IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 May 2025, accepted 20 July 2025, date of publication 1 August 2025, date of current version 7 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3594963

RESEARCH ARTICLE

Applying the FAIR Principles to Open Educational Resources: A Semantic Similarity Approach to Improve Resource Discovery

FARZANEH SADEGHI[1,2], **AURIOL DEGBELO**[3], **CARSTEN KESSLER**[2,1], **AND REZA ZOLNOURI**[4]

[1]Department of Sustainability and Planning, Aalborg University, 2450 Copenhagen, Denmark
[2]Department of Geodesy, Bochum University of Applied Sciences, 44801 Bochum, Germany
[3]Chair of Geoinformatics, TUD Dresden University of Technology, 01069 Dresden, Germany
[4]Chair of Mathematics of Information Processing, RWTH Aachen University, 52062 Aachen, Germany

Corresponding author: Farzaneh Sadeghi (farsad@plan.aau.dk)

This work was supported by German Research Foundation (DFG) under Grant 460036893.

ABSTRACT Open educational resources (OER) are teaching, learning, or research resources freely available for use and reuse. Despite their potential, OER uptake in existing education systems remains low, primarily due to challenges in locating suitable resources. This study addresses this challenge by proposing and implementing a workflow applying the FAIR (Findable, Accessible, Interoperable, and Reusable) principles to OER. We demonstrated this framework within the Earth System Sciences as an application domain. We constructed a knowledge graph of approximately 500 FAIR OER, each annotated with structured metadata using the Schema.org vocabulary and made accessible through a SPARQL endpoint. To bridge the gap between making resources queryable and enabling their practical reuse, we employed a transformer-based language model (Sentence-BERT). The model was fine-tuned using few-shot learning on a domain-specific dataset of course-description pairs. This specialized model was then used to map the OER collection against over 200 university courses across five academic programs at a German university, based on semantic similarity between OER descriptions and university course descriptions. Expert evaluation of the model's recommendations demonstrated 74% accuracy in identifying reusable OER for university courses. Notably, even with limited training data, fine-tuning the Sentence-BERT model significantly improved performance, resulting in a 16% reduction in mean squared error compared to the base model. This study provides both a generalizable methodology and a practical demonstration of how FAIR principles can streamline OER discovery, potentially accelerating OER uptake in higher education.

INDEX TERMS FAIR principles, higher education, knowledge graph, LLM, metadata, open educational resources, reusability.

I. INTRODUCTION

In 2001, the Massachusetts Institute of Technology (MIT) launched its OpenCourseWare project [1], making course materials from nearly all its undergraduate and graduate courses freely available online. The term ''Open Educational Resources'' was first introduced during the UNESCO Forum on the Impact of Open Courseware for Higher Education. This event marked the official recognition of the need for accessible educational resources on a global scale [2]. Over the past two decades, the definition of OER has evolved. What began as a way to share textbooks and lecture notes has expanded to include a wide variety of materials, including multimedia content, entire courses, assessments, and research materials. These resources are either in the public domain or released under open licenses, allowing users to freely access, adapt, reuse, and redistribute them. This flexibility has made OER a powerful tool to

The associate editor coordinating the review of this manuscript and approving it for publication was Wojciech Sałabun.

facilitate access to education and foster innovation in teaching practices [3].

The adoption of OER offers a variety of benefits for institutions, educators, and learners, enhancing the accessibility of education and providing opportunities for quality improvement through broader collaboration, peer review, and adaptation. OER can enable greater curriculum flexibility. The resources can be continuously updated and adapted to suit local contexts, ensuring that educational content remains relevant and of high quality. OER help eliminate the need for expensive textbooks and other costly educational materials, thereby reducing the financial burden on students and making education more inclusive. In a broader societal context, OER play a pivotal role in promoting educational equity. By providing open access to knowledge, OER empower individuals to improve their circumstances through education, contributing to the creation of a more just society. OER also foster collaboration and innovation within academic and professional communities, as educators can easily get an overview of other instructors' approaches when teaching a given subject. The open sharing of educational resources encourages cross-disciplinary engagement, which can lead to innovative solutions and technological advancements. On a personal level, sharing knowledge through OER can bring a sense of fulfillment to the educator/instructor and enhance an individual's professional reputation [3], [4], [5], [6], [7].

Despite the convincing arguments in favor of OER and their numerous potential benefits, widespread adaptation has been slow. While open licenses are fundamental to the OER movement, enabling legal reuse and adaptation is not sufficient for the successful reuse of OER [6], [8]. Several challenges have hindered their widespread adoption, including not only access barriers but also structural barriers. First, the vast number of openly licensed resources is spread across many platforms, repositories, and websites, making it difficult for educators and learners to find specific materials efficiently. This vast distribution complicates the search process, as users must sift through a multitude of sources to find relevant content. Additionally, the absence of standardized metadata worsens these difficulties. Without uniform descriptors and consistent tagging, search engines and repositories struggle to index and retrieve resources effectively [6], [8], [9], [10], [11]. Second, although OER provide a vast array of freely accessible materials for educators and learners, not all of these resources consistently meet the quality standards or seamlessly align with a specific educational context. Consequently, users of OER must invest significant time and effort in assessing and selecting appropriate materials that effectively address their unique training needs and goals [5], [7], [12], [13], [14]. Lastly, the lack of institutional frameworks and pedagogical support for self-directed learning with OER implies that individuals must manage their educational activities independently, which may not be feasible for all learners. The OER movement has often assumed that learners possess the inherent ability to self-direct their educational activities and navigate educational resources independently without institutional structures, overlooking the varying capacities and competencies required for self-directed learning [15], [16].

Despite significant research exploring the barriers to effectively reusing OER, there has been insufficient focus on establishing actionable criteria or standards that define their reusability. Although the FAIR principles (Findable, Accessible, Interoperable, and Reusable) are widely acknowledged as essential for enabling reuse and interoperability across various domains, their specific application to the OER landscape is not adequately addressed. Developed by a consortium of scholars and stakeholders to improve the reuse of research objects, the FAIR principles provide guidelines to enhance reusability by ensuring that resources can be easily found, accessed, and used by both humans and machines [17]. Implementing these principles to make data FAIR (FAIRification) has proven influential in enhancing data management and reusability [18], indicating a positive impact in fields where they have been applied. Both the FAIR data movement and the OER movement aim to reduce barriers to access, enhance reusability, and promote the sharing of knowledge; thus, applying the FAIR principles to OER could strengthen this alignment and potentially overcome existing challenges in finding and reusing educational materials; regardless, the application of the FAIR principles to OER has not received adequate attention and requires further investigation [9], [19].

To address the concerns surrounding the low reusability of OER and the absence of research on support systems that could automate and support the reuse process, this work set out to answer the following research questions:

- RQ1: How can the FAIR principles be applied to enhance the reusability of OER?
- RQ2: How can the discovery of OER relevant to specific educational contexts be improved?
- RQ3: How can OER be automatically integrated into existing educational frameworks to facilitate their effective reuse?

By addressing these questions through a case study, this article offers the following contributions:

- A workflow for the retrospective FAIRification of existing OER, enabling improved reusability and discoverability for educational resources accumulated over more than two decades. Its application yielded a proof-of-concept dataset comprising over 500 FAIR-compliant OER within a selected multidisciplinary field.
- A data-efficient methodology for fine-tuning general language models on domain-specific data to enhance OER discovery. By calculating semantic similarities, this approach refines the identification and retrieval of relevant resources, as demonstrated by successful application in the Earth System Sciences.
- A reproducible method for systematically integrating OER into existing educational contexts, promoting OER adoption and educational transformation. The method involves embedding curated OER recommendations

into curricula in an organized and sustainable manner, supporting easy replication by other educators or institutions, and ultimately broadening the reach of OER.

The remainder of this paper is organized as follows: Section II covers the related work, Section III details the workflow for making existing OER compliant with the FAIR principles and fine-tuning a language model for improved findability, including an exemplary case study that illustrates how OER can be integrated into existing university curricula along with the evaluation results. Our findings are discussed in Section IV, followed by the conclusion and future work in Section V.

II. RELATED WORK

The scientific community is becoming more collaborative, with increasing reliance on shared data, workflows, and experimental resources. As science increasingly relies on digital technologies, traditional scholarly publications fail to support the efficient exchange of complex, decomposable scientific outputs. To keep pace with modern scientific practices, the notion of Research Objects (ROs) was introduced. ROs aggregate data, methods, and context to create a self-contained unit of knowledge. The main goal of ROs was to offer a solution by bundling the diverse digital components of a scientific study into a reusable and shareable format [20].

The FAIR principles build upon the foundational ideas of transparency, reusability, and machine actionability that ROs pioneered, offering a broader, more formalized framework to encompass all forms of digital data. These principles provide guidelines to enhance the reusability of data by ensuring that they can be easily found, accessed, and used both by humans and machines. The FAIR principles, as the name implies, are organized into four main categories. Each of the four has specific objectives that need to be met. Findability requires that datasets be well-documented with extensive metadata, allowing both people and machines to find them easily. Accessibility ensures that the datasets can be retrieved using standard protocols, while interoperability focuses on the ability of different datasets to work together. Finally, reusability ensures that data can be used in other contexts by providing enough information about its origins and use conditions [17].

The FAIR principles are intentionally designed without specific technological mandates, allowing an adaptation by different communities to make their resources usable in the long term [17], [21]. However, the variability in interpretations of these principles can lead to inconsistent implementations. In response to the necessity for alignment toward common standards and methods, a set of considerations for FAIR implementation was introduced. The guidelines emphasize the need for unique identifiers, rich metadata, explicit references, and registration in searchable databases for enhanced findability. The accessibility guidelines focus on using standardized communication protocols, open access, and providing authentication for sensitive data. The interoperability principle highlights the importance of using shared, standardized languages, FAIR ontologies, and qualified references to other datasets. Lastly, the reusability principle stresses the need for detailed data descriptions, clear usage licenses, provenance information, and adherence to community standards. The implementation considerations emphasize that making data machine-actionable is a key aspect of optimizing resource usage by automated systems [21].

Complementary to implementation guidelines and to help different communities make their data FAIR, a generic FAIRification workflow has been designed. This workflow consists of several steps: identifying the FAIRification objective, analyzing the data and its accompanying metadata, defining semantic models for both, making the data and metadata linkable, hosting the FAIR data, and finally assessing how FAIR the data is [22].

Even though there are technical guidelines and a FAIRification workflow in place, implementing the FAIR principles still presents challenges. Different communities need to apply these principles to their resources, adapt them based on the type of resources and their specific characteristics, and continue refining and sharing their approaches to data management. By reusing existing solutions or developing new ones to fill gaps, communities can contribute to a cycle of continuous improvement [21], [22].

