



Electronic Communications of the EASST
Volume 86 (1) Year 2026

deRSE26 - Selected Contributions of the 6th Conference for Research Software Engineering in Germany (Part 1)

Edited by Bernd Flemisch, René Caspart, Bernadette Fritzsch, Florian Goth, Jean-Noël Grad, Florian Thiery

Metadata for DiscoRSE: An Open Educational Resource Catalogue for Research Software Engineers

Dhwani Solanki, Suhasini Venkatesh, Eckhard Kadasch, Christian Knüpfer, Bernd Müller, Philipp Schäfer, Florian Goth, Frank Löffler, Leyla Jael Castro

DOI: 10.14279/eceasst.v86.1.2788

License: This article is licensed under a CC-BY 4.0 License.

Electronic Communications of the EASST (https://eceasst.org).
Published by **Berlin Universities Publishing**
(https://www.berlin-universities-publishing.de/)

Metadata for DiscoRSE: An Open Educational Resource Catalogue for Research Software Engineers

Dhwani Solanki [1], **Suhasini Venkatesh** [1], **Eckhard Kadasch** [2], **Christian Knüpfer** [2], **Bernd Müller** [1], **Philipp Schäfer** [2], **Florian Goth** [3], **Frank Löffler** [2,4], **and Leyla Jael Castro*** [1]

[1] ZB MED - Information Centre for Life Sciences, Germany
[2] Friedrich Schiller University Jena, Germany
[3] Julius-Maximilians-Universität Würzburg, Germany
[4] de-RSE e.V. — Society for Research Software in Germany

Abstract: Research Software Engineers (RSEs) work at the intersection of research practice, software engineering, and domain-specific knowledge — a combination that few people enter the field with. The RSE community has responded to this skills gap by producing a large body of Open Educational Resources (OERs), but these resources are often scattered across institutional websites, poorly described, and hard to reuse beyond the institutions that created them. DiscoRSE addresses this gap by aggregating, curating, and enriching OERs for RSEs from multiple sources, with a shared metadata schema at its core to make them more findable, accessible, interoperable, and reusable (FAIR). In this paper, we report on the metadata work for DiscoRSE: using `schema:LearningResource` as the reference, we performed a property-level crosswalk against five existing schemas for OERs and training materials — the Bioschemas `TrainingMaterial` profile, the Allgemeines Metadatenprofil für Bildungsressourcen (AMB), the DINI/nestor Metadata Schema for RDM Training Materials, Skills4EOSC, and the MoDALIA ontology used in DALIA. The analysis identifies a small core of properties consistently supported across all five targets, a wider set supported by four targets, five candidate additions not yet covered by `LearningResource`, and a set of open discussion points for the community. On this basis we propose an initial core profile for DiscoRSE and outline next steps, i. e., ETL pipelines, AI-assisted metadata generation, and selection of an OER platform.

Keywords: Research Software Engineering (RSE), Open Educational Resources (OER), metadata, training materials, schema.org, FAIR, DiscoRSE.

1 Introduction

Research software is now an indispensable part of the research process in all disciplines [G+24]. The people who design, develop, maintain, and document such software are increasingly recognised as a distinct community of Research Software Engineers (RSEs) [CAHJ25, B+22]. RSE sits at the intersection of three domains: the research questions of a specific scientific field, software engineering practice, and the technical tooling and culture of the broader research data ecosystem. Few people enter the field with a balanced skill set across all three, most arrive either as discipline-specific researchers who have accumulated programming experience, or as software engineers who join a research environment. To fill this gap, as a starting point, high-quality, openly available training materials are required [G+24].

The RSE community has responded with a large and growing body of open educational resources (OERs): tutorials, workshops, slide decks, lecture series, and self-paced courses, often shared in individual Git repositories or institutional websites. While the supply of materials is healthy, their discoverability is not. OERs created in one institution rarely surface in searches performed elsewhere, are described using ad-hoc terminology, and frequently lack structured metadata that would allow them to be aggregated, filtered, or recommended automatically. The result is that resources exist but cannot be reliably found, evaluated, or reused: they are not FAIR [KKV+25, VMS+26].

DiscoRSE is an initiative that aims to provide a discovery layer for RSE-oriented OERs through a metadata layer. The main goal is to aggregate existing training materials from multiple resources and enrich them with FAIR metadata [KKV+25]. The DiscoRSE schema takes reference points using `schema:LearningResource` and aligned metadata properties with existing OER such as the MoDALIA ontology [Dat], AMB [P+24], DINI/nestor [DIN23] Metadata Schema for RDM Training and other mentioned platforms. This schema is built by extending the Bioschemas: `TrainingMaterial` [CPB+23] and can be visualised in the Data Discovery Engine (DDE) platform [CTZ+23, DDP]. To encourage adoption, the project will provide lectures and workshops and integrate with the communication channels already established within the RSE community, initially within Germany and subsequently across international RSE communities [KKV+25].

This paper focuses on one specific aspect of the DiscoRSE roadmap: the design of a metadata schema for training materials. i) A structured comparison of five existing metadata schemas for educational and training resources, anchored against `schema:LearningResource`; ii) a property-level crosswalk and an analysis of where the schemas converge and where they differ; iii) an initial proposal of core common properties for the DiscoRSE, together with candidate properties not yet covered by LearningResource, and iv) a discussion of practical challenges and the next steps planned for 2026.

