers and students asynchronously to carry out a didactic learning process. Among these tools, Learning Management Systems (LMSs) are acquiring great importance in online education, since they offer flexible integration online platforms for organising a vast amount of learning resources, as well as for establishing effective communication channels between teachers and learners, at any direction.

Consequently, LMSs are attracting an increasing number of users that continuously access, download/upload resources and interact each other during their teaching/learning processes. This entails the generation of large volumes of learning-related data that can be analysed to support teachers in lesson, course or faculty degree planning, as well as administrations in university strategic level. For example, it is possible to extract how students' interactions in the LMS are related to the grades they obtain, which somehow enables teachers to establish an expected performance classification based on the interactions in the system. This undoubtedly provides professors with

* Corresponding author at: Khaos Research, ITIS Software, Universidad de Málaga, Arquitecto Francisco Peñalosa 18, 29071, Málaga, Spain.
E-mail addresses: mpaneque@uma.es (M. Paneque), mmar@lcc.uma.es (M.d.M. Roldán-García), jnieto@lcc.uma.es (J. García-Nieto).

https://doi.org/10.1016/j.eswa.2022.118892
Received 16 November 2021; Received in revised form 16 September 2022; Accepted 20 September 2022
Available online 27 September 2022

access to new knowledge that, together with their experiences, can led them to get insights about the behaviour of students beforehand. Interestingly, this context is suitable, for example, for predictive machine learning algorithms based on time-series data to predict the number of visits in the LMS. These algorithms have been previously applied in different domains (Altan & Karasu, 2019; Karasu & Altan, 2019; Karasu, Altan, Saraç, & Hacioğlu, 2018). Nevertheless, managing such amount of data, usually coming from multiple heterogeneous sources and with attributes sometimes reflecting semantic inconsistencies, constitutes an emerging challenge, so they require common definition and integration schemes to easily fuse them, with the aim of efficiently feeding machine learning models. In this regard, semantic web technologies arise as a useful framework for the semantic integration of multi-source e-learning data, allowing the consolidation, linkage and advanced querying in a systematic way. The development of new ontologies and their use for data integration is widely documented in the existing literature in different domains of application, as worked in Brochhausen bibetal (2022), del Mar Roldán García, García-Nieto, and Aldana-Montes (2016), del Mar Roldán-García, Uskudarli, Marvasti, Acar, and Aldana-Montes (2018), McGlinn, Rutherford, Gisslander, Hederman, Little, and O'Sullivan (2022), Sobral, Galvão, and Borges (2020), Thaddeus, Jeganathan, and Leema (2011). These ontologies guide the creation of knowledge graphs that semantically represent integrated data and are the input of analytics, as done in del Mar Roldán García et al. (2016), or semantic reasoning tasks, as also developed in Aldana-Martín, García-Nieto, del Mar Roldán-García, and Aldana-Montes (2022), Delgoshaei, Heidarinejad, and Austin (2022).

Semantic Web technologies in e-learning are analysed in the current literature in two recent surveys: Rahayu, Ferdiana, and Kusumawardani (2022) and Heiyanthuduwage (2022), the former oriented to recommendation systems in e-learning driven by semantics, the latter identifying current trends in e-learning ontologies. In this sense, a set of issues are still identified in these works that require new proposals to be approached, mainly related to the data interoperability, linkage, enrichment and analysis.

With this motivation, the e-LION (e-Learning Integration ONtology) semantic model is proposed in this work to operate as data consolidation approach of different e-learning knowledge-bases, hence leading to enrich data analysis. It consists in an OWL 2 (Ontology Web Language, explained in Section 2.1) ontology that enables development of semantic mappings to the source schema, to transform the original raw data into standard RDF (Resource Description Framework) creating a knowledge graph. In this way, data from heterogeneous sources are stored and integrated within a common RDF repository, which can now be easily queried. The main objective is to feed artificial intelligence algorithms capable of analysing implicit interaction patters in LMSs registered by a given e-learning community.

To validate the proposed semantic model, a series of mapping functions and SQL dump-loading processes are conducted to populate e-LION with private and public data sources from different LMSs. In concrete, these sources consist in the Moodle space of the Software Engineering degree of the University of Malaga (Spain), which are enriched with the integration of the Open University Learning repository presented in Kuzilek, Hlosta, and Zdrahal (2017), as well as the COCO semantic-enriched collection of online courses data proposed by Dessì, Fenu, Marras, and Reforgiato Recupero (2018). The resulting semantic approach allows the advanced querying of data concerning the students' interactions and their academic performances to efficiently feeding predictive models and visualisations. Moreover, thanks to the semantic integration, a series of reasoning tasks are conducted to induce new implicit knowledge to classify different student's behaviours.

The main contributions of this study are outlined as follows:

- The e-Learning Integration ONtology (e-LION) semantic model is proposed to operate as data consolidation approach of different e-learning knowledge-bases, which enables to enrich machine learning analysis. It is development as an OWL 2 ontology that enables development of semantic mappings to the source schema. e-LION is online available.[1]
- Semantic model population is carried out with a series of SQL dump processes from the Moodle platform of the Software Engineering degree (University of Malaga), together with the Open University dataset (Kuzilek et al., 2017) and the COCO (Dessì et al., 2018) collection. In overall, the activity interactions of a number of 43,228 subjects and 2,466,712 students over several years of operation are mapped to the same knowledge graph and stored in the RDF repository, enabling SPARQL Endpoint for querying.
- The proposed semantic approach is validated by means of four case studies comprising predictive modelling and time-series forecasting of students' interactions with regards to final grades, as well as the generation of SWRL reasoning rules for student's behaviour classification.

The remaining of this article is organised as follows. A review of background concepts and related work is provided in Section 2. The proposed semantic model is described in Section 3, giving details of the e-LION ontology design and model implementation. In Section 4, a series of validation tasks are carried out throughout four different case studies. Section 5 includes discussions. Finally, in Section 6, the main concluding remarks and future work are commented.

2. Background concepts and related work

This section is devoted to explain background concepts of Semantic Web technologies for knowledge representation, structure and reasoning. A review of related works in the literature is conducted to position our proposal within the current state of the art.

2.1. Background concepts

In the ecosystem of semantic web technologies, ontologies are key elements that can be defined as formal descriptions of knowledge, comprising a set of concepts and the possible relationships among them (Gruber, 1993). The main components of an ontology are classes (or concepts), relations (or properties), instances (or individuals) and axioms. Logical class constructors (and, or, not) and property restrictions can be used to built complex classes. In addition to ontology elements, rules provide a mechanism to define more complex knowledge. Rules are described in terms of ontology elements (classes, properties, and instances). An ontology data model can be populated with a set of individuals using a knowledge graph, i.e., an interlinking collection of entities, where nodes and edges represent entities (things or concepts) and semantic relationships between them, respectively.

The Ontology Web Language (OWL) is a semantic markup language used to define ontologies. OWL is built on top of RDF (Resource Description Framework), being both standards by the W3C. RDF is a data format used for the representation of information in the Web (Schreiber & Raimond, 2014), which offers a common framework where information can be shared between applications without loosing their meaning. In RDF, resources are identified by URIs (Uniform Resource Identifier), so they are organised in form of triples with: subject, predicate and object. W3C recommends the use of RDF in those applications where the data are going to be processed by other applications instead of only being shown to users.

Therefore, in a given knowledge graph, RDF triples are linked and used to populate the ontology, then stored in a repository. To access

[1] e-LION OWL Ontology available at URL http://ontologies.khaos.uma.es/e-lion.

Table 1
Summary of proposals' main features selected in related works in comparison with e-LION ones.

Ontology/Feature	Main purpose	Target audience	Language	Machine learning	S.R.	Avail.
Suguna et al. (2016)	Information retrieval	Learner	OWL	NPL	No	No
Hssina et al. (2017)	Data annotation	Teacher	OWL	No	No	No
Taurus et al. (2017)	Recommender system	Learner	OWL	Seq. Patter Mining	No	No
Makwana et al. (2018)	Recommender system	Learner	Unknown	Fuzzy C-Means	No	No
Ouf et al. (2017)	Data annotation	Learner/Teacher	OWL	No	SWRL	No
Obeid et al. (2018)	Recommender system	Learner	Unknown	Unknown	No	No
Ham L. (2018)	Data annotation	Teacher	XML	Unknown	No	No
Dessì et al. (2018)	Dataset	Learner/Teacher	JSON	KNN, NPL	No	Yes
Bouihi and Bahaj (2019)	Data annotation	Teacher	OWL	No	No	No
Joy et al(2021)	Recommender system	Learner	OWL	kMeans	No	No
e-LION	Data integration and analysis	Learner Teacher	OWL 2	KNN, DT, SVM, RF, GNB, MLP, SARIMAX	SWRL	Yes

these data in form of RDF graphs, SPARQL query language (Harris & Seaborne, 2013) can be used to retrieve the set of triples in the RDF repository that match. SPARQL allows querying several linked data graphs in different repositories, so it is potentially used to perform federated queries throughout the semantic web.

Once a knowledge graph is developed according to a given ontology scheme, it is possible to infer implicit semantic relationships between individuals (Horrocks, Patel-Schneider, Bechhofer, & Tsarkov, 2005), providing OWL-based ontologies with reasoning capabilities. To do so, the SWRL (Semantic Web Rule Language) standard is used to construct rule expressions in form of "Antecedent ⇒ Consequent" to represent those semantic relationships. Both, Antecedent and consequent are formulated as conjunctions of elements associated to one or more attributes. They are denoted as question mark and a variable (e.g., $?x$) in the rule.

