
| | |
|---|---|
| Project Title | Fostering FAIR Data Practices in Europe |
| Project Acronym | FAIRsFAIR |
| Grant Agreement No | 831558 |
| Instrument | H2020-INFRAEOSC-2018-4 |
| Topic | INFRAEOSC-05-2018-2019 Support to the EOSC Governance |
| Start Date of Project | 1st March 2019 |
| Duration of Project | 36 months |
| Project Website | www.fairsfair.eu |

# Harmonizing Metadata for Exchange of FAIR Training Materials

| | |
|---|---|
| Work Package | WP3/6 - Harmonizing Metadata for Exchange of FAIR Training Materials |
| Lead Author (Org) | Phil Barker, CETIS LLP |
| Contributing Author(s) (Org) | Angus Whyte (DCC) |
| Due Date | Not applicable (Informal report) |

| | |
|---|---|
| Date | 21.12.2020 |
| Version | 1.0 |
| DOI | https://doi.org/10.5281/zenodo.4382677 |

Dissemination Level

| | |
|---|---|
| X | PU: Public |
| | PP: Restricted to other programme participants (including the Commission) |
| | RE: Restricted to a group specified by the consortium (including the Commission) |
| | CO: Confidential, only for members of the consortium (including the Commission) |


## Abstract

This is an informal (unscheduled) FAIRsFAIR report commissioned by DCC and written by Phil Barker of CETIS LLP. The focus of the report is on options for deploying standards established for OER (Open Educational Resources). It aims to help DCC and the FAIRsFAIR project consider options for creating a minimal metadata schema that may be used on a trial basis in their work to establish a shared catalogue, in collaboration with various providers of training materials. The context is a trial exchange of metadata between the FAIRsFAIR project and catalogues provided by related projects. These projects have shared interests in providing seamless discovery of learning resources across a number of catalogues that are currently provided by different communities involved in establishing the European Open Science Cloud. The contents will inform later work including FAIRsFAIR D6.3.


## Versioning and contribution history

| Version | Date | Authors | Notes |
|---|---|---|---|
| 0.1 | 24.11.2020 | Phil Barker | Outline in response to brief |
| 0.2 | 01.12.2020 | Phil Barker | Draft complete |
| 0.3 | 10.12.2020 | Phil Barker, Angus Whyte | Revised in response to comments |
| 1.0 | 21.12.2020 | Angus Whyte | Introduction added |
| | | | |



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

## Disclaimer

FAIRsFAIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558 The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

## Abbreviations and Acronyms

| | |
|---|---|
| FAIR | Findable, Accessible, Interoperable, Reusable |
| EOSC | European Open Science Cloud |
| OER | Open Educational Resources |
| API | Application Programming Interface |
| LOM | Learning Object Metadata |
| RDF | Resource Description Format |



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

## Executive Summary

This report aims to help DCC and the FAIRsFAIR project consider options for creating a minimal metadata schema that may be used on a trial basis in their work to establish a shared catalogue in collaboration with various providers of training materials.

In November 2020 DCC sought expert advice to inform a draft proposal for trial exchange of metadata between the FAIRsFAIR project and catalogues provided by related projects. These projects have shared interests in providing seamless discovery of learning resources across a number of catalogues that are currently provided by different communities involved in establishing the European Open Science Cloud. EOSC projects and other training providers have a common interest in making training materials FAIR, especially where these relate to training about enabling FAIR principles to be implemented in research practice. The focus of the report is on options for deploying standards established for OER (Open Educational Resources).

The OER standards may be used to make training materials more findable, by enabling training providers and cataloguers to use community-agreed discipline-agnostic terms in the metadata they use to annotate these training materials, so that catalogue users have more standardised ways to find and potentially reuse them. This is in pursuit of a recognised need to fill gaps in training capacity, expanding the availability of resources that may be used to by research communities, data stewards, and others who support research, in learning how to produce more FAIR data.

The advice sought by DCC was provided by Phil Barker of CETIS LLP, and follows below as an informal (unscheduled) report from FAIRsFAIR to inform its tasks 3.3. (Embedding FAIR data practice in research culture) and 6.3 (Knowledge bases for competence centres).

The report briefly reviews abstract models for data about training materials, and the specification of their semantic aspects as standard schemas. Considering the potential of application profiles to specify terms drawn from different standards in order to facilitate their exchange between catalogues, the report characterizes the various relevant specifications and schema. The report then suggests an approach to harmonization, first through establishing a minimal metadata application profile, and then by setting up a metadata store that would exploit linked data standards to make this metadata available in machine-actionable form.



FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

# Table of contents


Executive Summary 4

1. Introduction 6

1.1. Aims and focus of the report 6

1.2 Context 6

2. Overview of Metadata for Training Materials in Terms of Harmonization 8

2.1 Abstract models 8

2.2 Metadata semantics 10

2.3 Vocabularies 11

2.4 Application Profiles 11

2.5 Syntaxes 12

2.6 Summary: characterization of metadata schema for training materials 12

2.7 Harmonization suggestions 13

3. Towards a Minimal Application Profile 15

3.1 Use cases 15

3.2 Existing data 16

3.3 Expanding on the minimal metadata by exploiting linked data 17

4. Suggested further work 18

5. Bibliography 18




FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

# 1. Introduction

## 1.1. Aims and focus of the report

## 1.2 Context

The Digital Curation Centre (DCC) is one of the main partners in the FAIRsFAIR Project (2019-22). The project's overall objectives are to accelerate the realization of the goals of the EOSC by opening up and sharing all knowledge, expertise, guidelines, implementations, new trajectories, courses and education on FAIR matters.

The topic of this informal report is particularly relevant to DCC's role in work packages WP3 Policy and Practice and more especially WP6 (FAIR Competence Centre). The Competence Centre objectives include:

- supporting and organising core community knowledge bases of FAIR competence centres for different thematic areas and disciplines
- avoiding duplication of effort, by identifying opportunities for synergies

To pursue these objectives FAIRsFAIR is investigating options for the exchange of metadata about training resources, aiming for a trial exchange between at least two of the existing catalogues of these resources already set up by other EOSC projects. The training resources will come from various FAIRsFAIR workpackages that produce guidance and training materials to feed into the FAIR Competence Centre, and from other sources including catalogues in other EOSC projects. This Competence Centre will consist of various elements described in the project work plan as "knowledge bases (KBs), with general and community-specific parts;[which] will comprise resources with suitable metadata, managed so as to address the needs of communities and their stakeholders".

An outline design of the Competence Centre is available in FAIRsFAIR project reports D6.1 and D6.2. These envisage that the knowledge bases will be delivered partly through an online forum and partly through links from that to learning resources that will be deposited, classified and catalogued by other projects. WP6 has to report in February 2021 on "the development, maintenance and harmonisation of generic and community-specific knowledge bases".

In parallel to this, the EC has established under its INFRAEOSC funding calls five EOSC 'clusters', each including research infrastructures in one of five broad disciplinary categories. In addition, the INFRAEOSC-5 funding call established five regional and thematic projects, also involving research infrastructures, research producing organisations, and other stakeholders.



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Key to the FAIRsFAIR objectives is to promote harmonisation of standards and practices across the other EOSC projects, in particular the INFRAEOSC-5b projects with whom FAIRsFAIR has a formal Collaboration Agreement. A number of Task Forces have been established for that collaboration, including the EOSC-5 TF Training and Skills. Establishing a shared catalogue for training resources is one of its primary aims. In practice however, the EC has funded another project EOSC Futures, whose scope is likely to include implementation of a catalogue of training resources.[1] This is expected to start in 2021.

Preparatory work for the EOSC catalogue has also been carried out by a Training Resource Catalogue Task Force, a constituent part of the EOSC Board Working Group on Training and Skills. This overlaps substantially with the EOSC-5 TF Training & Skills, alongside Member State and other representatives. It reports in December 2020. The EOSC-5 TF Training & Skills will continue in 2021 to build on initial steps towards a harmonised approach on interoperability of training resource catalogues. These were discussed in a workshop FAIRsFAIR co-organised with the TF in October 2020.[2]

The Task Forces mentioned above are also drawing on work being conducted by an RDA Interest Group (ETHRD-IG). This has a number of 'focus groups' including one on 'minimal metadata for training resources'.[3] Its approach is to identify those properties of such resources that are considered necessary to fulfil a number of user stories across different groups of data management specialists/professionals, e.g. a learner wanting to find training resources on specific data management skills. The minimal list is being narrowed down from a long-list of terms used by existing catalogue services to describe training resources according to published OER standards (currently LRMI/DCMI, IEEE LOM, and the bioschemas training materials profile).

