OER State of art and outlook

Study on the aggregation infrastructures for OERs

Table of contents

Scope of the document

The main scope of the document is to present the technical recommendations, reference design, and necessary documentation of the metadata aggregation engine to be implemented by the pilot project.

State of the art for aggregating educational resources

Major projects and initiatives

Nowadays, there are several initiatives related to the aggregation of Open Educational Resources (OERs). The main projects and initiatives are presented in the following paragraphs.

Ariadne foundation

The ARIADNE Foundation (http://www.ariadne-eu.org/) is a not-for-profit association that aims to:

- Carry out basic and applied research that will improve creation, sharing and reuse of knowledge through the use of technology
- Develop and deploy methodologies and software that will provide flexible, effective and efficient access to large-scale knowledge bases
- Apply the results of its research and development activities to help preserve multicultural and multilingual knowledge assets and collections
- Explore how these research and development results can be adopted and sustained so that they support educational and research communities

Ariadne members have developed a number of tools, standards and specifications that can be used to build an OER federation service. A demonstrator of the online federation services that can be set up using the Ariadne can be found here http://ariadne.cs.kuleuven.be/finder/ariadne/.

ARIADNE is a member of the Global Learning Objects Brokering Exchange (GLOBE) Alliance and contributing towards the development of a global learning infrastructure that can be accessible from all.

The Globe Alliance

GLOBE (Global Learning Objects Brokering Exchange) is a one-stop-shop for learning resource broker organizations, each of them managing and/or federating one or more learning object repositories. GLOBE makes a suite of online services and tools available to its members for the exchange of learning resources.
GLOBE provides an discovery service for more than 800.000 learning resources. The discovery service is based on an aggregation engine that is built using the Ariadne tools.

OER Commons

OER Commons (https://www.oercommons.org/) developed by ISKME (http://www.iskme.org/) is a freely accessible online library that allows teachers and others to search and discover open educational resources (OER) and other freely available instructional materials. It provides access to resources that can be used in several educational levels including the Higher Education. It supports the aggregation of content from several sources through a metadata ingestion process connected to the OER Commons repository.

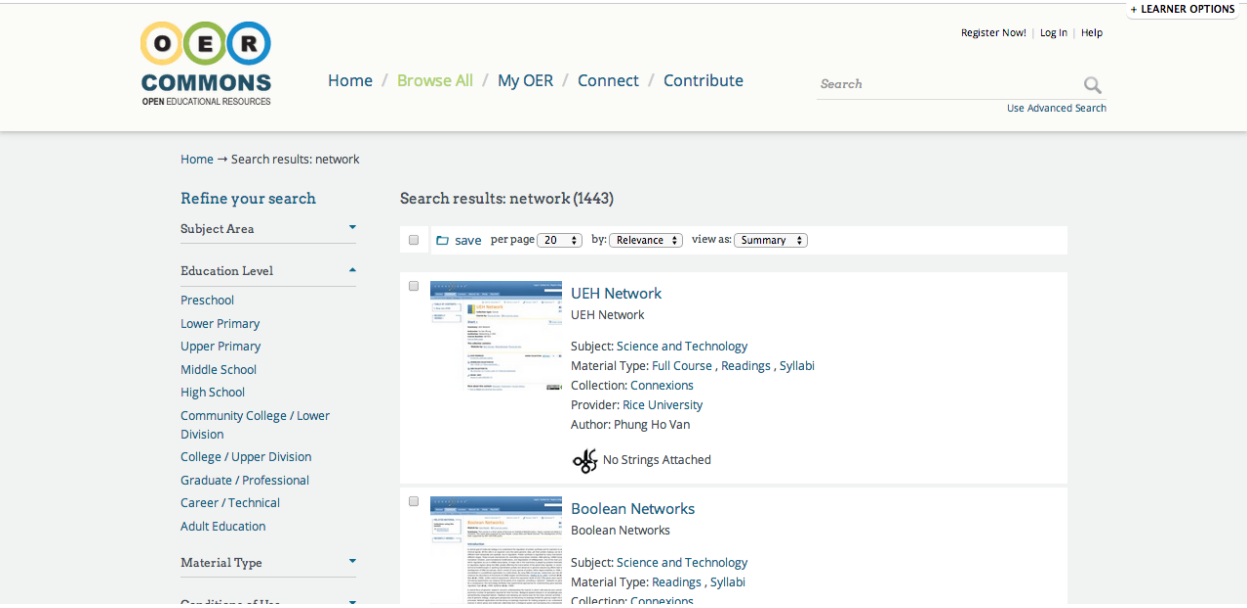

Figure xxxxx: OER Commons online discovery service

The online service of the OER Commons is a very useful resource to study the good practices in terms of user experience, tools and methods that can be used to build OER federation services.

Learning Resource Exchange of the European Schoolnet

The Learning Resource Exchange (LRE) from European Schoolnet is a service that enables schools to find OERs from many different countries and providers. The principle upon which the LRE is based is very simple. The LRE collects descriptions (i.e., metadata) of OERs and compiles them into a searchable catalog that can be consulted by users of connected e-learning platforms *[Unlocking Open Educational Interaction Data, http://www.dlib.org/dlib/may13/massart/05massart.html]*.