2.2. Related work

The use of semantic web technologies in e-learning and in particular the conceptualisation of knowledge with ontologies in this domain, has been widely studied in past literature reviews of Al-yahya, George, and Alfaries (2015), Pereira, Siqueira, Nunes, and Dietze (2018), from which a series of recent proposals have been appearing, covering the last five years. More recently, in K., Poscic, and Jaksic (2020) a categorisation of studies is conducted according to the ontology usage in the context of learning and education, namely: curriculum modelling and management, learning domain description, learner data description, and e-learning services.

Within these categories, the use of ontologies together with data analytic techniques is gaining in importance in e-learning, as it is supported on digital platforms, which enable the generation of new sources of data regarding learners' activities and behaviours. This aspect has been widely considered in two recent surveys of George and Lal (2019) and Rahayu et al. (2022), although with special focus on ontology-based recommender systems in e-learning in both of them.

To mention chronologically a representative set of related contributions to the current proposal, in 2016 an ontology based e-learning information retrieval system is proposed in Suguna, Sundaravadivelu, and Gomathi (2016), where authors analysed the importance of handling natural processing language-based concepts with tools such as Wordnet or Hownet. Also in 2016, Hssina, Bouikhalene, and Merbouha (2017) developed a semantic annotation platform to assess the skills of learners on an e-learning platform using semantic web technologies. This comprised a manually annotated OWL ontology for the exploitation of learner data to predict their performance in training. Lately, in 2017 a hybrid method for recommendation based on learners-resources ontology and sequential pattern mining is proposed in Tarus, Niu, and Yousif (2017) to identify the learner's historical patterns from weblogs. With similar focus, in Makwana, Patel, and Shah (2018), a knowledge-base system is created under an ontological scheme that enable item to item mapping for a collaborative filtering recommender. It is used to personalise user's search in the web by means of a weblog file to record user's clicks. The weblog is in turn used for feeding the knowledge-base. Also in this line, Ouf, Abd Ellatif, Salama, and Helmy (2017) proposed smart e-learning ecosystem based on an ontological model with SWRL reasoning rules. This model consists of four ontologies for learning objects, learning activities and teaching methods. The main aim is to provide learners with a personalising learning environment.

Another recommendation system is proposed by Obeid, Lahoud, El Khoury, and Champin (2018), which is enriched with machine learning methods to orient students in higher education. It is an ontology-based approach to annotate student's requirements, interests, preferences and capabilities, with the aim of recommending higher educational levels. Also in this year, based on e-learning domain ontology, an interdisciplinary intelligent teaching model is proposed by Han (2018) to enhance the cognitive ability of students, while supporting the teachers to understand the students' learning level. As argued by authors, this model evaluates the domain ontology abstraction layer and provides the basis for improving the teaching plan. This proposal is indeed validated thorough some use cases. In this sense, Dessì et al. (2018) presented COCO, a semantic-enriched collection of online courses that aims at supporting experimentation and design of services in online learning. COCO dataset includes information collected from Udemy[2] platform for online courses, enabling the generation of use cases oriented to e-learning data analysis.

From a different perspective, Bouihi and Bahaj (2019) proposed a methodology to build an ontology based on the Moodle database schema for social network analysis. This proposal models the semantics of relationships influences from the user's interaction graph topology. The ontology is built by directly mapping the UML Moodle structure of the Mount Orange School[3] demo source.

Joy, Raj, and V. G. (2021) presented an ontological framework used to address the pure cold-start problem for content recommendation. In this model, the proposed ontology is designed to cover the contextual domain of learners and Learning Objects (LOs). It also includes a multivariate k-means clustering to evaluate the learner similarity computation accuracy. Interestingly, the learner satisfaction achieved by 40 participants was measured when using this proposal.

Recently, two complete surveys have appeared in the current literature that cover different aspects in the intersection of web semantics with e-learning. The review presented by Rahayu et al. (2022) is oriented to recommendation systems in e-learning driven by ontology, which considers 28 journal articles that combine semantics with artificial intelligence, computing technology, education, education psychology, and social sciences. Secondly, Heiyanthuduwage (2022) discussed a series of current trends in e-learning ontologies, and identified the data interoperability as a key issue that should be faced in new approaches, not only in systems belonging to the same institutions,

[2] Online available at URL https://www.udemy.com/.
[3] Online available at URL https://school.moodledemo.net/.

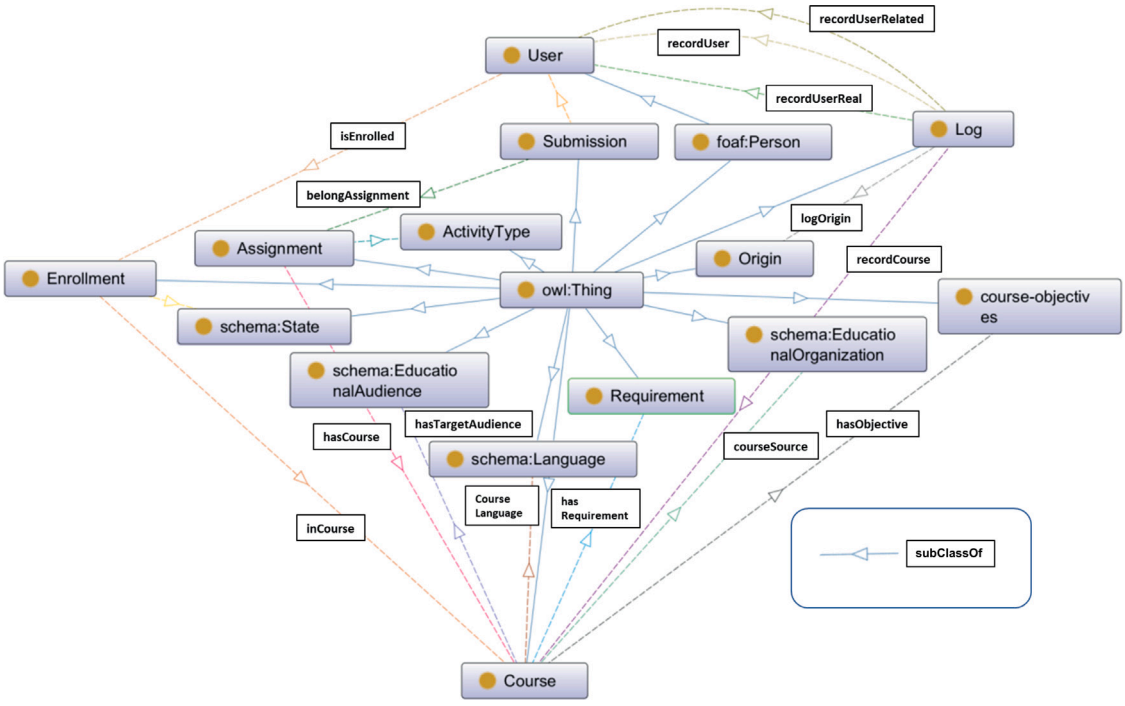

Fig. 1. General overview of the e-LION ontology. Continuous arrows refer to subclasses, whereas dotted ones refer to properties.

Table 2
Basic OWL-DL semantic syntax used to formally define the proposed ontology. It is organised by Operators (O), Restrictions (R) and Class Axioms (A).

	Abstract syntax	DL syntax
O	$intersection(C_1, C_2, \dots, C_n)$	$C_1 \sqcap C_2 \sqcap \cdots C_n$
	$union(C_1, C_2, \dots, C_n)$	$C_1 \sqcup C_2 \sqcup \cdots C_n$
R	for at least 1 value V from C	$\exists V.C$
	for all values V from C	$\forall V.C$
	R is Symmetric	$R \equiv R^-$
A	$A\ partial(C_1, C_2, \dots, C_n)$	$A \sqsubseteq C_1 \sqcap C_2 \sqcap \cdots C_n$
	$A\ complete(C_1, C_2, \dots, C_n)$	$A \equiv C_1 \sqcap C_2 \sqcap \cdots C_n$

but also in the context of different ones, where ontologies would be required to handle policies and discrepancies. This specific issue is approached by the e-LION proposed here, together with other different dimensions.

In this sense, Table 1 contains a summary of proposals' main features selected in related works in comparison with e-LION ones. In concrete, the main purpose, the target audience and the design language are reported, in addition to the machine learning techniques used (when applicable), and whether they have performed Semantic Reasoning (S.R.) and they have the resources online available, or not.

Much of these approaches are geared towards the generation of semantic models, that use ontologies to drive the development of recommendation systems in different aspects of the e-learning domain of knowledge. Nevertheless, to the best of our knowledge, none of them are conceived for the special task of multi-source data integration in e-learning environments to enrich data analytic processes and visualisations. The e-LION semantic model proposed in this work aspires to constitute a step forward in that direction.

3. Semantic approach

One of the main objectives of this work is to capture, clean, consolidate and integrate data from different e-learning LMS platforms and repositories. For this reason, we opted to design a semantic approach to share and unify the data involved, through an ontology that models the domain in which the system operates. Specifically, we have defined an OWL 2 ontology to describe the main characteristics of e-learning platforms as recommended in the Ontology 101 development process (Noy & McGuinness, 2001):

1. *Determine the domain and scope of the ontology*. As starting point, to limit the scope of the ontology, the type of variables that most of the e-learning systems usually store have been selected, for example: record of interactions, student attributes, attributes of assignment and submissions. Further formalism for describing interoperable model components and data could be linked from the BIGOWL ontology (Barba-González, García-Nieto, Roldán-García, Navas-Delgado, Nebro, & Aldana-Montes, 2019), which is devoted to data analytic workflow annotation. For simplicity, it has been omitted to just focus on the e-learning LMSs domain of knowledge.
2. *Consider reusing existing ontologies*. As studied in Section 2, there are no public ontologies that fully model user's interactions and their grades on tasks and assignments. Nevertheless, two related ontologies have been partially considered: first, the ontology proposed in Firdausiah Mansur and Yusof (2013) shows a basic model of e-learning knowledge-base, while the approach in Zeng, Zhao, and Liang (2009) takes into account the relationships between assignments and courses. These ontologies have not been directly reused by e-LION. They have served as an inspiration for the modelling of the proposed ontology. Our ontology aims to cover the information needs relevant to facilitate data mining and analytic in the scope of e-learning. Existing ontologies such as LOM (Learning Object Metadata),[4] CRSW (ReSIST Courseware Ontology),[5] Scorm and Tin Can API,[6] focus on a specific area of the learning process, i.e. e-learning resources, without containing the diversity of classes and metrics

[4] https://lov.linkeddata.es/dataset/lov/vocabs/lom
[5] https://lov.linkeddata.es/dataset/lov/vocabs/crsw
[6] https://xapi.com/

Table 3
Course class: object and data properties.