In DCC's view, the catalogue should operate as a federated and machine-actionable layer of metadata, aggregated from the providers of the resources, who are best placed to create, and deliver the training resources and have shared interest in pooling the curation and management of information about them to better support discovery.

The default access point for searching across the common catalogue will be the EOSC Portal, and the main use case is for learners and trainers to be able to search for resources relevant to 'digital skills', with a focus on open science, data science, and on making the outputs of these FAIR. As a federated resource, the catalogue should be available for use by any provider and should thereby stimulate the development of new discovery services that offer added value.

---

1 https://www.eosc-portal.eu/infraeosc-03-2020-integration-and-consolidation-existing-pan-european-access-mechanism-public

2 https://drive.google.com/drive/folders/10x_K4xQ_F50SnX1QX6WVJyFICjuhaDxj

3 https://rd-alliance.org/group/education-and-training-handling-research-data-ig/wiki/ethrd-ig-focus-group-materials



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DCC and FAIRsFAIR partners will outline in the D6.3 report a plan for exchanging metadata. We envisage collaboration with the RDA ETHRD-IG, the EOSC WG Training & Skills, and the EOSC-5 TF Training & Skills.

The report starts with a discussion of which factors in modelling and encoding metadata affect the ability to build systems based on combinations of different metadata specifications. The metadata specifications in use are then analysed according to these factors; some of these specifications are shown to lend themselves to use in combination more than others. On the basis of this a suggestion is made for harmonizing the metadata available into a knowledge graph that would form the basis of a shared catalogue. To illustrate what might be held in this knowledge graph some suggestions are made for what might be included in the definition of a schema for a minimal metadata application profile describing the data held.

## 2. Overview of Metadata for Training Materials in Terms of Harmonization

The FAIRsFAIR project objectives involve promoting the harmonisation of standards and practices across projects, considering that a number of existing projects use different metadata standards for describing learning materials. We can work on the assumption that it is not realistic to mandate that projects change the metadata that they use in order to contribute to a new shared catalogue, and so we should investigate what harmonization is possible given the metadata schema in use. Nilsson (2010) investigated such harmonization, which he defined as "interoperability of combinations of metadata specifications" and found it to depend on a number of factors, namely: the abstract model, the metadata semantics, vocabularies, and application profiles; he found syntax to be less important. The brief overview of metadata in use for learning materials provided below is framed by these factors.

### 2.1 Abstract models

An abstract model for metadata defines "the nature of the components used and how those components are combined to create information structures" (Powell et al 2007), providing "a common framework for the interpretation of metadata" (Nilsson 2010, p13)—common at least among schema that use the same abstract model. Two types of abstract model particularly concern us in this work, one hierarchical and record-like and other entity-oriented.

Hierarchical record-like models are used (but not often explicitly described) by the IEEE standard 1484.12.1 for Learning Object Metadata (LOM) and many APIs that return data in JSON. (JSON encodes data as an object comprising key-value pairs, in which other objects may be nested to form what is in effect a hierarchical record.) The Lifewatch catalogue, for example, combines both approaches, with records available via an API in both standard XML and JSON. Taking the LOM as our exemplar for hierarchical record-like abstract models we see elements arranged under category headings such as "General", "Life Cycle", "Technical", "Educational" etc (Barker, 2005). Elements with the same name have different purposes depending on their position in the hierarchy and, in some cases, the value of sibling elements. So the element "5.1 description" under the Educational



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

category has different purpose to the element "1.4 description" under General category; element "9.3 description" is different again and what it is a description of depends on the value of element "9.1 purpose". Similarly, to find the author of a resource in LOM metadata one must find the value of an instance of element "2.3.2 Entity" which is paired with an instance of "2.3.1 Role" that has the value of "creator". In general, few elements from the LOM can be interpreted independently from other elements, making LOM schema self-consistent but difficult to use with elements from other specifications (see application profiles, below), unless efforts are made to map the LOM elements to some self-contained non-hierarchical equivalents.

Frequently a record will include information about other things than the resource itself, for example the names and affiliations of people who created the resource. It is a weakness of many record-based approaches that there is no unique identifier provided for these "secondary entities", and so information is duplicated across many records, often imperfectly. The lack of unique identifiers and mismatched information about them makes it very difficult group records based on secondary entities such as authors. For example, are the authors with the name J Smith, John Smith and John Smith all the same person? (Aside: the common solution to this issue is to create an "authority file" for such entities, and to include an identifier for their entries in this file in the information provided in the record. This is a step in the direction of the entity-based approach.)