2 Background and Related Work

2.1 Why metadata matters for training materials

OERs without structured metadata are hard to integrate into any kind of automated workflow. Even when a resource is openly licensed and technically accessible, the absence of standard descriptions for its author, language, target audience, prerequisites, learning outcomes, or licence forces every potential user to inspect the resource manually before deciding whether it is relevant. For learners, this translates into wasted time searching across institutional sites; for trainers, it means duplicated effort to recreate materials that already exist somewhere else; for funders and infrastructure providers, it means difficulty assessing coverage and quality at scale. Structured FAIR metadata are therefore a prerequisite for building any kind of meaningful discovery service [C+20].

2.2 RSE-relevant OER platforms

Several platforms already aggregate educational resources of interest to RSEs, but none of them is tailored to the RSE community specifically. WirLernenOnline [Wir] aggregates German-language OERs across school and higher-education contexts, drawing from a wide range of public and institutional sources and exposing them through a faceted search interface. The Open Educational Resources Search Index (OERSI) [iHE] is the most directly comparable project to DiscoRSE in design and philosophy. Provided jointly since 2020 by the TIB — Leibniz Information Centre for Science and Technology and the hbz (University Library Center of North Rhine-Westphalia), OERSI is a central search entry point for OERs in higher education. It does not host resources itself, but harvests their metadata from distributed national

and international providers and homogenises them into a common schema — the Allgemeines Metadatenprofil für Bildungsressourcen (AMB). OERSI exposes an open API and downloadable metadata dumps, and all development is openly documented on GitLab. For DiscoRSE, OERSI is therefore both an architectural reference as its metadata-driven aggregation pattern is exactly the model DiscoRSE adopts. Where DiscoRSE differs, is in scope. OERSI is general higher education with broad facets (language, learning resource type, audience, license), whereas DiscoRSE is RSE-specific and needs finer-grained axes such as programming language, software ecosystem, and prerequisite skills.

In adjacent domains, ELIXIR's TeSS [B+20] provides a richer experience for the bioinformatics training community, demonstrating the value of a domain-specific aggregator with a tailored metadata profile (Bioschemas'sTrainingMaterial). The cross-domain platform DALIA [Dat] focuses on training addressing research data competencies and uses a tailored metadata model (the MoDALIA ontology and its subset the DALIA Interchange Format [GSDL24]). Skills4EOSC [Ski] takes a complementary approach. As part of the European Open Science Commons effort, it focuses on training materials and learning paths for open science skills, including research data management, FAIR principles, and computational skills relevant to RSEs. These platforms show how a structured metadata schema together with a domain-aware user interface provide a user-friendly discovery experience.

DiscoRSE positions itself in this landscape as an RSE-focused aggregator that builds on existing platforms and schemas to improve the findability of RSE-related OERs on specific topics. While existing schemas such as Bioschemas `TrainingMaterial`, AMB, and DINI/nestor provide a solid foundation for general OER description, none fully addresses RSE-specific needs such as linking to source repositories, describing the application domain, and capturing versioning information. DiscoRSE therefore proposes a targeted extension to `schema:LearningResource` that fills these gaps while remaining interoperable with existing schemas.

2.3 Metadata schemas considered

For the schema analysis, we selected sources that are either widely used in the German and European OER landscape, or that are explicitly aligned with schema.org and therefore directly comparable. The reference point is the schema.org type `LearningResource` [Sch], which is itself a specialisation of `CreativeWork` and thus of `Thing`. Choosing schema.org as the underlying schema for DiscoRSE is a pragmatic decision as schema.org is easy to embed in web content for search-engine indexing, and four of the five other schemas considered are explicitly aligned to it. In addition, schema.org `LearningResource` [Sch] includes properties that represent learning, teaching, skill development, and other educational purposes. It consists of properties such as *teaches*, *assesses*, *educationalLevel*, *competencyRequired*, and inherited properties from `CreativeWork` (e.g. *author*, *license*, *inLanguage*). Against this reference, we compared five existing schemas:

- **Bioschemas TrainingMaterial profile** [CPB+23] — an extension of `LearningResource`, developed by the Bioschemas community for life science training, used in production by TeSS.
- **MoDALIA ontology** [GSDL24] — underpinning the DALIA Interchange Format [Dat], a metadata profile aligned with schema.org and developed within the DALIA project for cross-domain, research data competence training in German.
- **Skills4EOSC** [FGM+23] — aims to create shared methodologies, coordinated activities, and training resources that bring together the existing training landscape into a collaborative and trustworthy ecosystem, while offering community-specific support to help researchers make effective use of EOSC for open and data-intensive research. We have referred their FAIR-by-Design Methodology stages for creating FAIR training materials for our crosswalk mapping.

