A Survey on Linked Data and the Social Web as facilitators for TEL recommender systems

Stefan Dietze[1], Hendrik Drachsler[2], Daniela Giordano[3]

[1] L3S Research Center, Leibniz University, Hannover, Germany
dietze@l3s.de
[2] Open University of the Netherlands, CELSTEC, Heerlen, The Netherlands,
hendrik.drachsler@ou.nl
[3]DIEEI, University of Catania, Catania, Italy
dgiordan@diit.unict.it

Abstract. Personalisation, adaptation and recommendation are central features of TEL environments. In this context, information retrieval techniques are applied as part of TEL recommender systems to filter and recommend learning resources or peer learners according to user preferences and requirements. However, the suitability and scope of possible recommendations is fundamentally dependent on the quality and quantity of available data, for instance, metadata about TEL resources as well as users. On the other hand, throughout the last years, the Linked Data (LD) movement has succeeded to provide a vast body of well-interlinked and publicly accessible Web data. This in particular includes Linked Data of explicit or implicit educational nature. The potential of LD to facilitate TEL recommender systems research and practice is discussed in this paper. In particular, an overview of most relevant LD sources and techniques is provided, together with a discussion of their potential for the TEL domain in general and TEL recommender systems in particular. Results from highly related European projects are presented and discussed together with an analysis of prevailing challenges and preliminary solutions.

Keywords. Linked Data, Education, Semantic Web, Technology-Enhanced Learning, Data Consolidation, Data Integration

1 Introduction

As personalisation, adaptation and recommendation are central features of TEL environments, TEL recommender systems apply information retrieval techniques to filter and deliver learning resources according to user preferences and requirements. While the suitability and scope of possible recommendations is fundamentally dependent on the quality and quantity of available data, e.g., data about learners, and in particular metadata about TEL resources, the landscape of standards and approaches currently exploited to share and reuse educational data is highly fragmented.

This landscape includes, for instance, competing metadata schemas, i.e., general-purpose ones such as Dublin Core[1] or schemas specific to the educational field, like IEEE Learning Object Metadata (LOM) or ADL SCORM[2] but also interface mechanisms such as OAI-PMH[3] or SQI[4]. These technologies are exploited by educational resources repository providers to support interoperability. To this end, although a vast amount of educational content and data is shared on the Web in an open way, the integration process is still costly as different learning repositories are isolated from each other and based on different implementation standards [4].

In the past years, TEL research has already widely attempted to exploit Semantic Web technologies in order to solve interoperability issues. However, while the Linked Data (LD) [2] approach has established itself as the de-facto standard for sharing data on the Semantic Web, it is still not widely adopted by the TEL community. Linked Data is based on a set of well-established principles and (W3C) standards, e.g. RDF, SPARQL [6] and use of URIs, and aims at facilitating Web-scale data interoperability. Despite the fact that the LD approach has produced an ever growing amount of data sets, schemas and tools available on the Web, its take-up in the area of TEL is still very limited. Thus, LD opens up opportunities to substantially alleviate interoperability issues and to substantially improve quality, quantity and accessibility of TEL data.

In particular, we expect LD to facilitate TEL community with relevant datasets in order to gain more knowledge about personalisation of learning and build better recommender systems. So far the outcomes of different recommender systems and personalisation approaches in the educational domain are hardly comparable due to the diversity of algorithms, learner's models, datasets and evaluation criteria [43]. A kind of reference dataset is needed for the TEL recommender systems field, as is the MovieLens dataset[5] in the e-commerce field. Initial characteristics of such a reference dataset for TEL have been described in [43]. Recently, some initiatives like LinkedEducation.org and the Special Interest Group dataTEL of the European Association of TEL started to collect representative datasets that can be used as a main set of references for different personalisation approaches within TEL [46]. Data driven companies like the Mendeley reference systems[6] are pioneers with this respect as they provided a reference dataset for Science2.0 research [57]. Similar, initiatives for TEL are highly needed to stimulate data driven research for education. Recently, the SOLAR foundation for Learning Analytics presented a concept paper that also contributes to this idea, and outlines an Open Learning Analytics platform for online data-driven studies [44]. At the Learning Analytics and Knowledge Conference 2012, the first workshop on Learning Analytics and Linked Data (LALD12) has raised the idea to use LD sources as reference dataset for these kinds of research [45]. The

[1] http://dublincore.org/documents/dces/

[2] Advanced Distributed Learning (ADL) SCORM: http://www.adlnet.org

[3] Open Archives Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html

[4] Simple Query Interface: http://www.cen-ltso.net/main.aspx?put=859

[5] http://www.grouplens.org/node/73

[6] http://www.mendeley.com/

workshop was inspired by the FP7 *LinkedUp* project[7] that aims to provide a data pool of linked educational datasets that can be used for developing and testing advanced TEL recommender systems and other data driven educational tools. Using LD as the foundation for the TEL references datasets provides various advantages due to two main reasons: (a) LD and the Social Web offer vast amounts of often publicly available data and resources of high relevance to educational contexts; and (b) LD techniques offer solutions for fundamentally improving quality and interoperability of existing data by, for instance, allowing to match schemas and interlink previously unrelated datasets. To this end, LD and the Social Web show high potential to alleviate data sparseness and interoperability problems towards Web-scale application of recommender systems.

In this article, we first provide a state of the art review of approaches to TEL resource data sharing on the Web, and of educational datasets relevant for TEL recommender research, including those that are available in the Linked Data landscape (section 2). Afterwards, in section 3, we describe the challenges that currently hinder the use of LD as data repository. In section 4, we outline a set of principles that need to be considered to overcome these challenges and create a suitable LD repository. To this aim we show how some of these challenges are being addressed by some key past and on-going European projects. Section 5 describes suitable data formats for dealing with data generated in the social web and from the tracking of user's activities. Section 6 describes how these data sources can be exposed to the general LD cloud, providing some examples of social and linked data sources integrated for recommendations. Finally, we summarise the article and outline the main aspects to develop a LD repository for TEL recommender systems.

2 TEL resource data sharing on the Web – State of the Art

Open Educational Resources (*OER*) are educational material freely available online. The wide availability of educational resources is a common objective for universities, libraries, archives and other knowledge-intensive institutions raising a number of issues, particularly with respect to Web-scale *metadata interoperability* or legal as well as *licensing aspects*. Several competing standards and educational metadata schemata have been proposed over time, including IEEE LTSC LOM[8] (*Learning Object Metadata*), one of the widest adopted, IMS[9], Ariadne, ISO/IEC MLR - ISO 19788[10] Metadata for Learning Resources (MLR) and Dublin Core (see also [23]). The adoption of a sole metadata schema is usually not sufficient to efficiently characterize learning resources. As a solution to this problem, a number of taxonomies, vo-

[7] LinkedUp: Linking Web Data for Education Project – Open Challenge in Web-scale Data Integration (http://www.linkedup-project.eu)
[8] http://ltsc.ieee.org/wg12/par1484-12-1.html
[9] http://www.imsglobal.org/metadata/
[10] http://www.iso.org/iso/

cabularies, policies, and guidelines (called *application profiles*) are defined [21]. Some popular examples are: UK LOM Core[11], DC-Ed[12] and ADL SCORM.

Due to the diversity of exploited standards, existing *OER repositories offer very heterogeneous datasets*, differing with respect to schema, exploited vocabularies, and interface mechanisms. Examples are the MIT Open Courseware[13] (OCW), and OpenLearn,[14] the UK Open University's contribution to the OER movement (OpenLearn is also member of the MIT OCW Consortium). Video material from OpenLearn, distributed through iTunes U has reached more than 40 million downloads in less than 4 years[15]. One of the largest and diverse collections of OER can be found in the GLOBE[16] (Global Learning Objects Brokered Exchange) where jointly, nearly 1.2 million learning objects are shared. KOCW[17], LACLO[18] and OUJ[19] expose a single collection of metadata instances with a common provenance. Other repositories, such as ARIADNE, LRE[20], OER and LORNET[21] expose the result of the aggregation of several metadata collections that have different provenance.

Regarding the presence of *educational information in the linked data landscape*, two types of linked datasets need to be considered: (1) datasets directly related to educational material and institutions, including information from open educational repositories and data produced by universities; (2) datasets that can be used in teaching and learning scenarios, while not being directly published for this purpose. This second category includes, for example, datasets in the cultural heritage domain, such as the ones made available by the Europeana project[22], as well as by individual museums and libraries (such as the British Museum[23], who have made their collection available as linked data, representing more than 100 Million triples, or the Bibliothèque Nationale de France[24], who made available information about 30,000 books and 10,000 authors in RDF, representing around 2 Million triples). It also includes information related to research in particular domains, and the related publications (see PubMed[25] which covers more than 21 Million citations, in 800 Million triples), as well as general purpose information for example from Wikipedia (see DBPedia.org).