Object properties	Description logic
courseSource	∃ courseSource Thing ⊑ Course ⊤ ⊑ ∀ CourseSource schema:EducationalOrganization
Data properties	**Description logic**
courseId	∃ courseId Datatype Literal ⊑ Course ⊤ ⊑ ∀ courseId Datatype string
coursePresentationLength	∃ coursePresentationLength Datatype Literal ⊑ Course ⊤ ⊑ ∀ coursePresentationLength Datatype int
courseUrl	∃ CourseUrl Datatype Literal ⊑ Course ⊤ ⊑ ∀ CourseUrl Datatype string
courseDescription	∃ courseDescription Datatype Literal ⊑ Course ⊤ ⊑ ∀ courseDescription Datatype string
courseClicksAVG	∃ courseClicksAVG Datatype Literal ⊑ Course ⊤ ⊑ ∀ courseClicksAVG Datatype int

Table 4
User class: object and data properties.

Object properties	Description logic
isEnrolled	∃ isEnrolled Thing ⊑ User ⊤ ⊑ ∀ isEnrolled Enrollment
Data properties	**Description logic**
userId	∃ userId Datatype Literal ⊑ User ⊤ ⊑ ∀ userId Datatype int
userBiography	∃ userBiography Datatype Literal ⊑ User ⊤ ⊑ ∀ userBiography Datatype string
userProfileUrl	∃ userProfileUrl Datatype Literal ⊑ User ⊤ ⊑ ∀ userProfileUrl Datatype string
userJobTitle	∃ userJobTitle Datatype Literal ⊑ User ⊤ ⊑ ∀ userJobTitle Datatype string

that e-learning systems store. On the other hand, some general-purpose ontologies, such as schema.org[7] and foaf,[8] include e-learning related classes. Section 3.1 describes how e-LION reuses some classes, mainly those related to Activity type, Educational audience and e-learning users.

3. *Enumerate important terms in the ontology*. Important terms in the ontology have been extracted in a previous phase of specification of requirements. In this phase, we defined the minimum set of variables that needed to be stored. Some examples of these terms are: *assignment*, *submission*, *user*, *course*, *log* and *enrollment*, among others.
4. *Define classes and class hierarchy*. From the list of the most important terms, a series of ontology classes have been selected. Fig. 1 shows the main set of classes in the hierarchy from the top *Thing* class. These main classes are related to other classes to model the relationships between the information they contain.
5. *Define the properties of classes*. To relate classes and define attributes, object and data properties are defined based on the minimum set of variables. Examples of object properties are: a *Submission* belongs to an *Assignment*, a *Submission* belongs to a *User*, an *Assignment* belongs to a *Course*, a *User* is enrolled in a *Course*, etc. Examples of data type properties are, the role of a user in a course, the score of a submission, the timestamp of a delivery, etc. Tables 3–8 describe a representative subset of object and data properties for a selection of the main classes.
6. *Define the facets of the slots*. This step includes defining cardinality and value constraints. Value constraints are used in e-LION to specify the data type values in all their properties. For example, the range of the *logTimeCreated* property is restricted to dateTime, while in the case of the *assignmentWeight* property the range is restricted to int.
7. *Create instances*. Instances (individuals in OWL) correspond to the specific data obtained from the interactions of the students from LMSs and e-learning datasets. Individuals will be obtained by mapping data from Moodle (SQL dumps) and other e-learning systems to RDF according to the e-LION ontology scheme.

3.1. Ontology model

After applying the previous methodology, the e-LION ontology has been developed comprising a total of 15 classes (groups of individuals with the same attributes), 17 object properties (represent binary relationships between individuals), 78 data properties (individual attributes), and 193 constraint axioms.

For simplicity, we describe here a representative subset of the main data and object properties of the classes that make up the ontology, which are formalised by means of OWL-DL description logic semantic syntax.[9] A summary of this syntax is shown in Table 2, which can be used for supporting the interpretation of the following tables defining data and object properties. These classes are: *Assignment*, *Course*, *Enrollment*, *Log*, *Submission*, *Origin* and *User*. In addition, classes *xapi: ActivityType, courseware: course-objectives, EducationalAudience, foaf: Person, schema: EducationalAudience, schema :EducationalOrganization, schema:Language, schema: State* and *Requirement* have been included to also consider other existing semantic repositories such as the COCO collection, hence enabling linked data. Each class requires a set of properties to be modelled, i.e., an individual who satisfies those properties is considered a member of that class. The complete ontology is developed in the OWL file "`e-LION.owl`", available in the link.[10] A description of the main e-LION classes are given below:

- ***Course***. This class represents the set of courses that are registered in the LMS e-learning system. It defines three main properties (among others) as described in Table 3, namely: *courseId*, to uniquely identify each course; *coursePresentationLength*, that represents the duration of the course in days; and *courseSource*, that registers the source of data associated with the course.
- ***User***. It is devoted to define the set of users registered in the LMS e-learning platform. This class has a data property *userId* to identify each user, independently s(he) is teacher or student, as well as an object property *isEnrolled* that represents that a user is part of a subject. Table 4 contains the description logic of these properties.
- ***Assignment***. It is an important class that represents the assignments proposed by the professors and delivered by the students. The object property *hasCourse* connects an assignment with a given course where it is created. Class *Assignment* defines a total of 29 data properties to model the task configuration, such as: *assignmentDueData* to consider the delivery due date of a task; *assignmentName*, to gather the name and description of a task; *assignmentMaxAttempts* to set the maximum number of allowed attempts; and *assignmentWeight*, to define the importance of the assignment in the course. Table 5 contains a summary of the properties of class *Assignment*, so the complete list can be extracted from the OWL ontology file.
- ***Submission***. This class connects the previous ones, as it represents the submissions made by a user with regards to a given assignment raised in a course. The *Submission* class largely records interactions and activities performed by the users, so it is also

[7] https://schema.org/docs/developers.html
[8] http://xmlns.com/foaf/spec/
[9] OWL Web Ontology Language Overview https://www.w3.org/TR/owl-features/.
[10] e-LION OWL Ontology https://github.com/KhaosResearch/e-lion.

Table 5
Assignment class: object and data properties.

Object properties	Description logic
hasCourse	∃ hasCourse Thing ⊑ Assignment ⊤ ⊑ ∀ hasCourse Course
Data properties	**Description logic**
assignmentAllowSubmissionsFromDate	∃ assignmentAllowSubmissionsFromDate Datatype Literal ⊑ Assignment ⊤ ⊑ ∀ assignmentAllowSubmissionsFromDate Datatype dateTime
assignmentDueDate	∃ assignmentDueDate Datatype Literal ⊑ Assignment ⊤ ⊑ ∀ assignmentDueDate Datatype dateTime
assignmentName	∃ assignmentName Datatype Literal ⊑ Assignment ⊤ ⊑ ∀ assignmentName Datatype string
assignmentTimeModified	∃ assignmentTimeModified Datatype Literal ⊑ Assignment ⊤ ⊑ ∀ assignmentTimeModified Datatype dateTime
assignmentWeight	∃ assignmentWeight Datatype Literal ⊑ Assignment ⊤ ⊑ ∀ assignmentWeight Datatype int

Table 6
Submission class: object and data properties.

Object properties	Description logic
belongAssignment	∃ belongAssignment Thing ⊑ Submission ⊤ ⊑ ∀ belongAssignment Assignment
belongsUser	∃ belongsUser Thing ⊑ Submission ⊤ ⊑ ∀ belongsUser User
Data properties	**Description logic**
submissionId	∃ submissionId Datatype Literal ⊑ Submission ⊤ ⊑ ∀ submissionid Datatype string
submissionScore	∃ submissionScore Datatype Literal ⊑ Submission ⊤ ⊑ ∀ submissionScore Datatype float
submissionTimeCreated	∃ submissionTimeCreated DatatypeLiteral ⊑ Submission ⊤ ⊑ ∀ submissionTimeCreated Datatype dateTime
submissionTimeModified	∃ submissionTimeModified Datatype Literal ⊑ Submission ⊤ ⊑ ∀ submissionTimeModified Datatype dateTime

an interesting class that considers information about users' interactions. Table 6 contains the description logic of the two object properties defined for this class: *belongAssignment*, that indicates the assignment in which the submission is made, and *belongsUser* that refers to the user who makes the submission of the task. In addition, this class considers a set of data properties (also described in Table 6) to cover information about a submission, such as: *submissionAttempNumber* to indicate the attempt number of the submission, *submissionId* is the identifier assigned to the submission by the platform, *submissionLatest* indicates whether this is the last submission attempt by that user in that assignment, *submissionStatus* to denote the submission status (draft or submitted), *submissionTimeCreated* to gather the timestamp of the first submission in this assignment, and *submissionTimeModified* to indicate the timestamp in case the submission is modified.