An illustrative example of extending the FAIR principles is adapting them for research software, addressing its unique requirements. New considerations like documenting software dependencies and ensuring the availability of executable versions of the software were identified as necessary for FAIR research software [23]. This effort led to the development of FAIR4RS, a formalized set of principles tailored explicitly for research software [24].

Building on the adoption of FAIR principles for research software, there is a growing recognition of the necessity to apply these principles to other complex digital resources. For instance, Degbelo [25] explored the application of the FAIR principles to geovisualizations on the Web. It was pointed out that a holistic discussion on FAIRness needs to go beyond a computer-centric perspective to embrace the perspective of both the consumer and the producer of the resources to be made FAIR. While the current article agrees with the need for a holistic discussion on the FAIRness of educational resources, it focuses particularly on the requirements for realizing FAIR OER from a computing perspective as a first step. A discussion of FAIRness implications from the perspectives of user interfaces for OER use and the authoring of OER is deferred to future work.

Extending the FAIR principles to educational resources represents a natural progression, further advancing FAIR beyond its application to traditional research outputs. The customization of the principles for educational resources is relatively unexplored. While there are existing guidelines for applying the FAIR principles to training materials, they tend to be somewhat simplified, offer limited details on practical implementation, and are primarily tailored to life sciences [19]. Further advancements in this area have acknowledged the challenges associated with the lack of standardized markup and have led to Bioschemas, a proposed solution that is a set of metadata standards. Bioschemas aims to improve the findability of life sciences resources, including training materials, by extending the capabilities of Schema.org [26] with new types and properties and defining usage profiles specifically designed for life science resources [9]. Bioschemas offer valuable frameworks for enhancing the findability of training resources, yet they are limited in scope, focusing primarily on the life sciences and not addressing the labor-intensive task of aligning retrieved resources with specific reuse requirements. Despite these limitations, they do highlight the role of metadata in improving the discoverability of educational resources.

Learning resource metadata includes details like the resource's title, subject, description, and technical requirements [27]. To standardize how metadata is used across the Web, initiatives like Schema.org have been developed. Schema.org is a metadata vocabulary that has significant implications for how information surfaces in search results and how content is structured across the Web [26], [28]. It standardizes how information is labeled on websites, making it easier for search engines to understand and retrieve relevant data. Building on this, the Learning Resource Metadata Initiative (LRMI) extends Schema.org specifically for educational content. LRMI is designed to describe the educational characteristics and relationships of learning resources [29], [30].

OER metadata plays a crucial role in assisting users in efficiently evaluating whether it meets their learning objectives without the need to review the entire content. Educational materials often comprise extensive texts or lengthy video content, which can be time-consuming to assess solely based on their content. When multiple resources are returned for a single search query, the significance of metadata becomes even more evident. It allows users to navigate through the options and select the most suitable resource. Research indicates that the quality of metadata, particularly the quality of title, description, and language fields, is closely associated with the overall quality of OER content [4], [11], [31], [32]. However, even with rich metadata, users may still face challenges finding the most suitable OER. OER metadata fields are descriptive in nature and can be quite lengthy and filled with jargon, making it necessary to invest significant effort in reading and comprehending them. When users have specific learning objectives, which may also involve complex, multi-paragraph texts, the task of comparing these against numerous OER descriptions becomes quite daunting. Navigating through extensive and technical content to align one's educational needs with the right resource is challenging, particularly when presented with a wide range of options [16], [33].

Recently, exploring the use of Large Language Models (LLMs) to accelerate labor-intensive and time-consuming FAIRification processes, like extracting complex data from large volumes of scientific publications, has emerged in the FAIR data community [34]. Inspired by these advancements, we apply LLMs to process and semantically compare learning objectives, aiming to automate the integration of OER into existing educational frameworks and address the need for efficient alignment.

While making OER FAIR inherently enhances their findability and accessibility, reusing them effectively in existing educational frameworks requires ensuring interoperability with their educational plan, which may not always be available as machine-actionable data. For effective integration, existing educational plans must also adopt formats and standards that facilitate interoperability with FAIR OER, ensuring that both resources can be aligned. Recent studies have focused on constructing educational knowledge graphs from resources such as syllabi, textbooks, and student assessments to represent instructional concepts and their relationships in a structured, machine-readable format [35], [36], [37], [38]. By employing these developments, our work aims to address the challenge of efficiently integrating OER into existing frameworks. Leveraging LLMs for the semantic comparison of learning objectives, we aim to enrich existing educational plans with relevant resource suggestions.

III. WORKFLOW FOR FAIR OER GENERATION

This section details the systematic workflow developed to make OER FAIR-compliant. Our efforts to consolidate dispersed OER into a centralized access point, aimed at addressing resource discovery challenges, led to the development of this workflow (Fig. 1). This OER FAIRification workflow was initially developed and validated using existing OER within the Earth System Sciences (ESS). ESS proved to be an ideal case study because it represents a rapidly evolving, multidisciplinary field where the constant need to update content and learn across disciplines drives a high demand for reusing educational resources. Furthermore, current developments in the field are fueled by data-intensive research and new technologies, meaning educators must frequently update their teaching materials to reflect the latest science. This workflow incorporates established guidelines (Section II) and adapts the general FAIRification workflow outlined by Jacobsen et al. [22] for the specific context of OER.

A. PRE-FAIRIFICATION

This section describes the essential "Pre-FAIRification" phase, which lays the groundwork for the subsequent FAIRification process. This phase focuses on the identification, categorization, and initial collection of OER, along with an analysis of their existing metadata.

1) PRE-FAIRIFICATION: IDENTIFY OER PROVIDERS

The pre-FAIRification phase focuses on identifying and categorizing existing OER. For our case study, the initial collection was manually compiled, comprising a total of 80 educational resources within the ESS domain from

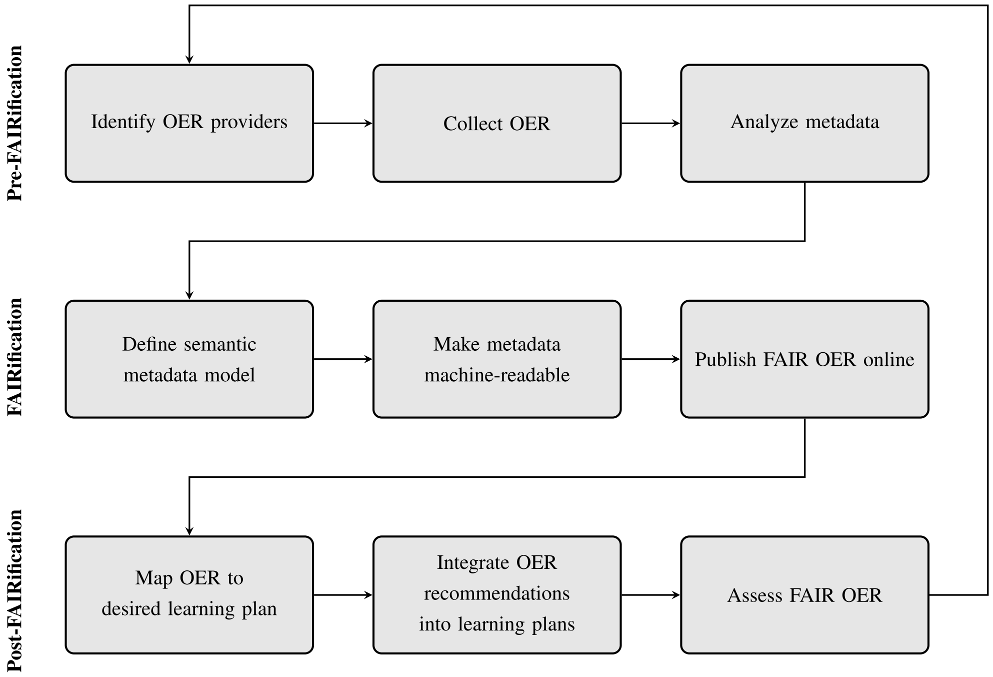

Pre-FAIRification
Identify OER providers
Collect OER
Analyze metadata
FAIRification
Define semantic metadata model
Make metadata machine-readable
Publish FAIR OER online
Post-FAIRification
Map OER to desired learning plan
Integrate OER recommendations into learning plans
Assess FAIR OER

FIGURE 1. Workflow for making OER FAIR adapted from the "Generic Workflow for the Data FAIRification Process" introduced by Jacobsen et al. [22]. All steps shown in the figure were implemented during the work.

31 providers. These providers were categorized into two distinct groups:

- Individual contributors: This group includes high-caliber content released under open licenses on platforms such as GitHub, YouTube, and educators' personal websites. These resources offer specialized content but require careful evaluation of metadata quality and consistency.
- Institutional entities: Organizations that consistently provide substantial, quality-controlled content. Examples include the EarthLab learning portal [39] and MIT OpenCourseWare [40].

2) PRE-FAIRIFICATION: COLLECT OER

Providers within the Institutional Entities category were subjected to automatic metadata harvesting to ensure the continuous expansion of the OER collection. For instance, metadata harvesting for resources provided by the EarthLab learning portal was initiated using text-scraping techniques. Raw materials available on GitHub were provided in Markdown format, featuring a structured YAML preamble containing information such as the title, objectives, URL, and author of each resource. This process resulted in a structured dataset of 376 resources from EarthLab.

Similarly, MIT OpenCourseWare offers filtering options across different disciplines. Web scraping techniques were employed on resources categorized under "Earth, Atmospheric, and Planetary Sciences" to extract course information, including title, description, URL, instructor, and topics. This effort delivered a structured dataset of 106 resources from MIT OpenCourseWare.

3) PRE-FAIRIFICATION: ANALYZE METADATA

After duplicates and empty entries were removed, the final collection of OER comprised 502 unique educational resources, which included both manually and automatically collected resources. Prior to initiating the FAIRification process, an analysis of metadata availability was conducted. The investigation revealed that while metadata and resource information were generally available, they were not consistently organized using structured metadata schemas. In instances where existing standards such as Schema.org were used, structured data often described resources such as WebPage, VideoObject, or Person rather than Course or LearningResource. Additionally, in many cases, structured data provided minimal information and lacked comprehensive details about courses, instructors, or learning objectives.

For all 502 resources, the following information was collected: title, description, publisher, URL, language, prerequisites, and license information. Many resources also included keywords detailing the subject area and the tools and data used. Additional information, such as educational level, date, type, and course author, was gathered where available. This comprehensive collection of metadata was used as a basis for subsequent FAIRification steps.