[11] http://zope.cetis.ac.uk/profiles/uklomcore/
[12] http://www.dublincore.org/documents/education-namespace/
[13] http://ocw.mit.edu/index.htm
[14] http://openlearn.open.ac.uk/
[15] http://www.bbc.co.uk/news/education-15150319
[16] http://globe-info.org/
[17] http://www.koreabrand.net/
[18] http://www.laclo.org/
[19] http://www.ouj.ac.jp/eng/
[20] http://lreforschools.eun.org/
[21] http://www.lornet.org/
[22] http://www.europeana.eu/
[23] http://collection.britishmuseum.org/
[24] http://data.bnf.fr/
[25] http://www.ncbi.nlm.nih.gov/pubmed/ and http://thedatahub.org/dataset/bio2rdf-pubmed

Regarding category (1), initiatives have emerged recently using linked data to expose, give access to and exploit public information for education. The Open University in the UK was the first education organization to create a linked data platform to expose information from across its departments, and that would usually sit in many different systems, behind many different interfaces (see http://data.open.ac.uk which includes around 5 Million triples about 3,000 audio-video resources, 700 courses, 300 qualifications, 100 Buildings, 13,000 people [27][28]). Many other institutions have since then announced similar platforms, including in the UK the University of Southampton (http://data.southampton.ac.uk) and the University of Oxford (http://data.ox.ac.uk). Outside the UK, several other universities and education institutions are joining the Web of Data, by publishing information of value to students, teachers and researchers with linked data. Noticeable initiatives include the Linked Open Data at University of Muenster[26] and the LODUM[27] project in Germany or the Norwegian University of Science and Technology exposing its library data as linked open data[28]. In addition, educational resources metadata has been exposed by the mEducator project [24][3]. A more thorough overview of educational Linked Data is offered by the Linked Education[29] platform and in [4].

In the TEL field many research projects are working with rather small internal datasets which cannot be shared with other research institutes [48][49]. Therefore, the EATEL Special Interest Group *dataTEL* was founded [43] with a focus on the analysis of issues around the development, sharing and using of TEL datasets for research. Recently, the dataTEL project published an initial list of 20 available TEL datasets for research and compared the different datasets according to certain criteria (see Table 1) [46]. With this initiative the amount of available TEL datasets has increased and initial comparison study's are emerging that use the same dataset for different personalisation techniques [47][48][49][50]. The overall aim of the dataTEL initiative is to make different personalisation approaches more comparable to gain a body of knowledge about the effects of personalisation on learning. Still, there are several issues as described in [53] that need to be resolved before the uptake and usage of such datasets can become standard practice as in other domains [51].

The emergence of several Linked Open Data initiatives is promising to overcome these issues by providing: 1) A vast and increasing amount of data, 2) An established set of exchange principles and standards, and 3) Standardised publication and licensing approaches for TEL datasets.

[26] http://data.uni-muenster.de
[27] http://lodum.de
[28] http://openbiblio.net/2011/09/08/ntnu/
[29] http://linkededucation.org

		Environment/ application	Collection Period	Statistics	Access rights	Educational context
dataTEL	**Mendeley**	Web portal	1 year	200.000 users 1.857.912 items 4.848.725 actions	Open access	Science
	APOSDLE	PLE	3 months	6 users 163 items 1500 actions	Open access	Workplace learning
	ReMashed	PLE/Mash-up environment	2 years	140 users 960.000 items 23.264 actions	Legal protection	Computer science
	Organic. Edunet	Web portal	9 months	1.000 users 11.000 items 920 actions	Legal protection	Agriculture
	MACE	Web portal	3 years	1.148 users 12.000 items 461.982 actions	Legal protection	Architecture
	Travel well	Web portal	6 months	98 users 1.923 items 16.353 actions	Open access	Various
	ROLE	PLE	6 months	392 users 11.239 items 28.554 actions	Legal protection	Computer science
	SidWeb	LMS	4 years	4.013.208 users 35.041 items 4.009.292 actions	Legal protection	Various
	UC3M	Virtual machine LMS	3 months	284 users 8.669 items 49.000 actions	Legal protection	Computer science
	CGIAR	LMS	6 years	841 users 14.693 items 326.339 actions	Legal protection	Agroforestry
PSLC DataShop	**Algebra 2008-2009**	ITS	1 year	3.310 users 8.918.055 actions 206.597 items	Legal protection	Math/ Algebra
	Bridge to Algebra	ITS	1 year	6.044 users 20.012.499 actions 187 items	Legal protection	Math/ Algebra
	Geometry Area	ITS	1 year	59 learners 139 items 6.778 actions	Open access	Math/ Geometry
	Electric Fields - Pitt	ITS	1 month	25 learners 139 items 5.347 actions	Open access	Math
	Chinese Vocabulary Fall 2006	ITS	4 years	101 learners 9.884 items 107.910 actions	Open access	Language learning
	Handwriting 2/Examples Spring 2007	ITS	2 months	54 users 11.162 items 20.016 actions	Open access	Math
Mulce	**Virtual Math Team (VMT)**	Chat	10 days	13 users 2.488 actions/ items	Open access	Math
	mce-simu	Forum Email Chat	10 weeks	44 users 12.428 actions/ items	Open access	Language learning
	mce-copeas	Video conferencing	10 weeks	14 users 37 videos	Open access	Language learning

Table 1: Overview of datasets from dataTEL project [46]

3 Challenges for using LD as references datasets for TEL research

While there is already a large amount of educational data available on the Web via proprietary and/or competing schemas and interface mechanisms, the main roadmap for improving impact of TEL recommender systems includes (a) start adopting LD principles and vocabularies while (b) leveraging on existing educational data available on the Web by non-LD compliant means. Following such an approach, major research challenges need to be taken into consideration towards Web-scale interoperability [4]:

(C1) **Integrating distributed data from heterogeneous educational repositories:** educational data and content is usually exposed by heterogeneous services/APIs such as OAI-PMH or SQI. Therefore, interoperability is limited and Web-scale sharing of resources is not widely supported yet.

(C2) **Metadata mediation and transformation:** educational resources and the services exposing those resources are usually described by using distinct, often XML-based schemas and by making use of largely unstructured text and heterogeneous taxonomies. Therefore, schema and data transformation (into RDF) and mapping are important requirements in order to leverage on already existing TEL data.

(C3) **Enrichment and interlinking of unstructured metadata:** existing educational resource metadata is usually provided based on informal and poorly structured data. That is, free text is still widely used for describing educational resources while use of controlled vocabularies is limited and fragmented. Therefore, to allow machine-processing and Web-scale interoperability, educational metadata needs to be enriched, that is transformed into structured and formal descriptions by linking it to widely established LD vocabularies and datasets on the Web.

(C4) **Integration of personal and social data:** While the above mentioned challenges focus on educational resource data and metadata, the user perspective has to be considered by integrating personal as well as social data into the data environment. In particular, the LD cloud is populated mainly with content driven information and less data available via the social web. Hence, knowledge obtained via the LD approach has to be complemented with data obtained from the social Web. This results in additional challenges with regards to integration of such diverse data sources in order to make them available as resources for recommender systems and other social web applications.

Our work builds on the hypotheses that Linked Data offers high potential to improve take-up and impact of TEL recommender systems and introduces key past and ongoing projects which serve as building blocks towards *Linked Education*[30], i.e. educational data sharing enabled by adoption of Linked Data principles.

[30] http://linkededucation.org: an open platform to share results focused on educational LD. Long-term goal is to establish links and unified APIs and endpoints to educational datasets.

In particular, we focus on three projects which address the aforementioned challenges by providing innovative approaches towards (a) integration of heterogeneous TEL data (as part of the *mEducator*[31] project), (b) exploitation of large scale educational open data addressed by the *LinkedUp*[32] project, and (c) exploitation of social data as linked data (as part of the Open Discovery Space[33] project). In the next section we focus on approaches to address challenges C1, C2 and C3, whereas in section 5 we focus on challenge C4 and point (c), exploitation of social data.

4 Towards integration and exploitation of heterogeneous educational resource data

With respect to the key issue - integration of heterogeneous TEL data - we first identify a set of principles (see [3][7]) to address the above mentioned challenges:

(P1) **Linked Data-principles:** are applied to model and expose metadata of both educational resources and educational services and APIs. In this way, resources are interlinked but also services' description and resources are exposed in a standardized and accessible way.

(P2) **Services integration:** Existing heterogeneous and distributed learning repositories, i.e. their Web interfaces (services) are integrated on the fly by reasoning and processing of LD-based service semantics (see P1).

(P3) **Schema matching:** metadata retrieved from heterogeneous Web repositories, is automatically lifted into RDF, aligned with competing metadata schemas and exposed as LD accessible via de-referenceable URIs.