- Class ***Enrollment*** represents the registration of students in courses. To do so, the object property *inCourse* specifies the course in which the enrollment of the user is made. In addition, this class is defined with a set of data properties as shown in Table 7 with their formal descriptions logic. Some of the most interesting properties are: *enrollmentRole*, that indicates the role of the user in the course (student, teacher, administrator, etc.), *enrollmentFinalResult* to set the final grade of student in the course, *enrollmentGender* and *enrollmentDateRegistration*, this last indicating the date of enrollment in the course.
- Class ***Log*** is also an important class since it covers the log events performed by the users in the LMS (Moodle) platform. It counts with a set of object properties and data properties that described in Table 8. Among these properties, a subset of then are worthy to mention, such as: *logEduLevel* that represents what type of user the event belongs to, e.g., if the event was motivated by a professor then the field contains the value 1, while if was generated by a student, it contains the value 2; *logAction* describes the type of action the user has taken (the most common values for this property are "view" and "submitted"). Some systems do not record the data in an aggregated way, so this information is stored in *logSumlick* and *logTimeCreated* to indicate the timestamp at which the event was logged.

3.2. Data consolidation

Once the ontology model is designed, a data consolidation strategy is conducted to allow the integration of the different data sources, according to this model. Fig. 2 shows a general overview of this strategy, where the terminological box (TBox) defines the vocabulary with concepts and relationships in the domain of e-Learning. Within this TBox, the e-LION is developed in OWL 2 according to which, concepts and relationships are represented by classes and data properties or object properties, respectively. This ontology allows the linkage with other educational ontologies oriented to different aspects, such as: recommendation, curriculum, teaching material, MOOCs, bibliografics, etc. Dessì et al. (2018) Navarrete and Luján-Mora (2015), as well as the alignment with other external linked data[11] in different domains (DBPedia,[12] Geonames,[13] FOAF,[14] etc.).

At a different level, the Assertional Box (ABox) considers all the instances in the knowledge domain involving the e-Learning LMS related data. These instances are stored in RDF triple format in a Stardog[15] repository with persistence and reasoning capabilities. To do so, a series of mapping functions have been implemented to convert the data coming from the different sources into RDF, then following the same e-LION scheme.

Therefore, in the case of Moodle data, they are mapped from a set of SQL-dump operations on a relational database regarding the Software Engineering degree of the University of Malaga. These data are used for the first time in this study and contain the anonymised information of the interactions carried out by users in this LMS platform. This dataset contains data of 8524 students in 93 courses, 1,235,063 log records, 1342 assignments and 28,270 submissions. A second data source is integrated from the Open University (OULAD) e-learning system, which was published to support research of educational data mining (Kuzilek et al., 2017). This dataset contains data of the interactions of 32,593 students in 22 courses, 10,655,280 log records, 173,913 submissions and 206 assignments. It also considers demographic information, as well as interaction records of the students with the materials and grades, both of the assignments and of the final grade of the course.

[11] Open Linked Data Cloud https://lod-cloud.net/.
[12] https://wiki.dbpedia.org/
[13] https://www.geonames.org/
[14] http://www.foaf-project.org/
[15] http://www.stardog.com/

Table 7
Enrollment class: object and data properties.

Object properties	Description logic
inCourse	∃ inCourse Thing ⊑ Enrollment ⊤ ⊑ ∀ inCourse Course
Data properties	**Description logic**
enrollmentAgeBand	∃ enrollmentAgeBand Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentAgeBand Datatype string
enrollmentDateRegistration	∃ enrollmentDateRegistration Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentDateRegistration Datatype dateTime
enrollmentRole	∃ enrollmentRole Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentRole Datatype string
enrollmentRating	∃ enrollmentRating Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentRating Datatype float
enrollmentRatingDate	∃ enrollmentRatingDate Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentRatingDate Datatype dateTime
typeOfUser	∃ typeOfUser Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ typeOfUser DataRange {“Looker” Datatype string , “Passive” Datatype string, “Active” Datatype string}
enrollmentNumberOfClicks	∃ enrollmentNumberOfClicks Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentNumberOfClicks Datatype int
enrollmentNumberOfSubmissions	∃ enrollmentNumberOfSubmissions Datatype Literal ⊑ Enrollment ⊤ ⊑ ∀ enrollmentNumberOfSubmissions Datatype int

Table 8
Log class: object and data properties.

Object properties	Description logic
recordCourse	∃ recordCourse Thing ⊑ Log ⊤ ⊑ ∀ recordCourse Course
recordUser	∃ recordUser Thing ⊑ Log ⊤ ⊑ ∀ recordUser User
recordUserReal	∃ recordUserReal Thing ⊑ Log ⊤ ⊑ ∀ recordUserReal User
recordUserRelated	∃ recordUserRelated Thing ⊑ Log ⊤ ⊑ ∀ recordUserRelated User
logOrigin	∃ logOrigin Thing ⊑ Log ⊤ ⊑ ∀ logOrigin Origin
Data properties	**Description logic**
logAction	∃ logAction Datatype Literal ⊑ Log ⊤ ⊑ ∀ logAction Datatype string
logEduLevel	∃ logEduLevel Datatype Literal ⊑ Log ⊤ ⊑ ∀ logEduLevel Datatype int
logId	∃ logId Datatype Literal ⊑ Log ⊤ ⊑ ∀ logId Datatype string
logSumClick	∃ logSumClick Datatype Literal ⊑ Log ⊤ ⊑ ∀ logSumClick Datatype int
logTarget	∃ logTarget Datatype Literal ⊑ Log ⊤ ⊑ ∀ logTarget Datatype string
logTimeCreated	∃ logTimeCreated Datatype Literal ⊑ Log ⊤ ⊑ ∀ logTimeCreated Datatype dateTime

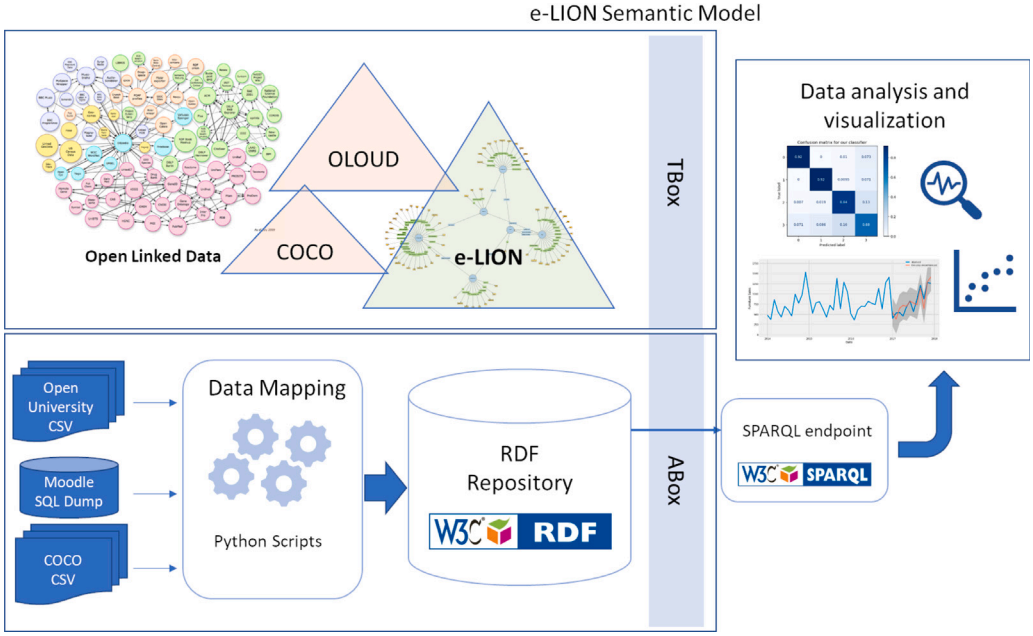

Fig. 2. General overview of the e-LION semantic model.

The OULAD dataset is provided in CSV tabular files, so the mapping functions are adapted to this kind of format. Similarly, the COCO dataset (Dessì et al., 2018) is also provided in CSV format, although consisting in different attributes that have been also adapted to the e-LION scheme. As commented before, COCO dataset includes information collected from Udemy platform for on-line courses, enabling

Table 9
Sample results obtained from Query 1.

userid	courseid	count_sub
629507	BBB2014B	10
629081	CCC2014B	3
2689210	FFF2014B	11
9485	73	2
11273	68	2

the generation of use cases oriented to e-learning data analysis. This third data source comprises 43,113 courses in two-level categories and languages, each course containing an average of 43 lessons. It also includes a number of 2,436,677 students who interacted with 4,584,313 ratings and 2,453,800 comments.

The main reason of integrating these three data sources is to constitute a first proof-of-concept for the validation of the proposed e-LION semantic model, since they constitute heterogeneous e-learning platforms, comprising Moodle private data, Open University public data and COCO academic dataset from Udemy on-line courses. Secondly, the resulting data repository can be then extended with other different e-learning related datasets, by semantically annotating and mapping their attributes in accordance with the e-LION ontology structure.

At this point, having the data consolidated in the common RDF repository, it is now possible to query them from a SPARQL Endpoint, independently of the source of the data, their structure or the syntax of the original format. In this way, the machine learning models used to perform exploratory and predictive analysis in the use cases are fed with the required information concerning student's interactions, user's views, number of deliveries, grades, so the resulting data can be grouped by date, subject, etc., with specialised SPARQL queries. An example of this can be observed in Query 1, which is executed to unify data accesses corresponding to the total number of submissions made by each student in a subject. Table 9 partially shows a sample of results obtained from this query.