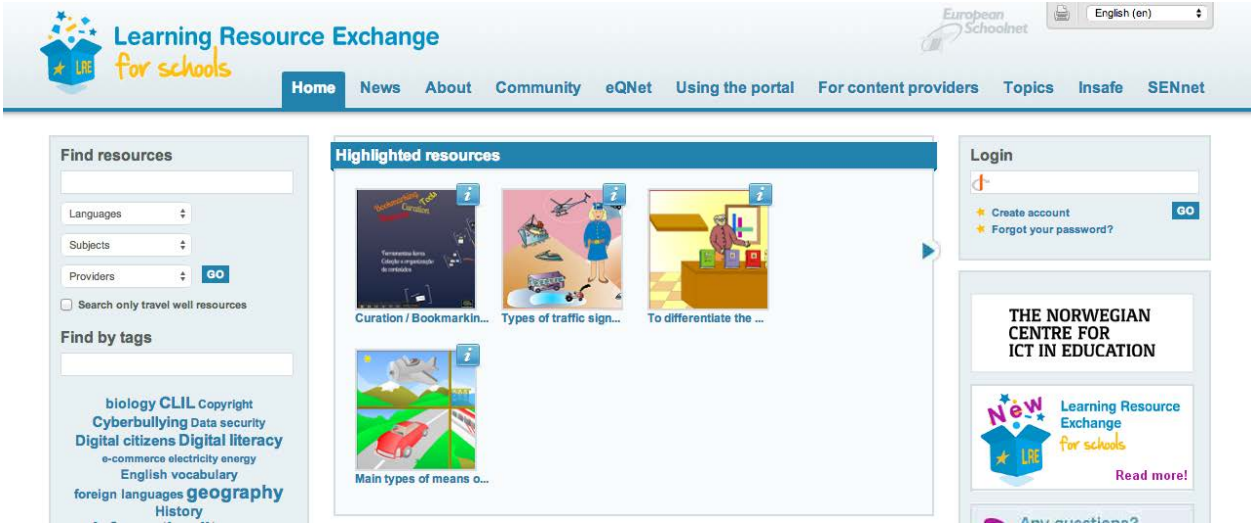

Figure: The OER online service of the LRE

The LRE aggregates metadata following several approaches to obtain metadata. Metadata is acquired from existing metadata repositories, automatically generated, or manually produced by human indexers. The LRE brings users to the providers and all further actions involving the use of the resource such as downloading, interacting with applets or playing videos or games occurs on the content providers' or users' environments.

TAACCCT project

TAACCCT is a $2 billion program of the US Department of Labor that provides community colleges and other eligible institutions of higher education with funds to expand and improve their ability to deliver education and career training programs that can be completed in two years or less, are suited for workers who are eligible for training under the TAA for Workers program, and prepare program participants for employment in high-wage, high-skill occupations. Through these multi-year grants, the Department of Labor is helping to ensure that our nation's institutions of higher education are helping adults succeed in acquiring the skills, degrees, and credentials needed for high-wage, high-skill employment while also meeting the needs of employers for skilled workers. The Department is implementing the TAACCCT program in partnership with the Department of Education.

The TAACCCT program is the largest OER initiatives in the world. A discovery services for all the OERs published in the context of the TAACCCT program is provided here http://open4us.org/find-oer/.

Learning registry

The Learning Registry is an effort jointly funded by the Department of Education and the Department of Defense. The effort began in 2010 and creates a set of technical protocols for the exchange of data in support of educational goals by multiple providers. It is not a portal, but a platform for aggregating metadata and other data about learning objects.

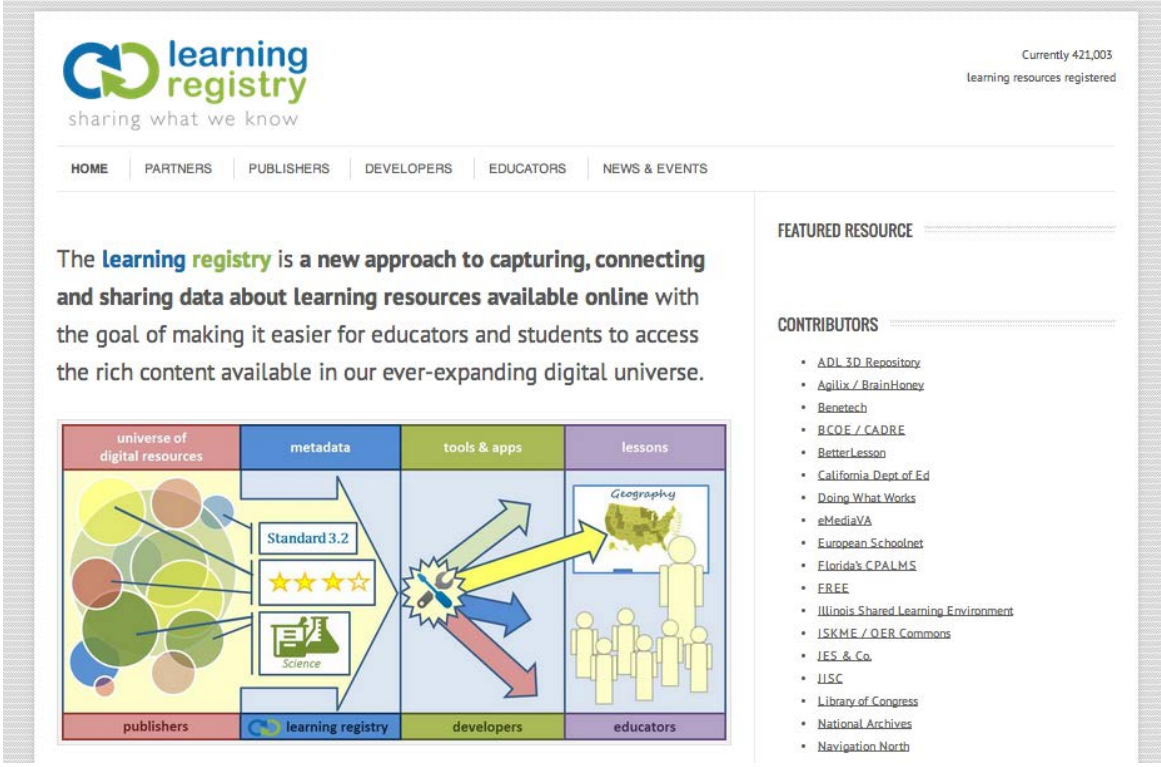

Figure xxxxx: The learning registry online service

The learning registry provides access both to commercial and OERs. However, the framework and technical approach are of great interest for OER initiatives.