(P4) **Data interlinking, clustering and enrichment:** Automated enrichment and clustering mechanisms are exploited in order to interlink data produced by (P3) with existing datasets as part of the LD cloud.

In the following we provide examples of how the above principles can be applied, starting from the conversion of data into RDF and touching on various approaches to harmonize educational metadata and on the available tools and techniques to achieve metadata enrichment and dataset interlinking.

4.1 Integration of educational resources data

The problems connected to the heterogeneity of metadata can be addressed by converting the data into a format that allows for implementing the Linked Data principles [2]. Most often this means that the data which is provided as part of RDBMS or in XML format – or, on occasion, in other formats – are converted into RDF. The data model of RDF is a natural choice as it allows for unique identification, interlinking to related data, as well as enrichment and contextualization. Therefore, general-purpose

[31] http://www.meducator.net

[32] LinkedUp: Linking Web Data for Education Project – Open Challenge in Web-scale Data Integration (http://www.linkedup-project.eu)

[33] http://www.opendiscoveryspace.eu/

tools such as D2R[33], Virtuoso[34] and Triplify[35] are often used to convert proprietary datasets into RDF.

It is common to use DBpedia or other big datasets as "linking hubs" [1]. One of the advantages of such an approach is that such datasets are commonly used by other datasets, which automatically leads to a plurality of indirect links. In the case of more specialized applications it is beneficial if domain specific datasets or ontologies can be found and linked to. This has been successfully demonstrated by specialized projects such as Linked Life Data[34] in the biomedical domain, Organic.Edunet[35] in organic agriculture and agroecology [30], and mEducator[36] in medical education [26][3].

The approaches applied for creating links between datasets can be fully automatic, semi-automatic and fully manual. A lot of tasks required for interlinking and enhancing (enriching) metadata can be automated by analyzing textual content using Information Extraction (IE) and Natural Language Processing (NLP) techniques. Most commonly this includes the detection of sentences, named entities, and relationships, as well as disambiguation of named entities. However, quality control implies that the process has to be supervised at some point. The links can be created manually; alternatively the automatically detected links can be approved manually. NLP has its roots in machine learning which implies the use of learning algorithms which are trained on large textual corpora which eventually are domain-specific. Public services such as DBpedia Spotlight[37] and OpenCalais[38] offer NLP services relevant for linking data and also provide their output in RDF. In addition to these services which are ready to use, frameworks such as Apache Stanbol[39] can be easily integrated and provide solutions for the most common tasks involved in the creation of Linked Data, such as textual analysis and metadata extraction. A RESTful API allows for easy integration which should help projects dealing with metadata management using semantic technologies to hit the ground running.

Traditional ways of managing metadata often take a document-centric approach and use XML as it is an established standard for expressing information. Transformation of metadata into other formats requires a thorough mapping to be crafted, which often involves an analysis of the exact semantics of the involved standards. If such heterogeneous formats are to be transformed into Linked Data, good knowledge of existing standards is required, as it is good practice to reuse established terms from other RDF-based standards [14] whenever possible. There are situations where the conceptual model of the origin data cannot be cleanly mapped to the RDF model and information may be lost. To avoid such situations, RDF should be considered as a

[33] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[34] http://virtuoso.openlinksw.com/
[35] http://triplify.org/
[34] http://www.linkedlifedata.com
[35] http://www.organic-edunet.eu
[36] http://www.meducator.net
[37] http://dbpedia.org/spotlight
[38] http://www.opencalais.com
[39] http://incubator.apache.org/stanbol/

basis for metadata interoperability [14] – a common carrier – when adapting existing or creating new metadata standards.

The joint working group from IEEE LTSC and Dublin Core made an attempt to address heterogeneity of educational metadata by developing a mapping of IEEE LOM into the Dublin Core Abstract Model. This work resulted in a draft report in 2008, but the uptake has not been overwhelming. To date, the only known project to implement this draft[40] is the Organic.Edunet project, whose achieved goal was to build a federation of learning repositories with material on organic agriculture and agroecology. The EntryStore backend[41] (the basic concepts behind it are described in [29] and [30]) is used across all Organic.Edunet repositories and stores all information in RDF. This requires that all metadata that are harvested for enriching in the Organic.Edunet repositories are converted from LOM/XML (which is the primary format in most of the source repositories) to an RDF representation. This makes it also possible to freely combine different standards and vocabularies, resulting in enriching LOM metadata with more specific terms from vocabularies such as EUN's LRE and blending in some FOAF and relational predicates from OWL and DC to create interlinkage between resources.

A similar yet even more exhaustive approach was followed by the mEducator project addressing two central challenges for educational data integration: integration at the repository-level facilitated by repository-specific APIs and integration at the (meta)data-level [3]. The former aims at integrating educational services and APIs in order to facilitate repository-level integration. To this end, it is concerned with resolving heterogeneities between individual API standards (e.g. SOAP-based services vs. REST-ful approaches) and distinct response message formats and structures (such as JSON, XML or RDF-based ones) where details are described in [31]. In order to enable integration of such heterogeneous APIs, Linked Data principles were used to annotate individual APIs in terms of their interfaces, capabilities and non-functional properties. This enables the automatic discovery and execution of APIs for a given educational purpose (for instance, to retrieve educational metadata for a given subject and language) while it resolves heterogeneities between individual API responses. All metadata of educational content retrieved from these services are transformed from their native (standardized or proprietary) formats into RDF. The second step deals with the actual integration of the retrieved heterogeneous educational (meta)data by exposing all retrieved educational (RDF) metadata as well-interlinked Linked Data. As starting point, all generated RDF is stored in a dedicated, public RDF store[42] which supports two main purposes: to expose existing educational (non-RDF) data in a LD-compliant way and allow content/data providers to publish new educational resource metadata. Automated interlinking of dataset as well as clustering and classification is employed to enrich and interlink the educational data. Transformation of heterogeneous metadata into RDF is indeed a substantial step towards integration, however, mere transformation does not improve metadata quality. Thus, it is even

[40] The reference implementation is part of EntryStore which is Free Software
[41] http://code.google.com/p/entrystore/
[42] http://ckan.net/packages/meducator

more challenging to enrich descriptions by automated data enrichment techniques to establish links with established vocabularies available on the LD cloud. Enrichment takes advantage of available APIs such as the ones provided by DBpedia Spotlight or Bioportal[43], which allow access to a vast number of established taxonomies and vocabularies. This way, unstructured free text is enriched with unique URIs of structured LD entities to allow further reasoning on related concepts and to enable the formulation of queries by using well-defined concepts and terms. In addition, automated clustering and classification mechanisms are exploited in order to enable data and resource classification across previously disconnected repositories.

Another attempt to harmonize educational metadata is currently carried out by the Learning Resource Metadata Initiative[44] (LRMI) whose goal is to build a common metadata vocabulary for the description of educational resources. LRMI is led by both the Association of Educational Publishers and the Creative Commons[45]. The applied approach is based on schema.org and has the declared goal of providing mappings to the most common standards for describing education resources, such as LOM and DC.

4.2 Large scale exploitation of educational open data

An issue complementary to the integration of heterogenous educational data is the large scale exploitation of open educational data, is addressed by the *LinkedUp* project, setting up to push forward the exploitation of the vast amounts of public, open data available on the Web, in particular by educational institutions and organizations. This will be achieved by identifying and supporting highly innovative large-scale Web information management applications through an open competition (the *LinkedUp Challenge*) and a dedicated evaluation framework. The vision of the LinkedUp Challenge is to realise personalised university degree-level education of global impact based on open Web data and information. Drawing on the diversity of Web information relevant to education, ranging from OER metadata to the vast body of knowledge offered by the LD approach, this aim requires overcoming substantial issues related to Web-scale data and information management involving Big Data, such as performance and scalability, interoperability, multilinguality and heterogeneity problems, to offer personalised and accessible education services. Therefore, the LinkedUp Challenge provides a focused scenario to derive challenging requirements, evaluation criteria, benchmarks and thresholds which are reflected in the LinkedUp evaluation framework. Information management solutions have to apply data and learning analytics methods to provide highly personalised and context-aware views on heterogeneous Web data. Building on the strong alliance of institutions with expertise in areas such as open Web data management, data integration and Web-based education, key outcomes of LinkedUp include a general-purpose evaluation framework for Web-data driven applications, a set of quality-assured educational datasets, innovative applications of large-scale Web information

[43] http://www.bioontology.org/wiki/index.php/BioPortal_REST_services
[44] http://wiki.creativecommons.org/LRMI
[45] http://creativecommons.org/

management, community-building and clustering crossing public and private sectors and substantial technology transfer of highly innovative Web information management technologies.

