QualityLink

Quality Data Exchange Architecture

Tasks 3.2 & 3.5 – January 2025

Authors

Tiago Simões, Colin Tück

Layout

Nuša Karo

Copyright

© 2024-2025 QualityLink consortium

This work is licensed under Creative Commons Attribution-ShareAlike 4.0 International.

Project reference

2023-1-DE01-KA220-HED-000161253

Work Package

3 - Development of Technical Standards to support sharing of Quality and Course information

Version History

v0.1	Initial version for discussion	05/09/2024
v0.2	Revised version after Standards Consultation Board kick-off	11/11/2024
v1.0	Revised after SCB face-to-face meeting	13/01/2025

This project has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Table of Contents

Introduction

The QualityLink project aims to empower stakeholders, including students, institutions, employers, and recognition information centres, by providing them with comprehensive and relevant quality data on courses and micro-credentials. This should help improve recognition decisions and allow learners to follow flexible learning pathways.

The project envisions contributing to a seamless and interoperable environment where quality information on study programs and micro-credentials is easily accessible and interoperable across diverse sources.

The project's guiding principles are:

- **Quality**: ensuring that all project endeavours adhere to the highest standards.
- **Competition**: as a positive force for driving continuous improvement and innovation.
- **Student-centric approach**: empower learners to navigate the diverse landscape and make personalised choices.
- **Transparency**: avoiding compound indicators and allowing the end user to make their own informed decisions.
- **Openness and inclusivity**: aggregating, sharing, and disseminating data.
- **Democratisation of quality data**: making it universally accessible and beneficial to a broad spectrum of stakeholders.

Challenges

Quality information on study programs and other courses is often dispersed throughout many different platforms and sites, and they are not interoperable. Through creating open standards and collaboration, the project aims to establish the infrastructure for aggregating quality information from a wide range of sources.

In particular, the project aims to address the following current challenges:

- Trust data is provided by existing tools, such as ETER, DEQAR or the EWP Registry, but only some of them are open and fully interoperable.
- Data might go to programme level, but this is not necessarily the case (e.g. only institutional accreditation). The module or course level is generally not covered.
- In existing data sources the programme level is less structured/managed than institutional level. It is thus difficult to link/connect data on the same programme/module/course from different sources.
- Institutions do generally not publish their own quality data (e.g. student survey results, grade distribution, …) in a structured and open format.
- Other quality data (from third-party sources) is often published through their own tools/websites, but rarely as open data.

Objectives

The overall objective of the QualityLink project is to address gaps in the standardisation landscape for higher education which prevent the comparison of information on courses, by proposing and developing standards to meet these gaps.

The proposed standards aim to facilitate the exchange of quality data between higher education institutions (HEIs), quality assurance agencies (QAAs), ministries, recognition information centres (ENIC-NARICs), ranking bodies, professional associations, course directories and other stakeholders. The components should allow these stakeholders to publish their lists of courses and quality indicators in a standardised format, and to aid their scraping by web-tools and their inclusion in search and other comparison tools.

This document focuses on the technical architecture that is being designed and later tested by the QualityLink project. See the overview of quality domains and indicators for details on the type of data and indicators the project focuses on.

Use cases

The QualityLink project and the technical architecture have been designed with the following overarching use cases in mind:

- Allow (potential) learners to discover and compare learning opportunities, based on reliable information on them. Provide information to which courses they can enroll/register under which conditions.
- Provide universities/HEIs a central database to which they can deliver information on learning opportunities so that different partners can retrieve it easily, without having to create interfaces to every partner's system. Joint catalogues of learning opportunities by European University alliances are a specific scenario of this use case.
- Allow HEIs and other data providers to move back and forth between standards relatively easily and in a semi-automated manner.
- Enable data providers (HEI or other) to compare themselves to other HEIs in order to identify areas of improvement.
- Supply relevant information to HEIs, recognition information centres (ENIC-NARICs) and employers for their recognition decisions, e.g. when evaluating a potential student's credentials.
- Enable the aggregation of relevant crowdsourced data in the future[1].

[1] The architecture would allow any aggregator to crowdsource data themselves or include crowdsourced data, but this is not being explored in practice by this project.