Open Discovery Space

Open Discovery Space is a EU funded project that aims to serve as an accelerator of the sharing, adoption, usage, and re-purposing of the already rich existing educational content base. It will demonstrate ways to involve school communities in innovative teaching and learning practices through the effective use of eLearning resources. It will promote community building between numerous schools of Europe and empower them to use, share and exploit unique resources from a wealth of educational repositories, within meaningful educational activities. In addition, it will demonstrate the potential of eLearning resources to meet the educational needs of these communities, supported by European Web portal: a community-oriented social platform where teachers, pupils and parents will be able to discover, acquire, discuss and adapt eLearning resources on their topics of interest. Finally, it will assess the

impact and document the whole process into a roadmap that will include guidelines for the design and implementation of effective resource-based educational activities that could act as a reference to be adopted by stakeholders in school education.
One of the main outcomes of the ODS project will be an online OER discovery service with community functionalities (portal.opendiscoveryspace.eu) that will provide access to more than 0.5 million educational resources. The OER aggregation service is built using an evolution of the Ariadne tools. The front end is developed using the open source CMS Drupal.

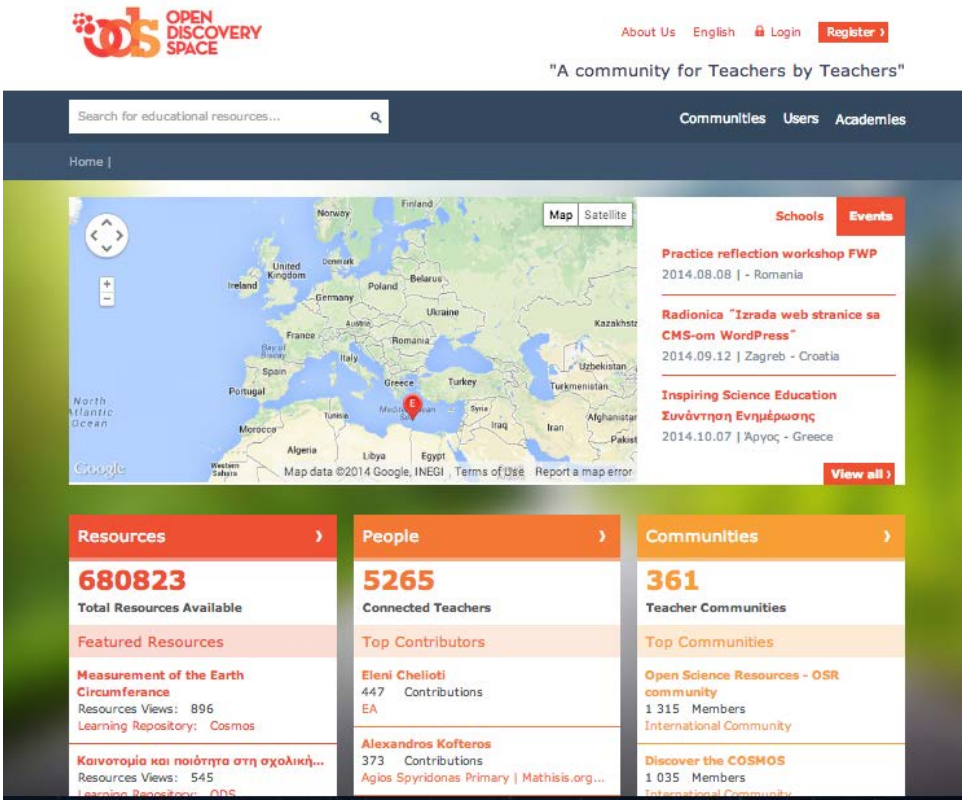

Figure xxxxxx: The OER discovery service of the ODS project

Gooru

The Gooru platform (http://goorulearning.org/) is a free, teacher curator multimedia search engine for K-12. It provides teachers and students with the ability to search for free and open educational resources and develop collections of multimedia resources, digital textbooks, videos, handouts, games, and quizzes, specifically around science, math and social studies and language arts. The data model that is used is based on the suggestions of the Learning Resource Metadata Initiative (http://www.lrmi.net/).
In essence Gooru provides three main funcions:

- users can search for resources on Gooru by entering a keyword and filtering by subject, grade level, standard, and/or source. Gooru has indexed millions of K-12 free and open education resources from providers such as Learnzillion, New Global Citizens, Autodesk, and NASA. Gooru has several libraries where users can easily find organized

courses from these sources: the Community Library curated by users, Partner Libraries curated by content partners, and Standards Library with Common-Core aligned courses.

- Gooru provides a mechanism for teachers and students to gather and grow collections of content. Gooru users can put together “collections” and populate them by dragging and dropping resources and questions from the search page or uploading their own documents or web URLs. Collections can be made private or public, and if public, can be search for by the Gooru community (such as this collection on the Silk Road).
- Gooru allows a teacher to view collection analytics to track the students in their class as they move through the collections they’ve been assigned. Classes can be “open” shared via link or Class Code, or invite-only. When students join the class, they can give permission for the teacher to view their progress.