- **DINI/nestor Metadata Schema for RDM Training Materials** [DIN23] — a schema developed by the community and specifically targeting research data management training, with strong attention to learning context and assessment.
- **Allgemeines Metadatenprofil für Bildungsressourcen (AMB)** [P+24] — a general metadata profile for educational resources mainly built on schema.org and the Learning Resource Metadata Initiative (LRMI), with selected elements from the Simple Knowledge Organization System (SKOS). It is designed for JSON-LD, following a LOUD (Linked Open Usable Data) approach.

3 Methodology

We approached schema alignment as a structured crosswalking exercise, where a crosswalk is a table where each row corresponds to a property in a reference source. The motivation for this choice is discussed in Section 5: existing automated tools tend to align labels syntactically, but the semantic differences between OER schemas often require human interpretation. Our methodology has three phases: schema collection, normalisation, and property mapping against the metadata profile based on schema: `schema:LearningResource` [Sch], and consensus building through community review. The crosswalks used in this process were created during hackathons hosted at ZB MED — Information Centre for Life Sciences in June 2025 and at the BioHackathon Germany in December 2025 [SVF+26]. A hackathon corresponds to a highly participative and collaborative unconference where participants work together towards a common goal, in this case the creation of crosswalks between a selected group of resources related to training materials and the reference point schema: `schema:LearningResource`.

3.1 Schema collection and analysis

For each property of `schema:LearningResource` and its parent types, we collected the official description from schema.org, and then, for each of the five target schemas, manually inspected the published specification to identify whether a corresponding property exists. During the aforementioned hackathons, participants manually extracted properties from the latest published version of each schema and recorded the location of every property within the source document so that every entry in the crosswalk remains traceable to its origin. No automated schema-matching tool was used. The analysis was carried out manually because the source schemas differ substantially in format and level of formality, which prevents a single automated extraction from working uniformly across all of them. Each mapping between the reference point and the target source is collected in the form of a crosswalk.

3.2 Property mapping against the reference

For each property in `schema:LearningResource`, we recorded, per target schema, the following information whenever a mapping was identified: the mapping type, following the recommendations by SSSOM (same as, exact match, close match, narrower scope, broader scope, or related match), the corresponding property name in the target schema, its cardinality and requirement level in that target, any associated controlled vocabulary, and a free-text comment capturing remaining caveats. For target schemas that introduce properties not present in `schema:LearningResource` (or its parents), we captured these separately as candidate additions, together with a proposed action of whether to include them in the resulting profile.

In total, 123 schema.org properties were analysed: 7 defined directly on `LearningResource`, 104 inherited from `CreativeWork`, and 12 inherited from `Thing`. The crosswalk is maintained as a structured spreadsheet with one row per schema.org property and dedicated column groups per target schema. The mapping was first drafted by a team of participants in the hackathon on a voluntary basis

and then reviewed by part of the DiscoRSE team. The spreadsheet also records the contributors and the date of the latest revision for provenance.

3.3 Consensus building

Building consensus across schemas required iterative discussion rather than a single mechanical step. For each property we asked three questions: i) is the property meaningful in the context of an RSE-oriented OER catalogue, ii) is it supported, with the same intent, by multiple source schemas, and iii) what level of marginality should the DiscoRSE schema attach to it (mandatory, recommended or optional). Mandatory properties are intended to be enforced for any record entering the DiscoRSE catalogue. Recommended properties are encouraged and exposed in the user interface and optional properties are accepted when present but not actively prompted. It is important to note that this process did not aim to produce an entirely new schema. Rather, the goal was to identify a minimal set of additions to `schema:LearningResource` that address the specific needs of the RSE community, grounded in feedback gathered at the hackathons described in Section 3.

4 Results

4.1 Common properties across schemas

Anchoring on `schema:LearningResource` and its ancestors, the crosswalk identifies three groups of properties: those defined directly on `LearningResource`, those inherited from `CreativeWork`, and those inherited from `Thing`. Within each group we counted, for every property, how many of the five target schemas provide a corresponding mapping.

Across the 123 properties analysed, 8 are supported by all five targets, a further 8 are supported by four of five, and 8 more are supported by three of five. The remaining properties are either supported by one or two targets or by none. The full counts are summarised in Table 1 and Table 2 lists the properties at each coverage level.

Table 1: Coverage of `schema:LearningResource` properties (and inherited properties from `CreativeWork` and `Thing`) across the five target schemas analysed. Counts are derived from the crosswalk spreadsheet [SVF+26].

Coverage	LearningResource (7)	CreativeWork (104)	Thing (12)
Supported by 5/5 targets	2	5	1
Supported by 4/5 targets	2	3	3
Supported by 3/5 targets	0	8	0
Supported by 1–2/5 targets	1	26	3
Not mapped by any target	2	62	5
Total	7	104	12

4.2 Core common profile

Combining the property frequencies above with the requirements identified for the DiscoRSE use case, we propose an initial core profile. Table 3 summarises the candidate properties together with their proposed marginality level. The selection prioritises properties that i) are supported by at least four of the

Table 2: Properties mapped by at least three of the five target schemas, grouped by coverage level. Properties marked with their schema.org domain in parentheses.