The alternative relevant model is the entity-oriented approach of Dublin Core and RDF in general, where entities and their relationships to each other are described separately. In RDF this is expressed in the form of subject-predicate-object triples known as statements. Thus a Dublin Core description will be a description of a single identified entity, e.g. a Book, that will provide values for various properties of that book. Some properties will have Literal values (strings, numeric values, dates etc), for example "title" or "description"; other properties will link the Book to other identified entities, for example "author" will create a link between the Book and an entity (commonly a Person) that wrote the book, and that entity may be independently described. There need be no difference between how a Person is described as an author of a book and how the same person is described as, say, a member of an institution (some information may be more relevant in one context than the other, but much will be in common), thus making the reuse of data possible. Sets of statements about related resources can be collected together as a graph, and an application profile can specify what entities and statements should be in a graph for it to be adequate for a given purpose. This provides a set of data in a similar way to the record-oriented approach. Because each statement in an entity-oriented description is self-contained, the meaning of metadata elements like the properties that relate entities do not depend on other elements, and so they can be used across different schema.

(Aside: it's quite possible for data to be provided in a syntax that hides or hybridizes the underlying abstract model. For example, rdf data may be presented using JSON-LD, which allows for descriptions of resources which are not identified to be embedded as values for properties. While remaining entity-based this presentation of the data is very record-like.)



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

## 2.2 Metadata semantics

Whether they are elements in the LOM or RDF properties and classes, the terms used in metadata should be defined so that these terms convey meaning, i.e. semantics, not just the structure of a record. These definitions can be human or machine readable, but preferably both.

RDF includes terms for defining terms such as properties and classes that importantly allow for relationships to be made between terms, even when the properties are defined as parts of different standards. Thus Dublin Core defines a term "description" as an RDF property, and so does schema.org. The human readable definitions are similar enough for us to form the opinion that these are equivalent; but more importantly the machine readable definition of the schema.org term includes the information

schema:description owl:equivalentProperty dcterms:description .

which tells any program that understands the relevant RDF vocabularies that the two are equivalent. Thus such a program that uses Dublin Core terms can make use of descriptions provided in schema.org. Other relationships exist for less closely matching terms. Other semantics can be defined, for example terms may identify a class of entity (such as Person, Book, Organization), and the definition of a property may define the classes that it can link between (for example the author property may be defined so that it links between entities of class Book and those of Person or Organization). This can be useful in making inferences about the nature of otherwise unknown entities, but it does restrict the scope for which a term is useful (what if I want to use the author property to say that a computer wrote an essay?).

Clearly there is some relationship between the various description elements in the LOM, however the machine readable definition of the LOM is in XML Schema, which provides no mechanism for making such assertions. Indeed the XML Schema definition of the LOM provided by IMS Global only provides information on where in the hierarchy the description element may be used, it says nothing about the different semantics associated with the various positions of the element within the hierarchy. Similarly JSON schema used by APIs frequently provide only rules structuring objects, not semantics.

If someone were to create RDF properties with equivalent semantics for each of the elements in the LOM, it would be possible to make assertions such as

lom:educational-description rdfs:subPropertyOf dcterms:description .

This means that any value provided as a LOM description in the Education category, is also a valid value of the more generally defined Dublin Core description. Some information is lost, namely the distinction between the different descriptions in the LOM, but it may be that some information is better than none. This approach to mapping via RDF terms is being taken by the Domain Ecosystem Schema Mapping (DESM) project, which aims to build a tool to facilitate mappings such as that created by the T3 Innovation Network project on semantic mapping



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Of course, mapping LOM or other record elements to RDF entity properties is only useful if the data can be similarly transformed for an XML encoding to RDF statements, one approach to doing this is the RDF Mapping Language (RML).

## 2.3 Vocabularies

Terms are rarely defined in isolation, more commonly a specification defines a set of terms or vocabulary. In RDF and many XML specifications URIs are used to uniquely identify terms (APIs, however, generally do not provide globally unique identifiers for the names used). In order to avoid collisions between similarly named terms, each vocabulary may define a namespace. Namespaces are also commonly identified with a URI. Typically each metadata specification will define a namespace for its vocabulary, so namespace, specification and vocabulary get used somewhat interchangeably. It is common practice to abbreviate the namespace URI, for example dct: for Dublin Core Terms namespace identifier, http://purl.org/dc/terms/, so that term identifiers such as http://purl.org/dc/terms/description can be written as dct:description. This is the basis of QNames in XML and Curies in RDF.