Principles

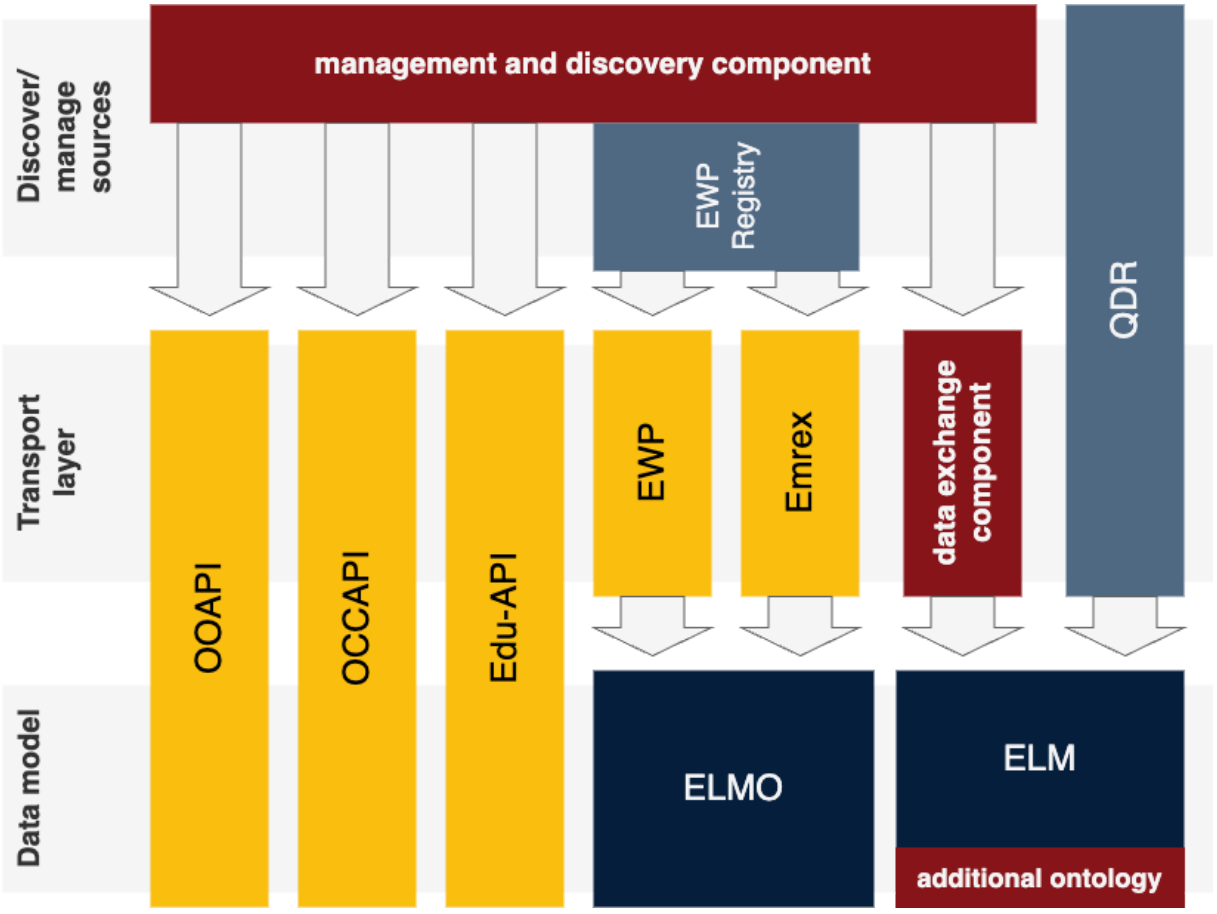
The proposed technical architecture should be guided by the following principles:

- **Seamless integration**: we aim to create an architecture that uses relevant existing standards and systems wherever possible, and integrates with them seamlessly.
- **Easy adoption**: we want to keep the bar low to promote adoption by HEIs and other data providers. That is, being minimalist when it comes to minimum requirements and, for example, allowing different routes for HEIs/other providers to expose their data where it is helpful.
- **Flexible**: we want to create a basic architecture that is not overly use-case-specific but supports a variety of current and possible future use cases.
- **Modular**: we want to design an architecture that is open to future extensions, e.g. to cover additional types of data or sources.
- **Scalable**: we want to design an architecture that is in principle scalable to thousands HEIs and other data providers.
- **Open source**: we will release all software developed as part of QualityLink as open source under the GPL, and all documentation (incl. standards, ontologies, etc.) under CC BY-SA.

Overview of components

Data on micro-credentials, courses and other learning opportunities is already covered by several established data standards. Most quality domains and indicators that QualityLink focuses on can be adequately represented by existing data standards, with the European Learning Model (ELM) offering the widest coverage. QualityLink will thus propose a **data model as an extension to ELM**, adding ontologies for indicators not covered yet.

Rather than focusing on a single data model or standard, QualityLink aims to use existing converters made by the community and to possibly create additional **converters to map/crosswalk** between data models. The project aims to design an interoperable architecture that functions with several widely used standards and data models.



Some of the existing standards provide APIs. For example, both the Open Education API (OOAPI) or the Edu-API include a data model and a defined data transport layer (in these cases a REST API).

Other standards, such as the European Learning Model (ELM), are a “pure” data standard and do not include such capabilities directly. Related to ELM, the Qualifications Dataset