B. FAIRIFICATION

In this section, we go through the process of transforming the OER collection into a format that adheres to the

FAIR principles. Given the complexity of implementing the FAIR principles across all aspects of OER, a fully comprehensive approach was neither feasible nor intended. Instead, this section aims to achieve a ''FAIR enough'' state that significantly improves the discoverability and reusability of the resources while acknowledging that some aspects of FAIR compliance may require further refinement in future work.

1) FAIRIFICATION: DEFINE SEMANTIC METADATA MODEL

The initial step in the FAIRification phase involved defining a semantic metadata model. This semantic metadata model acted as a template for transforming the existing OER metadata into a machine-readable format. Initially, attributes of OER, including their associated concepts and relationships, were mapped. This mapping was derived from analyzing the available metadata and the supplementary information we were able to collect to augment the existing metadata. As discussed in Section II, the course title, description, and language are critical fields that influence users' decisions to reuse resources.

The OER collection is restricted to English-language resources, and course titles were consistently available. In instances where descriptions were missing, text extraction methods were employed to retrieve course descriptions or learning objectives directly from the resource content. Additionally, summarization tools were utilized to generate descriptions where necessary. Table 1 outlines the attributes and relationships of the OER, providing a structured overview of the metadata schema implemented.

This minimalist approach to concept definition prioritizes the use of readily available information, thereby minimizing adaptation costs while enhancing findability. To enable machine-readability, we translated these concepts and relationships into a structured format using the Schema.org vocabulary. OER were represented using the `schema:LearningResource` class. The primary relationship between OER is defined by the `schema:coursePrerequisites` property, indicating that one resource explicitly requires prior knowledge from another. This seemingly simple relationship is crucial as it allows for the construction of directed learning pathways and facilitates effective knowledge acquisition. Table 2 presents a complete mapping of our defined attributes to the Schema.org vocabulary and the semantic metadata model for OER. As of this writing, the semantic model supports the annotation of five types of learning resources defined in a controlled vocabulary adapted from the IEEE Standard for Learning Object Metadata [41]: `exercise`, `slide`, `narrative text`, `quiz`, and `video`. Likewise, `educationalLevel` can take one of four values: `intro`, `beginner`, `intermediate`, and `advanced`. The `intro` level signifies resources providing a high-level overview or conceptual context for a topic, whereas `beginner` denotes resources that introduce specific foundational concepts, definitions, or basic tasks within the topic.

2) FAIRIFICATION: MAKE METADATA MACHINE-READABLE

Next, we operationalized the semantic metadata model by creating a linkable, machine-readable format for the OER metadata collection. We used the Resource Description Framework (RDF), which provides a common underlying model for data exchange, to facilitate metadata harvesting and interoperability. This involved instantiating the model with OER metadata using the Python RDFLib library, thereby generating a structured representation ready for practical implementation. The choice of RDF facilitated metadata harvesting and interoperability, promoting wider adoption and integration. The data was serialized in the Turtle format, a standard RDF syntax that ensures compatibility with a broad range of data processing and analysis tools. This approach enables the transformed dataset to be seamlessly incorporated into diverse systems, workflows, and future applications. An example of a *LearningResource* instance represented in Turtle format is shown below:

```
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:       <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema:     <http://schema.org/> .
@prefix xsd:        <http://www.w3.org/2001/XMLSchema#> .

<http://example.org/oer/123>
    rdf:type                    schema:LearningResource ;
    schema:author
   ↪  <https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/odx-course-v1-⌋
   ↪  NFDI4Earth-20230827PO-Self-Paced?jsonPointer=/author/1>
   ↪  ;
    schema:competencyRequired   "Basic understanding of Python, basic
   ↪  understanding of geographic information concepts" ;
    schema:coursePrerequisites  <http://example.org/oer/122> ;
    schema:datePublished        "2023-10-01" ;
    schema:description          "Dive into the world of spatial data analysis
   ↪  with our Python for spatial data analysis course. This advanced-level
   ↪  course focuses on exploring, mapping, and analyzing spatial data using
   ↪  Python, enabling tasks from desktop GIS to be more flexible and
   ↪  reproducible. By the end, you will understand core Python modules like
   ↪  Fiona and GeoPandas and gain sufficient knowledge to perform tasks
   ↪  independently. Enhance your interdisciplinary skills with this
   ↪  comprehensive course, and discover the power of Python in spatial
   ↪  analysis!" ;
    schema:educationalLevel     "Intermediate" ;
    schema:inLanguage
   ↪  <http://publications.europa.eu/resource/authority/language/ENG> ;
    schema:keywords             "tutorial" , "training" , "education" ,
   ↪  "Geoinformatics" , "Python" ;
    schema:learningResourceType "exercise" , "narrative text" ;
    schema:license              <http://spdx.org/licenses/CC-BY-4.0> ;
    schema:name                 "Python for spatial data analysis" ;
    schema:url                  "https://edutrain.nfdi4earth.de/courses/course-⌋
   ↪  v1:NFDI4Earth+20230828PFSDA+Self-Paced"^^xsd:anyURI
   ↪  ;
    schema:publisher            <https://ror.org/04x02q560> .
```

3) FAIRIFICATION: PUBLISH FAIR OER ONLINE

The final step in the FAIRification phase (Fig. 1) involves making the curated resources accessible to the target audience. We achieved this by registering the resources in the NFDI4Earth KnowledgeHub [42], an ESS-specific RDF triple store designed for FAIR compliance. NFDI4Earth is an initiative from the national research data infrastructure in Germany targeting the harmonization of services related to digital research products in the Earth System Sciences [43], [44]. The NFDI4Earth KnowledgeHub stores metadata in RDF format about datasets, repositories, standards, and software in the Earth System Sciences. Hence, publishing the OER through the NFDI4Earth KnowledgeHub not only ensures that they are findable online but also opens up the possibility for automated link discovery with other resources

TABLE 1. Conceptual metadata schema, attributes for describing OER based on the presence of information.

Concept	Explanation
OER	A freely accessible and openly licensed digital object designed specifically for learning, education, and skill-building purposes.
Title	An OER title refers to the name or heading given to an OER.
Description	A brief overview of the content and purpose of an OER. It typically outlines the topics covered and the learning objectives.
Publisher	The individual, organization, or institution responsible for creating, distributing, or making an OER publicly available.
Creator	The individual or group who develops or authors the content of an OER.
URL	The Web address or link that provides direct access to an OER online.
Language	The language in which the content of an OER is written or presented.
Prerequisite	A prior OER that learners are expected to have engaged with or understood before using the current OER.
Requirement	Any necessary knowledge, tools, technologies, or conditions that must be met to effectively use an OER.
License	A legal document that specifies how an OER can be used, shared, adapted, and distributed.
Type	The specific kind of material that an OER represents. Common types include textbooks, lecture notes, and videos.
Keywords	Subject-related terms, educational tools, methodologies, or technologies relevant to an OER.
Difficulty	The intended audience's proficiency level.
Date	Publication or release date of an OER.

TABLE 2. Semantic metadata model.

Attributes	LearningResource Property	Expected Type
Title	`schema:name`	Text
Description	`schema:description`	Text
Publisher	`schema:publisher`	Organization
Creator	`schema:author`	Person
URL	`schema:url`	URL
Language	`schema:inLanguage`	URL
Prerequisite	`schema:coursePrerequisites`	LearningResource
Requirement	`schema:competencyRequired`	Text
License	`schema:license`	URL
Type	`schema:learningResourceType`	DefinedTerm
Keywords	`schema:keywords`	Text
Difficulty	`schema:educationalLevel`	DefinedTerm
Date	`schema:datePublished`	Date

(e.g., relevant datasets or software for an OER). Fig. 2 provides a simple example of a SPARQL query (''Available learning resources and their types'') along with the results. The IRIs (Internationalized Resource Identifiers) for the OER are all dereferenceable, making it possible to access metadata about the OER online. Additional examples of queries possible through the KnowledgeHub include: ''All educational resources published by an organization'', ''Contact point for an educational resource'', and ''Licences under which an open educational resource has been published''.

C. POST-FAIRIFICATION

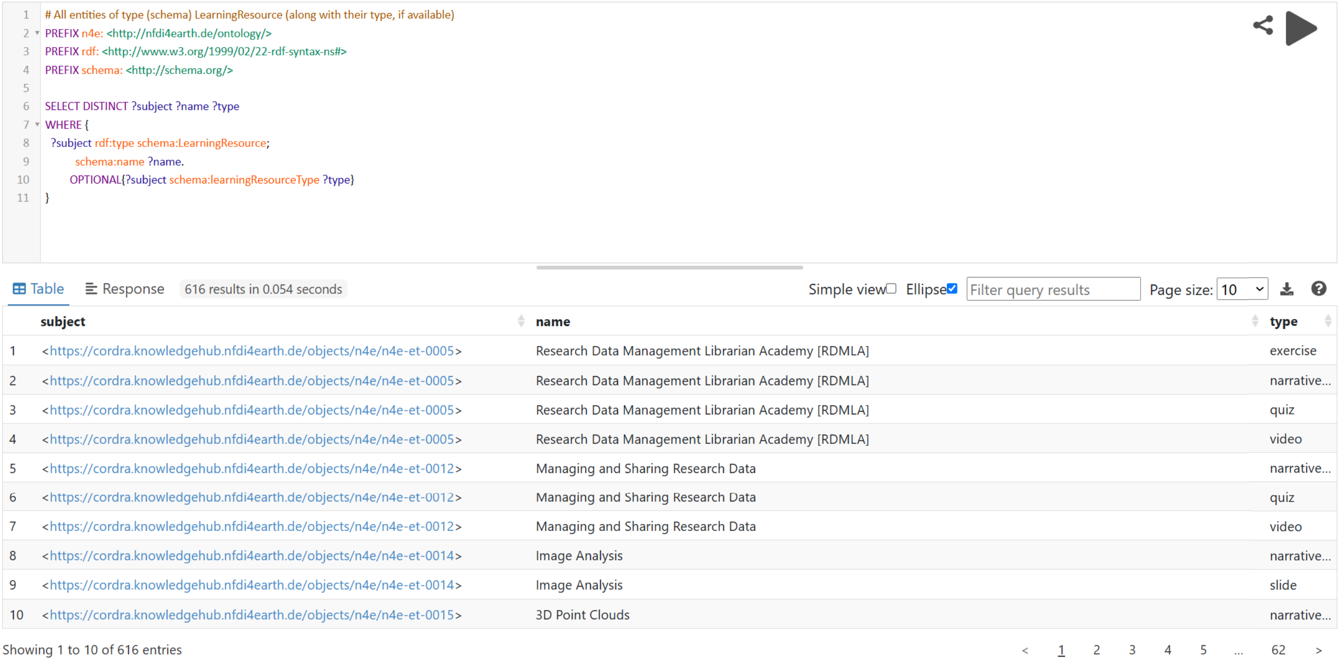
While FAIRification enhances the discoverability of OER, seamless integration into diverse learning plans requires further steps. This challenge arises from the descriptive nature of OER metadata, which often necessitates domain expertise and significant time investment to align resources with specific learning objectives. To address this, we present a practical example of mapping FAIR OER to existing educational programs using large language models. This automated approach analyzes OER metadata and matches resources with programs based on their description, including their learning objectives, reducing manual effort and promoting wider adoption of FAIR OER.

