e-Learning Object Ingestion in an Open Educational Environment

Enayat Rajabi, Kostas Vogias, Salvador Sanchez-Alonso, Ilias Hatzakis

***Abstract*— A large number of learning objects' metadata are available on the Web. Published by different sources, these metadata cannot be represented to the end users in their original formats, as they require some technical steps to be filtered, checked and cleaned due to several issues e.g. broken links. In this paper, we present an aggregation workflow followed in an open educational environment (the Open Discovery Space project) in which a large amount of metadata passed through several technical steps as a pre-filter to be integrated into this educational Web portal.**

***Index Terms*— Repository, Aggregation, Metadata, Open Discovery Space.**

I. INTRODUCTION

In the past decades, educational objects have been published by various suppliers of educational services on the Web [1]. These resources targeted to certain categories of learners, which can be students, teachers, employees, etc. On the other hand, researchers and repository owners in the educational domain have developed various e-learning systems to aggregate, publish and consume plenty of these e-learning resources, so that they can be discovered, navigated and reused by different kinds of applications and users on the Web [2]. However, low quality metadata can render a library or repository almost unusable, while ingesting metadata with high quality can lead to higher user satisfaction [3]. To this aim, data publishers utilize different approaches for filtering, enriching, and checking metadata. In this paper, we describe an aggregation approach in which a large number of metadata were harvested from several eLearning repositories, validated to be aligned to the defined metadata schema, and transferred to the proper application profile. In this experimentation, we filtered out the useless and low quality metadata as well. The final resources were imported to an eLearning repository as well.

The work presented in this paper has been part-funded by the European Commission under the ICT Policy Support Programme CIP-ICT-PSP.2011.2.4-e-learning with project No. 297229 "Open Discovery Space (ODS)".

Enayat Rajabi is a PhD researcher at the Computer Science Department of University of Alcalá, Spain. email: enayat.rajabi@uah.es

Kostas Vogias is a Software Engineer on metadata aggregation, processing and exposure at GRnet. email: gvog84@gmail.com

Salvador Sanchez-Alonso is an associate professor at the Computer Science Department of University of Alcalá, Spain. email: salvador.sanchez@uah.es

Ilias Hatzakis is a metadata aggregation related project manager, at GRnet. email: hatzakis@grnet.gr

The reminder of this paper is organized as follows. Section 2 describes the metadata aggregation in open learning environment and introduce the Open Discovery Space project as the case study we selected for this research. The methodology of the present study is outlined in Section 3. Section 4 presents the challenges we faced in the aggregations of metadata within the project. The final remarks are presented in section 5.

II. BACKGROUND

The exponential growth in the amount of digital learning objects is forcing architects, engineers and developers involved in creating digital repositories to face the harsh reality that their solutions need to handle an amount of e-objects that is orders of magnitude larger than originally intended. Optimizing, tuning, and tweaking the existing repository infrastructure can initially alleviate performance problems, but eventually limits are reached. At that point, a major redesign of the repository solution is an obvious option. An alternative is to move towards an environment that consists of parallel instances of the existing repository solution and to glue those together into a repository federation that behaves as if it were a single repository [4]. The desire to federate repositories in such a way typically emerges as a result of the understanding that no single repository hosts all e-learning objects that are relevant for a specific subject domain. Generally speaking, federation is a decentralized approach that emphasizes partial, controlled sharing among repositories and provides a means to share data and transactions using some protocols such as OAI-PMH [5] and providing the coordination of data exchange among them. A federation supports interoperability among the registered repositories and reduces dependency on any single metadata collection. Figure 1 illustrates in a simple way a federation of repositories.

Fig. 1. Harvesting a federation of repositories

In a federated repository metadata information is collected from many contributors or repositories to create, on top of all of them, a search service supporting simultaneous discovery of information resources residing in the collections of all the repositories. The way in which the metadata are collected is usually referred to as "harvesting", a computer software technique of extracting metadata information from external data sources by periodically accessing them, using a standard protocol agreed by the 2 parties (client and server). Collecting metadata through such standard protocol (e.g., OAI-PMH) has been utilized by a wide variety of projects including Open Discovery Project (ODS) [6].