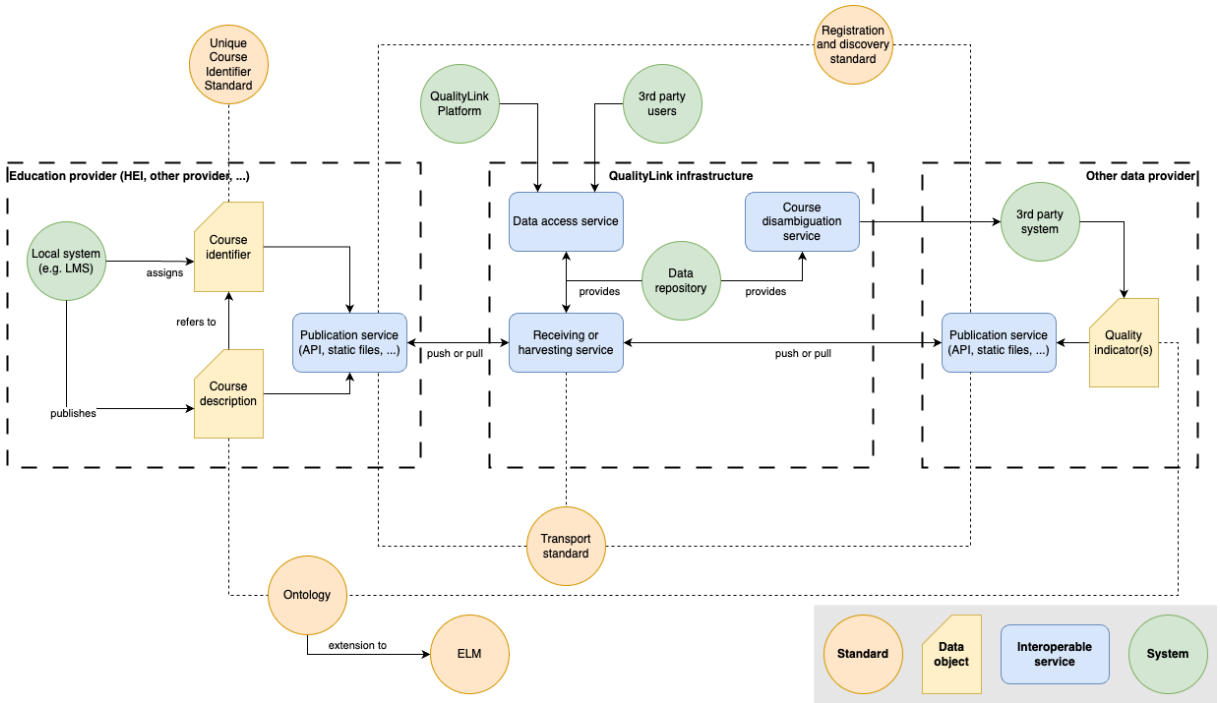
Register (QDR) defines serialisation formats as well as data transport options. These, however, mainly focus on data supplied by national authorities or their delegates.

Most existing standards do not include an approach for discovery and registration of data sources. That is, various APIs specify how higher education institutions can expose data through a standardised API or format, but do not specify how to find API endpoints that institutions expose.

The QDR aggregates data from ELM-based sources, but data sources themselves are managed manually; data can only be fed into QDR through the national level or by delegated authorities, not directly by HEIs. The EWP network includes a central registry of institutional APIs, but entries in the central registry are managed manually.

QualityLink aims to fill the gaps by specifying a **flexible data exchange/transport layer** for ELM-based data as well as a **scalable discovery mechanism** that allows higher education institutions to expose their data easily.

Although each existing standard accommodates a designated field for course identification, there is no harmonised system or format for identifying courses. In particular, there is no common understanding, terminology and business rules across standards. The QualityLink architecture thus includes a proposal for a **unique course identifier** that could be used across different systems and that is agnostic regarding data formats.



Policies

This discussion document focuses on technical architecture. Policies on key issues, however, influence some architectural choices and thus need to be defined beforehand to ensure that the architecture aligns with these policies.

The aim of QualityLink is not only to propose standards for data publication but also to pilot aggregation. Hence there is a need for a clear policy on data sources.

A policy on data use and sharing clarifies for education providers and other data sources how their aggregated data is available to the public and can be used.

Objectives

POL-1: As a user I want to be assured that the aggregated data is trustworthy and does not include flawed or rogue data.

POL-2: As a HEI I want to publish my course data without the need for endorsement or clearance by any other organisation.

POL-3: As a HEI I want to have clarity on how my data will be published and might be used.

POL-4: As the owner of a dataset I want clarity which data can be aggregated into the QualityLink dataset and platform.

POL-5: As a researcher or advanced user I want to access the QualityLink dataset under clear terms and conditions, in order to have legal certainty when copying, re-using or embedding QualityLink data in my own applications.

Data sources

The policy on data sources will determine the approach to managing data sources and providers. That is, it will clarify who will be allowed to publish what, or from which source the QualityLink aggregator will fetch which information.

Existing examples

Relevant existing initiatives handle this in different ways:

- The EU's QDR (see above) is based on national authorised representatives registering data sources. Registration of data sources is manual.
- The EWP network features a central registry of nodes and HEIs they cover. Registration of new nodes is manual.

- Data in DEQAR is provided by EQAR-registered agencies. Accounts for new agencies are manually created.
- Data in EHESO and OrgReg are aggregated from authoritative national sources manually.
- The US-based Credential Registry is based on organisations (e.g. HEIs) creating accounts which are manually verified by Credential Engine staff.

Approach

The policy will distinguish between three different types of data providers:

- **Authoritative sources** - may provide trusted data on any provider and their programmes/courses. This would include sources such as DEQAR, ETER or national authorities. Authoritative sources should be approved by the QualityLink consortium;
- **Education providers** (HEIs and others in the future) - may provide trusted data on their own programmes and courses. Any recognised education provider should be able to provide data without specific approval;
- **Other data sources** - may provide data on existing programmes/courses, but limited to specific domains. This would include ranking providers, etc.. For these data sources the QualityLink consortium should adopt an eligibility policy.

The technical specifications (see below) should ensure that education providers can decide to publish data and have it aggregated in the QualityLink platform without the need for manual intervention from the QualityLink consortium/admins or anyone else; other sources could be managed manually.