Record based metadata schemas typically define terms naming the elements of a record, the data types and controlled vocabularies that may be used to provide values. RDF schemas provide terms for properties, classes of entities, and concepts that may be referred to. As mentioned above, the semantic definition of an RDF property may include limits on the class of entity that can be used as the subject or object of a statement.

## 2.4 Application Profiles

Heery and Patel (2000) introduced application profiles as "schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application". As well as "mixing and matching" terms from different namespaces, an application profile may introduce additional constraints on value spaces such as the data type or class that may be used to provide a value for an element.

Application profiles are powerful tools for interoperability. They reduce the need to create new vocabularies. They also mean that constraints, such as those on the type of entity that a property can be used with, can be defined where they are needed without interfering with other uses of that property. These rules can be encoded using specifications such as ShEx or SHACL to allow for automatic data validation.

It has been possible to create application profiles of the LOM that constrain the number of elements that are used, limit the values that may be used for the elements chosen and add a few discrete additional elements. It has also been possible to reuse some terms from the LOM in RDF vocabularies—though the mismatch between XML Schema and RDF Schema does not allow for these to be declared as equivalent in a machine-readable way, so doing so does not make them automatically interoperable. However, wholesale mixing and matching of elements from several record-oriented schema involving the merging of different hierarchies is often impractical. The best that can be achieved is a block of data in one schema added to a block of data in another, i.e. the



FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

use of two schema not the creation of a new one from individual terms drawn from different vocabularies.

## 2.5 Syntaxes

The syntax used to express metadata is the least important factor affecting interoperability, however it does have an impact as many developers will view a specification through the syntax. Thus many developers will implement a specification through the XML Schema Definition (XSD) or JSON Schema provided. If a developer is familiar with web APIs they will be familiar with dealing with data in XML and JSON. These lend themselves to record-oriented views of data. RDF data can be expressed in many formats, each with its own strengths and weaknesses. For example, N3 is closest to the triple structure of RDF statements; Turtle is more human readable; RDFa embeds the RDF into an HTML representation of the same data. Most web developers will not be familiar with these formats unless they specialize in RDF. RDF can be expressed in XML, however the result is verbose and difficult (for a human) to read. JSON-LD (Javascript Object Notation for Linked Data) hits a sweet spot of using a syntax familiar to developers and being reasonably readable; and it includes mechanisms ("syntactic sugar") that can be used to hide some of the unfamiliar aspects that come from Linked Data being RDF.

## 2.6 Summary: characterization of metadata schema for training materials

Viewing the various specifications and schema relevant to the exchange of data about training materials through the lens of abstract models, semantics and vocabularies used should highlight how interoperability is enabled or hindered.

**IEEE LOM / IMS Learning Resource Metadata** (these are essentially the same) is a record-oriented specification which encodes metadata for everything related to a learning resource into a single, hierarchical XML document.

**Zenodo API** uses JSON and is implicitly record based, however it is relatively flat, and terms are mostly independent which will help in mapping to standard schema.

**Zenodo OAI PMH**, record based XML using datacite or simple Dublin Core schema.

**Dublin Core Metadata Terms**, broadly applicable vocabulary properties and classes expressed in RDF, relevant to information resources, "intended to be used in combination with metadata terms from other, compatible vocabularies in the context of application profiles".

**Schema.org**, a large vocabulary of properties and classes in RDF, that cover pretty much anything that can be on or have a web page. Several terms are related to Dublin Core equivalents. Has reasonable coverage of characteristics relevant to training, education and learning for with classes for courses, learning resources and educational events among others. Not explicitly designed for use with other vocabularies (the intent is that it is large enough to be used on its own), but some features of the semantics mean that it is increasingly used this way.



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

**Learning Resource Metadata Initiative** (LRMI), a collection of classes, properties and concept schemes in RDF for markup and description of educational resources. Many of the terms in the schema.org vocabulary for describing properties of a learning resource originate from LRMI. Some of these terms have close analogues in the Educational Category of LOM elements. As well as being defined in schema.org they are defined in a Dublin Core namespace in such a way as to facilitate use in with other vocabularies in application profiles.