Similar simple queries allow to obtain informative attributes, such as the number of visits made by the students in different periods. In this regard, Fig. 3 shows a time-series of the accumulated visits of students by weeks. In this plot, it can be observed that there are temporal patterns in the visits, decreasing in holiday periods (Christmas and summer), while increasing with certain peaks in February, related to the final evaluation dates of the courses.

Therefore, it is now possible to monitor the interactions of students on the integrated data platforms, as well as their possible correlation with the development of the students in the courses and the grades finally obtained. It can also be useful in order to understand what strategies work, as well as to detect when a student is deviating from the follow-up of the course.

Query 1: Query example of number of submissions per user and course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT (COUNT(?submission) AS ?count_sub) ?userid
?courseid
WHERE{
    ?course elion:courseId ?courseid.
    ?user elion:userId ?userid.
    ?submission elion:belongsUser ?user.
    ?submission elion:belongAssignment ?assigment.
    ?assigment elion:hasCourse ?course.
}GROUP BY ?userid ?courseid
```

An additional advantage of using the proposed semantic approach is the ability to connect the RDF repository with other external linked data. This requires minimal adaptation to establish which classes and which properties have an equivalent semantic meaning, as done with OLOUD and COCO.

4. Validation

The proposed approach provides a broad set of attributes of the students, courses, submissions to assignments and user's interactions that take place in the LMS, together with the performance obtained in the courses, allowing an advanced analysis of the data.

In terms of validation, a series of case studies are carried out in this section, consisting in: student's grade prediction in continuous evaluation, student's final grade prediction, student's visits time-series forecasting, and reasoning tasks for the classification of students' behaviours. These case studies have been elaborated to cover important aspects in the proposed semantic model, such us enabling a series of SPARQL queries from integrated data of common and different sources, that are used as training and test sets for machine learning approaches in predictive modelling tasks, as well as to allow semantic rule-base reasoning to illustrate how to infer new knowledge.

4.0.1. Case study I: Student's grade prediction in continuous evaluation

Grade prediction is one of the main tasks involving data analysis in education, since it allows teachers to plan and monitor their courses beforehand, as well as to adopt correction activities in case of deviations. In this first case of study, a series of supervised classification models are trained to predict the students' grades in continuous evaluation mode. The main features feeding the models comprise the visits made by the students, the submissions performed and the difference in days between submission and the cutoff-dates. Once the prediction models are generated with past information of previous courses, they can be used to predict the grade of those courses in Moodle, which are still recorded without a grade.

Following the current tendency in online education, which is indeed increased by the coronavirus pandemic global situation, classes are graded on a pass/fail basis. This is focused on a binary grading system, meaning that no letter grade will be recorded, but students just earn credit depending on whether they did satisfactory work in the class. Therefore, the students' grades have been discretised in 2 classes (pass and fail), according to the features that characterise them, such as the number of submissions, the number of views and the difference in days from the delivery date of assignments, i.e., label "Pass" refers to students with a high delivery rate and a high rate of views, whereas label "Fail" means low delivery rate in assignments and a low level of views.

Query 2: Number of views per user and course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT (SUM(?numclick) AS ?sum_clicks) ?userid
?courseid
WHERE{
  ?x elion:logEduLevel ?edulevel.FILTER (?edulevel
  = 2)
  ?x elion:recordUser ?user.
  ?x elion:recordCourse ?course.
  ?x elion:logSumClick ?numclick.
  ?course elion:courseSource elion:openUniversity.
  ?course elion:courseId ?courseid.
  ?user elion:userId ?userid
}GROUP BY ?userid ?courseid
```

In concrete, for this analysis a series of attributes are selected from the RDF repository: *sum_click*, *id_student*, *code_module*, *code_presentation*, *weight*, *date_submitted*, *id_assessment*, and *score*; which are obtained from SPARQL queries comprising different class properties of the e-LION ontology. In this sense, Query 2 is used to calculate the number of views by user and course. This query selects the triples whose educational level is equal to 2, since this level corresponds to students' interactions. In addition, it filters the origin of the data to those of the Open University and finally groups the results by applying the sum to the number of clicks of a user in a course.

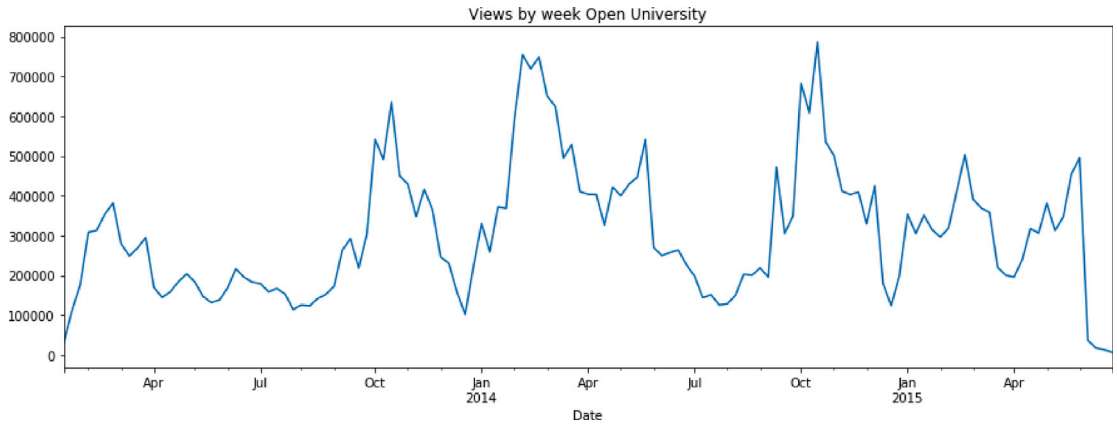

Fig. 3. Time-series plot of the student's views by week in a two year period from the Open University.

Query 3: Weight score per user and course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT (SUM(?w/100*?score/10)
       AS ?weight_score) ?userid ?courseid
WHERE{
  ?course elion:courseSource elion:openUniversity.
  ?course elion:courseId ?courseid.
  ?user elion:userId ?userid.
  ?assigment elion:hasCourse ?course.
  ?assigment elion:assignmentWeight ?w. FILTER
  (?w < 100)
  ?submission elion:belongAssignment ?assigment.
  ?submission elion:belongsUser ?user.
  ?submission elion:submissionScore ?score.
}GROUP BY ?userid ?courseid
```

Query 3 returns the data that will be discretised to later be used as a label (Pass/Fail). This query selects the triples whose origin is the Open University, the score obtained in the delivery and the weight that it has in the continuous grade. Those weights greater than 100 are filtered, since these tasks correspond to exams, which do not belong to the continuous evaluation. Finally, the query is grouped by user and course applying the sum of the grades taking into account their weightings.

Similarly, the number of submissions can be calculated with the SPARQL Query 4. It filters triples from the data source Open University and selects the submissions made by the users, to lately group the results by applying the count of submissions of a user in a course.

Query 4: Number of submissions per user and course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT (COUNT(?submission) AS ?count_sub) ?userid
?courseid
WHERE{
  ?course elion:courseSource elion:openUniversity.
  ?course elion:courseId ?courseid.
  ?user elion:userId ?userid.
  ?submission elion:belongsUser ?user.
  ?submission elion:belongAssignment ?assigment.
  ?assigment elion:hasCourse ?course.
}GROUP BY ?userid ?courseid
```