This section first details the selection and fine-tuning of an encoder language model using domain-specific data. The fine-tuning process aims to optimize the model specifically for assessing semantic similarity between university course descriptions and OER descriptions within a specific educational domain. This optimization relies on a training dataset consisting of course-description pairs, which were curated based on pre-existing associations established by material owners. Subsequently, we used the fine-tuned model to map OER to university study programs, integrate these mapping results, and evaluate the effectiveness of the generated OER recommendations.

1) DATASET

The foundation for our training data was an original dataset created using MIT resources within the ''Earth, Atmospheric, and Planetary Sciences'' discipline. The basis for each pairing was a clear equivalence or strong thematic overlap in the educational content covered by both descriptions. These pairs were generated by matching descriptions of the same MIT courses retrieved from two distinct sources: the MIT subject catalog [45] and MIT OpenCourseWare (OCW) [40]. OCW provides materials and descriptions from past and present MIT courses, while the subject catalog describes those currently offered.

We used Python's `selenium` library to fetch the HTML content and employed `BeautifulSoup` to parse the HTML and extract the course descriptions. The extracted data were structured and stored in CSV files for subsequent



```
# All entities of type (schema) LearningResource (along with their type, if available)
PREFIX n4e: <http://nfdi4earth.de/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX schema: <http://schema.org/>

SELECT DISTINCT ?subject ?name ?type
WHERE {
 ?subject rdf:type schema:LearningResource;
        schema:name ?name.
      OPTIONAL{?subject schema:learningResourceType ?type}
}
```

Table
Response
616 results in 0.054 seconds
Simple view☐ Ellipse☑ Filter query results Page size: 10

	subject	name	type
1	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0005>	Research Data Management Librarian Academy [RDMLA]	exercise
2	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0005>	Research Data Management Librarian Academy [RDMLA]	narrative...
3	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0005>	Research Data Management Librarian Academy [RDMLA]	quiz
4	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0005>	Research Data Management Librarian Academy [RDMLA]	video
5	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0012>	Managing and Sharing Research Data	narrative...
6	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0012>	Managing and Sharing Research Data	quiz
7	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0012>	Managing and Sharing Research Data	video
8	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0014>	Image Analysis	narrative...
9	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0014>	Image Analysis	slide
10	<https://cordra.knowledgehub.nfdi4earth.de/objects/n4e/n4e-et-0015>	3D Point Clouds	narrative...

FIGURE 2. **Example of available learning resources in the NFDI4Earth KnowledgeHub and their types.**

analysis. We then paired course descriptions based on matching course codes, resulting in 63 pairs, each containing a description from OCW and the MIT subject catalog. Although variations may exist between courses offered in different semesters or by different instructors, these pairs provide a realistic and diverse foundation for training our language model. This approach enhances the model's ability to handle the diversity of language and structure found in actual course catalogs and OER repositories by embracing real-world variations in course descriptions.

To further enhance the diversity and size of our dataset, we employed data augmentation techniques, generating slightly dissimilar, moderately dissimilar, highly dissimilar, and completely dissimilar pairs for each original pair. We identified domain-specific terms and jargon in the text, such as geological formations, climatology, and planetary names, and replaced them with other terms from the same domain. These replacements introduce subtle distinctions that are relevant within the specific domain but might be overlooked by a general-purpose language model. Slightly dissimilar pairs involved minor alterations, with only one key term influencing course similarity. Moderately dissimilar pairs had half of the key terms modified. Highly dissimilar pairs altered all key terms except for one feature, and completely dissimilar pairs involved altering all domain-specific features.

Assigning precise numerical similarity scores to text pairs is challenging due to the nuanced nature of language. To structure our dataset for analysis, we developed a heuristic scoring system based on the degree of alteration in domain-specific terms during data augmentation. To guide the model's learning process, we assigned a similarity score to each pair, ranging from 0.05 to 0.95. Original pairs, which contained equivalent course descriptions from both sources, were assigned a high similarity score of 0.95, reflecting minimal divergence. Augmented pairs received progressively lower scores corresponding to the level of alteration. The primary objective of this scoring system was to evaluate the model's ability to rank course description pairs accurately based on their degree of similarity. Heuristic scoring methods are successfully used in similar tasks where objective measures are difficult to define [46], [47].

This augmentation process resulted in a total of 215 pairs. We used the NASA Global Change Master Directory (GCMD) Keyword Viewer [48] as a reference for feature replacement during data augmentation.

2) METHOD

Our method quantifies the semantic similarity between the textual descriptions of educational resources by first generating sentence embeddings for each description and then computing the cosine similarity between these embeddings. This yields a similarity score between 0 and 1, with higher scores indicating greater similarity. We used a pre-trained Sentence-BERT (SBERT) model [49], which adapts the BERT architecture using siamese and triplet network structures to produce semantically meaningful sentence embeddings suitable for comparison via cosine similarity. Based on comprehensive evaluation metrics provided by the model developers [50], we selected the `all-mpnet-base-v2` model, which demonstrated strong performance across various sentence embedding and semantic tasks.

To assess the inherent ability of the pre-trained `all-mpnet-base-v2` SentenceTransformer model to capture the semantic similarity between course descriptions, we first employed a zero-shot learning approach. This involved encoding the MIT and OCW original course descriptions into sentence embeddings and computing the cosine similarity between the resulting vectors. We then evaluated the correlation between these similarity scores and the labels using Pearson correlation coefficients. The Pearson correlation of 0.545 indicates a moderate positive linear relationship, highlighting room for further refinement to capture the complex relationship between course descriptions and their corresponding OER.

To further enhance the model's performance, the `all-mpnet-base-v2` model was fine-tuned on the augmented dataset. The dataset was randomly partitioned into training (129 samples), evaluation (43 samples), and test (43 samples) sets using a 60/20/20 split. The model was trained using the `CoSENTLoss` function and `AdamW optimizer` with a learning rate of 2e-5 and a batch size of 256 for 107 epochs. Model performance was monitored during training using the evaluation set, and the best-performing model was selected for final evaluation on the held-out test set and mapping phase.

Following training, we evaluated the performance of both the zero-shot and fine-tuned models using a held-out test set. Evaluation metrics included the Pearson and Spearman correlation coefficients to assess the correlation between predicted and given similarity scores, as well as the Mean Squared Error (MSE) and R-squared (R^2) to quantify the model's accuracy in predicting the actual similarity scores. The results of this evaluation are presented in Table 3.

The fine-tuned model demonstrates improvements over the baseline in predicting similarity scores, with the Pearson correlation coefficient increasing from 0.679 to 0.741, Spearman correlation from 0.704 to 0.747, MSE decreasing from 0.084 to 0.046, and R^2 value rising from 0.059 to 0.482. These enhancements indicate that the model has developed a deeper understanding of the domain-specific terms and jargon critical to accurately assessing course-OER similarity within Earth System Sciences. Consequently, the model can now distinguish between closely related subdomains, a capability that general-purpose models lack.

3) POST-FAIRIFICATION: MAP OER TO DESIRED LEARNING PLAN

To demonstrate the practical application of our approach, we selected study programs at one university in Germany – the Technical University of Munich (TUM) – as a case study for mapping our FAIR OER dataset. We curated a dataset of course definitions by extracting information from publicly available module handbooks, which detail course structures and learning outcomes. These handbooks, available as PDF documents, were selected from TUM departments responsible for ESS-related education and included both bachelor's and master's programs:

- B.Sc. Geodesy and Geoinformation [51]
- M.Sc. Geodesy and Geoinformation [52]
- M.Sc. Earth Oriented Space Science and Technology [53]
- M.Sc. Cartography [54]
- M.Sc. Information Technologies for the Built Environment [55]

To enable computational analysis, we converted the PDF handbooks into a machine-readable format. We developed a custom Python function, `pdftocsv`, to parse these documents, extract relevant sections using predefined keywords, and convert them into structured CSV files. These individual course lists were then aggregated into a single dataset. Subsequent data cleaning involved removing duplicate entries, filtering out non-course offerings (e.g., thesis, seminars), and translating German descriptions to English using the `googletrans` library. We then performed text normalization to remove special characters, convert text to lowercase, and eliminate extra newlines. The final dataset comprised 224 distinct university course descriptions in CSV format, including course titles, content, intended learning outcomes, and recommended prerequisites.

To identify potentially relevant OER for each TUM course, we used the fine-tuned SentenceTransformer model developed previously. We employed the `AutoTokenizer` with a maximum sequence length of 512 tokens to accommodate lengthy descriptions. Embeddings were generated by passing the tokenized inputs through the fine-tuned SBERT model and applying mean pooling over the token embeddings. This process encoded both the TUM course descriptions and the OER descriptions into sentence embeddings. We then calculated the cosine similarity between each TUM course embedding and all OER embeddings, resulting in a similarity matrix with 112,448 values. This matrix quantifies the semantic similarity between each TUM course and all OER in our dataset.

To visualize the distribution of similarity scores, we generated a histogram (Fig. 3). The histogram exhibits a long-tailed pattern, with a high concentration of low similarity scores and a small proportion of higher scores. This indicates that most OER have limited similarity to the TUM courses, while a select few exhibit strong semantic alignment. This observation is further supported by the skewness (0.82) and kurtosis (1.36) of the distribution, which indicate a right skew and a higher prevalence of extreme values compared to a normal distribution.

The highest calculated similarity score (0.73) was observed between the TUM course "Satellite Navigation and Advanced Orbit Mechanics" and the OER titled "Modern Navigation." Table 4 presents the course descriptions for these two resources.

Given the observed long-tailed distribution of similarity scores, we focused on the highest-scoring pairs for integration

TABLE 3. Performance comparison of the fine-tuned model and the baseline model.