Data access and usage

The aggregated QualityLink dataset will be accessible through a prototype portal, as full download and through an API. The policy sets the terms and conditions for reusing the data. Related to this are also policies and conditions for using the API to access the QualityLink dataset.

Examples

The following examples can be found for similar datasets:

- DEQAR: data can be accessed and re-used without any conditions, pursuant to the Open Data Commons Public Domain Dedication and Licence (PDDL) - see DEQAR Terms & Conditions. API is subject to registration, free for basic use but reserves the right to charge a fee for extensive or commercial use.

- ORCID: full data is released under CC0, API can be freely used for any “non-commercial” use, i.e. data cannot be used in a paid service or revenue-generating product - ORCID Public API Terms
- CWTS Leiden Ranking: available under CC Attribution 4.0

Approach

In principle, it would be desirable to dedicate the aggregated QualityLink dataset to the public domain and impose no restrictions on using, copying or re-mixing.

At the same time, some HEIs and other upstream data providers impose constraints on the use of their published data, especially their detailed course descriptions.

Initially, to enhance compatibility with upstream data sources and maximise readiness to feed data into QualityLink, the terms of use should exclude commercial use and require attribution, i.e. release the dataset under CC Attribution-NonCommercial.

The QualityLink API should be available for public use with certain rate limits. In addition, non-commercial registered users (e.g. researchers, non-commercial platforms) could be granted access with higher rate limits.

In the long term, it could be discussed with HEIs and other data providers whether the aggregated data could be published for re-use under more permissive terms, such as CC0 or PDDL.

Technical specifications

Unique course identifier

Role

The unique course identifier should facilitate the identification and comparison of courses and allow for higher education courses to be tracked across multiple different systems and databases.

User stories

CID-1: As an education provider I want to publish and control authoritative identifiers for my courses.

CID-2: As a user I want to dereference a course identifier to a standardised description of the course, incl. other identifiers assigned to the same course.

CID-3: As a user I want to retrieve historical information (e.g. previous names and identifiers) or demographic information (e.g. course has been replaced by another course, course was discontinued, …).

Fundamental challenges

The most significant challenge in course matching has to do with identifying a course across multiple data sets. In known processes involving course data, such as Learning Agreements, information is replicated as if copied on paper (despite the paperless workflows currently in effect) and the notion of programmatically retrieving data in a digital format is no more than an afterthought.

As such, it is important to establish a methodology that would allow identifying any given course (to be understood as a Learning Opportunity) as offered by any given institution.

Course vs. Offering/Instance

Most existing data models and standards for course descriptions distinguish between a course as a long-lived specification of a learning opportunity and a course offering or instance, referring to the course being offered in one specific semester (or academic session more generally). Potentially, there could also be several offerings during a single period of time, referring for example to different languages in which the same course is offered or to different modes (e.g. on-line vs. on-site).

The following table illustrates how these concepts are named in different existing standards:

	Course/Specification	Offering/Instance
ELM v3	Learning Achievement Specification	Learning Opportunity
OOAPI v5	Course (Program)	Course (Program) Offering
Edu-API	Course Template	Course Offering
EWP Courses	Learning Opportunity Specification	Learning Opportunity Instance
OCCAPI	Course	*(not modelled)*
DEQAR	Programme	*(not modelled)*

In the context of the QualityLink project, the focus is on the course as a long-lived, stable unit, and not on the offering/instance.

Who offers the course?

Any institution offering a course must be reliably identifiable. In the case of Higher Education Institutions, particularly in Europe, there are many identifiers available, all serving any particular purpose (Erasmus codes, PICs, OIDs, etc.) often related to the Erasmus+ programme. However, a more generic way to identify an organisation offering courses would be ideal, so as to broaden the scope of learning opportunities beyond the classic study programme model.

Changes in identity

Although not a very common occurrence, institutions tend to change over time, leading to splits and mergers, rebranding and similar processes, which may lead to changes in the way they identify themselves. As such, keeping track of such changes would be paramount for a reliable method of identification over time.

This problem is not necessarily within the scope of the project, but it must be noted.

Key requirements

It should be assumed that any organisation offering learning opportunities will have their own internal system for their identification, and the particulars of such systems are not necessarily of interest to the current work. However, some core requirements must be met:

- Each course identifier must be unique within the organisation. Ideally they should “survive” mergers or other demographic changes, e.g. using mechanisms such as previous-schac in EWP.
- Course identifiers must not be recycled, i.e. assigned to different courses.

- Course identifiers may change over the lifetime of a course (e.g. as a result of mergers or internal reorganisation). The information obtained when dereferencing an identifier should/must(?) indicate the previous identifier(s) of the same course. The previous identifiers must remain dereferenceable and redirect/point to the new identifier.
- Each individual combination of workload and learning outcomes must be treated as a separate learning opportunity.

Existing standards

There is no existing universal (Europe-wide or cross-national) system for a standardised and resolvable notation of programme or course identifiers.

Naturally, several existing standards include some form of identification of programme and course objects:

- OOAPI and CTDL identify all objects by a UUID-based identifier, including courses and programmes. In addition, OOAPI for example allows the expression of existing codes for programmes as well as courses, but scoped to the offering institution and without a standardised format.
- OCCAPI recommends UUIDs as identifiers of courses, but allows any arbitrary identifier scoped to the offering institution.
- ELM allows the expression of arbitrary identifiers for programmes and courses.
- EWP expects unique identifiers scoped to an institution and also recommends UUIDs as surrogate keys.