5 Integration of social data

Social data can be defined in many ways when seen from different disciplines or perspectives. From the LD perspective we see social data as an end user added information that is publicly available on the Web and provides an indication of the quality of an artefact on the Web. We further distinguish between '*social data'* and '*paradata'*. The main difference between '*social data*' and '*paradata*' whether they have been contributed by the user intentionally or were tracked by the system in the background. The CIP ICT-PSP eContentPlus project *Open Discovery Space* (ODS) has a work package dedicated to develop a social metadata cloud that can contribute social activity data like ratings, tags, bookmarks and comments to the LD cloud. ODS represents large amount of data in the field of education with a critical mass of approximately 1.550.000 eLearning resources from 75 content repositories, as well as 15 educational portals of regional, national or thematic coverage connected to it that will be exposed as LD. This vast amount of data and the emerging social activities around it will be captured and exposed in an anonymised way to the LD cloud. A first design of this social data cloud has been already specified in one deliverable [52]. In this section we discuss suitable data formats that could be applied to store the social activities of the users and expose it as LD.
According to the ODS project social data and paradata are defined as follows [52]:

1. *Social Metadata* which refers to the direct interaction users have with an artefact or with other actors around the artefact. Interactions with artefacts can include the adding of keyword, ratings, tags, bookmarks, or comments.

2. *Paradata*, is another type of social data as it requires further processing of the data before it can be meaningful. Paradata consists of *automatic traces* of the interaction the user has with various artefacts together with appropriate contextual information.

The following four metadata schemas (1. CAM, 2. Organic.Edunet, 3. Learning Registry Paradata, 4.NSDL Paradata) have been investigated by the ODS project and are suitable to store social data in a database. So far there is no LD RDF schema available for these data formats but it is the intention of the ODS project to first select the most appropriate data format, and second design a suitable RDF schema to expose the social data as LD. In the following subsections we shortly introduce the different data formats and conclude by presenting an initial comparison of the candidate data formats. This analysis is mainly based on the findings of [52] (*Review of Social Data Requirements).* A more in depth analysis of the different formats and how they can be interconnected can be found in [13][55][56].

5.1 Contextualized Attention Metadata (CAM)

Contextualized Attention Metadata[46] (CAM) [16][18] is a format to describe *events* conducted by a human user, e.g. accessing a document or sending an e-mail. As little information as possible is stored in the CAM instance itself, e.g. the event type and the time stamp. All other information, e.g. metadata describing users or documents involved in the event, are linked. This way, every entity/session can be described in a different and suitable format and no information is duplicated.
The main element of each CAM instance is the *event* entry which comprises its *id*, the *event type*, the *timestamp*, and a *sharing level reference*. Examples for *event types* are "send", "update" or "select". CAM is used in a couple of European projects such as ROLE[47] and OpenScout[48] that already started to define a collection of various *event types*. Depending on the *event*, various *entities* with different *roles* can be involved. For example: When bookmarking a file at a social bookmarking service, there's a person with the role *sender*, and at least one person or a community with the role *receiver* and a *document* with the role *website*. Each event can be conducted in a *N:M relation*.

5.2 Organic.Edunet format

The Organic.Edunet portal[49] [9] is a learning portal that provides access to more than 10.000 digital learning resources on organic agriculture and agro-ecology hosted in a federation of external repositories. Regarding social data, Organic.Edunet relies on a representation model detailed in [10] which to some degree is based on CAM, since it stores data about which tags, reviews, ratings and recommendations were assigned to learning resources by which user. This conceptual model, not specific of any portal, context or particular application, was intended as a structured, reusable and interoperable way of representing the different types of user feedback and was used as a basis for the social module in the Organic.Edunet portal. The model by Manouselis and Vuorikari (2009) [10] is based on the concept of an *annotation schema*, a formal declaration of the type(s) of feedback (i.e. rating, review, tags, etc.) including the exact structure and value spaces of the collected feedback. For instance, ratings may be collected upon one or more attributes (criteria), and may use different rating scales, particularly in different application areas.

5.3 The Learning Registry format

The Learning Registry model [8] collects social data such as tags, comments, ratings, clicked and viewed data, shared data, data aligned to a standard, and any other data

[46] https://sites.google.com/site/camschema/
[47] http://www.role-project.eu/
[48] http://www.openscout.net/
[49] http://portal.organic-edunet.eu

about the usage of learning resources and shares this data in a common pool for aggregation, amplification and analysis.

By design, a loose format for the submission of metadata is defined without specifying what metadata schema should be used. The Learning Registry uses a Resource Data Description (RDD) document for submitting social metadata as a thin wrapper around the submitted metadata. The services built on top of the Learning Registry can provide extraction or crosswalk services across RDDs that use disparate standards, or can assemble metadata fields from different schemas into custom views.

5.4 National Science Digital Library format

National Science Digital Library (NSDL) is an online portal for education and research on learning in Science, Technology, Engineering, and Mathematics. NSDL's mission is to provide quality digital resources to the science, technology, engineering, and mathematics (STEM) education community, both formal and informal, institutional and individual. The STEM Exchange is a collaboration with a range of education partners that has been initiated for the implementation of an NSDL web service to capture and share social media-generated information and other networked associations about educational resources.

Collections and records stored in the NSDL repository are made available through the Search API and the NSDL OAI data provider. In addition, the Strand Map Service APIs provide access to Benchmarks, Maps and visualizations. Developers can use the Search API, SMS APIs and OAI data provider to build customized search and browse interfaces and other applications.

In creating the concept of the STEM Exchange, two different kinds of Item-Level Metadata evolved, i.e. the NSDL Annotation and NSDL Paradata. The main purpose of NSDL Annotation is to capture user comments, reviews, and teaching tips. It also allows annotations to include additional information, e.g. the metadata record contributor, annotator, or the subject. NSDL Paradata was defined to capture usage data about a resource, such as downloaded or rated [15].

5.5 Definition of the standards and a comparison

In order to evaluate the four described candidate schemes and for selecting the best suited format for social metadata and paradata recording we applied a social media use case called '*Irma'*. A detailed description of the use case can be found in [52]. Table 2 illustrates a feature comparison of the four mentioned data formats: CAM, Organic.Edunet, Learning Registry, and NSDL Paradata. Each of the formats are either rated with a (+) to indicate it supports a requirement derived from Irma, or with a (-) meaning it does not support this requirement. The first nine requirements relate to social metadata, the second nine requirements show support of paradata.

Nr.	Social metadata requirements from Irma	CAM	Organic Edunet format	Learning Registry paradata	NSDL paradata
1	Rate	+	+	+	+
2	Tag	+	+	+	+
3	Bookmark	+	+	+	+
4	Share (FB, twitter, e-mail)	+	-	+	+
5	share count	+	+	-	+
6	Comment	+	+	+	+
7	Join groups	+	-	+	-
8	Posts (discussion, blog, etc.)	+	-	+ (Google discussion)	-
9	following/followers	+	-	+	-
	Social data sum (+)	**9**	**5**	**8**	**6**
Nr.	**Paradata requirements from Irma**	**CAM**	**Organic Edunet format**	**Learning Registry paradata**	**NSDL paradata**
10	Login / logout	+	+	+ (Google)	-
11	Access learning object metadata	+	+	+	+
12	Navigation history of users	+	+	+	-
13	Search history of users	+	-	-	-
14	History of LO (new upload or edit)	+	+	-	-
15	IP location of user	+	+	+	-
16	Language of LO (and of browser of the user)	+	+	+	+
17	Language of user browser	+	+	+	+
18	Group metadata to extend user profile (new interests)	+	+	-	-
	Total sum (+)	**18**	**13**	**14**	**9**

Table 2: Overview comparison of suitable data formats to store social data [52]

The comparison expressed in Table 2 emphasises that any of the described formats can store common social activities like rating, tagging and commenting in the social web.

Differences in applicability of the various schemata appear between the formats when we consider paradata aspects. The most promising data format therefore is CAM, as it covers all 18 requirements from the Irma use case, while NSDL supports only 9 of them. Learning Registry and Organic.Edunet have also 14 and 13 points in Table 2 respectively. All the mentioned formats use application specific models and services for implementing social services for their users.

From the analysis, it appears that Organic.Edunet might be a suitable candidate for collecting social metadata in a database and become exposed as Linked Data. This would also be aligned with privacy aspects, as Organic.Edunet mainly focuses on social metadata that is publicly available on the web whereas CAM, for instance, first needs to be filtered to not expose all private data of a user. CAM, on the other hand, clearly turns out to be the most suitable data format for tracking and storing paradata. Both formats have a strong European community behind them and some ready-to-use services that are already applied in different EU projects.

6 Bridging the gap between Linked Data and the Social Web

In this section we focus on some current efforts that are relevant to the integration of linked data and the social web, with the goal of providing more sophisticated recommender systems. In the previous section some schemas meant to capture social data and paradata have been discussed. Here we turn to ontologies and vocabularies written in RDF/OWL that can be instrumental to exposing social data and paradata to the Linked Data Cloud and survey some first applications that demonstrate their potential.