**Data Catalog Vocabulary** (DCAT) an RDF schema for description of data sets and related entities, which is part new terms in the DCAT namespace and part application profile of these terms with terms from other vocabularies, notably Dublin Core Metadata Terms.

**BioSchemas** several application profiles of schema.org, some which introduce new terms that have been suggested for inclusion in schema.org. There are relevant drafts application profiles for Training Materials, Courses and Course Instances.

**LifeWatch Catalogue** uses selected elements from the IEEE LOM, with an API to disseminate records in XML or JSON format.

The following two are noted because they illustrate more recent entity-based approaches to the LOM, but are less relevant in terms of metadata found in the wild.

**ISO/IEC Standard 19788 Metadata for Learning Resources** (MLR), largely based on LOM elements redefined in an entity-oriented abstract model that is compatible with RDF. The spec documents are long and expensive, which inhibits use of this standard.

**IEEE P2881 Project for Standard for Learning Metadata** (LM) A nascent project which is not a new versioning of the LOM, but which may recast something like the LOM into RDF.

Looking across these we can see harmonization across entity-based RDF schema: for example, equivalent terms in Dublin Core and schema.org, including those from LRMI; application profiles drawing on terms from these vocabularies for data catalogues (DCAT) and training materials (BioSchemas); and various attempts at providing versions of elements from IEEE LOM that can be included.

## 2.7 Harmonization suggestions

This analysis of the relationships between various metadata schema used to describe training materials suggests an approach to harmonization that would allow the creation of a shared catalogue. A minimal metadata application profile (minimap?) should be created comprising those elements/terms that are required for the catalogue and available from partners' metadata. The flexibility of the entity-based approach and ability to use RDF to express semantic relationships between terms in the different vocabularies used point to this application profile being RDF based. This application profile should be based on a broad vocabulary such as schema.org or Dublin Core Metadata Terms, as these are used directly be several specialized schemas in used (e.g. schema.org by BioSchemas, Dublin Core by DCAT), however semantic mapping between terms in these vocabularies means that the choice of which one to use is less critical. The application profile



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

schema should contain the mappings required for the underlying data store to infer equivalence between properties from schema.org and Dublin Core when required. These mappings will mean that if (say) schema.org is used as in the application profile, any resource described in DCAT containing dct.title and dct.description will automatically fulfil the need for a schema:name and schema:description. The combination of RDF schema and data held in a suitable datastore creates a knowledge graph from which applications can draw the data they require via an API. An RDF triple store with a sparql interface would fulfil this.

It may also be useful to allow partners to provide the required metadata directly via JSON-LD. This should be done via a developer-friendly JSON-LD encoding, which means providing a JSON context file to map the JSON keys to RDF properties, and optionally a JSON Schema file to assist in generating conforming data.

Metadata provided in record-oriented formats such IEEE LOM XML and the zenodo API JSON objects will need to be mapped and converted to RDF statements. The ease of this will depend on which elements are required, but hopefully many relevant properties already exist in the generic terms (like title and description) or specialist educational terms from LRMI. Technologies like RML may help with this—even if only as a way of documenting the transforms used—or the extraction and transformation of the relevant data could be handled with simple bespoke programming.

(Aside: there is no reason why the metadata gathered need be limited to what is in the application profile. One of the advantages of RDF data stores such as triple stores is that the data held does not have to conform to a fixed schema. It may be the data that isn't described in the schema for the application profile can still be displayed even if it cannot be used for semantic search).

The pain points in this approach are likely to come from metadata that is difficult to map (e.g. getting authors from LOM records), metadata that is not available (e.g. identifiers for entities that are only described in record fields), and mis-matches between value spaces. This last one is easiest to overlook when the focus is on vocabularies for properties and classes, but is critical to some aspects of a functioning catalogue, such as consistent indexing. It may, for example, be desired to produce an index of resources by subject, however this will be difficult if different projects providing data are using different subject classification schemes.

A useful format for developing such an application profile, the Tabular Application Profile (TAP) is being developed by the Dublin Core Application Profile Interest Group which allows spreadsheets to be used to document the properties used and any constraints on value spaces imposed. This information can then be used to create more formal representations, for example in ShEx or SHACL, useful for validating whether the metadata held is sufficient.

In summary, the harmonization approach suggested involves:

- A minimal metadata application profile, including semantic mappings of the terms included, encoded in RDF.
    - The current work from Dublin Core on TAP may help in developing the application profile.