As regards national systems, Q to experts: which should we investigate as possible role models?

Suggested approach

In order to uniquely identify a course, the most reliable path would be to tackle the issues in two ways:

- identify the course within the organisation;
- identify the organisation itself.

This mirrors the approach of the European Student Identifier for identifying students.

Identifying a course within an organisation

Whatever method a particular organisation may use to identify its courses internally, it is sufficient that this method guarantees that identifiers are kept unique, i.e. not “recycled” to identify different courses at different times.

In principle, a universal approach to course identification must allow for any kind of internal identification system within an organisation, as imposing changes to already existing systems would cause too much friction and hinder adoption.

Identifying an organisation

An organisation can be identified in many different ways, depending on the context (for example in the Erasmus+ context one may find several types of institutional identifiers), so ultimately it would be desirable to adopt a method that would remain consistent across different contexts.

Various supra-national standards exist for identifying institutions:

- SCHAC (used in European Student Identifier, EWP)
- ETER ID (used in EHESO, OrgReg, DEQAR)
- WHED ID
- Erasmus code

In this sense, the concept of a SCHAC code as used in higher education is very interesting. The SCHAC code is essentially a domain name, which is necessarily unique and owned by the organisation itself. It is currently used in Erasmus Without Paper (EWP) but its application can go well beyond that environment. The HEI API provided by EUF gives access to information on SCHAC codes currently in use in the EWP network or in eduGAIN, as well as other (hypothetical) SCHAC codes of HEIs holding an Erasmus charter.

One challenge of SCHAC codes is that they may change if the institution's domain name changes. Even though not very frequent, these changes happen and are currently not available in the HEI API (partial records exist, but are currently not exposed in a practical way). Moreover, mergers of institutions or other demographic events are currently not tracked by the HEI API.

In this regard it could be helpful to additionally leverage OrgReg and the unique entity IDs it maintains for institutions. These remain stable regardless of name, domain name/website or other changes. OrgReg also systematically tracks demographic events, such as mergers of institutions, although the dataset may not be complete.

Proposed format: URN

From Wikipedia:

A Uniform Resource Name (URN) is a Uniform Resource Identifier (URI) that uses the urn scheme. URNs are globally unique persistent identifiers assigned within defined namespaces so they will be available for a long period of time, even after the resource which they identify ceases to exist or becomes unavailable. URNs cannot be used to directly locate an item and

need not be resolvable, as they are simply templates that another parser may use to find an item.

Inspired by the European Student Identifier, part of the European Student Card Initiative.

The URN format `urn:<NID>:<NSS>` requires defining:

1. A namespace (NID) registered with IANA: https://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml
    a. Q: would this fit within the SCHAC namespace? https://www.rfc-editor.org/rfc/rfc6338.html
2. A syntax for the namespace specific string (NSS)
    a. Including the SCHAC code of the institution at the time of ID assignment
    b. Including a unique identifier for the resource within the institution
    c. With all non ASCII characters percent-encoded

Other avenues to explore: DID

The Decentralized Identifiers (DIDs) approach was suggested in relation to course identification. The DID specification goes beyond identifiers in a narrow sense and standardises a DID Document, which must be resolvable for every DID.

The DID Document can notably include public key material and service endpoints related to the entity identified by the DID. The latter could be of interest for a course identifier, e.g. the DID Document could describe the service endpoint that provides data on the course.

DIDs are based on a registry of DID methods, each of which specifies how a DID is resolved to a DID Document. A popular method for institutional DIDs, i.e. where no privacy matters are involved, is the did:web Method Specification. A did:web is simply resolved by retrieving a DID Document from a URL under the domain that is part of the DID.

While most features of DIDs are not directly needed for course identification and using DIDs as a native identifier might be overly complex, it would be thinkable that each course identifier could *also* be expressed as and be resolvable as a DID.

For example, `urn:schac:uniqueCourseId:uni-lj.si:ABCD1234` could be expressed as `did:web:did.quality-link.eu:uni-lj.si:ABCD1234`. This could be resolved by a simple service implemented at https://did.quality-link.eu, generating DID Documents for known courses.

Course disambiguation service

Role

Existing datasets will not necessarily use the unique course identifier immediately. In order to facilitate their integration into the ecosystem it could be desirable to offer a disambiguation or matching service that would allow discovery of identifiers for existing lists.

User stories

CDS-1: As a data provider I want to match my existing dataset and discover the official identifiers for courses identified by name or another identifier.

CDS-2: As a user or data provider I want to use a reference list of courses including their identifiers. (For example, when building an application.)

Approach

It should be discussed during the course of the project whether a disambiguation service will be necessary and feasible.

Ontology

Role

A standardised ontology would be at the heart of the proposed architecture. It should cover basic course data as well as the quality indicators to be covered within the project.

User stories

ONT-1: As a HEI or data provider I want to publish data in a clearly defined and understandable way.

ONT-2: As a user, I want to obtain clear definitions for the data I can access.