Coverage	Properties
5/5 targets	*learningResourceType*, *teaches* (`LearningResource`); *audience*, *inLanguage*, *keywords*, *license*, *timeRequired* (`CreativeWork`); *identifier* (`Thing`)
4/5 targets	*competencyRequired*, *educationalLevel* (`LearningResource`); *about*, *author*, *contributor* (`CreativeWork`); *description*, *name*, *url* (`Thing`)
3/5 targets	*abstract*, *dateCreated*, *dateModified*, *datePublished*, *hasPart*, *isAccessibleForFree*, *isPartOf*, *mentions* (`CreativeWork`)

five analysed schemas and ii) directly contribute to one of DiscoRSE's functional goals: identification, discoverability, audience matching, accessibility, and learning-chain navigation.

Table 3: Initial DiscoRSE core common profile derived from the 5-source crosswalk.

Property	Role	Marginality
name	Human-readable title of the resource.	Mandatory
identifier / url	Persistent or stable identifier (DOI, ORCID-style ID, or URL).	Mandatory
description	Free-text description sufficient for a learner to assess relevance.	Mandatory
license	Reuse conditions; SPDX-style identifier preferred.	Mandatory
inLanguage	BCP 47 language tag.	Mandatory
author	Person or organisation responsible for the resource.	Recommended
datePublished	Date of first publication.	Recommended
audience	Intended audience (e.g. RSE, researcher, student).	Recommended
educationalLevel	Beginner / intermediate / advanced (controlled vocabulary).	Recommended
learningResourceType	Format (presentation, hands-on tutorial, video, notebook, etc.).	Recommended
teaches	Skills or knowledge taught by the resource.	Recommended
competencyRequired	Prerequisite skills or knowledge.	Recommended
keywords	Free or controlled keywords supporting full-text search.	Recommended
about	Subject of the resource (linked to a controlled vocabulary where possible).	Optional
version	Version of the material (especially for materials hosted in a VCS).	Optional
assesses / educationalUse	Learning outcome measured / intended educational use.	Optional

The full marginality levels in Table 3 should be read as a starting point for community discussion, not as a final specification. In particular, the boundary between *recommended* and *minimum* for prop-

erties such as *author* and *datePublished* is informed both by the analysis (4/5 support) and by practical considerations: many materials hosted on GitHub do not have these properties recorded explicitly in machine-readable form, and demanding them as mandatory at the input layer would currently exclude a substantial share of community resources. For such properties, our intention is to enforce them at the catalogue level by deriving them with extraction pipelines in Section 6 when not provided directly. Note that *datePublished* and *version* should be advanced together whenever there is a dependency between them. However, since a training resource can be published without a version, and this still commonly happens, *version* remains optional.

4.3 Properties proposed beyond `schema:LearningResource`

In addition to the properties already available through `schema:LearningResource` and its ancestors, the target schemas exposed five recurring needs that are not currently covered. Based on the crosswalk we propose these as candidate additions to the Bioschemas `TrainingMaterial` profile, and therefore to DiscoRSE:

- **applicationDomain** (range: `DefinedTerm` or `Text`) — the discipline, area, or research domain to which the training material belongs. Useful for cross-disciplinary discovery and also relevant for adjacent research artefacts such as research software.
- **contactPoint** (range: `ContactPoint`) — contact information for inquiries about the material, mirroring the corresponding `schema:contactPoint` property used elsewhere in schema.org.
- **sourceRepository** (range: `URL`) — a link to the repository used to develop and host the material, recognising that training materials are increasingly maintained on platforms such as GitHub or GitLab.
- **previousVersion** and **nextVersion** (range:`CreativeWork`) — immediate predecessor and successor of the resource, supporting versioning and learning-chain navigation. The two are symmetric, so one can be inferred from the other.

These properties are derived from the analysis but are not yet part of any standardised profile. So we will bring them forward to the Bioschemas community for further discussion. These candidate additions are directly motivated by RSE community practices, reflecting the common use of Git-based platforms, the cross-disciplinary nature of RSE work, and the versioning culture inherited from software engineering.

4.4 Bioschemas draft proposal

Based on the discussions and analysis of properties, we defined i) a new type proposal that extends `schema:CreativeWork`, corresponding to the properties beyond `schema:LearningResource` (see Subsection 4.3) and ii) an updated profile draft for `bioschemas:TrainingMaterial` (Table 4), corresponding to commonalities found across the 5-source crosswalks (see Subsection 4.1 and Subsection 4.2). We opened two pull requests on Bioschemas Specifications: PR720 for the type and PR721 for the profile. To make future references to our proposal, we also published the corresponding files [SVC26].

Table 4: Bioschemas `TrainingMaterial` profile draft proposal with properties grouped by marginality level, including expected types, descriptions, and cardinality.

Note: Descriptions are taken directly from schema.org and Bioschemas specifications.