Although it is not focused on the higher education, the Gooru case is of interest to TERENA OER pilot due to the tools and APIs that are provided to the developers (http://developers.goorulearning.org/).

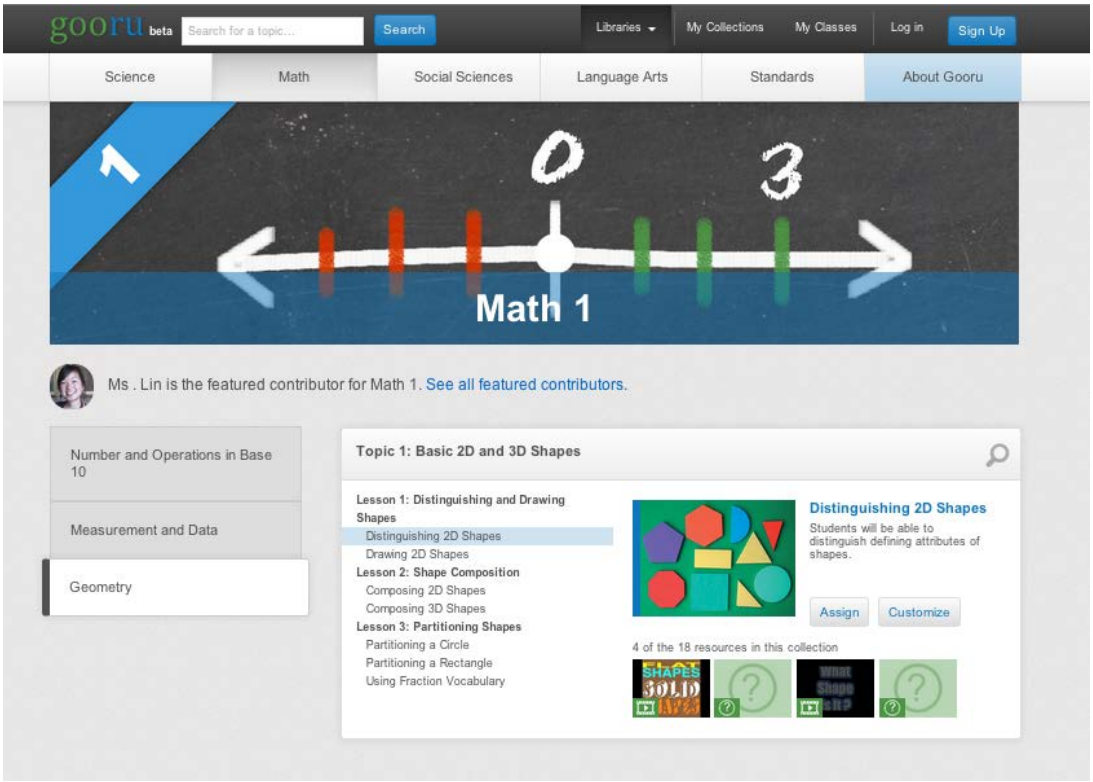

Figure xxxx: OER discovery service provided by Gooru

MERLOT

It is mainly repository so we should not include it in our state of the art study

Technologies and standards used for metadata aggregation

Technologies that are currently used in the various initiatives for aggregating OERs' metadata are the following:

- NoSQL approaches for storage of metadata such as MongoDB
- DSpace with custom editor for educational metadata
- Content management and authoring systems for the annotation of OERs such as the Agricultural Learning Repository tool (http://wiki.agroknow.gr/agroknow/index.php/AgLR)
- Search engines for the indexing of metadata records based on Apache Lucene
- REST architectural style for building the web services

The main standards and protocols used to develop educational aggregation services

- OAI-PMH harvesting protocol
- IEEE LOM standard for learning object metadata
- The simple publishing protocol (http://www.dlib.org/dlib/september10/ternier/09ternier.html)

Open source software

There is a big number of open sources tools related to the aggregation of metadata for several types of data. However, a small number of these tools focus on the educational domain. An example of an effort that is focused on the educational domain is the Ariadne Foundation. Ariadne members have developed open source software that can be used and adapted in order to build an aggregation engine for educational content. More specifically the main tools that are provided as open source software are:

- Metadata harvester based on the OAI-PMH protocol
- Metadata repository based on the IEEE LOM metadata standard for learning objects
- Metadata registry that stores the information for the organizations and collections that are connected to a federation service
- Metadata validation service for the IEEE LOM based application profiles

All the tools that can be found at the technical wiki of the Ariadne Foundation (http://wiki.ariadne-eu.org/dev/). The Ariadne tools have been evolved in the context of initiatives like the Open Discovery Space.

The learning registry provides also a set of tools to developers to allow the publishing and management of educational resources (https://github.com/LearningRegistry/).

As regards the social and usage data generated in the educational applications the AGL initiative provides the Experience API, which is set of services that can be used to store, manage and publish the social and usage data (http://www.adlnet.gov/tla/experience-api/).

Requirements for building an aggregation infrastructure for TERENA OER pilot

The generic requirements for the OER aggregation engine are:

- Be open
- Be modular
- Be scalable

The specific requirements for the TERENA OER aggregation engine are summarized in the following table. These requirements were identified in previous EU initiative e.g. Organic.Edunet, Open Discovery Space, LRE of European Schoolnet *[References]* based on real problems that organizations managing educational federation services are facing.