Key requirements

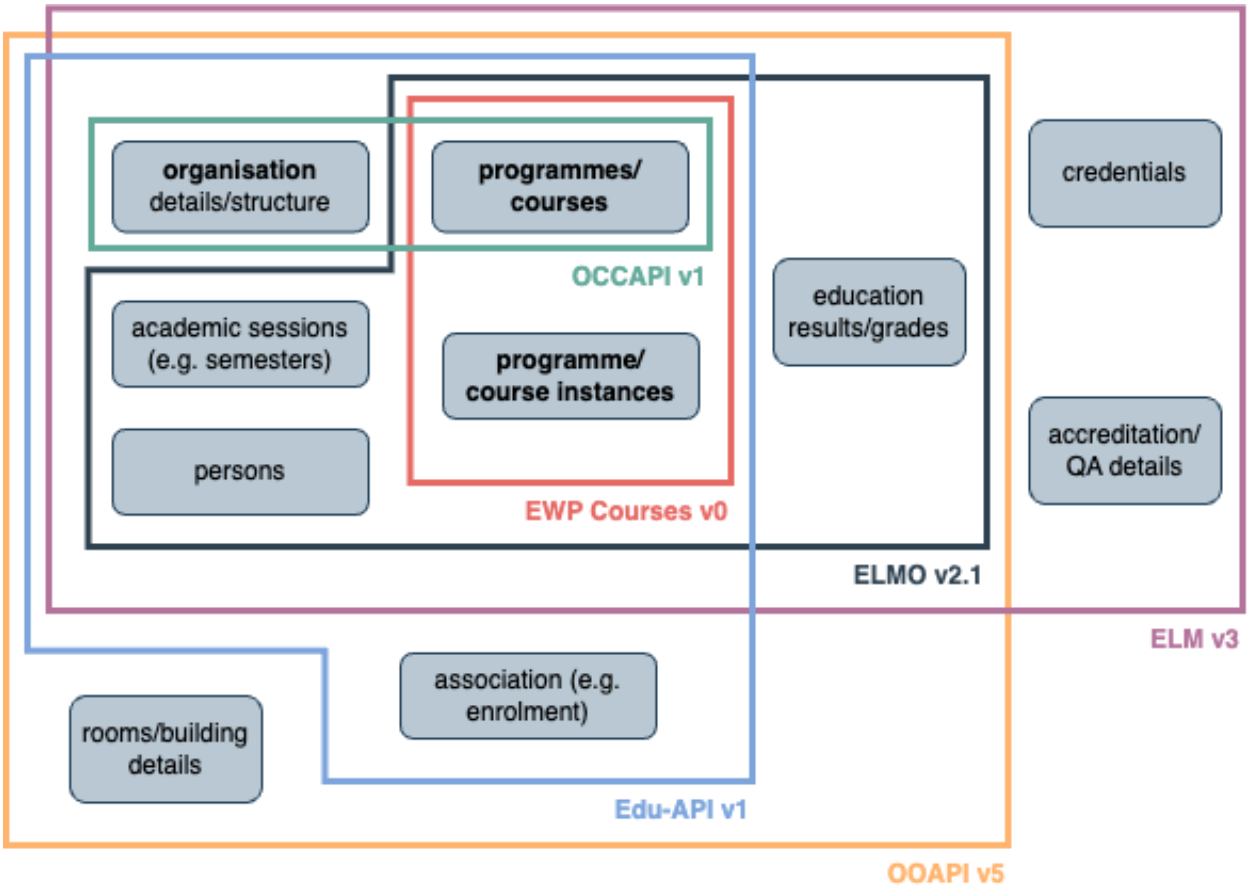
It would not be reasonable to expect existing data sources to adopt a new or different ontology as their native data model. Hence the ontology should be designed such that existing data can be mapped to it easily.

To that end the ontology should be as simple as possible for the purposes envisaged by the project; it should not be overly detailed or nuanced on matters that are not of key importance.

The ontology should be easily extendable to cover new domains and types of information.

Existing standards

There are several existing data models and API standards that specify descriptions of basic course data. The following illustration and table provide an overview of the different domains covered by those:



Name	Description	Maintained by	Technical approach
European Learning Model (ELM)	used in the context of the European Digital Credentials for Learning (EDC) or the Qualification Dataset Register (QDR)	European Commission, DG Employment	RDF ontology and application profiles

ELMO	Based on the CEN standard EN 15981-2011 EuroLMAI and used in Emrex.	Emrex	XML schema
EWP Courses API		EWP consortium	XML schema
Open Course Catalogue API (OCCAPI)	inspired by EWP APIs	EUF	REST API + JSON schema
Open Education API (OOAPI)	specification developed by Dutch HEIs and used in some European University alliances; significant overlap/similarity to Edu-API but covers additional domains	SURF & Community	REST API + JSON schema
Edu-API	candidate specification inspired by different existing 1EdTech standards	1EdTech	REST API + JSON schema
Database of External Quality Assurance Results (DEQAR)	aggregates standardised information on quality assurance results from European QA agencies	EQAR	REST API + JSON schema
European Higher Education Sector Observatory (EHESO, previously ETER) & OrgReg	maintains basic institutional descriptors and aggregates statistical data on HEIs from national sources	EHESO consortium on behalf of European Commission	REST API + JSON schema
Credential Transparency Description Language (CTDL)	developed by a US-based think tank and used for its Credential Registry	Credential Engine	RDF ontology

A detailed comparison which basic data and indicators are modelled by the most relevant existing standards is available in a separate spreadsheet.