The ODS project aims to support open access to digital educational resources and practices from members of school communities (that is teachers, students and parents) in Europe. This project, which is the result of collaboration between 51 partners from 23 European countries, exploits the elements of eLearning resources (i.e., educational objectives, pedagogical models, learners' personal characteristics and needs, etc) collected from many educational repositories and federations across the Europe. This ongoing project has promoted community building between numerous schools of Europe (At the time of this research 2,833 schools with around 7,600 teachers participated) and empowered them to use, share and exploit unique resources from a wealth of educational repositories. ODS uses an Open Linked Learning Content Infrastructure and has recently exposed around 800,000 eLearning metadata (from 25 eLearning repositories) through a Web portal.

In ODS, most of the harvested repositories conform to an Application Profile, which is based on an IEEE LOM schema and created within the project, and provided their eLearning metadata according to it. However, some repositories have their own custom schema and a mapping between their schema and ODS Application profile was carried out after harvesting by means of producing a XSLT file (specific of each repository) and running a transformation on the harvesting infrastructure side by using transformation/alignment tools. The production of these XSLT files was the responsibility of each repository assisted by some technical experts in alignment and integration mentioned in the description of work (Figure 2).

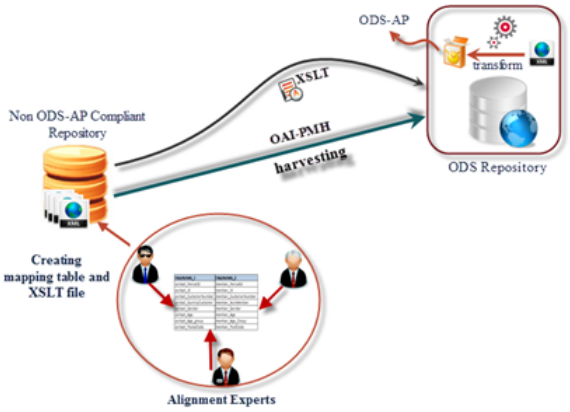

Fig. 2. harvesting process

III. METHODOLGY

The harvested metadata in ODS have passed the following steps to be imported in the ODS portal (consider Figure 3):

A. Repository Catalog

The Repository Catalog is a CRUD software developed for maintaining repositories' related information. This information is used by the aggregation workflow software in order to configure the various constituent steps.

B. Harvesting

This is the first step in the metadata aggregation workflow. The harvesting protocol used is OAI-PMH.

C. Transforming

In this step, each repository metadata were transformed to the ODS AP using the XSLT file they sent. Each repository generated its own XSLT file and sent it to the ODS repository owner to perform the transformation step. As a result of this step, all the repositories metadata were transformed to the ODS schema as well.

D. Identification

In this step, each learning object and its metadata were identified by a global identification approach.

E. Validation

In the validation step, each repository data were validated against the ODS AP and thus the invalid records were filtered out. The main reasons for a record to be considered as invalid are the following:

- Mandatory elements absence
- Incorrect vocabulary mapping

F. Filtering

Those metadata that did not include any text in the mandatory elements (e.g learning object's title and location) were filtered out.