Requirement name/id	**Description**
Content workflow support	it should support for a number of different content workflows for the creation, curation and publishing of metadata records
Standards based	Support educational metadata and standards e.g. IEEE LOM
OER suggestion	it should support the the suggestion of resources by the users through a mechanism that can be integrated at the data discovery application
Metadata transformation	XSL based transformation of the metadata
Metadata ingestion/harvesting	Ingest metadata for educational content of various format
Metadata enrichment	the enrichment of the ingested metadata using manual, semi-automated and automated methods,
Metadata filtering	it should support the filtering of metadata in order to keep from each source only these metadata records that are relevant to the scope of the TERENA OER pilot
Publishing metadata	it should publishing of high quality metadata records to external systems so they can get the latest versions of the aggregated metadata
Search metadata	it should allow the search of the aggregated metadata records through a RESTful and well documented Search API
Automated ingestion and	it should support the automatic ingestion and processing of

processing	metadata records from diverse sources,
Content dissemination	it should allow the development of numerous front end applications/local interfaces and thus the high dissemination of the content.
Taxonomies/classifications support	it should support the organization of content based on the taxonomies and mapping of the different taxonomies used by the content providers
Multilinguality	it should support multilingual vocabularies and id based management of vocabularies in the db
Machine interfaces	it should allow the development of several apps based on the OER content

The above table can be used as a checklist at the time of delivery of the aggregation engine.

TERENA OER aggregation engine architecture

The following diagram shows the proposed architecture of the aggregation engine for the TERENA OER pilot.

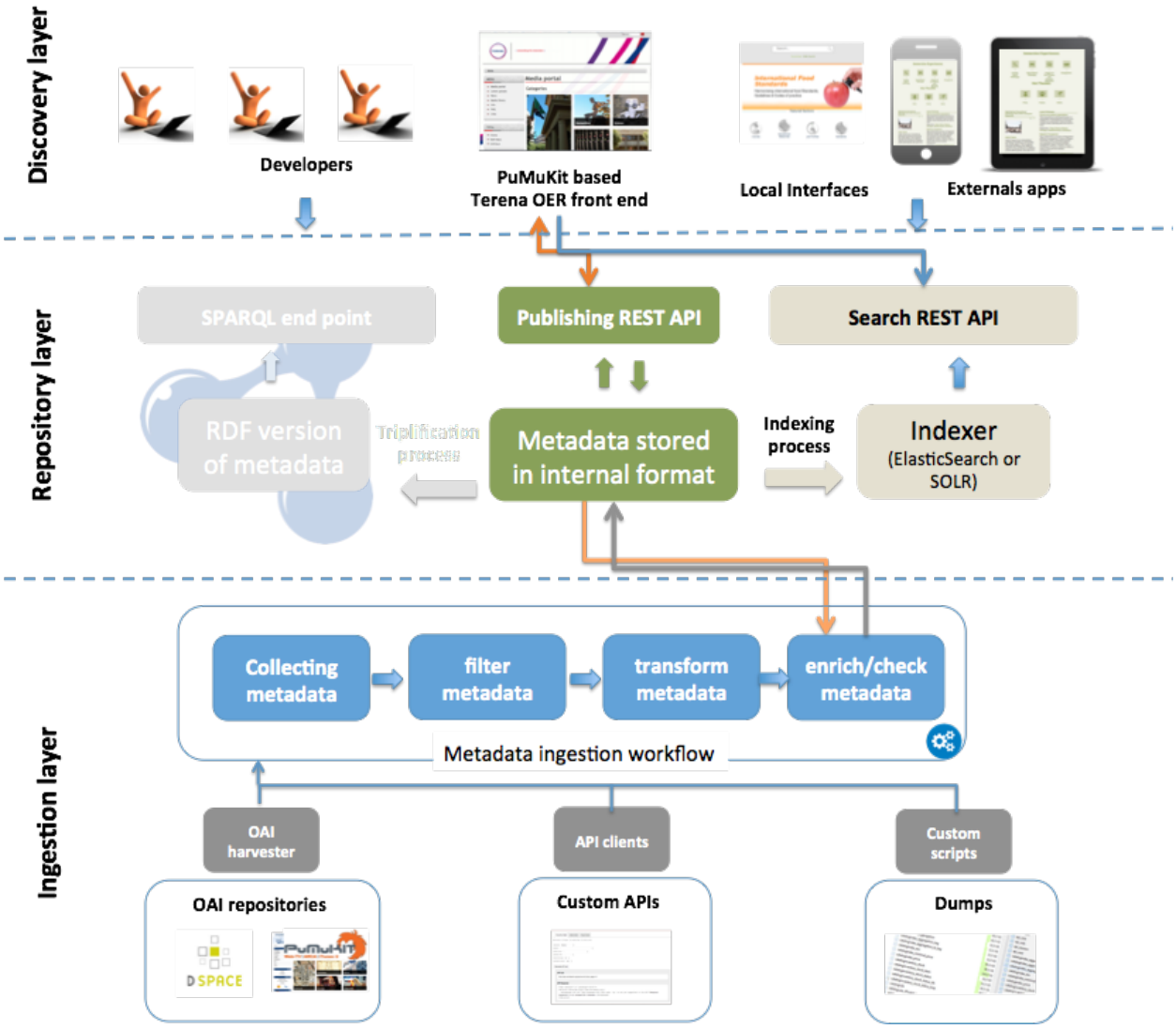

Figure 1: TERENA OER aggregation engine architecture

The proposed architecture for the TERENA OER aggregation engine is composed of three layers:

- **the data ingestion**, responsible for the aggregation and processing of metadata
- the **data repository,** responsible for the storage, indexing and publishing of the metadata records
- and the **data discovery layer** that allows the discovery of the content.

The layers are analysed in the following sections of this document.