Query 5: Difference of days per user and course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT (DAY(?date_diff_with_hours) AS ?diff_days)
       ?userid ?courseid
WHERE{
  SELECT (SUM(?fdate - ?timecreated) AS
  ?date_diff_with_hours)
          ?userid ?courseid
  WHERE{
   ?course elion:courseSource elion:openUniversity.
   ?course elion:courseId ?courseid.
   ?user elion:userId ?userid.
   ?assigment elion:hasCourse ?course.
   ?assigment elion:assignmentAllowSubmissionsFrom
   Date
   ?fdate.
   ?submission elion:belongAssignment ?assigment.
   ?submission elion:belongsUser ?user.
   ?submission elion:submissionTimeCreated
   ?timecreated.
  }GROUP BY ?userid ?courseid
}
```

Table 10
Classification report of all the methods used (KNN, DT, SVM, RF, GNB, and MLP) in continuous evaluation. Computed metrics comprise the global Accuracy (Acc.), Precision (Prec.), Recall (Rec.), F1-Score (f1-Sc.), and Support (Sup).

Method	Acc.	Class	Prec.	Rec.	f1-Sc.	Sup.
KNN	0.90	Pass	0.86	0.92	0.89	3135
		Fail	0.92	0.87	0.89	3441
DT	0.90	Pass	0.87	0.92	0.89	3135
		Fail	0.92	0.88	0.90	3441
SVM	0.90	Pass	0.84	0.95	0.89	3135
		Fail	0.95	0.84	0.89	3441
RF	0.90	Pass	0.87	0.93	0.90	3135
		Fail	0.93	0.87	0.90	3441
GNB	0.89	Pass	0.82	0.96	0.88	3135
		Fail	0.95	0.81	0.88	3441
MLP	0.90	Pass	0.87	0.92	0.90	3135
		Fail	0.93	0.87	0.90	3441

Query 5 is used to calculate the difference in days from the assignment opening date to the submission date of students. It filters the Open University triplets and group the results by applying the sum of delayed dates of a user in a course. Finally, it extracts the total number of days.

All these computed feature values are scaled in the interval $[0, 1]$ for each course to homogenise numeric ranges. These values, as well as those of the grade label attribute are joined in a dataset to be used in the supervised classification tasks. The dataset is indeed split (randomly) into training and testing subsets with percentages of 75% and 25%, respectively.

For prediction modelling, a series of well-known classification algorithms have been used to check the consistency of data on different

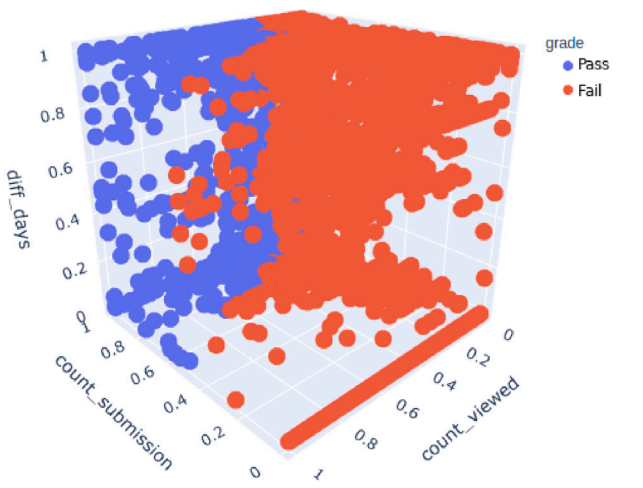

Fig. 4. Plot of prediction grades in continuous evaluation with regards to the Moodle (University of Malaga) source dataset.

learning procedures, namely: k-Nearest Neighborhood (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Gaussian Naive Bayes (GNB), and MultiLayer Perceptron Neural Network (MPL). These methods have been tuned throughout a grid search cross-validation task for hyper-parameter setting in training phase. With the obtained models, a set of predictions are conducted in validation phase with regards to the test set, so the computed results are organised in Table 10, for all the classifiers.

From this table, it is worth noting that successfully accuracies (Acc.) close to 90% are reached for all the methods, which indeed show balanced precision (Prec.) and Recall (Rec.) metrics for the two classes (Pass, Fail). Therefore, the obtained models can be used for grade prediction of new student's activity data.

In this sense, a last step in this study consists in continuous grade prediction, but in this case with the Moodle source dataset (University of Malaga), according to the input features defined for this model. As shown in Fig. 4, the resulting grade predictions can be visually distinguished for the two classes, so it could provide the teacher with an informative tool before the final evaluation of students.

4.0.2. Case study II: Student's final grade prediction

Similar to the previous case study, a set of prediction models are trained in this case, although considering the final grade of students in a course according to the interactions they made in the LMS. Therefore, in addition to the minimum set of features that characterise student's interactions (number of submissions, number of visits and the delivery time of assignments), the grade obtained of each learner in continuous evaluation is also included in the dataset as an extra feature for training the models. The final grade (also Pass/Fail) is considered as the response feature to be predicted in testing time.

In this regard, the SPARQL Query 6 is used to obtain the data that will be later grouped as the classification label (final grade). It selects the triples filtered by the Open University, as well as those for which the final grade is available. Properties *courseid* and *userid* of enrollment are also selected.

Query 6: Final grade per user and course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT ?userid ?courseid ?final_result
WHERE{
  ?course elion:courseSource elion:openUniversity.
  ?enroll elion:inCourse ?course.
  ?enroll elion:enrollmentFinalResult ?final_result.
  ?user elion:isEnrolled ?enroll.
  ?user elion:userId ?userid.
  ?course elion:courseId ?courseid.
}
```

Table 11
Classification report of all the methods used (KNN, DT, SVM, RF, GNB, and MLP) in final evaluation. Computed metrics comprise the global Accuracy (Acc.), Precision (Prec.), Recall (Rec.), F1-Score (f1-Sc.), and Support (Sup).

Method	Acc.	Class	Prec.	Rec.	f1-Sc.	Sup.
KNN	0.74	Pass	0.81	0.92	0.86	5058
		Fail	0.51	0.28	0.36	1518
DT	0.74	Pass	0.81	0.92	0.86	5058
		Fail	0.52	0.27	0.36	1518
SVM	0.74	Pass	0.77	0.99	0.87	5058
		Fail	0.62	0.03	0.06	1518
RF	0.76	Pass	0.85	0.79	0.82	5058
		Fail	0.44	0.55	0.49	1518
GNB	0.70	Pass	0.80	0.82	0.81	5058
		Fail	0.36	0.33	0.34	1518
MLP	0.73	Pass	0.79	0.97	0.87	5058
		Fail	0.55	0.14	0.22	1518

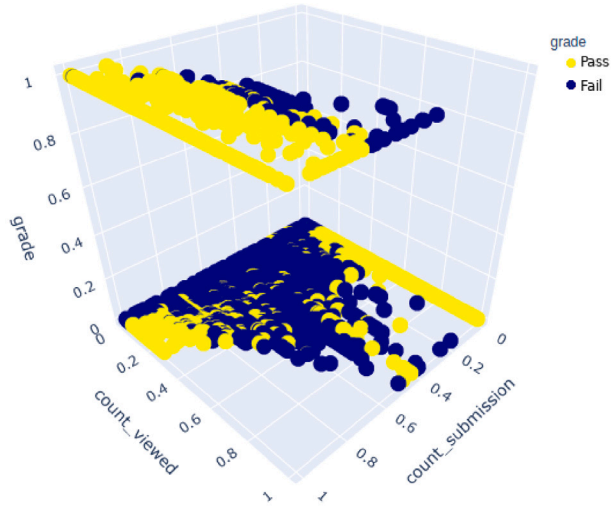

Fig. 5. Plot of prediction grades in final evaluation with regards to the Moodle (University of Malaga) source dataset.

The resulting dataset is used for feeding the classification models after splitting it for training (75%) and testing (25%). Again, grid search cross-validation is performed for hyper-parameter setting of classification methods in training phase. The resulting metric values obtained for the testing dataset are organised in Table 11, for all the methods. In this case, the global accuracies are lower than in continuous evaluation, with percentages between 76% (obtained by RF) and 70% (obtained by GNB). This is mostly due to low precision and recall values when predicting class "Fail", which shows certain bias (probably produced by grades in exams) in final grading of students.

Accordingly, Fig. 5 plots the final grades of students predicted for the Moodle source dataset of the University of Malaga, which clearly visualises the bias produced for the class "Fail". This is probably caused by a certain unbalancing of the resulting dataset because of a higher percentage of samples with label "Fail", which could be mitigated with undersampling until reaching balance of classes. Besides the moderate classification performance in this case of study, it is worth noting that integrating data from different LMSs (Open University) allows to generate useful predictors able to reproduce similar results in other sources (University of Malaga), so the e-learning semantic model proposed here is a useful contribution in this direction.

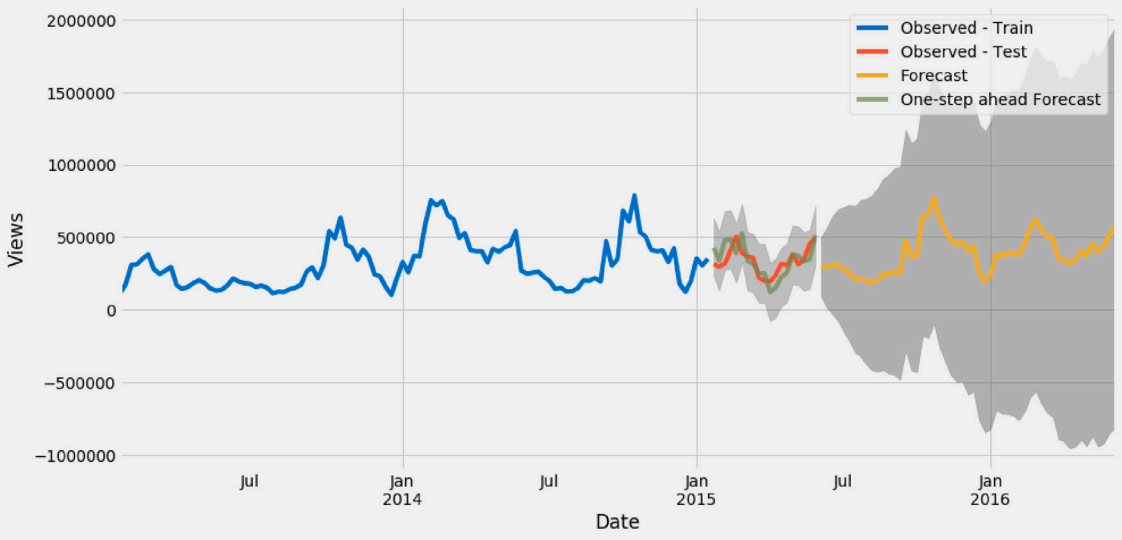

Fig. 6. Time-series plots of student's views of the LMS, where observed data is represented with regards to the forecast obtained by SARIMAX.

4.0.3. *Case study III: Student's visits time-series forecasting*

Another interesting use case consists in forecasting the tendency of students' visits in the LMS over time, then to warn of possible decreases in activity in certain periods, which would help decide on specific days in the semester for updating contents or activities.

To do so, a first step is to obtain the required data concerning the visits that students make in the e-learning system, together with the dates of such visits. These data can be selected with SPARQL Query 7, which aggregates the number of clicks and filters triples for the Open University source of data. The resulting dataset is a time-series of visits grouped by weeks to get homogeneous sample periods. A subset of 15% of these data are used for testing.

At this point, a previous analysis consists in checking whether the temporal series is stationary, or not, in order to decide which kind of algorithm to use (and how to tune it) for forecasting. Therefore, the Augmented Dickey–Fuller test is used on the dataset, resulting 95% confidence interval, with Test-statistic −3.22 and p-value 0.018, so it can be stated that the time-series is stationary.

For time-series training and forecasting two auto-regressive popular methods have been used: SARIMAX[16] and Prophet.[17] The former is an extension of the classical Autoregressive Integrated Moving Average (ARIMA) that supports Seasonal component, while the later implements a procedure for forecasting time-series data based on an additive model where non-linear trends are fitted with yearly, weekly, and daily seasonality, plus holiday effects. Therefore, these two methods are well-adapted to characterise the different learning periods in University courses.