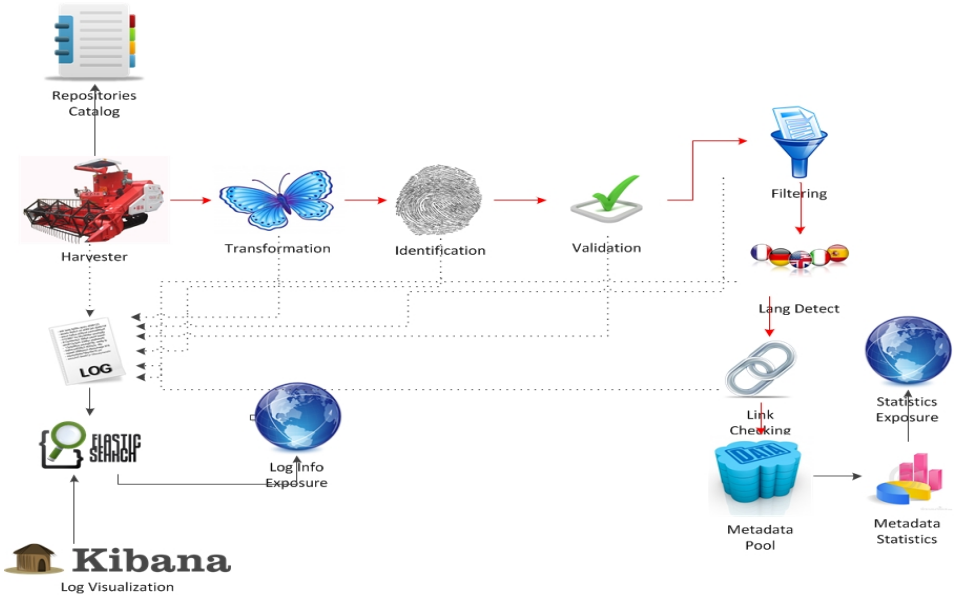

Fig. 3. Aggregation workflow

G. Language enrichment

In this step, if there are some metadata elements that it is important from ODS Portal perspective to contain a language attribute(title, description, keywords etc), then they are enriched with it using a language detection software [7].

H. Link checking

Figure 4 represents the link checking process in ODS, which is part of the metadata aggregation system. Overall, the ODS harvester (1) collects repositories' metadata based on ARIADNE-powered infrastructure. The harvested metadata are stored in a file system, separated by their harvested source in different folders. The link checker engine checks (2) each learning resource individually by testing the URL contained in the respective resource metadata, as mentioned in the previous paragraphs. Information regarding to broken links (e.g. File name, file path, status, timestamp) is stored in a log file (3). Non-broken links, what we call usable resources, are moved in different folders (4). Resources that contain broken links are checked (7) periodically (definite periodicity is still under discussion) and they will be moved to live folder (8) for joining to usable resources, if the link is live again. The recovered resources will be re-indexed in the next harvesting cycle (5). The link checking results are also exposed with a variety of REST APIs. The whole idea is these APIs to be consumed by ODS Portal in order this to be informed about the link checking results and perform the necessary updates to its data base.

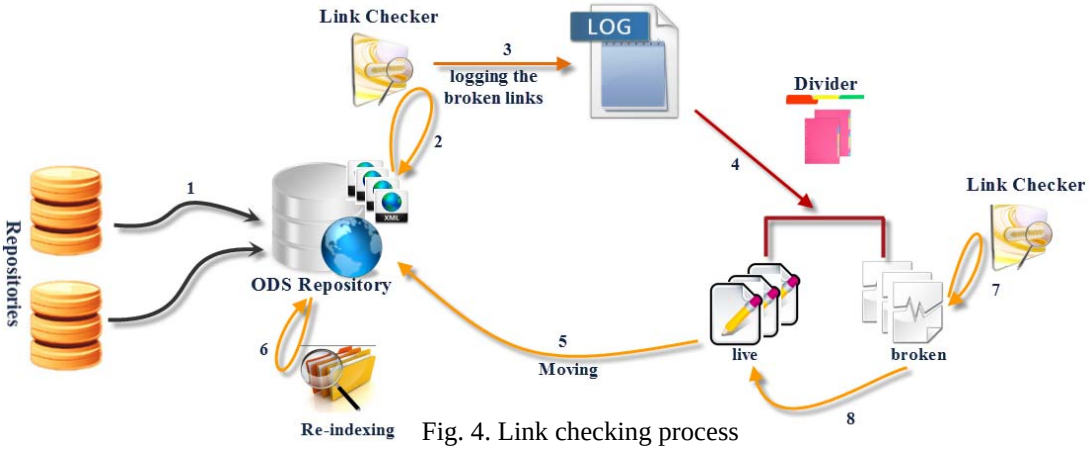

Fig. 4. Link checking process

I. Metadata Pool

The metadata pool is a file system folder where all the usable metadata records are placed.(The name is an abstraction of the file system folder that contains the final usable metadata.)

J. Metadata Statistical Analysis and visualizations

The aggregation workflow is accompanied with a metadata analysis tool. This tool was developed using JAVA and its purpose is to analyze XML documents and export the results to a human readable format (CSV) [8]. s a first step, it performs a per repository analysis and as a second step it performs the analysis of the repositories at the aggregation level. The statistical measures that are being calculated are: element frequency, element completeness, element dimensionality and element content relative entropy [9]. It is also possible for the user to additionally choose a specific element for vocabulary usage analysis (frequency). Finally, an attribute based value analysis (attribute value frequency) is also implemented that can be used to study the multilinguality of the free text metadata elements. This tool proved to be very helpful for aggregation specialists and data providers to gain a

deeper knowledge of the quality and content anomalies of the aggregated metadata.