The proposed architecture is open and scalable and will be developed using several well adopted open source tools and standards. The different layers are decoupled in order to support modularity and high availability of the TERENA OER front end services and open APIs. More specifically, the proposed architecture enables:

- the support for a number of different content workflows for the creation, curation and publishing of metadata records e.g. the operating organization can introduce a validation step in the workflow without changing radically the architecture by just introducing a new open source tool that will connect to the repository layer.
- automatic ingestion and processing of metadata records from diverse sources,
- the suggestion of resources by the users through a mechanism that can be integrated at the data discovery application,
- the enrichment of the ingested metadata using manual, semi-automated and automated methods,
- publishing of high quality metadata records to external systems,
- it allows the development of numerous front end applications/local interfaces and thus the high dissemination of the TERENA OER content.

As depicted in the architectural diagram, the architecture can be extended to also support the publishing of educational metadata in linked data format and their exposure through a SPARQL end point that can be used by external systems and developers. This is not at the core of the scope of the TERENA OER pilot but it just shows how the architecture could evolve in to this direction.

Data ingestion layer

The **data ingestion layer** which is responsible for the ingestion of content from various diverse sources that are publishing the metadata through custom API, OAI-PMH protocol and dump files. This layer will also include scraper for getting metadata from web sources. This is a layer totally decoupled from other layers of the architecture in order to ensure the efficiency and the high availability of the TERENA OER Service. The ingestion layer constitutes a workflow consisted of several steps for the processing of metadata records. More specifically, it will include

- a step for the transformation of metadata records from the input format to the IEEE LOM based AP of the TERENA OER,
- a step for filtering the metadata records,
- a step for the enrichment of the metadata records (e.g. adding the missing lang attributes and assigning semantic tags) and
- a step for the url checking of the metadata records.

At each step of the ingestion workflow log files will be stored and indexed and a visual dashboard will be provided to the IT team to easily monitor the process and diagnose issues.

Data repository layer

The **data repository layer** which a) stores all the processed and ingested metadata in the IEEE LOM based data model and b) exposes the processed metadata through RESTful APIs so the external systems can consume the metadata records. The metadata records of the repository layer are accesses frequently by tools of the ingestion layer to check the quality of the metadata records. All the metadata records that will be stored at the metadata repository will be indexed using an open source search engine and will be published through a Search API. The search API will be presented and documented in a specific site of the TERENA OER pilot. External systems and developers will be able to access the Land Library data either using the search and/or the metadata publishing interface.

The data ingestion and repository layers are working in two modes, namely the data acquisition and the data maintenance mode. The primer is responsible for the acquisition of the data (new, updated) while the later is responsible for the checking and enrichment of the metadata records.

The repository layer will support multilingual indexing of the records by using multilingual vocabularies and taxonomies. To support coherent content discovery the mappings of the different taxonomies to a common taxonomy will be used to extend the index.

Data discovery layer

The **data discovery layer** that includes all the front end applications that are consuming the publishing and search APIs of the data repository layer. The following modules can be developed in a standard front end systems (e.g. PuMukit or Drupal) be developed to interact with the repository layer:

- A data discovery module that will support keywords search and faceted browsing. This module will use the elasticsearch based search API provided by the repository layer.
- A data publishing module that allow the publishing of web resources suggested by the users of the Land portal

The discovery layer will use the multilingual extension of the indexes in order to support searching and browsing of content in the various languages of the interface. The usage of automatic translation mechanism for the free text elements of the metadata records can be explored at a later stage of the TERENA OER project.

OER aggregation engine Specifications

Metadata aggregation workflow

The metadata aggregation workflow of the ingestion and repository layer includes a number of steps for the acquisition and maintenance of the metadata records from different content providers and from UGC.

--

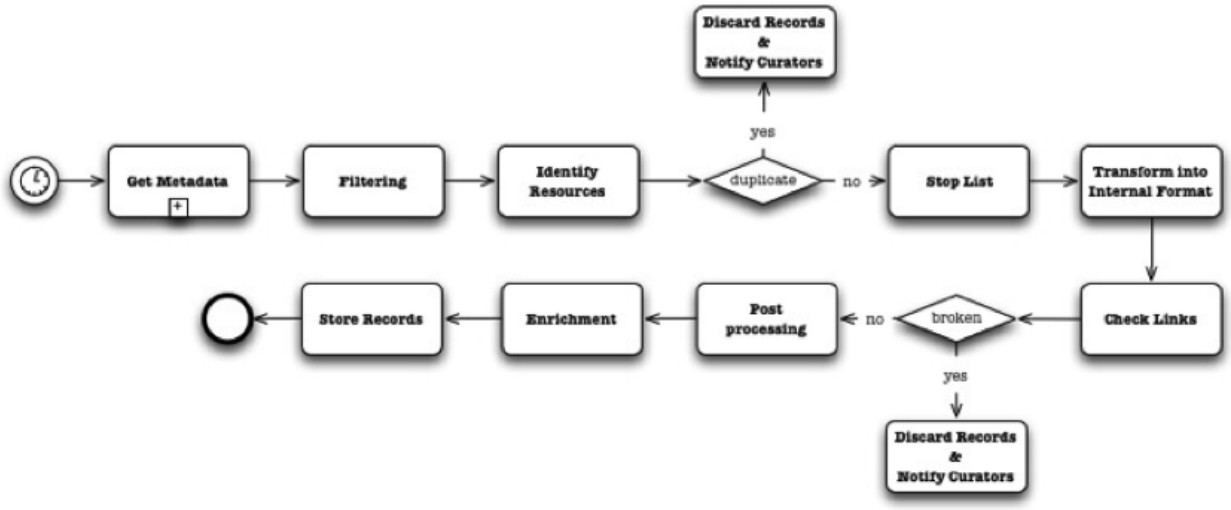

Figure XXXXXXXX. The TERENA OER metadata aggregation workflow

The various steps of the aggregation workflow are presented in figure XXXXX. More specifically the workflow for metadata acquisition includes:

- **The harvesting step:** the first step consists of harvesting all the metadata records from a remote site of a content provider. An open source tool for harvesting that was developed based on the software proposed Ariadne Foundation (http://www.ariadne-eu.org/). The target is first validated and then the metadata are records are harvested in the original metadata schema and stored in the file system. In several cases selective harvesting is enabled and only particular sets of metadata records of a target are being harvested.
- **The filtering step:** filtering is a step consisting of discarding incoming records considered as inappropriate either because the object it describes is inappropriate (e.g., in a collection of educational resources, discarding metadata describing resources covering topics not related to TERENA OER e.g. OERs that are only for the primary school) or because the record is syntactically incorrect. The latter can be seen as a light form of validation that focuses on detecting errors that can potentially compromise the correct functioning of the aggregation service. The filtering step can be “set”-specific in the sense of that each set potentially requiring a different filter
- **The identification and deduplication step:** during this step, a software component can be used to compare new metadata records to the existing ones to see if the objects they describe are already referenced in the catalogue. Each new learning object receives a unique identifier, and the corresponding metadata file is renamed accordingly and saved in a new directory. The organisation of this directory is based on the providers’ name. Duplicate records (proved or suspected) are stored in a separate directory where they can be examined by the curators who can decide how to handle them. The options are the following ones:
    1. Keep the existing record
    2. Replace the existing record by the new one

3. Update the existing record with some of the information available in the new record

By default, the workflow protects the existing records by discarding the new ones.

- **Stop list:** this step is being used to stop records that are not appropriate for the audience, they are systematically wrong or include errors in the metadata that cannot be identified in the filtering process. Uniquely identifying resources makes it possible to maintain a list of resources that have been considered as not suitable for the catalogue by the collection curators and to make sure these resources are not reintroduced in the catalogue.
- **Transform into internal format:** this step is used to transform the XML versions of the metadata records to JSON files that follows the principles of an abstract data model for describing educational resources but it also includes additional facets that can support better multilingual requirements such as to have two language versions of the same learning object which is not currently supported by standard such as IEEE LOM. This step requires transformers capable to convert the various formats and application profiles of the metadata records collected at step 1 into the internal format. In order to keep this step as simple as possible, we ensure that the metadata records belonging to the same set have all the same format. At the end of this step, identified records in the internal format are stored on the file system.
- **Link checking:** this step is responsible for checking if the URL for accessing the learning object is broken or not. For all learning objects for which the location included in the metadata record has been recognised as broken, the index is updated accordingly in an automatic way. Also the metadata records with broken URLs are stored in a separate directory and are being re-checked in a more frequent basis in the context of maintenance mode. In case of URLs that redirect to other URLs these can be considered to be broken and they have to be evaluated by curators. In order to facilitate such kind of tasks after each run of the link checker the numbers of broken links per domain are stored into log files so that it is clear in which domain there are several problems.
- **Post processing:** there are cases in which there is a need to normalize the metadata records in order to avoid problems in the front-end applications. Such example is the normalization of language attributes values for title in English which may be provided either using “en” or “eng”. In this case the post processing step will normalize all the values so they can use the correct ISO code for the language.
- **Enrichment:** this step can be used to enrich the metadata elements of some collections. For instance if a collection is not using a specific thematic classification, then an automatic annotation tool like Maui (https://code.google.com/p/maui-indexer/) can be used to extract the classification concepts from the description and keywords elements.
- **Store and publish records:** the final step of the metadata aggregation workflow is the storage at the repository layer of all the new metadata records that have successfully passed the deduplication and URL checking step. They are stored on the file system where they are organized by sets. This consolidated metadata store is exposed to a web server so that records can be easily access online. A typical URL is of the form http://catalogue-name/FORMAT/set/identifier.extension,

e.g. http://terena-metadata /LOM/GREENOER/12345.xml. Also, this step consists of the metadata publishing through APIs.

Figure 9. Indexing logs with elastic search and visualizing them with Kibana

The log files from each step of the metadata aggregation workflow can be parsed with Logstash and then indexed with Elastic Search, that is a distributed, RESTful, free/open source search server based on Lucene and developed in Java. Finally logs can be visualised with Kibana tool, an open source (MIT License), browser based interface on Logstash and ElasticSearch. Therefore, it is easy for the metadata curators to get an overview of the results from each step of the workflow e.g. of the broken URLs and the redirections through visualizations of the outputs of the link-checking step.

Figure XXXXXX: Visualizing the workflow steps and the results of link-checker

At this point we should also mention that we can take advantage of the technology of Elastic Search indexer and Kibana for providing to metadata curators dashboards with metadata analysis according to dimensions such as the frequency of languages, completeness, frequency and entropy. Figure XXXXX depicts a sample dashboard.

Figure XXXX. Sample of the aggregation workflow dashboard

Internal data model

The repository layer will store all the identified and processed metadata records in a internal data model that is based on a standard metadata schema for educational applications, namely the IEEE LOM *[Standard for Learning Metadata Object, http://grouper.ieee.org/groups/ltsc/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf]*. A JSON binding of the IEEE LOM based AP will be used internally in the repository to facilitate the processing and indexing of metadata. This will be done mainly for efficiency reasons as parsing data in JSON format is much more faster compared to XML format. An example of such internal metadata schema is presented in the following diagram.