Query 7: Time-series of views.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT (SUM(?numclick) AS ?count_viewed) ?timecreated
WHERE
  { ?x elion:logEduLevel ?edulevel. FILTER (?edulevel
  = 2)
    ?x elion:logTimeCreated ?timecreated.
    ?x elion:recordCourse ?course.
    ?x elion:logSumClick ?numclick.
    ?course elion:courseSource elion:openUniversity.
}GROUP BY ?timecreated
```

[16] Available in URL https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html.

[17] Available in URL https://github.com/facebook/prophet.

Table 12
Metrics error results of time-series predictions for SARIMAX and Prophet models.

Error measure	SARIMAX	Prophet
MAE	6.81e+04	6.91e+04
MSE	6.45e+09	6.04e+09
RMSE	8.03e+04	7.77e+04
MSLE	6.00e−02	5.00e−02

The hyper-parameters of the SARIMAX model have been tuned with a grid search procedure, choosing the one with the lowest Akaike Information Criterion (p=P=1, d=D=1, q=Q=0, AIC=412.55). Once the two algorithmic models are trained, the testing partition is predicted and a forecast is performed for the following 53 weeks. An illustrative plot of these results is shown in Fig. 6, where the time-series obtained by SARIMAX are represented with regards to the observed data.

Table 12 contains the results obtained by SARIMAX and Prophet in terms of commonly used error metrics for regression models, where the observed test data is compared with the forecast ($y - \hat{y}$). These metrics are: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Root Mean Square Log Error (RMSLE). In general, it can be checked that Prophet shows lower error values than SARIMAX, although both methods are successfully fitted with moderate error values, e.g., RMSE close to 80,000 in 53 weeks (1132 views per week), with an average of 318,577 students' views per week in observed data.

4.0.4. *Case study IV: Reasoning tasks*

In order to cover as much as possible the potentials of using the proposed semantic model, this last case study consists in the generation of SWRL semantic rules to perform reasoning tasks on the knowledge-base. To this end, a classification task has been specified for students according to their activities and the number of assignments delivered.

SWRL Rule 8: Looker.

```
elion:Course(?c)
^ elion:enrollmentNumberOfClicks(?e, ?nclicks)
^ swrlb:greaterThan(?nclicks, ?avgclicks)
^ elion:enrollmentNumberOfSubmissions(?e, ?nsub)
^ elion:courseClicksAVG(?c, ?avgclicks)
^ elion:enrollmentRole(?e, ?r) ^ swrlb:equal(?nsub, 0)
```

```
^ swrlb:stringEqualIgnoreCase(?r, "Student")
^ elion:User(?u) ^ elion:isEnrolled(?u, ?e)
^ elion:inCourse(?e, ?c)
-> elion:typeOfUser(?e, "Looker")
```

The SWRL rules have been defined by using several data properties of classes *Course* and *Enrollement*, this last connecting each user to the course (s)he is enrolled. Concretely, for class *Course* (Table 5), the data properties *CourseClicksAVG* and *CourseNumberOfSubmissions* have been used, which denote the average number of clicks done by the students and the number of assignments declared for the course, respectively. For class *Enrollement* (Table 7), SWRL rules are defined with regards to data properties: *enrollmentNumberOfClicks*, to register the number of clicks that a student has done on this course, and *enrollmentNumerOfSubmissions* that stores the number of deliveries the student has performed for this course. The values of these properties are calculated throughout queries, so they are not explicitly stored a-priori in the RDF repository.

The data property *TypeOfUser* defines, for each course and student, the kind of student according to his/her behaviour by means of a categorical label: *Looker* (high number of clicks, but low number of deliveries), *Active* (high number of clicks and deliveries) and *Passive* (low number of clicks and deliveries). Therefore, the values of *TypeOfUser* property are those induced by the reasoner (Stardog).

SWRL Rule 9: Passive.

```
elion:Course(?c)
^ elion:enrollmentNumberOfClicks(?e, ?nclicks)
^ swrlb:greaterThan(?nclicks, ?avgclicks)
^ swrlb:multiply(?per, ?rate, 100)
^ swrlb:lessThan(?rate, 50)
^ elion:User(?u)
^ elion:courseClicksAVG(?c, ?avgclicks)
^ elion:enrollmentNumberOfSubmissions(?e, ?nsub)
^ swrlb:stringEqualIgnoreCase(?r, "Student")
^ elion:courseNumberOfAssignments(?c, ?nassig)
^ swrlb:divide(?rate, ?nsub, ?nassig)
^ elion:enrollmentRole(?e, ?r)
^ elion:isEnrolled(?u, ?e)
^ elion:inCourse(?e, ?c)
-> elion:typeOfUser(?e, "Passive")
```

In this way, each professor could analyse the data of previous years to define classification rules according to their own parameters, i.e., the numeric thresholds to discriminate among the student's behaviours. For example, code snippet in SWRL Rule 8 defines a student who clicks more than the average in a course, but without performing deliveries, so (s)he is classified as a *Looker*. SWRL Rule 9 denotes a student who clicks more than the average and performs 50% of required deliveries in a given course, therefore (s)he is classified as *Passive*. Similarly, the *Active* student would be one who does more clicks than the average and performs more than 80% of required deliveries, in a given course.

Query 10: Example of course.