Property	Expected Type	Description	Cardinality
Marginality:	Minimum		
description	Text	A description of the item.	ONE
keywords	DefinedTerm or Text or URL	Keywords or tags used to describe this content. Multiple entries in a keywords list are typically delimited by commas.	MANY
name	Text	The name of the item.	ONE
Marginality:	Recommended		
about	Thing	The subject of this Training Material. Use the DefinedTerm type to add a controlled vocabulary term to describe the topic (such as from the EDAM ontology) The subject matter of the content. Inverse property: subjectOf.	MANY
applicationDomain	DefinedTerm or Text	The discipline, area, or research domain to which this item aligns or belongs to.	MANY
audience	Audience	A succinct description of the intended target audience for your materials: e.g., graduates, postgraduates, clinicians. An intended audience, i.e., a group for whom something was created. Supersedes serviceAudience.	MANY
author	Organization Person	Those involved in the preparation, creation and/or presentation of the published work, specifically writing the initial draft The author of this content or rating. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.	MANY
competencyRequired	DefinedTerm or Text or URL	Knowledge, skill, ability or personal attribute that must be demonstrated by a person or other entity in order to do something such as earn an Educational Occupational Credential or understand a LearningResource.	MANY
educationalLevel	DefinedTerm or Text or URL	The students level of ability in the topic being taught. Examples of skill levels include 'beginner', 'intermediate' or 'advanced'. The level in terms of progression through an educational or training context. Examples of educational levels include 'beginner', 'intermediate' or 'advanced', and formal sets of level indicators.	ONE
identifier	PropertyValue or Text or URL	An identifier for this resource such as a DOI or compact URI The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.	MANY
inLanguage	Language or Text	Defaults to English if not specified. Please choose a value from IETF BCP 47 standard. You can add multiple languages if the Training Material offers different translations The language of the content or performance or used in an action. Please use one of the language codes from the IETF BCP 47 standard. See also availableLanguage. Supersedes language.	MANY
learningResourceType	DefinedTerm or Text	This may include things such as video lecture, e-Learning module, or tutorial. The predominant type or kind characterising the learning resource. For example, 'presentation', 'handout'.	MANY
license	CreativeWork or URL	If there is a licence it must be added. A license document that applies to this content, typically indicated by URL.	MANY
mentions	Thing	Datasets, tools, technologies, entities etc, which are referred to by this training material or actively used in this training material. Indicates that the CreativeWork contains a reference to, but is not necessarily about a concept.	MANY

Property	Expected Type	Description	Cardinality
sameAs	URL	URL of a reference Web page that unambiguously indicates the item's identity. E.g., the URL of the item's Wikipedia page, Wikidata entry, or official website.	MANY
teaches	DefinedTerm or Text	The item being described is intended to help a person learn the competency or learning outcome defined by the referenced term.	MANY
timeRequired	Duration	The estimated time it takes to work through this resource. Please specify in ISO 8601 duration format. Approximate or typical time it takes to work with or through this learning resource for the typical intended target audience, e.g. 'PT30M', 'PT1H25M'.	ONE
url	URL	The preferred URL of the Training Material. You must provide this value if it is known. URL of the item.	ONE
Marginality:	Optional.		
abstract	Text	An abstract is a short description that summarises a CreativeWork.	ONE
accessibilitySummary	Text	A human-readable summary of specific accessibility features or deficiencies, consistent with the other accessibility metadata but expressing subtleties such as "short descriptions are present but long descriptions will be needed for non-visual users" or "short descriptions are present and no long descriptions are needed."	ONE
assesses	DefinedTerm or Text	The item being described is intended to assess the competency or learning outcome defined by the referenced term.	MANY
associatedMedia	MediaObject	A media object that encodes this CreativeWork. This property is a synonym for encoding.	MANY
conditionsOfAccess	Text	Conditions that affect the availability of, or method(s) of access to, an item. Typically used for real world items such as an ArchiveComponent held by an ArchiveOrganization. This property is not suitable for use as a general Web access control mechanism. It is expressed only in natural language. For example "Available by appointment from the Reading Room" or "Accessible only from logged-in accounts".	ONE
contactPoint	ContactPoint	Contact details for further information relevant to this item.	MANY
contributor	Organization or Person	Contributors are those that made non-authorship contributions, e.g., critical review, commentary or revision A secondary contributor to the CreativeWork or Event.	MANY
creativeWorkStatus	DefinedTerm or Text	The status of a training material. If this is not filled in it will be regarded as Active. Options are 'Active', 'Under development', and 'Archived'. The status of a creative work in terms of its stage in a life cycle. Example terms include Incomplete, Draft, Published, Obsolete. Some organisations define a set of terms for the stages of their publication life cycle.	ONE
creditText	Text	Text that can be used to credit person(s) and/or organisation(s) associated with a published Creative Work.	ONE
dateCreated	Date or DateTime	The date on which the CreativeWork was created or the item was added to a DataFeed.	ONE
dateModified	Date or DateTime	The date on which the CreativeWork was most recently modified or when the item's entry was modified within a DataFeed.	ONE
datePublished	Date or DateTime	Date of first broadcast/publication.	ONE
educationalUse	DefinedTerm or Text	The purpose of a work in the context of education; for example, "assignment", "group work".	MANY
funding	Grant	A Grant that directly or indirectly provide funding or sponsorship for this item. See also ownershipFundingInfo. Inverse property: fundedItem	MANY
hasPart	CreativeWork	A sub-training material or externally referenced training material Indicates an item or CreativeWork that is part of this item, or CreativeWork (in some sense). Inverse property: isPartOf.	MANY
isAccessibleForFree	Boolean	A flag to signal that the item, event, or place is accessible for free. Supersedes free.	ONE