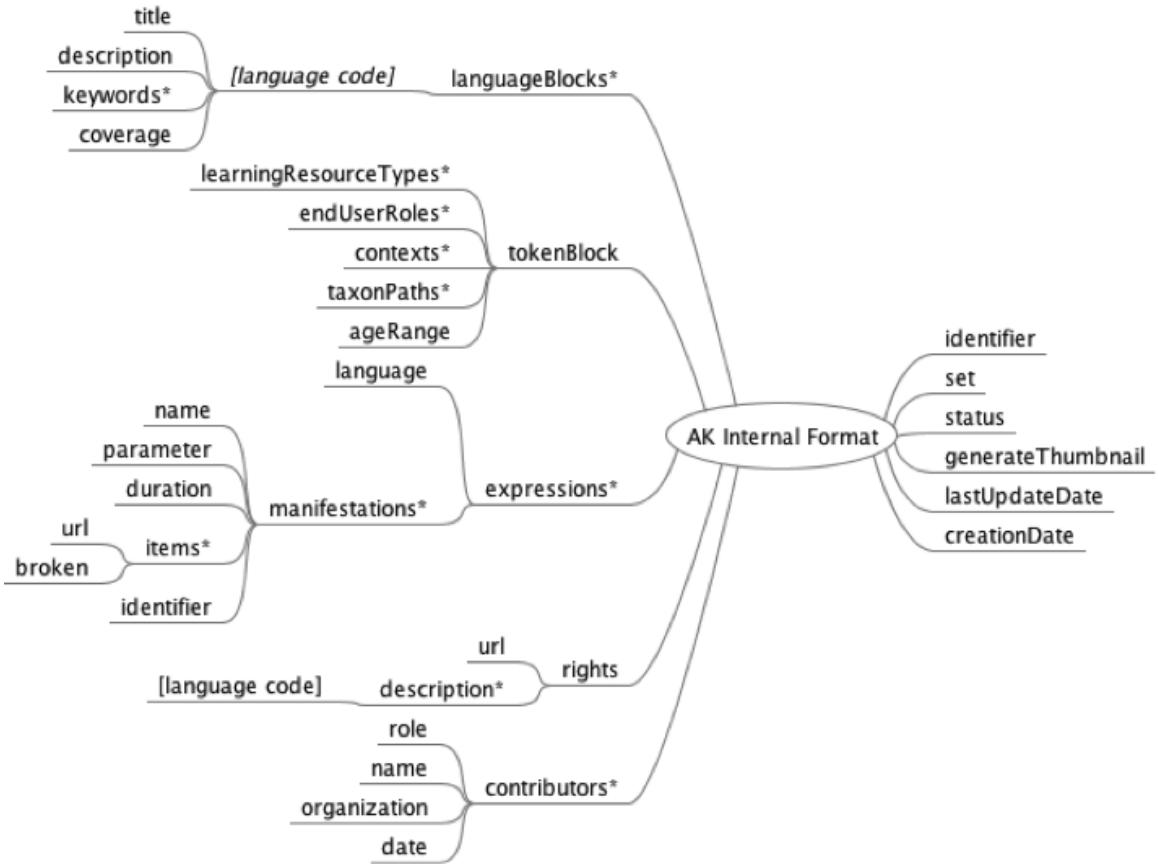

Figure XXXX. An example of an internal format representation used for the educational metadata

Software interfaces for the interaction with external systems

OAI-PMH

The metadata aggregator will support the publishing of all the metadata through the OAI-PMH protocol. The metadata are exposed in IEEE LOM format. The OAI-PMH target can be used by

- other components of the TERENA OER architecture such as the TERENA OER portal to get all the metadata records..
- By external referatories of learning resources such as the Open Discovery Space (http://www.opendiscoveryspace.eu/) and the Open Education (http://www.openeducationeuropa.eu/) patform of the EU.

An example of provided metadata through the OAI-PMH interface is presented later in this document (Section 8.2).

Specialized Search API

In order to provide a more flexible and scalable solution for covering more information needs and search options a REST service on top of Elastic Search indexer that allows easier metadata integration will be developed. The specialized Search API will be RESTful API that will allow several search options over the indexed metadata records (JSON files) following the internal

data model of the aggregation engine. In specific it will allow the user or application to make: 1) Simple search, 2) Searching within specific fields, 3) Temporal, 4) Fetching specific items, 5) Complex queries.

Publishing of metadata records status

As already mentioned the metadata aggregation workflow is an automated process that can aggregate and process metadata records for OERs from diverse sources. The periodicity of the process can be set by the administrator of the workflow per each collection. In order to inform external systems about the status of the metadata aggregation repository a simple REST interface will be developed. More specifically, the OER aggregation workflow will publish the status of the metadata records that have been ingested and processed through a simple REST interface which will inform external systems which metadata records have been recently ingested, which have been updated and which have been deleted from the metadata aggregation repository. According to this information the external system can get all the updated or new records using either directly the catalogue of the repository layer or the Search API.

An example of such manifest file that could provide this information is shown below. Using this file the remote system can get from the metadata aggregation repository these records that should be ingested or updated in the local database.

```
{
   "params": {
      “service”: “TERENA OER”,
      "changed": 30,
      "date": “date_of_manifest_creation”,
      “previousDate”: “date_of_previous_manifest"
   },
   "resources": [
      {
         "resource": "resource_id_number",
         "location": "resource_location",
         "action": "(DELETE|UPDATE|HIDE|ADD)"
      },
      {
          "resource": "resource_id_number",
         "location": "resource_location",
         "action": "(DELETE|UPDATE|HIDE|ADD)"
      }
   ]
}
```

OER metadata publishing API

In order to support the suggestion of OERs through the TERENA OER portal, the aggregation engine needs to provide a simple publishing REST interface that allows the submission of metadata records to the repository layer.

OER aggregation engine software License

All the components of the TERENA OER aggregation engine will be available as an open source software under the licence of MIT license or LGPL.