	Pearson Correlation	Spearman Correlation	Mean Squared Error	R-squared
Fine-Tuned Model	0.7410	0.7473	0.0460	0.4821
Baseline Model	0.6787	0.7036	0.0835	0.0591

TABLE 4. Course descriptions for TUM's "Satellite navigation and advanced orbit mechanics" and the OER "Modern navigation".

Satellite navigation and advanced orbit mechanics [53]	The module consists of three lectures; Introduction to Satellite Navigation, Carrier Phase Positioning, and Advanced Celestial Mechanics. The first two lectures are blocked, Introduction to Satellite Navigation takes place in the first part of the semester, Carrier Phase Positioning takes place in the second part of the semester. Carrier Phase Positioning and Advanced Celestial Mechanics are accompanied by exercise labs. Introduction to Satellite Navigation: Principles of satellite navigation, Space segment and ground segment, Code and phase observations, Observation equations, Propagation error sources, Differences and linear combinations. Carrier Phase Positioning: Differential GPS and GNSS, Carrier smoothing, Augmentation systems for aeronautical and other critical navigation tasks, Carrier phase ambiguity resolution. Advanced Orbit Mechanics: Orbit determination, Selected problems: Lagrange points, swing-by, Analytical orbit theory: Hill-theory After the successful conclusion of the module, the students are able to: • to understand the concept and algorithms of satellite navigation and precise positioning, • to apply them to practical problems of satellite navigation and precise positioning, • to assess the impact of error sources on signal propagation, • to compute linear combinations and evaluate tracking data quality, • to apply processing strategies to analyze GNSS data for precise positioning applications, • to understand the principles and concepts of differential navigation and augmentation systems, • to apply phase ambiguity resolution strategies, • to understand the concepts of orbit determination, analytical orbit theories and the origin of Lagrange points, • to apply these concepts to practical problems and to analyze and assess the results.
Modern Navigation [56]	This course introduces the concepts and applications of navigation techniques using celestial bodies and satellite positioning systems such as the Global Positioning System (GPS). Topics include astronomical observations, radio navigation systems, the relationship between conventional navigation results and those obtained from GPS, and the effects of the security systems, Selective Availability, and anti-spoofing on GPS results. Laboratory sessions cover the use of sextants, astronomical telescopes, and field use of GPS. Application areas covered include ship, automobile, and aircraft navigation and positioning, including very precise positioning applications.

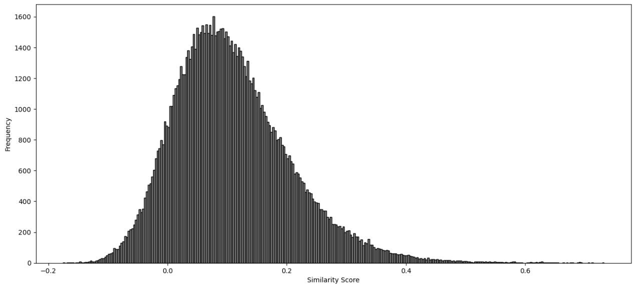


FIGURE 3. Similarity scores distribution between OER descriptions and the course descriptions from TUM used in the case study.

into the TUM curricula and subsequent evaluation. This strategy prioritizes the most promising candidates, where the model exhibits the highest confidence in identifying relevant matches between TUM courses and OER. Specifically, we set a threshold at the 99.5th percentile (similarity score $\geq$ 0.44), effectively focusing on the top 0.5% of the pairs (563 pairs) with the strongest indication of semantic alignment. By concentrating on these high-confidence matches, we aim to maximize the likelihood of successful OER integration and minimize the effort required for manual evaluation.

4) POST-FAIRIFICATION: INTEGRATE OER RECOMMENDATIONS INTO LEARNING PLANS

To represent the relationships between TUM courses and potentially relevant OER, we constructed a knowledge graph using the `RDFlib` library in Python. We employed the Schema.org vocabulary (Table 2) to define entities and relationships within the graph. TUM courses and OER were represented as nodes, with edges connecting courses to their identified OER matches based on the similarity threshold ($\geq$ 0.44). Prerequisite relationships between TUM courses, extracted from the module handbooks, were also incorporated. To ensure a comprehensive representation of the educational landscape, we interconnected the TUM courses based on their recommended prerequisites. Most prerequisites referred directly to other courses within the dataset. In cases where prerequisites suggested topics instead of courses, we matched them to courses covering those topics, assuming that the university offers a self-contained curriculum. This process resulted in a richly interconnected graph of courses and their relationships.

Each TUM course was represented as an instance of `schema:Course`, with properties for `schema:name` and `schema:description`. High-similarity OER were linked to their corresponding TUM courses using the `schema:material` property. These OER, represented

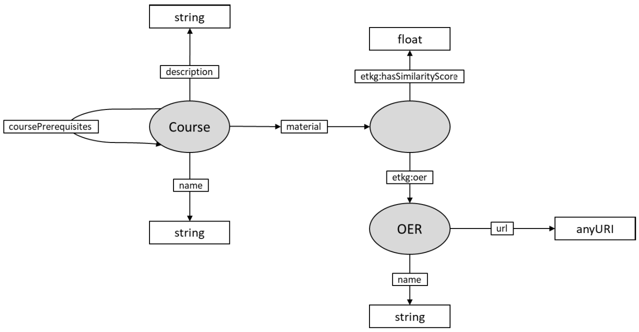


FIGURE 4. **An example of a data model for constructing learning plans interoperable with FAIR OER.**

as instances of `schema:LearningResource`, included properties for `schema:name`, `schema:description`, and `schema:url`. The `schema:url` property linked each OER to its structured metadata stored on KnowledgeHub (see Section III-B). Additionally, we annotated the OER-course connections with their respective similarity scores using a custom property, `etkg:hasSimilarityScore`. This knowledge graph, serialized in Turtle format (3,447 triples), provides a machine-readable representation of the identified connections between educational resources, enabling further analysis and exploration. Fig. 4 illustrates a conceptual data model for this graph-based enhanced curriculum, illustrating the entities, relationships, and properties used to structure the learning plan and ensure interoperability with FAIR OER. The node without a label stands for a blank node.

5) POST-FAIRIFICATION: ASSESS FAIR OER

To assess the effectiveness of our semantic similarity model in suggesting reusable OER for real-world educational settings, we conducted a survey involving domain experts. The primary objective was to assess the relevance, coverage, and appropriateness of the recommended OER for specific TUM courses.

Given the long-tailed distribution of similarity scores and our focus on high-confidence mappings, we randomly selected 36 OER-course pairs from those integrated into the learning plan, which were above the 99.5th percentile threshold (similarity score ≥ 0.443). To assess the presence of false negatives and further analyze the characteristics of pairs across different similarity levels, we also randomly sampled 36 pairs below the threshold.

Eighteen domain experts participated in the survey, including professors, lecturers, and research associates with an average of 13 years of experience in ESS-related domains, including geodesy, cartography, and geoinformatics. All experts were affiliated with German or Danish higher education institutions and research centers. To capture comprehensive feedback, we developed a structured questionnaire comprising both quantitative and qualitative questions. Each expert received four unique course-OER pairs. The detailed survey questions presented to the experts were as follows:

Question 1: Does the OER cover the topics included in the course syllabus?
□ Covers all major topics
□ Covers most topics
□ Covers some topics
□ Covers few of the topics
□ Covers none of the topics

Question 2: Is the level of depth and complexity of the OER appropriate for the course level?
□ The OER is too basic for this course
□ The OER is appropriate for this course
□ The OER is too advanced for this course
□ I cannot judge based on the information provided

Question 3: On a scale of 1–5, how relevant is this OER to the course content?
□ 1 □ 2 □ 3 □ 4 □ 5

Question 4: Would you recommend this OER as a resource for this course?
□ Yes □ No □ Maybe

Question 5: Do you have any additional comments or suggestions regarding this OER-course pair?

Surveys were distributed electronically, and responses were collected anonymously over 3 weeks and entered into a structured format for statistical analysis. To assess the model's performance, we used the responses to Question 4, which directly addressed the suitability of the OER for the corresponding course. We categorized the responses into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), as defined below:

- True Positive (TP): an OER is correctly identified as potentially reusable for a university course.
- False Positive (FP): an OER is incorrectly identified as potentially reusable for a university course.
- True Negative (TN): an OER is correctly identified as not reusable for a university course.
- False Negative (FN): an OER is incorrectly identified as not reusable for a university course.

In cases where experts responded with "Maybe" to Question 4, we adjusted the classification to "Yes" or "No" based on their responses to Questions 1 and 2, which probed the OER's coverage and appropriateness for the course. This adjustment provided a more definitive assessment of the OER's suitability. We employed standard performance metrics to evaluate the models' ability to detect semantically related course topics: recall, precision, and accuracy. These metrics enabled a quantitative comparison between the OER-course pairings detected by the models and those delineated by expert evaluations.

To assess the relationship between the experts' ratings and the similarity scores generated by our model, we conducted correlation analyses, first normalizing the expert responses regarding relevance (Question 3) using linear scaling normalization. To provide a comprehensive performance benchmark, we evaluated our fine-tuned SBERT model against three reference points: the original base model (all-mpnet-base-v2), OpenAI's GPT-4o, and DeepSeek R1. We selected GPT-4o and DeepSeek R1 as representative state-of-the-art baselines. GPT-4o was chosen as arguably the most popular and widely used advanced language model at the time this research was conducted. Concurrently, DeepSeek R1 was selected as one of the strongest-performing open-source language models available during the research period, demonstrating top-tier results on public leaderboards. For the LLM benchmarks, GPT-4o and DeepSeek R1 were prompted to assign similarity scores to the same set of 72 OER-course description pairs previously rated by the experts. The scores generated by all four models were then compared against the ground-truth labels derived from the expert survey using evaluation metrics, including Pearson and Spearman correlations, Mean Squared Error (MSE), and R-squared (R^2). This allowed for quantification and direct comparison of performance across the different models.

Finally, we conducted a thematic analysis of the qualitative feedback provided by the experts in response to Question 5. This analysis aimed to identify common themes, suggestions, and concerns raised by the experts, providing valuable insights for further refinement of our approach.

Analysis of the expert survey responses revealed that our fine-tuned model achieved an accuracy of 0.736 in identifying relevant OER for TUM courses, with a recall of 0.840 and a precision of 0.583. This indicates that the model effectively identified a majority of the relevant OER (high recall) while maintaining a moderate level of accuracy in its predictions.