Approach

Considering that it describes/models the widest range of data and indicators already, the ELM is the most suitable starting point. Moreover, it is a genuinely European initiative, sponsored officially by the EU and used for various EU-sponsored initiatives.

From a technical perspective, as an RDF ontology the ELM is open to extend where needed.

The QualityLink architecture should thus use and extend ELM: that is, use ELM terms where they exist and define additional ones where needed. This would roughly imply that ELM can be used as is for the following information:

- Basic course data - fully covered by ELM
- Quality assurance data - covered by ELM (if needed, define additional terms or nuances based on DEQAR)
- Quality indicators - already be covered by ELM:
    - Grade distribution
    - Recognition agreements/history

One important dimension that is not covered by ELM is information to whom a certain course is available for enrolment, e.g. whether it is available to specific learner groups only. This information is especially relevant if data is used to populate a joint catalogue of learning opportunities for a European University alliance, for example. A standardised way of describing this should be part of the QualityLink ontology.

The ELM was not designed to model statistical data on courses or study programmes. Hence it is not surprising that several quality indicators are not modeled by the ELM and an additional ontology would need to be developed. This should include an abstract class of a quality indicator as well as specific models for some the quality indicators considered in the project:

- Demand for skills
- Stackability
- Platform QA
- Active learning methodologies used
- Availability of tutoring or mentoring
- Student/staff ratios
- Assessment methods used
- Virtual learning environment
- Make-up/diversity of the student body
- Recognition of prior learning
- Learner support services
- Eligibility for grants / loans
- Student ratings/satisfaction
- Graduation rate
- Student/graduate performance
- Expertise of lecturers
- Rankings
- Network memberships

Within the QualityLink project it will not be feasible to develop an ontology for all those areas. The focus should be on those indicators that are found to be most relevant and feasible in the ongoing feasibility analysis and stakeholder ranking.

For any additions developed, relevant existing systems/standards (see above) and their data models should be taken into account, also where not published as RDF ontology.

For those additions that would fit into the scope of the ELM, the ontologies developed by the QualityLink project should also be submitted as proposals for future extensions of the ELM.

Conversion and mapping

Role

In an ideal world, each data source/provider would map their data to the “ELM+QualityLink” ontology. For existing data standards in wide use it would, however, be helpful to have a conversion/mapping service or library available so that different providers using the same data model do not have to re-do each other’s work.

User stories

MAP-1: As an education provider or data source I want to be able to use my existing tools and standards to publish/send my data in a widely-used format.

MAP-2: As an aggregator I want to be able to consume data available in a variety of formats.

Existing examples

The T3 Innovation Network created a Data Ecosystem Schema Mapper (DESM: GitHub repo - project page). Credential Engine host a DESM instance and provide a mapping for various micro-credential-related data standards, including ELM, CTDL and Open Badges among others. These could be used to build a conversion service.

OERSI is an initiative to index and aggregate metadata from various repositories of open educational resources. The project has created a substantial number of data importers and mappings from different data sources. These could be used to inspire a data conversion and mapping service, or even be partially re-used.

Approach

For pragmatic reasons, conversion/mapping for some highly relevant data standards should be developed by QualityLink itself. We should namely consider the following existing data models and standards:

- OOAPI
- OCCAPI

- Edu-API
- EWP Courses API
- DEQAR
- ELMO (check existing converter)
- Schema.org

NB: As ELM would be the reference ontology used in QualityLink, no mapping is needed.

The detailed comparison spreadsheet (see also under Ontology above) can be used as a basis to inform the mapping, next to DESM or OERSI where applicable.

In addition, it should be possible and encouraged that third parties create additional converters or mappers, and make them available to the community.

Management and discovery of data sources

Role

A standard for data discovery should specify how HEIs and other data providers can indicate from where their data can be harvested for aggregation.

User stories

DIS-1: As a HEI or education provider I want to advertise my data source to a Europe-wide aggregator in a secure way, but without the need for additional action or manual intervention by another organisation.

DIS-2: As a QualityLink administrator I want to manage trusted sources and other data providers whose data should be integrated in the QualityLink platform.

DIS-3: As a data provider I want to access a dashboard or similar to manage my data sources and get insights/stats on how my data was harvested and processed.

Existing standards

- ELM itself does not address the transport layer. The QDR uses RDF serialised as XML, which can be manually uploaded, hosted by the data provider and regularly pulled, or pushed by the data provider.
- Edu-API, OOAPI and the OCCAPI define a REST API to expose the information in JSON format.
- The EWP Courses API defines an XML schema in which a list of courses should be returned.

- DEQAR and the Credential Registry each allow submission of data through manual entry, upload of a CSV file or use of a JSON-based REST API.