```
PREFIX elion: <http://ontologies.khaos.uma.es/
e-lion/>
SELECT ?userid ?type WHERE {
  ?user elion:isEnrolled ?e.
  ?user elion:userId ?userid.
  ?e elion:inCourse ?course.
  ?e elion:typeOfUser ?type.
  ?course elion:courseId ?id_course. FILTER
  (?id_course='66')
  ?course elion:courseSource elion:
  UniversidadDeMalaga.
}
```

Once these rules are executed in the reasoner, it is possible to obtain all the students' ids of a given course together with their classifications referring behaviours. The SPARQL Query 10 shows an example in this regard, to classify learners (by userid) of course '66' of the University of Malaga.

5. Discussions

In light of the previous use cases and results, it can be argued that the proposed semantic model is able to constitute a multi-source integration knowledge-base for efficiently feeding data analysis and visualisations. This is a clear step ahead to approach the data interoperability challenge identified in current surveys e.g., Heiyanthuduwage (2022), Rahayu et al. (2022), in the context of ontologies for e-leaning management systems.

From a practical perspective, the e-LION ontology can be used at the core of a linked open data consolidation strategy, where current LMSs and other academic data sources are systematically queried to support teachers with analysis on the course evolution. In addition, the underpinning semantic model is useful when analysing students' activities in the context of the overall performance of a given grade, which would provide them with a global overview of their progresses, hence leading to their own learning plans.

In concrete, a series of technical remarks can be extracted in form of lessons learned, as follows:

- The definition of a semantic model, based on an OWL2 ontology, on top of Moodle data, is helpful to integrate data from other e-learning platforms or datasets. OWL2 ontologies enable the unambiguous identification of entities and assertion of named relationships that connect these entities. Furthermore, the OWL2 language provides the mechanisms defining logical constraints on integrated data if needed.
- The OWL 2 ontology proposed in this work can be easily aligned with existing e-learning ontologies and vocabularies thanks to the ontology alignment mechanisms. Therefore, other data related to e-learning, i.e. e-learning resources, can be easily consolidated within the same ABox structure.
- In this regard, the implementation of a common RDF repository integrating heterogeneous data in a standard formalism simplifies the user queries and facilitates obtaining the input data to algorithms.
- The SWRL language allows defining rules in the context of e-LION, hence providing a reasoning mechanism to infer new knowledge from the integrated data, automatically classify users or subjects, etc.

Together with these comments, it is worthy to mention that e-LION could be used for the ontology alignment with many other ones, not only in the educational domain of knowledge, but also in different domains, such as: social networks, health related COVID-19 user's behaviours, demographic and social evolution. This would enable more advanced data integration and querying mechanisms and to nourish multi-dimensional analysis.

6. Conclusions

In this work, a semantic approach for multi-source e-learning data integration is proposed, which comprises the generation of a new OWL 2 ontology called e-LION. A series of mapping functions are defined to consolidate data sources from different LMSs (Moodle, COCO Udemy, Open University) to RDF in a common repository, that can be now used for feeding advanced analysis by means of SPARQL queries, to monitor student's and teacher's interactions. A set of use cases have been developed for validation, which deal with grade prediction, time-series forecasting of student's views and semantic reasoning with SWRL rules for the classification of students' behaviours.

The proposed semantic approach is shown to properly integrate e-learning data, enabling the advanced querying and constituting a well-grounded knowledge-base to enhance informative analysis in the context of online learning management systems. This leads the proposed e-LION ontology to provide scientific added value, which in the context of the current state of the art (as explained in Section 2), it allows the semantic connection with other related ontologies and vocabularies, hence promoting the generation of extensive linked data in the domain of e-learning.

Therefore, the proposal can be used at the core of a data consolidation strategy for future applications, where current LMSs and other academic data sources are systematically queried to support teachers with advanced analytics and visualisations. Similarly, for students, it allows analysing students' activities in the context of the overall performance of a given grade, which would provide them with a global perspective of their course performances, thus promoting their proactive learning.

As future work, we plan to include more data from other learning management systems, as well as to update the e-LION ontology to incorporate new relevant attributes from different e-learning perspectives. In this regard, another future activity is the ontology alignment of many other not only in the educational domain of knowledge, but also in different domains, such as: social networks, health related COVID-19 user's behaviours, demographic and social evolution.

CRediT authorship contribution statement

Manuel Paneque: Conceptualization, Investigation, Methodology, Software, Data curation, Writing – original draft. **María del Mar Roldán-García:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Project administration, Validation, Funding acquisition. **José García-Nieto:** Conceptualization, Methodology, Formal analysis, Supervision, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Innovation, Spain via Grant PID2020-112540RB-C41 (AEI/FEDER, UE) and Andalusian PAIDI program, Spain with grant P18-RT-2799. It has been developed in the context of PIE-17-166: Advanced Analysis of Students in Virtual Campus, and we specially thank to Carlos Romero and Rafael Gutierrez from the Virtual Campus Service of the University of Malaga for their technical support and data availability. Funding for open access charge: Universidad de Málaga / CBUA.

References

Al-yahya, M., George, R. P., & Alfaries, A. A. (2015). Ontologies in e-learning: Review of the literature. *International Journal of Software Engineering and Its Applications*, *9*(2), 67–84.

Aldana-Martín, J. F., García-Nieto, J., del Mar Roldán-García, M., & Aldana-Montes, J. F. (2022). Semantic modelling of earth observation remote sensing. *Expert Systems with Applications*, *187*, Article 115838. http://dx.doi.org/10.1016/j.eswa.2021.115838, https://www.sciencedirect.com/science/article/pii/S0957417421012008.

Altan, A., & Karasu, S. (2019). The effect of kernel values in support vector machine to forecasting performance of financial time series. *The Journal of Cognitive Systems*, *4*(1), 17–21.

Barba-González, C., García-Nieto, J., Roldán-García, M. M., Navas-Delgado, I., Nebro, A. J., & Aldana-Montes, J. F. (2019). BIGOWL: Knowledge centered big data analytics. *Expert Systems with Applications*, *115*, 543–556.

Bouihi, B., & Bahaj, M. (2019). An UML to OWL based approach for extracting Moodle's Ontology for Social Network Analysis. *Procedia Computer Science*, *148*, 313–322.

Brochhausen, M., Whorton, J. M., Zayas, C. E., Kimbrell, M. P., Bost, S. J., Singh, N., et al. (2022). Assessing the need for semantic data integration for surgical biobanks—A knowledge representation perspective. *Journal of Personalized Medicine*, *12*(5), http://dx.doi.org/10.3390/jpm12050757, https://www.mdpi.com/2075-4426/12/5/757.

Delgoshaei, P., Heidarinejad, M., & Austin, M. A. (2022). A semantic approach for building system operations: Knowledge representation and reasoning. *Sustainability*, *14*(10), http://dx.doi.org/10.3390/su14105810, https://www.mdpi.com/2071-1050/14/10/5810.

Dessì, D., Fenu, G., Marras, M., & Reforgiato Recupero, D. (2018). COCO: Semantic-enriched collection of online courses at scale with experimental use cases. In A. Rocha, H. Adeli, L. P. Reis, & S. Costanzo (Eds.), *Trends and advances in information systems and technologies* (pp. 1386–1396). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-77712-2_133.

Firdausiah Mansur, A. B., & Yusof, N. (2013). Social learning network analysis model to identify learning patterns using ontology clustering techniques and meaningful learning. *Computers & Education*, *63*, 73–86.

George, G., & Lal, A. M. (2019). Review of ontology-based recommender systems in e-learning. *Computers & Education*, *142*, Article 103642. http://dx.doi.org/10.1016/j.compedu.2019.103642.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, *5*(2), 199–220.

Han, L. (2018). An interdisciplinary intelligent teaching system model based on college students' cognitive ability. In *2018 int. conf. on virtual reality and intel. sys.* (pp. 259–262).

Harris, S., & Seaborne, A. (2013). SPARQL 1.1 query language. https://www.w3.org/TR/sparql11-query/.

Heiyanthuduwage, S. R. (2022). A review: Status quo and current trends in e-learning ontologies. In M. E. Auer, H. Hortsch, O. Michler, & T. Köhler (Eds.), *Mobility for smart cities and regional development - challenges for higher education* (pp. 114–125). Cham: Springer International Publishing.

Horrocks, I., Patel-Schneider, P. F., Bechhofer, S., & Tsarkov, D. (2005). OWL rules: A proposal and prototype implementation. *Web Semantics: Science, Services and Agents on the WWW*, *3*(1), 23–40.

Hssina, B., Bouikhalene, B., & Merbouha, A. (2017). An ontology to assess the performances of learners in an e-learning platform based on semantic web technology: Moodle case study. In A. Rocha, M. Serrhini, & C. Felgueiras (Eds.), *Europe and MENA Cooperation advances in information and communication technologies* (pp. 103–112). Cham: Springer International Publishing.

Joy, J., Raj, N. S., & V. G., R. (2021). Ontology-based e-learning content recommender system for addressing the pure cold-start problem. *J. Data and Information Quality*, *13*(3), http://dx.doi.org/10.1145/3429251.

K., S., Poscic, P., & Jaksic, D. (2020). Ontologies in education – state of the art. *Education and Information Technologies*, (25), 5301–5320. http://dx.doi.org/10.1007/s10639-020-10226-z.

Karasu, S., & Altan, A. (2019). Recognition model for solar radiation time series based on random forest with feature selection approach. In *2019 11th international conference on electrical and electronics engineering* (pp. 8–11). http://dx.doi.org/10.23919/ELECO47770.2019.8990664.

Karasu, S., Altan, A., Saraç, Z., & Hacioğlu, R. (2018). Prediction of bitcoin prices with machine learning methods using time series data. In *2018 26th signal processing and communications applications conference* (pp. 1–4). http://dx.doi.org/10.1109/SIU.2018.8404760.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*, http://dx.doi.org/10.1038/sdata.2017.171.

Makwana, K., Patel, J., & Shah, P. (2018). An ontology based recommender system to mitigate the cold start problem in personalized web search. In S. C. Satapathy, & A. Joshi (Eds.), *1, Inf. and comm. tech. for intel. sys.* (pp. 120–127). Cham: Springer.

del Mar Roldán García, M., García-Nieto, J., & Aldana-Montes, J. F. (2016). An ontology-based data integration approach for web analytics in e-commerce. *Expert Systems with Applications*, *63*, 20–34. http://dx.doi.org/10.1016/j.eswa.2016.06.034.

del Mar Roldán-García, M., Uskudarli, S., Marvasti, N. B., Acar, B., & Aldana-Montes, J. F. (2018). Towards an ontology-driven clinical experience sharing ecosystem: Demonstration with liver cases. *Expert Systems with Applications*, *101*, 176–195. http://dx.doi.org/10.1016/j.eswa.2018.02.001.

McGlinn, K., Rutherford, M. A., Gisslander, K., Hederman, L., Little, M. A., & O'Sullivan, D. (2022). FAIRVASC: A semantic web approach to rare disease registry integration. *Computers in Biology and Medicine*, *145*, Article 105313. http://dx.doi.org/10.1016/j.compbiomed.2022.105313, https://www.sciencedirect.com/science/article/pii/S0010482522001056.

Navarrete, R., & Luján-Mora, S. (2015). Use of linked data to enhance open educational resources. In *2015 international conference on information technology based higher education and training* (pp. 1–6). http://dx.doi.org/10.1109/ITHET.2015.7218017.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. *Tech. Rep.*, Stanford University Knowledge Systems Laboratory KSL-01-05, http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html.

Obeid, C., Lahoud, I., El Khoury, H., & Champin, P.-A. (2018). Ontology-based recommender system in higher education. In *Companion proceedings of the the web conference 2018* (pp. 1031–1034). International World Wide Web Conferences.

Ouf, S., Abd Ellatif, M., Salama, S., & Helmy, Y. (2017). A proposed paradigm for smart learning environment based on semantic web. *Computers in Human Behavior*, *72*, 796–818.

Pereira, C. K., Siqueira, S. W. M., Nunes, B. P., & Dietze, S. (2018). Linked data in education: A survey and a synthesis of actual research and future challenges. *IEEE Transactions on Learning Technologies*, *11*(3), 400–412. http://dx.doi.org/10.1109/TLT.2017.2787659.

Rahayu, N. W., Ferdiana, R., & Kusumawardani, S. S. (2022). A systematic review of ontology use in E-learning recommender system. *Computers and Education: Artificial Intelligence*, *3*, Article 100047. http://dx.doi.org/10.1016/j.caeai.2022.100047, https://www.sciencedirect.com/science/article/pii/S2666920X22000029.

Schreiber, G., & Raimond, Y. (2014). RDF 1.1 primer. https://www.w3.org/TR/rdf11-primer/.

Sobral, T., Galvão, T., & Borges, J. (2020). An ontology-based approach to knowledge-assisted integration and visualization of urban mobility data. *Expert Systems with Applications*, *150*, Article 113260. http://dx.doi.org/10.1016/j.eswa.2020.113260, https://www.sciencedirect.com/science/article/pii/S0957417420300853.

Suguna, S., Sundaravadivelu, V., & Gomathi, B. (2016). A novel semantic approach in e-learning information retrieval system. In *2016 IEEE international conference on engineering and technology* (pp. 884–889). http://dx.doi.org/10.1109/ICETECH.2016.7569374.

Tarus, J. K., Niu, Z., & Yousif, A. (2017). A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Generation Computer Systems*, *72*, 37–48. http://dx.doi.org/10.1016/j.future.2017.02.049.

Thaddeus, S., Jeganathan, A., & Leema, G. T. (2011). Semantic integration of classical and digital libraries. In *Multimedia information extraction and digital heritage preservation* (pp. 51–65). http://dx.doi.org/10.1142/9789814307260_0003.

Zeng, Q., Zhao, Z., & Liang, Y. (2009). Course ontology-based user's knowledge requirement acquisition from behaviors within e-learning systems. *Computers & Education*, *53*(3), 809–818.