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- A RDF data store for this schema and data provided by partners that performs semantic search and provides an API
    - must also include interfaces for data ingest
- The ability to transform LOM metadata and data from APIs to RDF that conforms to the application profile schema.
- A developer-friendly JSON-LD format of the applications profile for those who want to use it to provide data.
    - This means a JSON-LD context file and optionally a JSON Schema file for the application profile.
- ShEx and SHACL representations of the profile for data validation.
- An application through which users may search and browse the data held.

# 3. Towards a Minimal Application Profile

Any metadata application profile should be built to meet the requirements of the application or community that is going to use the data. These requirements can be derived from use cases, or simple stories about what activities should be enabled. It is also pragmatic to consider the data that is actually available when considering these requirements.

This section briefly presents some simple use cases and the data available with the aim of illustrating how the harmonization suggestions above would work in practice. They also include hints at how the amount of metadata required can be minimized.

## 3.1 Use cases

The simplest use case is discovery by free text search. This can be supported by a description of the resource, a list of keywords and a title (and possibly other textual fields). Ideally it would also be possible to search the full text of the resource, but this may not be feasible in a metadata-based catalogue. Division of the summary text into description, keywords and title allows higher weights to be given to text matches in more significant fields.

Title and description are also required when displaying results from a search or browse query.

A more informed searcher might want to search or browse for a more specific topic. If there is a specific competency framework that is used across many partners, which enumerates the skills required in a domain this may be used to support such resource discovery through the LRMI `teaches` property that associates a learning resource with a learning outcome. If no formal competency framework is widely used it may still be useful to provide textual descriptions of the learning outcomes. Similar information, but less specific to learning outcomes may be provided as the subject of the resource, again either using a controlled concept scheme or text descriptions.

It may be useful to filter results returned by various characteristics. These may include whether the resource is intended for beginners, intermediate or advanced learners; the type of resource, e.g. a



FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

video tutorial, online course, simulation, an assessment; and possibly the length of time required may be important. The relevant properties for these are the LRMI properties `educationalLevel`, `learningResourceType` and `timeRequired` (these also have analogues in IEEE LOM).

When displaying results it may be relevant to show the origin of a resource, e.g. information about the author and their affiliation, or the organization that published the resource. If authors can be unambiguously identified, it can be useful to group all the resources created by the same person.

To access a resource the location of the resource will be required. Ideally the identifier for the resource will be a URI that resolves to the resource.

To facilitate legal reuse, the licence terms for the resource should be included.

**In summary**, terms for the following characteristics seem most relevant:

- title, for: free text search, and displaying query results.
- keywords, for: free text search.
- description, for: free text search, and displaying query results.
- subject, for: focussed search or browse.
- learning outcome, for focussed search or browse.
- author, for showing the origin of a resource.
- publisher, for showing the origin of a resource.
- learning resource type, for filtering results by educational characteristic.
- typical learning time, for filtering results by educational characteristic.
- educational level, for filtering results by educational characteristic.
- licence, for legal reuse.
- location, for access.

Many other properties may be usefully displayed if available: date published, alternative identifiers such as DOIs, indicators of quality, file format (pdf, docx, mp3 etc), play times for some formats, page counts for others, and so on.

## 3.2 Existing data

There is not time for this report to include detailed examination of what data is available from all the potential sources. We can look at some of the schemas used by some of the projects, but this still leaves questions about how many relevant resources are available, to what extent do the existing descriptions go beyond minimal requirements, what value spaces are being used and so on.

The Google sheet, MAP contains a summary of which of the characteristics are covered by what term, for several of the relevant schema. The entries below discuss any missing characteristics and other issues.

**Zenodo** has an API that publishes metadata in JSON. Publisher is covered only for some resource types (imprint_publisher: publisher of a book/report/chapter). The education-specific properties—



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

learning outcome, typical learning time, educational level—are not covered; learning resource type is partially covered by a generic typology. The summary sheet does not include how authors and subjects may be described, but there is scope for identification with ORCIDs and URI respectively

**IEEE LOM** the details of the mapping will depend on how the LOM has been implemented. Many of the mappings are quite complex. Entities for authors and publishers are provided as vCards, with no standard way for encoding an identifier such as an ORCID. Licence may be covered in section 6 rights, but there is no standard means for doing so.