The correlation analysis further demonstrated the model's ability to capture the semantic similarity between OER and TUM courses, as evidenced by the moderate to strong correlations between the predicted similarity scores and the expert ratings. The fine-tuned model performs strongly relative to the others. It achieves a Pearson correlation of 0.598, the highest linear correlation in this set, indicating its predicted similarity scores track closely with actual scores in a linear sense. The Spearman correlation is 0.584, which is roughly on par with MPNet and GPT-4o, showing that its ranking of similarities is also in line with human judgment. Notably, the fine-tuned model has the lowest MSE at 0.063, indicating its similarity score predictions are quite close to the true values. Correspondingly, it has the highest R^2 at 0.353, meaning it explains about 35.31% of the variance in human similarity ratings. This substantially higher R^2 suggests that fine-tuning improved the model's alignment with actual similarities. Overall, the fine-tuned model demonstrates the best combination of high correlation and low error, making it a top performer in this task, as detailed in Table 5.

Fine-tuning the model achieved a 16.0% reduction in MSE compared to the base model, suggesting that fine-tuning the SentenceTransformer model significantly enhances its ability to capture semantic similarity in line with human perception, approaching the performance level of a sophisticated language model like GPT-4o, even with limited training data and without access to human-annotated labels. Furthermore, the lower MSE and higher R^2 values for the fine-tuned model indicate its superior accuracy in predicting the actual similarity scores compared to the other models. This is particularly noteworthy considering the SBERT model's smaller scale and reduced complexity compared to a sophisticated large language model like GPT-4o.

Finally, we conducted a thematic analysis of the qualitative feedback provided by the experts in response to Question 5. Thematic analysis revealed that:

- OER descriptions were sometimes too brief or lacking detail, causing experts to consider them not reusable.
- The difficulty levels of the OER were often perceived as too basic or unclear based on the descriptions provided.
- The model sometimes struggled to differentiate between courses with similar titles but distinct learning objectives.

Summarizing the FAIR assessment phase, we conducted an expert-driven assessment to evaluate the attainment of our initial objectives. This evaluation allowed us to assess the effectiveness of our approach in enhancing the reusability of OER, and identify areas for potential refinement. The assessment highlighted the importance of aligning technical solutions with the needs and expectations of educators to ensure the successful adoption and integration of FAIR OER.

IV. DISCUSSION

This study addressed the need for enhanced reusability of OER by proposing and implementing an OER FAIRification workflow. Through a case study within the ESS domain, we demonstrated the practical application of this workflow, identifying key challenges and opportunities in making educational resources FAIR. While we focused on a particular domain, core aspects of our methodology and the lessons learned are readily transferable to other disciplines. Specifically, the overall workflow (Fig. 1) structure for identifying, describing, and exposing OER according to the FAIR principles can serve as a general model. The use of widely adopted, domain-agnostic vocabularies such as Schema.org for metadata annotation enhances cross-disciplinary interoperability (Table 2). Furthermore, the technical infrastructure employing a knowledge graph, SPARQL endpoint, and the data model for constructing interoperable learning plans (Fig. 4) provides a scalable and standardized method for resource access applicable in diverse fields. Employing language models to assess semantic similarity for mapping OER to study programs is, in itself, a powerful and generalizable technique. Furthermore, this work highlights that the effectiveness of the proposed approach can be further improved in specific domains by adopting the data-efficient fine-tuning concept developed in this study. Finally, many of the practical challenges we documented, such as ensuring metadata quality and interpreting FAIR principles, in particular reusability for educational resources, are shared across different educational domains. These transferable components and insights offer a foundation for other disciplines aiming to improve the FAIRness and utility of their OER collections. We now discuss our findings in relation to the research questions posed at the outset of this study and the challenges and limitations encountered throughout the process.

TABLE 5. Correlation and error metrics for similarity scores, performance comparison of the fine-tuned model against the baseline model, GPT-4o, and DeepSeek R1.

	Pearson Correlation	Spearman Correlation	Mean Squared Error	R-squared
Fine-Tuned Model	0.5977	0.5844	0.0631	0.3531
Baseline Model	0.5649	0.5959	0.0751	0.2302
GPT-4o	0.5957	0.6125	0.0654	0.3298
DeepSeek R1	0.5377	0.5550	0.0733	0.2494

RQ1: How can the FAIR principles be applied to enhance the reusability of OER?

Applying the FAIR guiding principles [17] effectively to OER presents unique considerations and challenges compared to traditional data FAIRification. Our FAIRification workflow (Fig. 1) offers a tailored implementation approach for OER, taking into account the specific nuances of educational content as digital objects. Table 6 provides a detailed summary of this application, outlining for each principle both the specific implementation approach and the distinct challenges encountered when working with educational resources.

While there are limitations to OER FAIRification, the effort toward making OER FAIR remains crucial. Creating effective educational content demands considerable time and specialized expertise, a fundamentally human endeavor often more complex than automatable data collection. Given this investment in creation, maximizing the discovery and reuse of existing OER holds great significance. Ultimately, our experience demonstrates that applying FAIR principles to OER is a feasible and necessary approach that provides a foundation for enhancing downstream discovery and integration, aligning with the broader recognition of FAIR principles as an effective guide for resource management aimed specifically at maximizing reuse.

RQ2: How can the discovery of OER relevant to specific educational contexts be improved?

The post-FAIRification phase (Section III-C) is crucial for realizing the full potential of OER. While FAIRification enhances resource findability through structured metadata and accessibility protocols, it does not inherently solve the practical challenge of efficiently identifying the most relevant resources for specific learning contexts. Users still face the time-consuming task of evaluating the suitability of potentially numerous OER for their specific learning objectives. The evaluation process is often complicated by the characteristics of OER metadata itself. First, OER metadata is descriptive, demanding time-consuming human interpretation rather than enabling quick, objective comparison. Second, comparing these often nuanced descriptions across multiple resources to gauge subtle differences in suitability remains inherently challenging. Furthermore, since individual resources may not fully cover all learning goals, users often need to perform this demanding evaluation process for multiple OER.

The challenge of interpreting and comparing descriptive OER metadata can now be partially addressed by leveraging the sophisticated capabilities of modern LLMs. State-of-the-art models available as of early 2025 possess advanced natural language understanding, holding the potential to support users by comparing resource descriptions and evaluating their alignment with learning objectives, offering a valuable enhancement over purely manual evaluation. Acknowledging that finding a perfect OER match for specific learning objectives is often unrealistic, such automated approaches primarily aim to suggest the best available fits and reduce discovery time. Building upon the potential of LLMs, our study investigated the added value of domain-specific fine-tuning, particularly using a data-efficient approach. While general LLMs provide considerable baseline assistance, this fine-tuning step was pursued to further enhance the domain-specific accuracy of OER recommendations. Our assessment results are promising, demonstrating that our fine-tuned model can effectively identify relevant OER, even with the challenges we faced, such as the limited scope of our training data and the absence of human-annotated labels. Notably, a smaller language model surpassed the performance of a more complex model, like GPT-4o, in identifying domain-specific needs, despite being fine-tuned on a small amount of data. Although this automated semantic similarity approach provides an efficient way to identify potentially relevant OER, expert evaluations are still essential for ensuring that the recommendations meet real-world educational needs.

RQ3: How can OER be automatically integrated into existing educational frameworks to facilitate their effective reuse?

While using LLMs to evaluate OER suitability addresses a key challenge, gathering and applying these resources in the teaching and learning process still requires effort. To fully realize OER's potential, they need to be integral to educational frameworks, such as study programs and course syllabi. This visible placement enables educators and learners to access relevant resources without the need to query and evaluate them, eventually encouraging their reuse.

Making individual educators responsible for achieving such integration is impractical due to the necessary technical skills and time investment required. Our study demonstrated a practical example of how technology can streamline this large-scale integration. By reconstructing university programs as an interconnected knowledge graph, which included OER suggestions linked to specific university courses, we presented a method for explicitly modeling relationships between formal study programs and relevant external resources (Fig. 4). A reliable assessment of semantic similarity is essential for creating such an integrated view, as the calculated scores provide relevance indicators between specific courses and suggested resources within the knowl-

TABLE 6. Overview of translating FAIR principles to (open) educational resources [17].

Principle	Action	Limitations
Principle F1: (meta)data are assigned a globally unique and persistent identifier	Assigning unique identifiers to each OER through FAIR-compliant infrastructure. Example: NFDI4Earth KnowledgeHub	Limited availability of FAIR-compliant infrastructure for OER, especially to ensure findability by the target audience.
Principle F2: data are described with rich metadata	Identifying and collecting essential metadata for describing educational resources. Example: use of Web and text scraping techniques for information collection	Variability in metadata quality and completeness across OER providers; absence of machine-readable metadata in many cases; limited use of education-specific metadata.
Principle F3: metadata clearly and explicitly includes the identifier of the data it describes	Linking metadata to corresponding OER using the `schema:url` field	Potential for broken links if ordinary OER URLs change; absence of persistent identifiers can reduce long-term reliability; need for automated URL checks and dataset updates.
Principle F4: (meta)data are registered or indexed in a searchable resource	Registering OER collection and their metadata in FAIR-compliant infrastructure. Example: NFDI4Earth KnowledgeHub	Limited availability of FAIR-compliant infrastructure for OER, especially to ensure findability by the target audience.
Principle A1: (meta)data are retrievable by their identifier using a standardized communication protocol	Allowing users to query and retrieve OER metadata. Example: Public SPARQL endpoind	Limited availability of persistent-identifier infrastructure for OER.
A1.1: the protocol is open, free, and universally implementable	Providing public API. Example: NFDI4Earth SPARQL endpoint	Potential language barriers (e.g., English proficiency may be required).
A1.2: the protocol allows for an authentication and authorization procedure, where necessary	Removing authentication requirements for OER or using infrastructure that does not require those	None identified.
Principle A2: metadata are accessible, even when the data are no longer available	Storing OER metadata separately from the OER in relevant databases	None identified.
Principle I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	Using RDF for metadata representation. Example: turtle syntax	Requires technical expertise; need for support infrastructure to facilitate broader adoption by non-technical users.
Principle I2: (meta)data use vocabularies that follow FAIR principles	Using `LearningResource` type from Schema.org , a FAIR vocabulary.	Limited uptake of `LearningResource` type by OER providers.
Principle I3: (meta)data include qualified references to other (meta)data	Using controlled vocabulary and identifiers to connect different types of resources. Example: use of `schema:coursePrerequisites` to connect OER with each other.	Potential for incomplete or inaccurate linking due to variations in metadata quality and identifier consistency.
Principle R1: (meta)data are richly described with a plurality of accurate and relevant attributes	Making the following metadata attributes mandatory: title, description, language, URL, and license.	Descriptions that include not only the course overview but also explicit learning objectives are needed.
R1.1: (meta)data are released with a clear and accessible data usage license	Ensuring license information is a mandatory metadata field.	License information can be difficult to locate for some OER.
R1.2: (meta)data are associated with detailed provenance	Collecting author and publisher information when available. Example: use of `schema:Person` and `schema:Organization` to connect resources with relevant entities.	Capturing complete provenance, such as the creation workflow, source datasets, and version history, can be difficult when historical records are missing or scattered.
R1.3: (meta)data meet domain-relevant community standards	Using domain-specific vocabularies, and tools. Example training LLM with domain-specific knowledge for integrating OER in different learning scenarios for enhanced reusabilty	No specific community standards currently exist for OER reuse; however, the community recognizes the need for further research and the exploration of technologies like AI.

edge graph, thereby informing adoption decisions. It also implies that, alongside FAIR OER, associated educational data such as study programs should move toward digitalization to ensure interoperability and integration and provide the flexibility needed to evolve and adopt the programs as fast as new resources emerge.