Property	Expected Type	Description	Cardinality
isPartOf	CreativeWork or URL	The Course this Training Material was/will be used in. Or a training material this training material is a part of (for example, if this is a module in a book, isPartOf can describe the book). Inverse property: hasPart. If this varies in CourseInstances, use the workFeatured property Indicates an item or CreativeWork that this item, or CreativeWork (in some sense), is part of. Inverse property: hasPart.	MANY
nextVersion	CreativeWork	The immediate next version of this item	ONE
previousVersion	CreativeWork	The immediate previous version of this item	ONE
publisher	Organization or Person	The publisher of the article in question.	ONE
recordedAt	Event	The course instance or event where this training material was or will be featured. Use isPartOf to refer to a Course, unless this training material is unique to a specific Course Instance. The Event where the CreativeWork was recorded. The CreativeWork may capture all or part of the event. Inverse property: recordedIn.	MANY
sourceOrganization	Organization	The Organisation on whose behalf the creator was working.	MANY
sourceRepository	URL	Link to the repository used to host this item while in development. Similar to codeRepository but for all sort of creative works	ONE
version	Number or Text	If this training material is versioned, its strongly recommended you use this property to list the version being displayed The version of the CreativeWork embodied by a specified resource.	ONE
workTranslation	CreativeWork	A work that is a translation of the content of this work. Inverse property: translationOfWork.	MANY

4.5 Open discussion points

Several properties surfaced during the crosswalk where the analysis identified an issue but no consensus could be reached within the working group. We record these explicitly here because they are the natural starting point for the community workshop in the future (see Section 6).

- **about.** Currently used in Bioschemas, originally as a workaround when *keywords* did not yet accept `DefinedTerm` in schema.org. With *keywords* now accepting `DefinedTerm`, *about* becomes redundant. Following discussion in the Bioschemas community meetings, this property will be removed. The open question is migration cost: how many resources already rely on *about*?
- **Qualified references** (*citation*, *mentions*, `dc:requires` and similar). Bioschemas currently uses *mentions* as a catch-all for any other resource referenced by a training material. *citation* would be a more accurate match for references in the bibliographic sense, especially given that *creditText* now exists for self-citation. `dc:requires` captures hard dependencies (e. g., a dataset that must be used). The question is which qualified references should be introduced, and whether changing the existing *mentions* usage is feasible.
- **Notes for learners and trainers.** DINI, Skills4EOSC, and MoDALIA all include a property for additional notes from authors to learners or trainers. DINI mapped this to `schema:comment`, but schema.org's definition of *comment* is closer to user comments than to author notes. A dedicated property may be needed.
- **Copyright.** *copyrightHolder* is not used by any analysed schema, but *copyrightNotice* (which carries the copyright owner) is used by two. Whether a copyright property should be added, and which, remains open.

- **Community membership.** A way to link a training material to the community or consortium it belongs to (for example, a Zenodo community or an NFDI consortium) is wanted, but no agreed property exists yet.

5 Discussion: Challenges in Aligning OER Metadata Schemas

The crosswalk exercise made several recurring challenges visible. They are worth recording explicitly, both because they shape our roadmap and because they apply to most OER metadata-alignment work, not just to DiscoRSE.

Same words, different meanings. Even when two schemas use identical or near-identical property names, their intended semantics can differ in subtle but consequential ways. For example, *audience* is interpreted in some schemas as a free-text descriptor of the target reader, in others as a structured reference to a controlled vocabulary of educational levels, and in others as a combination of role and prior knowledge. The crosswalk format encourages curators to record these differences explicitly rather than collapsing them into a binary "match / no match", because the differences directly affect data normalisation downstream.

Granularity. Schemas differ in the granularity at which they describe a single concept. For example, "contact" is sometimes modelled as a single property containing free-text contact information, sometimes split into a contact person and a contact email, and sometimes nested inside an organisation object. Mapping a more granular schema onto a less granular one is straightforward (concatenate or pick one); the other way requires either lossy mapping or additional inference.

Manual crosswalking and scalability. The crosswalk was carried out manually due to differences in format and adoption while mapping with other resources. While manual mapping ensures semantic accuracy, it is time-consuming and does not scale easily to a larger number of schemas. As the number of OER schemas grows, maintaining and updating the crosswalk will require sustained community effort that should be combined with (semi)automatic approaches.

Provisional marginality levels. The marginality levels proposed in Table 3 and Table 4 represent a pragmatic starting point rather than a final specification. There is an inherent trade-off between strictness and inclusiveness. Enforcing too many mandatory properties would exclude a substantial share of existing community resources, particularly those hosted on GitHub without machine-readable metadata, while being too lenient risks reducing the overall quality and consistency of the catalogue.

Open issues in property selection. Beyond the alignment challenges described above, the crosswalk revealed several unresolved property-level issues. These include the redundancy between *about* and *keywords*, the use of *mentions* as a catch-all for referenced resources, the lack of a dedicated property for author notes to learners and trainers, and open questions around copyright representation and community membership. These issues are recorded as starting points for the community workshop planned in Section 6.