**Dublin Core Terms** does not cover keywords or the education-specific characteristics of typical learning time and learning outcome. Learning resource type is covered with a generic `type` property. Related resources such as Agents (authors and publishers) and Concepts (e.g., the subject of a resource) will normally be identified with a resolvable URI. Likewise, the resource being described will normally be identified with a resolvable URI that will be the subject of statements made about the resource.

**LRMI** provides terms for education-specific characteristics to supplement other vocabularies such as Dublin Core Terms.

**DCAT** uses some Dublin Core terms and provides terms for some other characteristics—for the terms we are looking at this is "keywords". Most of the comments for Dublin Core Terms apply, DCAT does however provide more guidance on what data one should expect to be available for related entities such as Agents and Concepts.

**schema.org** covers all properties discussed (any many others that may be relevant), however there is a deal of leeway in how data may be provided.

**BioSchemas** uses the schema.org properties and provides further constraints and recommendations.

**Lifewatch** includes general information such as title, description, and keywords, and some educational properties such as learning time and resource type as well others supported by the LOM but not identified above. Does not include author, publisher, subject or learning outcome.

## 3.3 Expanding on the minimal metadata by exploiting linked data

One way by which the number of terms in an application profile can balloon is when secondary entities need to be described. For example, say we wish to provide some information about the origin of a training material by displaying who the author is. How do we know whether Kevin Ashley or K Ashley are the same person as Kevin Ashley who works at our favourite university? Should we be asking for information about the institution to which the author is affiliated? The best answer to this is to require identifiers that can be resolved to provide further information, and use these to gather whatever information we require. For a person, an ORCID URI is the obvious identifier to require, but others are available. Given the ORCID URI https://orcid.org/0000-0002-3426-6283 we find the information we want. Likewise for the subject of a resource: a string will not be enough to place the topic in a hierarchical browse tree (does "bridges" belong under civil engineering, music,



FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

dentistry..?), but a URL that links to wikidata may provide us with all we need to know. This linked data approach is a strength of the entity-based RDF abstract model.

Not all data providers will be able to provide resolvable URIs for such linked data, so it is sensible to have the option for them to supply further information. As a minimum this might just be a name; as an option this could be everything they know about the entity.

## 4. Suggested further work

Further work is required to validate the use cases from which the requirements were derived and to fully investigate what data is actually available (or likely to be available) from partners.

The minimal metadata application profile should be fleshed out. The vocabularies that are to form the basis of the application profile should be chosen, and the desired value spaces for properties should be specified. The Dublin Core Tabular Application Profile provides a framework for doing this. Since it covers most characteristics schema.org (sdo) seems a natural contender for the base vocabulary of the application profile. It may be worthwhile to create a developer-friendly JSON-LD format of the applications profile.

A schema should be created for the application profile that includes semantic relationships to other vocabularies encoded using the owl vocabulary. For example sdo:name should be shown to be equivalent to dct:title; sdo:keywords should be shown to be equivalent to dcat:keywords. Thus (if a suitable system is used to process the data and the schema) data that is provided as dcat will be recognised as equivalent to what is required for the application profile (and *vice versa*, stored data can be exported as dcat).

Data that is supplied from the Zenodo API or as LOM records should be transformed to whatever terms are used by the application profile. This will be easier for the Zenodo API than IEEE LOM.

If the imported data and the application profile schema are stored in a data store that makes inferences based on OWL (e.g. Ontotext GraphDB) then queries can validate whether this data meets the use case requirements. If the data store has an API then a more user-friendly application could be built to illustrate this.

So far little consideration has been given to what information should be provided about any entity other than the training material being described. This should not be neglected in further work. If, for example, subjects or learning outcomes are to be used for a browse interface, all materials will have to be categorised according to the same concept scheme for browsing to work coherently. Agreeing on common categorization schemes (or coping with different schemes in use) is a significant issue.

## 5. Bibliography

Barker, P. (2005) What is IEEE Learning Object Metadata / IMS Learning Resource Metadata? CETIS Standards Briefings. Online http://publications.cetis.org.uk/2005/10



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Heery, R and Patel, M. (2000) Application profiles: mixing and matching metadata schemas. Ariadne Issue 25. Online http://www.ariadne.ac.uk/issue25/app-profiles/

Nillson, M. (2010) From Interoperability to Harmonization in Metadata Standardization. (Doctoral Thesis) Stockholm. URL: http://www.diva-portal.org/smash/get/diva2:369527/FULLTEXT02.pdf



FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558