6.1 SIOC and its applications

A fundamental component towards the goal of exposing social data and paradata as LD is provided by the SIOC (Semantically-Interlinked Online Communities) initiative[50]. SIOC [32] is an ontology to describe user-generated content on forums, weblogs, and web2.0 sites, and link online communities. The goal of SIOC is to harness, across online communities, discussions on interrelated topics relevant to a post, either from similar members profiles of from common-topic discussion forums. By narrowing the scope of a search to a set of interlinked community sites a first advantage is that the problem of low precision of a query issued in the web can be addressed. Also, concepts such as Site, Forum, Post, Event, Group and UserAccount are described in the SIOC ontology with a focus on the relationships, sub-classes and properties of these concepts relevant to the arena of online discussion methods, in such a way to enable use cases not previously possible with other ontologies describing similar concepts. The SIOC community provides mappings and interfaces to commonly-used

[50] http://sioc-project.org/

ontologies such as Dublin Core, FOAF and RSS 1.0. and several tools to import and export data from SIOC.

In particular, SIOC is often used in conjunction with FOAF[51] (Friend of a Friend) vocabulary for describing users and their connections to interests and other users profiles. In such a way, the contribution of Social Web sites to the linked data cloud is explicated from two synergistic points of view: via the direct links from person to person and by the links arising from the notion of "object centred sociality" [33], i.e., people in a community are indirectly connected because they share objects of a social focus (e.g., a topic in a post or a link to the song of a band).

Currently, SIOC is generating much interest and is widely adopted, resulting in an active support for the development of tools, API and applications. For example, in [34] user-generated contents are lifted to SIOC by a method that extracts users comments directly from HTML pages, without requiring any a priori knowledge about the webpage. This approach circumvents the problem of SIOC data scarcity, which is a consequence of the fact that SIOC exporter plugins often are not enabled be the site administrators. One remarkable aspect of SIOC is that its high-level description of communities can be easily integrated with more specific ontologies to bridge the Social Semantic Web and more application domains. An example is the SWAN/SIOC project[52] that defines a coherent ontology capable of representing both high-level descriptions of communities (thanks to SIOC) and the argumentative discussions (using the SWAN ontology) taking place in that communities [35]. The goal of this alignment is to make the discourse structure and component relationships accessible to computation, so that information can be better navigated, compared and understood, across and within domains.

Another example of leveraging on SIOC extensibility is provided in [36], where MediaWiki integration is accomplished by resorting to SKOS ontology to model topic and categories and to other vocabularies that model user tagging and are helpful in alleviating issues such as ambiguity between tags. From a methodological point of view, the two examples above are representative of the type of efforts that can be pursued to create datasets that integrate social, user-centric, and linked data to generate novel types of recommendations that can certainly find a space in the TEL scenario.

A first example of the joint usage of SIOC and linked data in recommender systems is in the music domain [37], where social data encompasses the publishing and sharing of music-related data on the Web, whatever their format is (blog posts, wiki pages, community databases, mp3s or playlists). This work demonstrates how FOAF, SIOC and linked data can be used to provide a completely open and distributed social graph, where SPARQL queries can implement a simple collaborative filtering algorithm, and the wide range of interlinked data in multiple domains allow the user to get more data rich, justified recommendations. This latter aspect (justified recommendations) seems to play an important role in the acceptance and trust of end-users towards recommender systems, e.g., [38].

[51] http://www.foaf-project.org/

[52] http://www.w3.org/TR/hcls-swansioc/

6.2 The CAM-RDF binding and the Atom Activity Stream RDF mapping

An RDF binding of the Contextualised Attention Metadata (CAM) model discussed in the previous section has been recently proposed [39]. Among the advantages pointed out in [39] for the CAM-RDF binding with respect to the CAM-XML one, are the following ones: first, it facilitates the integration of CAM into RDF-based learning systems; second, the underlying graph-based representation may support more convenient ways of analyzing the observations, i.e., by resorting to graph algorithms. The binding has been tested for equivalence to the CAM-XML binding with respect to tasks such as creating statistics over the MACE dataset (consisting of learning resources in the architectural domain), and by developing, over the same learning data set and collected CAM data, a "find similar users" application. Another application of this binding in the learning domain has been done in the context of an exercise system that provides personalized help to learners in the form of hints [40]. Personalization is achieved in terms of the content of hints and of the appropriate hint-giving approach. The CAM-RDF binding is available at:
http://www.fit.fraunhofer.de/~wolpers/ontologies/cam/cam.owl

A representation of activity alternative to RDF-CAM, is provided by the Atom Activity Streams RDF mapping[53]. Atom Activity Streams[54] extends the Atom specification, which is a widely used syndication format to transmit various types of web content such as weblog posts, news headlines, as well as user activities within social sites. This extension provides the ability to express within existing Atom entries and feeds much of the activity-specific metadata in a machine-parseable format. Within the NoTube project[55] an RDF mapping of the Atom Activity Streams (AAIR) has been developed, in conjuction with the W3C Semantic Web Interest group. The specification is available at http://xmlns.notu.be/aair/. A typical expression in AAIR would have the form (Actor, Verb, Object, Context), where typical verbs include "Play" (open resource), "MarkAsFavorite", "Save" (download), "Rate", "Share" and so on.

AAIR was chosen, mostly due to its intuitiveness and fair coverage of both social data and paradata, as the starting reference point of a line of action pursued within the mEducator project, concerned with modeling data about activities on social learning resources and make them portable across learning platforms and provide resources useful for recommendations. To achieve this goal some extensions to AAIR were done to track the user activities. In particular, the proposed extensions were devised to deal with the need to model Search activities within the mEducator platforms, keeping track of the queries executed by the user, the results of the queries, and of the activites executed by the users on the results of a given query. This was accomplished by extending the lists of verbs by *ActivityVerb:Search*, by creating a recursive reference to Activity and by introducing the notion of session. These extensions are sketched in Fig. 1. In particular, *hasQueryString:* is the property that represents the user's ke-

[53] http://xmlns.notu.be/aair/
[54] http://activitystrea.ms/specs/atom/1.0/
[55] http://notube.tv/

yword sequence to describe the query search; *isRelatedTo:* is a property to bind an activity to another one, so that it is possible to model an action performed on an object returned by a query, and *activitySession:* is a property that binds an activity to a specific user session.

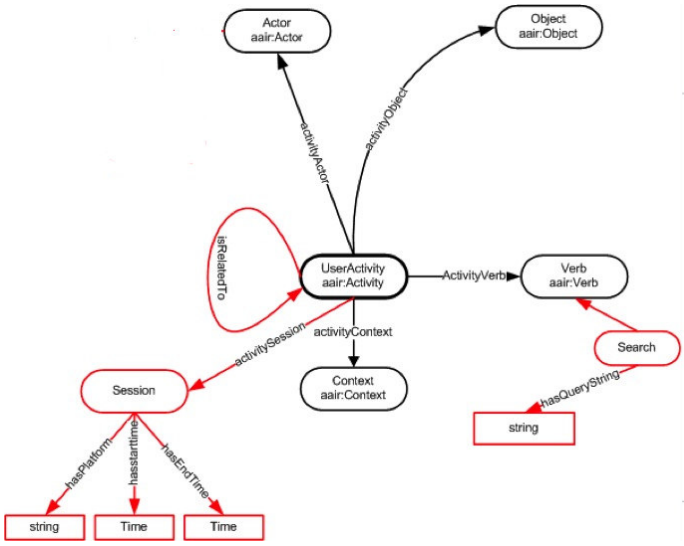

Fig. 1: AAIR extension adopted in mEducator to support recommendations

A proof of concept recommender systems architecture based on this extension has been developed and deployed in one of the mEductor platforms[56], where an Activity Monitor collects AAIR data and sends the activity to an external web service providing the recommendations.
In Fig. 2 is an example of how data generated by the Activity Monitor can be expressed as Linked Data. The example refers to a user "John Smith" who has an account in a mEducator platform, where he performs a search about "Magnetic Resonance Imaging" and saves one of the results of this search. In a subsequent session he comments on the saved resource. The example uses the AAIR extension, SIOC and FOAF vocabularies and the mEducator vocabulary to describe a learning resource.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix mdc: <http://www.purl.org/meducator/ns/> .
@prefix aair: <http://xmlns.notu.be/aair/> .
@prefix aairext: <http://www.mEducator2.net/aairext/> .