Controlled vocabulary. A recurring challenge identified during the crosswalk is the absence of a mandatory controlled vocabulary for the subject description. Without one, properties such as *keywords*, *about* and *mentions* remain loosely defined fields that limit precise filtering and discovery which is the core goal of DiscoRSE. Defining or adopting a suitable controlled vocabulary for RSE topics is therefore an important next step.

Community adoption. The practical value of the DiscoRSE schema depends on its adoption by the RSE community. The contribution of metadata is voluntary and therefore coverage may be uneven between institutional and research communities. To mitigate this, the AI-assisted metadata generation pipeline described in Subsection 6.2 aims to lower the barrier for contributors by automating the extraction of metadata. However, ensuring the quality and consistency of extracted metadata remains an open challenge and requires careful evaluation in future work.

The challenges and open issues identified above directly inform the roadmap described in Section 6, which outlines concrete next steps to address them.

6 Roadmap and Next Steps

Building on the identified challenges, the roadmap for DiscoRSE focuses on three priorities: refining the metadata schema through community evaluation, automating metadata extraction, and selecting an appropriate OER platform.

6.1 Reviewing and evaluating the metadata schema

The crosswalk and the proposed core profile are working artefacts. Two activities are planned to evaluate and refine them. First, a community workshop is planned in the future, with the explicit goals of i) gathering more feedback on the metadata schema, ii) testing and refining ETL pipelines, and iii) specifying the requirements of the OER platform. Second, the schema will be exercised against a small but representative set of real RSE training materials to surface any gaps that are not visible from the schema-only analysis.

6.2 Automated metadata extraction

It is not expected that the DiscoRSE catalogue will be primarily manually populated. Rather it will use the help of extract-transform-load (ETL) pipelines for several sources of training material, for instance:

- Training materials deposited in Zenodo, where structured DataCite-style metadata is already available.
- Bioschemas-annotated resources surfaced by the Glittr community [GHZP24], which already use a profile compatible with our reference schema.
- Push-based contributions from RSE training material hosted on GitHub or GitLab, where authors can opt in to have their repository indexed [VMS+26].

For each source, the pipeline will perform the same three steps: extracting metadata from the source, transforming it into the DiscoRSE core profile (using the crosswalk to drive the property mapping), and loading it into our DiscoRSE platform. The transformation step is the most complex; it is where the crosswalk decisions documented in Section 3 and Section 4 become operational rules.

For training materials without enough structured metadata at their source, a common situation for slide decks and lecture notes deposited directly on GitHub, we are exploring an AI-assisted generation step. We plan to use a large language model to use the raw content (e. g., slides, Markdown, Jupyter notebooks, HTML pages, video transcripts) to produce candidate values aligned to the metadata schema defined for DiscoRSE (at the moment corresponding to [SVC26]).

7 Conclusion

DiscoRSE addresses a concrete and well-known problem: open educational resources for Research Software Engineers exist in abundance but are difficult to discover, evaluate, and reuse. By anchoring the analysis on schema.org's `LearningResource` and crosswalking against five further schemas — DALIA Interchange Format, DINI/nestor RDM, AMB, skills4EOSC and Bioschemas TrainingMaterial — we have identified a small core of properties consistently supported across the OER metadata landscape, and proposed an initial core common profile for DiscoRSE that should be both practical to enforce and interoperable with existing platforms. Five additional candidate properties motivated by RSE community practices such as *applicationDomain*, *contactPoint*, *sourceRepository*, *previousVersion*, and *nextVersion* have been submitted as proposals to the Bioschemas community for further discussion. The challenges identified in Section 5, including semantic heterogeneity, granularity differences, and the absence of a mandatory controlled vocabulary, highlight that schema alignment is an ongoing community effort rather than a one-time exercise. The roadmap and next step addresses these challenges through three concrete next steps: community workshops to evaluate and refine the schema, ETL pipelines to automate metadata extraction, and AI-assisted metadata generation for resources that lacks structured metadata. Together, these steps lay the foundation for a findable, accessible, interoperable, and reusable catalogue of RSE-oriented training materials.

Acknowledgements:
This work has been partially funded by the Klaus Tschira Stiftung, as part of the support programme Scientific Software, under the scope of the project DiscoRSE (00.027.2024). FG acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the CRC 1170 "Tocotronics", project Z03, project number 258499086. The authors acknowledge participants of the hackathon organised by ZB MED Information Centre for Life Sciences on June 2025, supported by the NFDI4DataScience (part of the National Research Data Infrastructure -NFDI), Quadriga Competence Center, and DiscoRSE, and the participants in the hackathon project "On the path to machine-actionable training materials" part of the German BioHackathon on December 2025 and supported by ELIXIR Germany and deNBI. Crosswalks described in Section 3 were created as part of the activities of the mentioned hackathons.

Declaration on the use of AI: The authors used AI-based writing assistance tools to paraphrase and improve the readability and clarity of selected parts of the manuscript. The tools were used only for language refinement and did not contribute to the conceptual development, analysis, interpretation of results, or scientific conclusions of the paper. The authors reviewed and edited all AI-assisted text and take full responsibility for the final content of the manuscript.

Data availability: Crosswalks are available at https://doi.org/10.5281/zenodo.20190406. The type and profile draft proposal is available at https://doi.org/10.5281/zenodo.20188548.