V. CHALLENGES AND LIMITATIONS

During the FAIRification process, we encountered several challenges (Table 6). In our efforts to collect and harmonize OER, we noticed a significant lack of public APIs that could facilitate the harvesting of OER metadata. This absence of standardized APIs necessitates the development of custom scraping and parsing techniques, increasing the complexity and effort required for data acquisition. Additionally, although specific metadata schemas, such as `schema:LearningResource`, are available, their adoption by OER publishers remains limited. We found that many OER lacked structured metadata; even when present, this metadata often did not utilize education-specific vocabularies and schemas. Such inconsistency hinders interoperability and complicates the automated discovery and integration of OER. While the FAIRification workflow itself requires technical expertise, wider adoption of public APIs and standardized metadata schemas by OER providers would significantly simplify its implementation. FAIRification of OER demands both domain expertise and technical knowledge of metadata schemas and ontologies to accurately capture and structure metadata. This dual requirement can be a significant hurdle, particularly for educators wishing to share their resources in a FAIR manner but potentially lacking the necessary technical background or dedicated time. Furthermore, successfully publishing and managing FAIR OER depends heavily on the availability of FAIR-compliant infrastructure tailored for educational content, which is currently limited or often domain-specific. Therefore, increased investment in such infrastructure is crucial to enable wider adoption. This includes developing user-friendly interfaces that simplify metadata creation and harmonizing existing OER repositories to enhance discoverability and interoperability.

Further, our FAIRification approach, like many FAIR implementation strategies, relies heavily on the availability of comprehensive and high-quality metadata. Critically, when metadata is sparse, inaccurate, or missing, users may be forced to examine the full resource content directly. Assessing suitability by reviewing lengthy texts or potentially hours of video dramatically increases the time required to evaluate a single resource. This underscores the crucial need to embed FAIR principles from the outset of OER development, ensuring that metadata creation is an integral part of the resource creation process. High-quality metadata, including clear learning objectives, licensing information, and detailed descriptions, significantly enhances the reusability and findability of OER.

Expert feedback also highlights areas for improvement and underscores the importance of human validation in the OER integration process. Challenges include aligning OER with broader or interdisciplinary courses, the need for more detailed OER descriptions, and the inclusion of clear learning objectives. Addressing these challenges will require ongoing research and development. Nevertheless, the findings presented in this study provide a strong foundation and highlight a viable path towards realizing the full potential of OER.

VI. CONCLUSION AND FUTURE WORK

This study contributes to a growing body of knowledge that seeks to promote the widespread adoption and effective reuse of OER. We presented a comprehensive workflow for the FAIRification of OER, demonstrating its practical application through a case study in ESS. Our approach encompassed data acquisition, metadata enhancement, semantic modeling, and knowledge graph construction to enhance the discoverability, accessibility, interoperability, and reusability of OER. By addressing the challenges of aligning OER with specific learning objectives, we bridged the gap between the technical implementation of FAIR principles and their practical value in real-world educational settings. Our findings highlight the importance of combining automated methods, such as language models for semantic similarity matching, with expert knowledge to ensure the quality and appropriateness of OER recommendations. The development of user-friendly tools and interfaces, coupled with increased investment in FAIR-compliant infrastructure, is crucial to empower educators and content creators to actively engage in the development of new OER and the reuse of existing ones.

Future work will focus on several key areas:

- Expanding the OER dataset: We aim to continually expand our FAIRified OER dataset through automated metadata harvesting and community contributions, ensuring comprehensive coverage of diverse ESS-related topics.
- Enhancing semantic similarity matching: We will continue to enhance the accuracy and robustness of our semantic similarity matching, using our curated dataset of human-annotated OER-course pairs. We will also explore methods for integrating domain-specific knowledge into language models, such as fine-tuning on specialized corpora. Additionally, we will investigate the potential of larger language models to improve the accuracy of similarity matching.
- Developing user-friendly tools: We plan to develop intuitive tools and interfaces to facilitate metadata creation for the ESS community, validation, and knowledge graph exploration for educators and OER creators.

Looking ahead, the insights gained from this assessment can inform the development of interactive tools and platforms that empower educators to effectively explore, evaluate, and integrate OER into their curricula.

DATA AVAILABILITY STATEMENT

The data and code supporting the findings of this study are publicly available at https://doi.org/10.5281/zenodo.16602457. The fine-tuned model weights are available at https://doi.org/10.57967/hf/6086

REFERENCES

[1] C. Goldberg, "Auditing classes at M.I.T., on the web and free," *New York Times*, 2001. Accessed: Oct. 21, 2024. [Online]. Available: https://www.nytimes.com/2001/04/04/us/auditing-classes-at-mit-on-the-web-and-free.html

[2] *Forum on the Impact of Open Courseware for Higher Education in Developing Countries*, UNESCO, Paris, France, 2002. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000128515?posInSet=1&queryId=d552ced6-e312-4fe9-a51b-135b8638df20

[3] *The 2019 UNESCO Recommendation on Open Educational Resources (OER): Supporting Universal Access to Information Through Quality Open Learning Materials*, UNESCO, Paris, France, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000383205.locale=en

[4] P. Ingavelez-Guerra, V. E. Robles-Bykbaev, A. Perez-Munoz, J. Hilera-Gonzalez, and S. Oton-Tortosa, "Automatic adaptation of open educational resources: An approach from a multilevel methodology based on students' preferences, educational special needs, artificial intelligence and accessibility metadata," *IEEE Access*, vol. 10, pp. 9703–9716, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9669174/

[5] H. M. Adil, S. Ali, M. Sultan, M. Ashiq, and M. Rafiq, "Open education resources' benefits and challenges in the academic world: A systematic review," *Global Knowl., Memory Commun.*, vol. 73, no. 3, pp. 274–291, Jul. 2022. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/GKMC-02-2022-0049/full/html

[6] S. Mishra, "Open educational resources: Removing barriers from within," *Distance Educ.*, vol. 38, no. 3, pp. 369–380, Sep. 2017. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01587919.2017.1369350

[7] J. Hylén, "Open educational resources: Opportunities and challenges," in *Proc. Open Educ.*, vol. 4963, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:277065215

[8] S. Koseoglu and A. Bozkurt, "An exploratory literature review on open educational practices," *Distance Educ.*, vol. 39, no. 4, pp. 441–461, Oct. 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01587919.2018.1520042

[9] L. J. Castro, P. M. Palagi, N. Beard, T. K. Attwood, and M. D. Brazas, "Bioschemas training profiles: A set of specifications for standardizing training information to facilitate the discovery of training programs and resources," *PLOS Comput. Biol.*, vol. 19, no. 6, Jun. 2023, Art. no. e1011120. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1011120

[10] T. Luo, K. Hostetler, C. Freeman, and J. Stefaniak, "The power of open: Benefits, barriers, and strategies for integration of open educational resources," *Open Learn., J. Open, Distance E-Learn.*, vol. 35, no. 2, pp. 140–158, May 2020. [Online]. Available: https://doi.org/10.1080/02680513.2019.1677222

[11] C. Limongelli, M. Lombardi, A. Marani, and D. Taibi, "A semantic approach to ranking techniques: Improving web page searches for educational purposes," *IEEE Access*, vol. 10, pp. 68885–68896, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9807311/

[12] T. Richter and M. McPherson, "Open educational resources: Education for the world?" *Distance Educ.*, vol. 33, no. 2, pp. 201–219, Aug. 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01587919.2012.692068

[13] S. Henderson and N. Ostashewski, "Barriers, incentives, and benefits of the open educational resources (OER) movement: An exploration into instructor perspectives," *First Monday*, vol. 23, no. 12, Dec. 2018. [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/9172

[14] A. Mikroyannidis and A. Papastilianou, "Open educational resources in public administration: A case study in Greece," *Open Learn., J. Open, Distance E-Learn.*, vol. 39, no. 3, pp. 226–240, Jul. 2024. [Online]. Available: https://doi.org/10.1080/02680513.2021.1950526

[15] J. Knox, "Five critiques of the open educational resources movement," *Teaching Higher Educ.*, vol. 18, no. 8, pp. 821–832, Nov. 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/13562517.2013.774354

[16] C. G. Duran and C. M. Ramirez, "Integration of open educational resources using semantic platform," *IEEE Access*, vol. 9, pp. 93079–93088, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9464963/

[17] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018. [Online]. Available: https://www.nature.com/articles/sdata201618

[18] P. Rocca-Serra et al., "The FAIR cookbook-the essential resource for and by FAIR doers," *Sci. Data*, vol. 10, no. 1, p. 292, May 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-02166-3

[19] L. Garcia et al., "Ten simple rules for making training materials FAIR," *PLOS Comput. Biol.*, vol. 16, no. 5, May 2020, Art. no. e1007854. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007854

[20] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan, "Research objects: Towards exchange and reuse of digital knowledge," *Nature Precedings*, vol. 2010, p. 1, Jul. 2010. [Online]. Available: Available: https://www.nature.com/articles/npre.2010.4626.1

[21] A. Jacobsen et al., "FAIR principles: Interpretations and implementation considerations," *Data Intell.*, vol. 2, nos. 1–2, pp. 10–29, Jan. 2019. [Online]. Available: https://direct.mit.edu/dint/article/2/1-2/10-29/10017

[22] A. Jacobsen, R. Kaliyaperumal, L. O. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, and M. Thompson, "A generic workflow for the data fairification process," *Data Intell.*, vol. 2, nos. 1–2, pp. 56–65, Jan. 2020. [Online]. Available: https://direct.mit.edu/dint/article/2/1-2/56-65/9988