<http://www.mEducator2.net/Activities/Activity15>
          a aair:Activity ;
          aair:ActivityActor   [a aair:Actor; owl:sameAs
```

[56] http://www.meducator2.net/

```
                                                        [a sioc:UserAccount;
                    owl:sameAs <http://www.mEducator2.net/Account/JohnSmith>]
                                      ] ;
          aairext:ActivityVerb  <http://www.mEducator2.net/Verb/Search_50> ;
          aairext:ActivitySession <http://www.mEducator2.net/Session/Session1> .

<http://www.mEducator2.net/Verb/Search_50>
          a aairext:Search ;
          aairext:hasQueryString "Magnetic resonance imaging" .

<http://www.mEducator2.net/Session/Session1>
          a aairext:Session ;
          aairext:hasPlatform  <http://www.mEducator2.net> ;
          aairext:hasStartTime  "21/08/2010 12:27"^^xsd:date ;
          aairext:hasEndTime    "21/08/2010 12:32"^^xsd:date .

<http://www.mEducator2.net/Activities/Activity16>
          a aair:Activity ;
          aair:isRelatedTo    [a aair:Activity; owl:sameAs
                    <http://www.mEducator2.net/Activities/Activity15>] ;
          aair:ActivityActor  [a aair:Actor; owl:sameAs
                                                        [a sioc:UserAccount;
                    owl:sameAs <http://www.mEducator2.net/Account/JohnSmith>]
                                      ] ;
          aair:ActivityVerb   aair:Save ;
          aair:ActivityObject  <http://www.mEducator2.net/Resources/resource123> ;
          aairext:ActivitySession <http://www.mEducator2.net/Session/Session1> .

<http://www.mEducator2.net/Resources/resource123>
          a mdc:Resource owl:sameAs  aair:Object .

<http://www.mEducator2.net/Activities/Activity17>
          a aair:Activity ;
          aair:ActivityActor  [a aair:Actor; owl:sameAs
                                                        [a sioc:UserAccount;
                    owl:sameAs <http://www.mEducator2.net/Account/JohnSmith>]
                                      ] ;
          aair:ActivityVerb   aair:Post ;
          aair:ActivityObject  <http://www.mEducator2.net/Post/Comment1>  ;
          aairext:ActivitySession <http://www.mEducator2.net/Session/Session2> .

<http://www.mEducator2.net/Post/Comment1>
          a aair:Comment owl:sameAs sioc:Comment ;
          aair:Commenter <http://www.mEducator2.net/Account/JohnSmith> ;
          aair:Content "Provides one of the best explanation of MRI functioning" ;
          sioc:reply_of <http://www.mEducator2.net/Resources/resource123> .

<http://www.mEducator2.net/Session/Session2>
          a aairext:Session ;
          aairext:hasPlatform  <http://www.mEducator2.net> ;
          aairext:hasStartTime  "21/08/2010 18:15"^^xsd:date ;
          aairext:hasEndTime  "21/08/2010 18:42"^^xsd:date .

<http://www.mEducator2.net/Account/JohnSmith>
          a sioc:UserAccount ;
          sioc:email_sha1 "f4d5b3eaaff75fa981e626a3492d9030cb15191d" .