References

[B+20] N. Beard et al. TeSS: a platform for discovering life-science training opportunities. *Bioinformatics* 36(10):3290–3291, 2020.
doi:10.1093/bioinformatics/btaa047

[B+22] M. Barker et al. Introducing the FAIR Principles for research software. *Scientific Data* 9, 2022. doi:10.1038/s41597-022-01710-x

[C+20] L. J. Castro et al. Ten simple rules for making training materials FAIR. *PLOS Computational Biology* 16(5):e1007854, 2020. doi:10.1371/journal.pcbi.1007854

[CAHJ25] N. P. Chue Hong, S. Aragon, S. Hettrick, C. Jay. The future of research software is the future of research. *Patterns* 6(7):101322, 2025. doi:10.1016/j.patter.2025.101322

[CPB+23] L. J. Castro, P. M. Palagi, N. Beard, T. K. Attwood, M. D. Brazas. Bioschemas training profiles: A set of specifications for standardizing training information to facilitate the discovery of training programs and resources. *PLOS Computational Biology* 19(6):1–14, 06 2023. doi:10.1371/journal.pcbi.1011120

[CTZ+23] M. A. Cano, G. Tsueng, X. Zhou, J. Xin, L. D. Hughes, J. L. Mullen, A. I. Su, C. Wu. Schema Playground: a tool for authoring, extending, and using metadata schemas to improve FAIRness of biomedical data. *BMC Bioinformatics* 24(1), Apr. 2023. doi:10.1186/s12859-023-05258-4

[Dat] Data Literacy Alliance. DALIA – Data Literacy Alliance portal. https://dalia.education/. Accessed: 2026-07-03.

[DIN23] DINI/nestor working group. DINI/nestor Metadata Schema for RDM Training Materials. https://dini.de/, 2023. Accessed: 2026-07-03.

[FGM+23] S. Filiposka, D. Green, A. Mishev, V. Kjorveziroski, A. Corleto, E. Napolitano, G. Paolini, S. Di Giorgio, J. Janik, L. Schirru, A. Gingold, C. Hadrossek, I. Souyioultzoglou, C. Leister, G. Pavone, S. Sharma, E. M. Mendez Rodriguez, E. Lazzeri. D2.2 Methodology for FAIR-by-Design Training Materials. Aug. 2023. doi:10.5281/zenodo.8305540

[G+24] F. Goth et al. Foundational Competencies and Responsibilities of a Research Software Engineer. *F1000Research* 13:1429, 2024. doi:10.12688/f1000research.157778.1

[GHZP24] G. van Geest, Y. Haefliger, M. Zahn-Zabal, P. M. Palagi. Using Glittr.org to find, compare and re-use online materials for training and education. *PLOS ONE* 19(12):1–11, 12 2024. doi:10.1371/journal.pone.0308729

[GSDL24] J. Geiger, P. Steiner, A. A. Desouki, F. Lange. DALIA Interchange Format. 2024. doi:10.5281/zenodo.11521029

[iHE] O. E. R. in Higher Education. Open Educational Resources Search Index (OERSI). https://oersi.org/resources. Accessed: 2026-07-03.

[KKV+25] E. Kadasch, C. Knüpfer, S. Venkatesh, L. J. Castro, F. Goth, P. Schäfer, F. Löffler. DiscoRSE: an open educational resource for Research Software Engineers. Conference on Research Data Infrastructure (CoRDI 2025), 2025. doi:10.5281/zenodo.16846029

[P+24] A. Pohl et al. Allgemeines Metadatenprofil für Bildungsressourcen (AMB). https://dini-ag-kim.github.io/amb/draft/, 2024. Accessed: 2026-07-03.

[DDP] T. Data Discovery Project. Data Discovery Engine. https://discovery.biothings.io/. Accessed: 2026-07-03.

[Sch] Schema.org. schema.org LearningResource. https://schema.org/LearningResource. Accessed: 2026-07-03.

[Ski] Skills4EOSC. Skills4EOSC – Skills for the European Open Science Commons. https://www.skills4eosc.eu. Accessed: 2026-07-03.

[SVC26] D. Solanki, S. Venkatesh, L. J. Castro. Draft proposal on TrainingMaterial profile to Bioschemas. May 2026.
doi:10.5281/zenodo.20188548

[SVF+26] D. Solanki, S. Venkatesh, S. Filiposka, H. Schanaitter, N. Juty, A. Pohl, E. Samoilova, H. Schnitzer, P. Steiner, M. Voigt, L. J. Castro. Metadata Crosswalks on schema:LearningResource. May 2026.
doi:10.5281/zenodo.20190407

[VMS+26] S. Venkatesh, B. Müller, D. Solanki, E. Kadasch, C. Knüpfer, P. Schäfer, F. Goth, F. Löffler, L. J. Castro. Metadata for DiscoRSE, an Open Educational Resource for RSEs. Conference for Research Software Engineering in Germany (deRSE26), Stuttgart, 2026.
doi:10.5281/zenodo.18836950

[Wir] WirLernenOnline. WirLernenOnline. https://wirlernen.online/. Accessed: 2026-07-03.