K. Aggregation Results Visualization

A need that came up after the first harvesting cycle was the metadata aggregation results and statistics visualization. In this way the aggregation analysts, data providers and project managers became able to supervise the whole aggregation process.

L. Ingesting in the portal

Using an updater the finalized metadata are read and inserted into a relational database which is later used for the portal.

Fig. 5. Aggregation workflow results visualization

IV. Challenges

The challenges we faced in the aggregation steps are described below:

Harvesting: At this point the most usual problems occur on data providers' side and are mostly OAI-PMH protocol implementation issues:

- Bad timestamp implementation. This disables the harvester's ability to recognize the new and also the updated records thus incremental harvesting approach cannot work.
- Bad resumption token implementation. This makes the harvester unable to harvest all the metadata records exposed by data providers. The harvester freezes to a certain record and can't harvest the rest metadata.
- No deleted records policy: The absence of a deleted records policy makes the harvester and the whole aggregation workflow unable to recognize what records should be deleted from the respective repository.
- Identification: If a metadata record doesn't use any element that describes the learning objects' existence (in LOM the technical.location and the general.identifier elements) then the specific metadata describes nothing, therefore the respective metadata record is discarded.
- Language Detection: The language detection mechanism could be considered as second filtering step since it filters out all metadata records that contain elements the content of which can't be language detected. The most usual case here is a metadata element that contains various symbols as text (this record passed successfully validation and filtering steps but should not be presented to ODS Portal though).
- Link Checking: A very common issue at this step is the case when a Learning Object's link although it is live, accessible and well formed however it points to a login page or a web representation of the metadata that describes it instead of the actual Learning Object.

V. Conclusion

In this paper, we described a workflow in which a large number of eLearning metadata, were collected from several repositories, processed, cleaned, evaluated and finally imported into a learning portal. After defining a schema for structuring the metadata, a large number of metadata were harvested by ARIADNE harvester, and then we separated the healthy metadata by checking their schema and contents. Particularly, we filtered out the metadata with broken links and empty titles, and finally we imported the cleaned data into the portal. The aggregation workflow described above could be considered as generic enough, covering most of the learning object's metadata processing needs. Thus with very small adaptations (mostly at metadata schema level) the specific workflow could be used as a ruler for future learning object metadata repositories.

References

[1] M. Fernandez, M. d' Aquin, and E. Motta, "Linking data across universities: an integrated video lectures dataset," in *Proceedings of the 10th international conference on The semantic web - Volume Part II*, Berlin, Heidelberg, 2011, pp. 49–64.

[2] M.-C. Valiente, M.-A. Sicilia, E. Garcia-Barriocanal, and E. Rajabi, "Adopting the metadata approach to improve the search and analysis of educational resources for online learning," *Comput. Hum. Behav.*

[3] B. Stvilia, L. Gasser, M. B. Twidale, S. L. Shreeves, and T. W. Cole, "Metadata Quality for Federated Collections," Nov. 2004.

[4] H. Van de Sompel, R. Chute, and P. Hochstenbach, "The aDORe Federation Architecture," *ArXiv08034511 Cs*, Mar. 2008.

[5] "Open Archives Initiative: Protocol for Metadata Harvesting." [Online]. Available: https://www.openarchives.org/pmh/. [Accessed: 29-Apr-2015].

[6] "Open Discovery Space Project." [Online]. Available: http://www.opendiscoveryspace.eu/. [Accessed: 28-Apr-2015].

[7] "Language Detection Library for Java." [Online]. Available: https://code.google.com/p/language-detection/. [Accessed: 05-Jan-2015].

[8] K. Vogias, I. Hatzakis, N. Manouselis, and P. Szegedi, "Extraction and Visualization of Metadata Analytics for Multimedia Learning Repositories: the case of Terena TF-media network," in *Proceedings of the LACRO 2013 Workshop*, 2013.

[9] "GLOBE Metadata Analysis." [Online]. Available: https://sites.google.com/site/globemetadata/home.