Approach

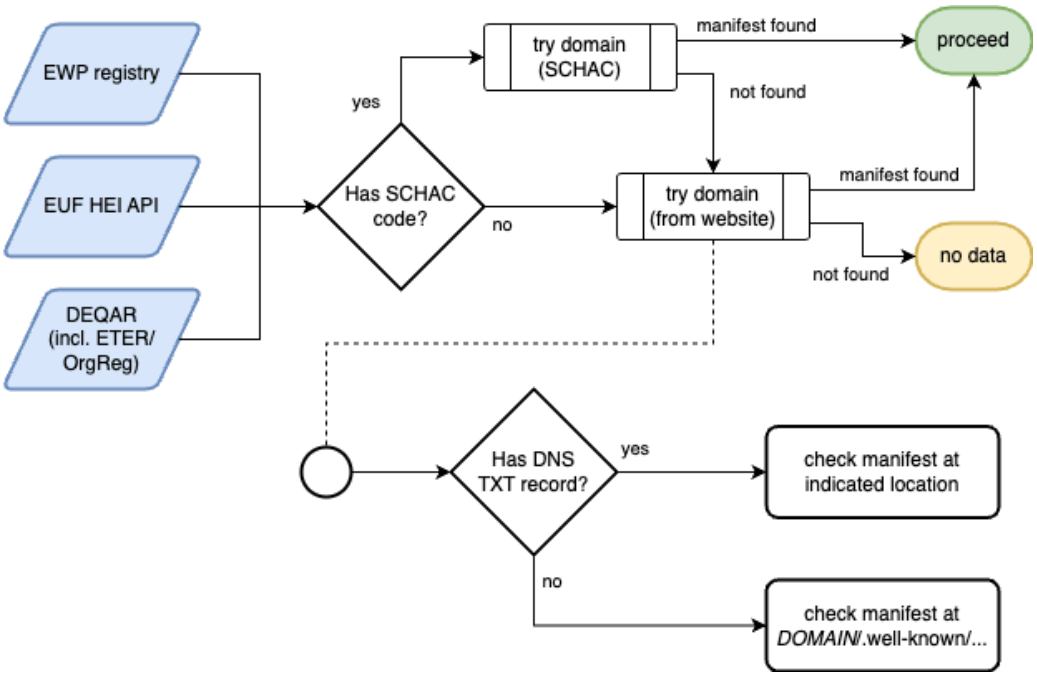
The approach should be as flexible and as easy to adopt as possible for HEIs and other data providers.

The approach should be scalable. Hence, data sources of HEIs need to be discovered automatically and without manual registration or intervention, in line with the policies defined above. (Authoritative sources and other data sources are managed manually, as each case will need to be screened against the applicable policies.)

Discovery of data from education providers/HEIs should be based on data from the EWP Registry, the EUF HEI API and DEQAR (which includes all HEIs listed in ETER and OrgReg).

The location of one or several data files or API endpoints (see Data exchange below) could be indicated in a simple, standardised "manifest" file published by the HEI. The specification for the manifest file should be extensible, so that additional options for making available course data or advertising of other APIs could be using the same spec.

For authoritative sources and other data sources, the location of the manifest file can be managed manually.



For HEIs, the location of the manifest file needs to be standardised or discoverable in a standardised way. Two options would be available to HEIs:

- a DNS TXT record containing a URL;

- a .well-known URL, e.g. of the form https://www.example.org/.well-known/quality-link-manifest.

The discovery mechanism should attempt to discover a manifest file under a known official domain of the HEI, for example in the following order:

- the domain/host name that is the HEI's SCHAC code (for HEIs part of the EWP network or appearing in the EUF HEI API);
- the host name of the HEI's website (as indicated in OrgReg or DEQAR);
- the highest level registrable domain name from that address (i.e. public suffix + one further DNS component).

Data exchange

Role

A standard for data transport should specify how HEIs and other data providers can make data available for aggregation.

User stories

TRL-1: As a HEI I want to feed my data into a Europe-wide aggregator easily and flexibly, in order to avoid unnecessary burden.

TRL-2: As a data aggregator I want to aggregate data from HEIs and other sources efficiently, i.e. avoiding too much overhead or redundant data transfers.

TRL-3: As a data provider (HEI or other) I want to ensure that only the QualityLink platform (or other specified aggregators) can harvest my data.

Existing standards

- ELM itself does not address the transport layer. The QDR uses RDF serialised as XML, which can be manually uploaded, hosted by the data provider and regularly pulled, or pushed by the data provider.
- Edu-API, OOAPI and the OCCAPI define a REST API to expose the information in JSON format. Edu-API also plans additional bindings, e.g. async and message queues.
- The EWP Courses API defines an XML schema in which a list of courses should be returned.
- DEQAR and the Credential Registry each allow submission of data through manual entry, upload of a CSV file or use of a JSON-based REST API.

Approach

In principle, it would be desirable to offer HEIs and other data providers to choose between a push or pull strategy for data transport. For simplicity, only a pull approach should be implemented within the QualityLink project. A push option could be realised later.

In order to be flexible, different ways of data transport should be provided. The following two should be explored during the project:

- Data provider exposes a standardised API (e.g. similar to OOAPI or OCCAPI)
- Data provider hosts a static file (e.g. similar to QDR)

Changes in data should be handled and indicated in such a way that repeated transport of unchanged data is avoided as far as possible. In a pull approach, each option should include a way of indicating last modification or listing/fetching only resources created/modified since a given time.

Data providers could implement access control in three ways:

- limit access **based on IP address**: this is straight-forward by publishing the IP address (range) used by the QualityLink aggregator;
- authentication based on **HTTPS client certificate**: the QualityLink aggregator could authenticate itself with a specified certificate;
- specify a **HTTP header** to be included in all requests (e.g. a token): this option requires the data provider to have an account in the QualityLink aggregator/data source registry to manage this.