[23] A. Lamprecht, L. García, M. Kuzak, C. Martínez-Ortiz, R. Arcila, E. M. de Pico, V. D. D. Angel, S. Van De Sandt, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J. L. Gelpí, N. C. Hong, C. Goble, and S. Capella-Gutiérrez, "Towards FAIR principles for research software," *Data Sci.*, vol. 3, no. 1, pp. 37–59, Jun. 2019. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/DS-190026

[24] M. Barker, N. P. C. Hong, D. S. Katz, A.-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos, J. Harrow, L. J. Castro, M. Gruenpeter, P. A. Martinez, and T. Honeyman, "Introducing the FAIR principles for research software," *Sci. Data*, vol. 9, no. 1, p. 622, Oct. 2022. [Online]. Available: https://www.nature.com/articles/s41597-022-01710-x

[25] A. Degbelo, "FAIR geovisualizations: Definitions, challenges, and the road ahead," *Int. J. Geographical Inf. Sci.*, vol. 36, no. 6, pp. 1059–1099, Jun. 2022. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/13658816.2021.1983579

[26] R. V. Guha, D. Brickley, and S. Macbeth, "Schema.org: Evolution of structured data on the web," *Commun. ACM*, vol. 13, no. 9, pp. 10–37, Jan. 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2844544

[27] L. E. Anido, M. J. Fernández, M. Caeiro, J. M. Santos, J. S. Rodríguez, and M. Llamas, "Educational metadata and brokerage for learning resources," *Comput. Educ.*, vol. 38, no. 4, pp. 351–374, May 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360131502000180

[28] A. Iliadis, A. Acker, W. Stevens, and S. B. Kavakli, "One schema to rule them all: How schema.org models the world of search," *J. Assoc. Inf. Sci. Technol.*, vol. 76, no. 2, pp. 460–523, Feb. 2023. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24744

[29] P. Barker and L. M. Campbell, "LRMI, learning resource metadata on the web," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, p. 687. [Online]. Available: https://dl.acm.org/doi/10.1145/2740908.2741745

[30] P. Ingavélez-Guerra, V. E. Robles-Bykbaev, A. Perez-Muñoz, J. Hilera-González, S. Otón-Tortosa, and E. Campo-Montalvo, "RALO: Accessible learning objects assessment ecosystem based on metadata analysis, inter-rater agreement, and Borda voting schemes," *IEEE Access*, vol. 11, pp. 8223–8239, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10007848/

[31] M. Tavakoli, M. Elias, G. Kismihók, and S. Auer, "Metadata analysis of open educational resources," in *Proc. 11th Int. Learn. Analytics Knowl. Conf.*, Apr. 2021, pp. 626–631. [Online]. Available: https://dl.acm.org/doi/10.1145/3448139.3448208

[32] M. Vai and K. Sosulski, *Essentials of Online Course Design: A Standards-Based Guide*. Evanston, IL, USA: Routledge, 2011.

[33] M. E. Ibrahim, Y. Yang, D. L. Ndzi, G. Yang, and M. Al-Maliki, "Ontology-based personalized course recommendation framework," *IEEE Access*, vol. 7, pp. 5180–5199, 2019.

[34] L. Patiny and G. Godin, "Automatic extraction of FAIR data from publications using LLM," *ChemRxiv*, 2023. [Online]. Available: https://chemrxiv.org/engage/chemrxiv/article-details/65570cb1dbd7c8b54b6ff36b

[35] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, "KnowEdu: A system to construct knowledge graph for education," *IEEE Access*, vol. 6, pp. 31553–31563, 2018.

[36] Z. Li, L. Cheng, C. Zhang, X. Zhu, and H. Zhao, "Multi-source education knowledge graph construction and fusion for college curricula," in *Proc. IEEE Int. Conf. Adv. Learn. Technol. (ICALT)*, Orem, UT, USA, Jul. 2023, pp. 359–363. [Online]. Available: https://ieeexplore.ieee.org/document/10260880/

[37] Y. Su and Y. Zhang, "Automatic construction of subject knowledge graph based on educational big data," in *Proc. 3rd Int. Conf. Big Data Educ.* NY, USA: ACM, May 2020, pp. 30–36. [Online]. Available: https://doi.org/10.1145/3396452.3396458

[38] M. A. M. Nieto, P. D. V. Mora, J. De La Calleja Mora, M. T. Vidal, E. L. Dominguez, D. A. Diaz, and I. E. B. Patino, "Web service to retrieve and semantically enrich datasets for theses from open educational repositories," *IEEE Access*, vol. 8, pp. 171933–171944, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9199864/

[39] EarthLab. (2022). *Earth Data Science: Free Online Courses, Tutorials and Tools*. [Online]. Available: https://www.earthdatascience.org/

[40] MIT OpenCourseWare. *MIT OpenCourseWare: Free Online Course Materials*. Accessed: Apr. 17, 2025. [Online]. Available: https://ocw.mit.edu/

[41] *IEEE Standard for Learning Object Metadata*, IEEE Standard Std 1484.12.1-2020, 2020, pp. 1–50.

[42] NFDI4Earth. *NFDI4Earth Knowledge Hub: Discover Semantic Resources From the Earth System Sciences*. Accessed: Apr. 17, 2025. [Online]. Available: https://knowledgehub.nfdi4earth.de/

[43] L. Bernard, P. Braesicke, R. Bertelmann, S. Frickenhaus, H. Gödde, C. Keßler, S. Lorenz, M. Mahecha, H. Marschall, D. Hezel, W. E. Nagel, M. Reichstein, M. Sester, H. Thiemann, C. Weiland, and A. Wytzisk-Arens, "NFDI Consortium Earth System Sciences—Proposal 2020 revised," Zenodo, tex.copyright: Creative Commons Attribution 4.0 International, Open Access, Nov. 2021. [Online]. Available: https://zenodo.org/record/5718944

[44] L. Bernard, C. Henzen, A. Degbelo, D. Nüst, and J. Seegert, "NFDI4Earth: Improving research data management in the Earth system sciences," in *Proc. 1st Conf. Res. Data Infrastruct. (CoRDI)*, vol. 1, Karlsruhe, Germany, Y. Sure-Vetter and C. Goble, Eds., Sep. 2023. [Online]. Available: https://www.tib-op.org/ojs/index.php/CoRDI/article/view/288

[45] Massachusetts Institute of Technology, *Earth, Atmospheric, and Planetary Sciences (Course 12)—MIT Course Catalog*. Accessed: Jul. 17, 2024. [Online]. Available: https://catalog.mit.edu/subjects/12/

[46] R. Mihalcea, C. D. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. 21st Nat. Conf. Artif. Intell.*, Jun. 2006, pp. 775–780.

[47] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "SemEval-2012 task 6: A pilot on semantic textual similarity," in *Proc. 1st Joint Conf. Lexical Comput. Semantics*, vol. 1, 2012, pp. 385–393.

[48] NASA. *GCMD Keyword Viewer NASA Earthdata*. Accessed: Apr. 17, 2025. [Online]. Available: https://gcmd.earthdata.nasa.gov/KeywordView

[49] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3980–3990. [Online]. Available: https://www.aclweb.org/anthology/D19-1410

[50] Sentence Transformers. *Pretrained Models—Sentence Transformers Documentation*. Accessed: Apr. 17, 2025. [Online]. Available: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#original-models

[51] Tech. Univ. Munich. (Dec. 2021). *Module Catalog of the Study Program B.Sc. Geodesy and Geoinformation*. [Online]. Available: https://collab.dvb.bayern/download/attachments/73390039/Module-Catalog_BScGeodesyGeoinformation_EN_01122021.pdf

[52] Tech. Univ. Munich. (Oct. 2022). *Module Catalog of the Study Program M.Sc. Geodesy and Geoinformation. English Version*. [Online]. Available: https://collab.dvb.bayern/download/attachments/73389896/Module-Catalog_MScGeodesyGeoinformation_EN_19102022.pdf

[53] Tech. Univ. Munich. (Sep. 2019). *Module Handbook of the Study Program M.Sc. Earth Oriented Space Science and Technology. Winter Semester 2019/20*. [Online]. Available: https://collab.dvb.bayern/download/attachments/73389806/ESPACE_ModuleHandbook_WS2019.pdf

[54] Tech. Univ. Munich. (Aug. 2022). *Module Catalog of the Study Program M.Sc. Cartography*. [Online]. Available: https://cartographymaster.eu/wp-content/documents/CARTOGRAPHY_Module_Handbook.pdf

[55] Tech. Univ. Munich. (Dec. 2023). *Module Catalog of the Study Program M.Sc. Information Technologies for the Built Environment*. [Online]. Available: https://collab.dvb.bayern/download/attachments/73389824/MSc-ITBE_Module-Catalog__20231219.pdf

[56] MIT OpenCourseWare. (2006). *Modern Navigation—Earth, Atmospheric, and Planetary Sciences*. [Online]. Available: https://ocw.mit.edu/courses/12-215-modern-navigation-fall-2006/

FARZANEH SADEGHI is currently a Doctoral Researcher at Aalborg University, Denmark, and a Research Associate at Bochum University of Applied Sciences, Germany. Her research focuses on the intersection of educational technology, spatial data science, and research data management, with an emphasis on integrating semantic web technologies and AI to enhance spatial data science education. She also coordinates an educational portal that offers OER. Prior to her Ph.D., her research involved applying generative AI and statistical analysis to rainfall-runoff and drainage data, resulting in an open-source workflow to identify flood hotspots accurately.

AURIOL DEGBELO is currently a Postdoctoral Research Associate at TU Dresden, Germany. His research interests include semantic integration of geospatial information, re-use of open government data, and interaction with geographic information. His past contributions include a theory of spatial and temporal resolution of sensor observations, mathematical models of map user experience, and a semi-automatic approach for the creation of thematic web maps.

CARSTEN KESSLER is currently a Professor of geographic information systems and spatial data analysis at Bochum University of Applied Sciences, Germany. His research interests include spatio-temporal analyses (e.g., for climate impact research), research data infrastructures, data protection and privacy in the context of geodata, and the semantic interoperability of geographic information. He is actively involved in several research projects, focusing on topics, such as digital public participation, research data infrastructures for Earth system sciences, and open educational resources for spatial data infrastructures.

REZA ZOLNOURI is currently a Doctoral Researcher at RWTH Aachen University, Germany. Before his Ph.D., he worked on mirror descent methods for policy optimization in reinforcement learning and applied machine learning, modeling, and statistical analysis to improve virtual reality through electrical muscle stimulation. His research interests include reinforcement learning from human feedback, mean field games, and transformer-based models, focusing on the theoretical and algorithmic foundations of decision-making in multi-agent systems.