<http://www.mEducator2.net/Person/JohnSmith>
          a foaf:Person ;
          foaf:name "John Smith" ;
          foaf:knows <http://www.mEducator2.net/Person/Ernest>  .
```

Fig. 2: Example of the Activity Monitor data, including social data and paradata generated during two sessions, and expressed as Linked Data (Turtle format).

6.3 Summary

The above approaches and applications point to a scenario where the first efforts and concrete demonstrations of the possibilites of bridging the Linked Data and the Social Web world are beginning to emerge. Beyond the sketching of the potentialities, there is also some early evidence of the tangible benefits. In particular, the early evaluations that have been perfomed, although not necessarily in the TEL domain, show some

advantages of the novel resulting data infrastructure. Evaluation research has shown that by using linked data to build open, collaborative recommender systems, the the "cold start" problem (related to to initial lack of data about new users and new items) is ameliorated, and it is possible to improve precision and recall with respect to simple collaborative filtering (CF) approaches [41]. In particular, the evaluation in [41] reports an improvement from an average precision of 2% and average recall of 7% of a simple collaborative filtering recommendation applied to a music streaming database to an average precision of 14% and average recall of 33% achieved by augmenting the initial data set with linked data from another music social platform (DbTune MySpace) and DBpedia. The use of social trust to improve the data sparsity problem of recommender systems has been investigated in [38], on the Movielens dataset. The results are reported in terms of F-score (harmonic mean of precision and recall), at various sparsity percentages, and show an improvement of of the F-score in the range 7%-18% with respect to the baseline obtained with a standard collaborative filtering approach. Interestingly, the peak if the advantage is achieved when it is most needed, i.e., at high data sparsity percentages (98,57%). Still, it is also clear that targeted, task-dependent strategies are needed to leverage on this wealth of data since, as it is demonstrated in the case of harnessing LOD evidence for profiling expertise and, accordingly, recommend experts [42]. This work, in particular, points out how, with respect to the expertise recommendation task, the LOD offers data that are decoupled from any specific hypothesis about what constitutes expertise, and, as such, are flexible and can serve multiple approaches in defining expertise, besides offering the clear advantage of harnessing richer, cross-platform evidence. On the other hand, it must be ensured that the type of data that are needed is available in the LOD with the necessary level of detail and that relevant datasets are accessible through effective interlinking, which are two current shortcomings that can be addressed by more informed data publishing strategies and better interlinking services. Whereas TEL related, task specific recommendation algorithms and relevant strategies to harness the LOD will be the object of future research, in the meantime some prior challenges related to fulfilling the vision of integrated social and linked data infrastructure are to be addressed, as pointed out in the next section.

7 Conclusions - Open Challenges and Scenarios

In the previous sections, we provided an overview of different efforts aiming at utilising Linked Data as well as social and user-centric data for recommender systems in TEL. While the accessibility of large-scale amounts of data is a foundation for TEL recommender systems, these efforts contribute to improvements in scope, quantity and quality of recommendations in TEL environments. This includes both TEL recommender systems in research, where data is required for evaluation and benchmarking, as well as in practice, where data is a core requirement for offering suitable recommendations to users.

There is still a range of shortcomings that need to be addressed. Social data is usually stored locally in the content management system of a single portal. Harvesting and aggregating such data from various learning object repositories will allow the generation of a social data cloud and will enable the provision of new services across multiple portals. For instance, more accurate recommendations can be generated by taking into account social data from more than one learning object repository or social environment, even non-TEL platforms. Collecting heterogeneous social data from different sources is not a trivial task and requires the adoption of efficient technologies and protocols. The main aspects that should be taken into account are therefore:

Data quality & trust

One fundamental issue in distributed data environment is related to diversity of quality, provenance and trustworthiness of data. While, for instance, the LD cloud has received a lot of attention due to its large quantities of data covering a wide variety of topics, take-up by data consumers is slow and usually focused on a small set of well-established datasets [4]. This can be attributed to the varied quality of the datasets and hence, the lack of trust on the data consumer side. Therefore, assessment of data, better and more structured approaches towards labeling and cataloging data and the exhaustive provisioning of provenance information are crucial for enabling a widespread take-up of distributed data.

Licensing and privacy issues

Licensing as well as privacy issues are related challenges which apply to educational resources metadata (licensing) and social data (privacy). Reuse of distributed datasets and exploitation by applications and data mashups have to consider and address the diversity of license models used by distributed datasets and the potential impact on any derived datasets. In addition, sharing of social and learner-centric data requires the consideration of privacy problems and how these can be addressed, in particular within distributed data environments such as the Web. Within the Open Discovery Space project a specific paragraph was written for the Terms-Of-Use of the platform to cover this aspect. This paragraph informs the users about the usage of their personal data within the ODS portal. If they sign-up for ODS platform they also agree to support certain personalization services with their personal data. The following services will be activate for all registered users to provide personalized access to the information of the platform:

- personalized recommendations for learning material
- bookmark items
- utilize personal history (i.e. on searches undertaken, objects viewed, etc.)
- upload and share learning material (publish)
- utilize upload library
- view user stats
- rate and comment items, follow discussions, comments and groups, etc.

If users do not agree with theses Term-Of-Use they are free to use the ODS platform without having a registered user account and by anonymised browsing of the educa-

tional resources. We believe that such legal solutions will be more frequently used in the close future.

Common schemas and vocabularies for social and attention data
Platforms usually deploy proprietary schemas and vocabularies for representing learner activities and social information. Common schemas are important to manage and process social and attention data. Potential options for such a schema are CAM for paradata in combination with Organic.Edunet for social metadata. Although these seem to be promising and most feasible, it needs to be analysed how Organic.Edunet can be aligned to events stored in CAM. The Organic.Edunet partners are preparing a new release of their social data schema by the end of the year 2012 that will address this issue and provide required adjustments to link CAM to Organic.Edunet.

Interoperability between different social & resource data formats
Complementary to unified schemas and vocabularies, LD approaches to representation of social and attention data (Section 6) can further alleviate interoperability issues. LD principles in particular provide standard query and interfacing mechanisms together with de-referencable URIs, which allows data consumers to easily interact with remote data repositories containing resource or social or activity data.

Scalability of Web data processing
Dealing with distributed Web data sources, in particularly graph- and reasoning-based environments such as the Semantic Web, poses challenges with respect to scalability and performance [3]. Performance issues arise from distributed processing, often requiring large quantities of HTTP-based message exchanges, lack of parallelisation techniques and the still often comparably poor performance of graph-based data storage. Previous work has shown [54] that still, with a limited amount of data sources acceptable performance can be achieved, also in distributed data settings. Additionally, techniques such as map/reduce, local replication of datasets or indexing are required to further alleviate this issue in actual large-scale data scenarios.

Towards federated recommendation
Very large, cloud-based data infrastructures like the one that Learning Registry is setting up for the US, are expected to provide a new perspective into the way that intelligent systems (in general) and socially-generated data-based services (in particular) will be developed [11][19]. Such global learning data infrastructures can help in scaling up the existing data-driven services, by allowing them to consume, process and use a rich variety of usage data streams, and thus enable novel forms of real time intelligence and cross-platform recommendations that can only become possible on extremely large data volumes.

Future work, in particular in highly related projects such as LinkedUp and ODS will address these issues in order to enabling the widespread adoption of data – resource metadata as well as learner-centric and social data – by TEL environments.

Acknowledgments

This work is partly funded by the European Union under FP7 Grant Agreement No 317620 (LinkedUp) and the CIP ICT PSP eContentPlus project Open Discovery Space.

References

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, C., Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. Proceedings of the 6th International Semantic Conference (ISWC2007).

[2] Bizer, C., T. Heath, Berners-Lee, T. (2009). Linked data - The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems.

[3] Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. and Taibi, D. (2012), "Linked Education: interlinking educational Resources and the Web of Data", *Proceedings of the 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications*, Riva del Garda (Trento), Italy, 2012.

[4] Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H., Giordano, D., Marenzi, I., Pereira Nunes, B. (2013) Interlinking educational Resources and the Web of Data – a Survey of Challenges and Approaches, accepted for publication in Emerald Program: electronic Library and Information Systems, Volume 47, Issue 1 (2013).

[5] IEEE (2002), "IEEE Standard for Learning Object Metadata", *IEEE Std 1484.12.1-2002*, pp.i–32. doi: 10.1109/IEEESTD.2002.94128.

[6] World Wide Web Consortium (2008). *W3C Recommendation, SPARQL query language for RDF*, 2008, available at:www.w3.org/TR/rdf-sparql-query/

[7] Yu, H. Q., Dietze, S., Li, N., Pedrinaci, C., Taibi, D., Dovrolls, N., Stefanut, T., Kaldoudi, E. and Domingue, J. (2011), "A linked data-driven & service-oriented architecture for sharing educational resources", in *Linked Learning 2011,* Proceedings of the 1st International Workshop on eLearning Approaches for Linked Data Age, May 29, 2011, Heraklion, Greece.

[8] Bienkowski M., Brecht J., Kio J. (2012). The learning registry: building a foundation for learning resource analytics. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC, Canada — April 29 - May 02, 2012.

[9] Kosmopoulos T., Kastrantas K., Manouselis N. (2009). Social Navigation Module Specification Web Services API for Organic.Edunet, eContentplus Deliverable Document D5.3 March 2009.

[10] Manouselis, N. and Vuorikari, R. (2009). What If Annotations Were Reusable: A Preliminary Discussion. In M. Spaniol (Ed.), Proceedings of the 8th International Conference on Advances in Web-Based Learning - ICWL 2009, Lecture Notes in Computer Science, Vol. 5686, pp. 255-264. Berlin Heidelberg: Springer-Verlag.

[11] Manouselis N., Drachsler H., Verbert K., Duval E. (2012). *Recommender Systems for Learning*, Springer Briefs, ISBN 978-1-4614-4360-5, 2012. http://www.springer.com/computer/information+systems+and+applications/book/978-1-4614-4360-5]

[12] Manouselis N., Kosmopoulos T., Kastrantas K. (2009). Developing a Recommendation Web Service for a Federation of Learning Repositories", in Proc. of International Conference on Intelligent Networking and Collaborative Systems (INCoS 2009), Barcelona, Spain, IEEE Computer Press, 2009.

[13] Niemann K., Scheffel M., Wolpers M. (2009). A Comparison of Usage Data Formats for Recommendations in TEL", in Manouselis N., Drachsler H., Verbert K., Santos O.C.

(Eds.), Proceedings of the 2nd Workshop on Recommender Systems in Technology Enhanced Learning 2012, 7th European Conference on Technology Enhanced Learning (EC-TEL 2012), CEUR Workshop Proceedings, ISSN 1613-0073, Vol. 896, 95-100, 2012. [http://ceur-ws.org/Vol-896/]

[14] Nilsson, M. (2010), *From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization*. PhD thesis, KTH Royal Institute of Technology, Sweden, 2010.

[15] NSDL Annotation, 2012. Retrieved from https://wiki.ucar.edu/display/nsdldocs/comm_para+%28paradata+-+usage+data%29

[16] Paradata Specification v1.0, Retrieved from https://docs.google.com/document/d/1IrOYXd3S0FUwNozaEG5tM7Ki4_AZPrBn-pbyVUz-Bh0/edit

[17] Schmitz H.C., Wolpers M., Kirschenmann U., Niemann, K. (2012). Contextualized Attention Metadata. Human Attention in Digital Environments, Eds: Claudia Roda, Cambridge University Press, Cambridge, US, 2012 [pdf]

[18] Wolpers M., Najjar, J., Verbert, K., Duval, E. (2007). Tracking Actual Usage: the Attention Metadata Approach, Journal of Educational Technology and Society, 10 (3), 106-121.

[19] Zhou L., El Helou S., Moccozet L., Opprecht L., Benkacem O., Salzmann C., Gillet D. (2012). "A Federated Recommender System for Online Learning Environments". Advances in Web-Based Learning - ICWL 2012. Lecture Notes in Computer Science, Volume 7558/2012, 89-98, 2012 [DOI: 10.1007/978-3-642-33642-3_10]

[20] Dietze, S., Linked Data as facilitator for TEL recommender systems in research & practice, in Proceedings of 2nd Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL2012), at 7th European Conference on Technology-Enhanced Learning (EC-TEL 2012), Saarbrücken (2012).

[21] Duval, E., Hodgins, W., Sutton, S. and Weibel, S. (2002), "Metadata Principles and Practicalities", D-Lib Magazine, Vol. 8 No. 4, doi:10.1045/april2002-weibel.

[22] IEEE (2002), "IEEE Standard for Learning Object Metadata", IEEE Std 1484.12.1-2002, pp.i–32. doi: 10.1109/IEEESTD.2002.94128.

[23] Koutsomitropoulos, D.A., Alexopoulos, A.D., Solomou, G.D. and Papatheodorou, T.S. (2010), "The Use of Metadata for Educational Resources in Digital Repositories: Practices and Perspectives", D-Lib Magazine. January/February 2010, Vol. 16 No. 1/2, available at: http://www.dlib.org/dlib/january10/kout/01kout.print.html#14

[24] Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P., Bratsas, C. and Woodham, L. (2011), "Connecting Medical Educational Resources to the Linked Data Cloud: the mEducator RDF Schema, Store and API", in Linked Learning 2011, Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, CEUR-WS, Vol. 717, 2011

[25] World Wide Web Consortium (2008). W3C Recommendation, SPARQL query language for RDF, 2008, available at:www.w3.org/TR/rdf-sparql-query/

[26] Yu, H. Q., Dietze, S., Li, N., Pedrinaci, C., Taibi, D., Dovrolls, N., Stefanut, T., Kaldoudi, E. and Domingue, J. (2011), "A linked data-driven & service-oriented architecture for sharing educational resources", in Linked Learning 2011, Proceedings of the 1st International Workshop on eLearning Approaches for Linked Data Age, May 29, 2011, Heraklion, Greece.

[27] Zablith, F., d'Aquin, M., Brown, S. and Green-Hughes L. (2011b), "Consuming Linked Data Within a Large Educational Organization", Proceedings of the 2nd International Workshop on Consuming Linked Data (COLD) at International Semantic Web Conference (ISWC), October 23-27, 2011. Bonn, Germany.

[28] Zablith, F., Fernandez, M. and Rowe, M. (2011a), "The OU Linked Open Data: Production and Consumption", in Linked Learning 2011, Proceedings of the 1st International Work-

shop on eLearning Approaches for the Linked Data Age, at the 8th Extended Semantic Web Conference (ESWC), May 29, 2011, Heraklion, Crete

[29] Ebner, H. and Palmér, M. (2008), "A Mashup-friendly Resource and Metadata Management Framework", in Wild, Kalz, and Palmér (Eds.), *Mash-Up Personal Learning Environments,* Proceedings of the 1st Workshop MUPPLE, European Conference on Technology Enhanced Learning (EC-TEL), Maastricht, The Netherlands, CEUR Vol. 388, available at: http://ceur-ws.org/Vol-388/

[30] Ebner, H., Manouselis, M., Palmér, M., Enoksson, F., Palavitsinis, N., Kastrantas, K. and Naeve, A. (2009), "Learning Object Annotation for Agricultural Learning Repositories", *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Riga, Latvia, 2009.

[31] Dietze, S., Yu, H. Q., Pedrinaci, C., Liu, D., & Domingue, J. (2011). SmartLink: a Web-based editor and search environment for Linked Services, In *Proceedings of 8th Extended Semantic Web Conference*. Heraklion, Greece.

[32] Breslin, G. C., Harth, A., Bojars ,U., and Decker, S. (2005). Towards semantically-interlinked online communities. In *Proceedings of the 2nd European Semantic Web Conference (ESWC '05)*, Heraklion, Greece, LNCS 3532, p 500-514

[33] Bojars, U., Passant, A., Cyganiak, R., and Breslin, J. (2008). Weaving sioc into the web of linked data. In *Linked Data on the Web Workshop*.

[34] Subercaze, J. and Gravier, C. (2012) Lifting user generated comments to SIOC. *Proceedings of the 1st International Workshop on Knowledge Extraction & Consolidation from Social Media (KECSM 2012)*, vol 895, CEUR-WS.org

[35] Passant, A., Ciccarese, P., Breslin, J.G., Clark, T. (2009)SWAN/SIOC: Aligning Scientific Discourse Representation and Social Semantics. *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, DC, USA, 2009, Vol. *523*, CEUR-WS. org.

[36] Orlandi, F., and Passant, A. (2009) Enabling cross-wikis integration by extending the SIOC ontology. In *Proceedings of the Fourth Workshop on Semantic Wikis (SemWiki2009)* Workshop at 6th European Semantic Web Conference (ESWC2009), 2009

[37] Passant, A, and Raimond, Y. (2008) Combining Social Music and Semantic Web for Music-related Recommender Systems. In *Proceedings of the First Workshop on Social Data on the Web (SDoW2008)*, Vol. 405, CEUR-WS.org. Workshop at International Semantic Web Conference, 2008

[38] Pitsilis, G., and Knapskog , S. J. (2009) Social Trust as a solution to address sparsity-inherent problems of Recommender systems. *ACM RecSys 2009, Workshop on Recommender Systems & The Social Web*, Oct. 2009, New York, USA.

[39] Muñoz-Merino, P.J., Pardo, A., Kloos, C.D., Muñoz-Organero, M., Wolpers, M., Katja Niemann, K., and Friedrich, M. (2010). CAM in the Semantic Web World . *Proc. iSemantics 2010* September 1-3, 2010 Graz, Austria. ACM

[40] Muñoz-Merino, P.J., Kloos, C.D., Wolpers, M., Friedrich, M., Muñoz-Organero, M. (2010) "An Approach for the Personalization of Exercises Based on Contextualized Attention Metadata and Semantic Web technologies," *Proceedings 10th IEEE International Conference on Advanced Learning Technologies*, 2010, pp.89-91,

[41] Heitmann, B. and Hayes, C. (2010). Using Linked Data to Build Open, Collaborative Recommender Systems. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, 2010. Association for the Advancement of Artificial Intelligence (www.aaai.org).

[42] Stankovic, M., Wagner, C., Jovanovic, J., Laublet, P. (2010) Looking for Experts? What

can Linked Data do for You? *LDOW2010*, April 27, 2010, Raleigh, USA.

[43] Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., et al. (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. (N. Manouselis, H. Drachsler, K. Verbert, & O. C. Santos, Eds.) *Procedia Computer Science*, 1(2), 2849–2858. Retrieved from http://www.sciencedirect.com/science/article/B9865-50YNHC8-B/2/6297391c895db31c1e10c1258443a201

[44] Gasevic, G., Dawson, C., Ferguson, S.B., Duval, E., Verbert, K. and Baker, R.S.J. d. 2011. Open Learning Analytics: an integrated & modularized platform.

[45] Drachsler, H., Dietze, S., Greller, W., D'Aquin, M., Jovanovic, J., Pardo, A., Reinhardt, W., Verbert, K. (2012). 1st International Workshop on Learning Analytics and Linked Data. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12* (p. 9). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?id=2330601.2330607

[46] Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-Driven Research to Support Learning and Knowledge Analytics. *Educational Technology & Society*, 15(3), 133–148. Retrieved from http://www.ifets.info/journals/15_3/10.pdf

[47] Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikar, R., & Duval, E. (2011). Dataset-driven Research for Improving Recommender Systems for Learning. *1st International Conference on Learning Analytics and Knowledge (LAK 2011)*. New York: ACM Press. Retrieved from http://dx.doi.org/10.1145/2090116.2090122

[48] Manouselis, Nikos, Vuorikari, R., & van Assche, F. (2010). Collaborative Recommendation of e-Learning Resources: An Experimental Investigation. *Journal of Computer Assisted Learning*, 26(4), 227–242. Retrieved from http://doi.wiley.com/10.1111/j.1365-2729.2010.00362.x

[49] Sicilia, M.-Á., García-Barriocanal, E., Sánchez-Alonso, S., & Cechinel, C. (2010). Exploring user-based recommender results in large learning object repositories: the case of MERLOT. *Procedia Computer Science*, 1(2), 2859–2864. Retrieved from http://dx.doi.org/10.1016/j.procs.2010.08.011

[50] Fazeli, S., Drachsler, H., Sloep, P. (submitted). Toward a trust-based recommender system for teachers. *13th International Conference on Knowledge Management and Knowledge Technologies.* 4-6 September 2013, Graz, Austria.

[51] Ekstrand, M. D., Ludwig, M., Konstan, J. A., & Riedl, J. T. (2011). Rethinking the recommender research ecosystem. *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11* (p. 133). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?id=2043932.2043958

[52] Drachsler, H., Greller, W., Fazeli, S., Niemann, K., Sanchez-Alonso, S., Rajabi, E., Palmér, M., Ebner, H., Simon, B., Nösterer, D., Kastrantas, K., Manouselis, N., Hatzakis, I., & Clements, K. (2012). D8.1 Review of Social Data Requirements. Open Discovery Space project. http://hdl.handle.net/1820/4617.

[53] Drachsler, H., Verbert, K., Manouselis, N., Vuorikari, R., Wolpers, M., Lindstaedt, S., (2012). Preface of the dataTEL Special Issue. *Int. J. Technology Enhanced Learning, Vol. 4, Nos. 1/2, 2012.*

[54] Hendrix, M., Protopsaltis, A., Dunwell, I., de Freitas, S., Petridis, P., Arnab, S., Dovrolis, N., Kaldoudi, E., Taibi, D., Dietze, S., Mitsopoulou, E., Spachos, D., Bamidis, P., Technical Evaluation of The mEducator 3.0 Linked Data-based Environment for Sharing Medical Educational Resources, World Wide Web Conference 2012, 2nd International Workshop on Learning and Education with the Web of Data, Lyon, France; 04/2012.

[55] Aggregating Social and Usage Datasets for Learning Analytics: Data-oriented Challenges. *Katja Niemann, Giannis Stoitsis, Georgis Chinis, Nikos Manouselis and Martin Wolpers. Proceedings of the 3rd International Conference on Learning Analytics and Knowledge - LAK '13* (p. XX). New York, New York, USA: ACM Press.

[56] Rajabi, E., Greller, W., Kastrantas, K., Niemann, K., Sanchez-Alonso, S., (accepted). Social data harvesting and interoperability in ER federations. Special Issue on Advances in Metadata and Semantics for Learning Infrastructures. (Eds. Nikos Palavitsinis, Joris Klerkx, Xavier Ochoa) IJMSO - International Journal of Metadata, Semantics and Ontologies. Inderscience Publisher.

[57] Jack, K., Hristakeva, M., Garcia de Zuniga, R., Granitzer, M. (2012). Mendeley's open data for science and learning: a reply to the DataTEL challenge. Special Issue dataTEL - Data Supported Research in Technology-Enhanced Learning. (Eds. Hendrik Drachsler, Katrien Verbert, Nikos Manouselis, Riina Vuorikari, Martin Wolpers and Stefanie Lindstaedt). Int. J. of Technology Enhanced Learning, 2012 Vol.4, No.1/2, pp.31 - 46. DOI: 10.1504/IJTEL.2012.048309