Towards analytics and collaborative exploration of social and linked media for technology-enchanced learning scenarios

Conference or Workshop Item

Towards Analytics and Collaborative Exploration of Social and linked Media for Technology-Enchanced Learning Scenarios

Sergej Zerr, Mathieu d'Aquin, Ivana Marenzi, Davide Taibi, Alessandro Adamou, and Stefan Dietze

L3S Research Center,
Appelstr. 9a, 30176 Hannover, Germany
{dietze,marenzi,zerr}@l3s.de
m.daquin@open.ac.uk, davide.taibi@itd.cnr.it, alexdma@apache.org
Knowledge Media Institute, The Open University,
Walton hall, Milton Keynes
MK76AA United Kingdom
m.daquin@open.ac.uk, alexdma@apache.org
Italian National Research Concil, Institute of Educational Technologies
Palermo, Italy
davide.taibi@itd.cnr.it

Abstract. Social Web applications such as "Flickr", "Youtube" and "Slideshare" offer a vast body of multimedial knowledge, discoverable through the appropriate search interfaces and API's. This extensive information source, however, is largely unstructured and the available metadata is typically limited to title, tags and description for a resource. On the other hand, Linked Web Data is both structured and well described through a variety of metadata. Combining those sources opens promising direction for knowledge discovery and, at the same time, new challenges for collaborative searching in various Technology-Enchanced Learning Scenarios. In this paper, we explore how to support (collaborative) search in such scenarios through an initial analysis of the Web data landscape and introduce early results from efforts on exploiting Linked Data techniques to solve critical issues in this context.

1 Introduction

Learning traditionally has been perceived as a process taking place in some form of gated communities, where the class-room is increasingly being supplemented by Technology-Enhanced Learning (TEL) environments [1] and Web-based yet closed educational platforms [2]. This has led to a vast body of research, comprising a wide range of dedicated educational metadata standards, frameworks for competency modeling or recommender systems in TEL [3] (see [4] for a more thorough overview). Fundamental challenges related to this work arise from its assumption that learning of individuals could be contained, described and supported within rather isolated environments and communities facilitated by dedicated TEL content and platforms.

The *open educational resources* (OER) movement already partially considered the liberalization of education and successfully generated an unprecedented range of standards and content, including examples such as MIT OpenCourseware[1] (OCW), GLOBE[2] (Global Learning Objects Brokered Exchange) or the more recent MOOC (Massively Open Online Courses) movement. Still, these initiatives tend to perceive learning as a process somewhat disconnected from non-educational activities, leading to a limited scope and take-up [5]. However, along with the emergence of the Web, particularly the Social Web, and its increasing ubiquity, learning has evolved into a multi-faceted, ubiquitous and continuous process of knowledge acquisition, which takes place in a wide variety of settings where the distinction between educational and not explicitly educational activities has become blurred and to a large extent obsolete. The notions of Web-based informal and pervasive learning [6] broaden this understanding, but support for Web-based informal learning still appears to be lacking behind its full potential.

Web-based knowledge acquisition is characterised by fine-grain and highly diverse knowledge items, ranging, for instance, from Wikipedia articles and YouTube videos to scholarly papers and semi-didactic material, such as user-generated slidesets. In this context, non-educational information Web resources, in particular user-generated content, gained similar importance as dedicated educational material. This development has been fundamentally driven by technological advancements, such as the emergence of social media together with a Web of Data which allowed an unprecedented body of knowledge to be reused and shared across the Web. Technological drivers of this evolution are, for instance, open APIs and interfaces, and more recently, the Linked Data (LD) [7] approach, which contributed successful principles based on established W3C standards such as RDF and SPARQL.

While the emerging new forms of (informal) learning require a new set of techniques and skills, prevalent Web Data present themselves also as an unprecedented resource for deriving and detecting new pattern and theories about Web-based informal learning activities. Web Science [8] offers the scientific and analytical toolset to deepen the understanding and detect and document such newly emerging pattern and behaviors.

This has recently led to a growing inter-disciplinary community of researchers from data engineering, Semantic Web, education and social sciences, converging in disciplines such as Learning Analytics [9]. Here, the Web of Data emerged as one important pillar by providing the data which allows the derivation and validation of theories about prevalent trends. Recent efforts on using LD for educational data sharing emerged, shown, for instance, by data released from the Linked Universities[3] movement (e.g., from The Open University UK[4] or Oxford

[1] http://ocw.mit.edu/index.htm
[2] http://globe-info.org/
[3] http://linkeduniversities.org
[4] http://data.open.ac.uk

University[5]) or the efforts documented by the Linked Education[6] community (see also [10]). However, these approaches so far were mainly dedicated to expose educationally related data, where its use in learning scenarios is often not well defined. Additionally, the exploitation of the vast body of knowledge not explicitly dedicated to learning has not yet been considered widely.

In this paper, we argue that Web Science and Linked Data provide resources and methods for (a) analysing, detecting and documenting the ongoing paradigm shift as well as the relevant data landscape and (b) acting as facilitator by resolving interoperability issues which naturally arise when adopting the broad, diverse and less restricted view sketched above. To underline the emerging Web-supported practices in learning, we first introduce results from a study analysing (social) Web information and data for its educational relevance, scope and diversity. We then propose early research results, emerging from the LinkedUp project[7], which aim at applying and exploiting LD (principles) together with data analytics approaches to facilitate new forms of Web-based learning by adopting a broader view on educationally relevant resources.

2 (Social) Web Data for Learning: Exploratory Investigations

Data and their analytics facilitate education on a variety of levels, for instance, by providing (a) knowledge and information resources as input to learning processes, (b) social and personal data which enable the investigation and detection of emerging patterns, (c) social and personal data for use in (learning analytics) and recommender system scenarios. In particular with respect to (a), it is assumed that Web data and content, in particular social media and user-generated content, provide valuable input for knowledge acquisition and informal learning [5]. Here we consider both content, often described with structured metadata and structured data as exposed, for instance, via the datasets being part of the Linked Open Data cloud[8] or registered in the DataHub[9]. To further investigate the relevance of different Web sources, we performed an initial study, taking advantage of usage data gained from LearnWeb[10], a social Web-based learning environment [11]. LearnWeb allows users to search a variety of data aggregators, ranging from generic search engines to social media sites (see full list in [11]), as well as offering the possibility to identify and re-annotate resources of use for instructional and learning purposes. Retrieved resources of any kind are tagged and rated by users leading to a knowledge base of 1730 *repurposed resources*. These were retrieved through 3439 unique queries from 337 distinct users (state 24/01/2013).

[5] http://data.ox.ac.uk
[6] http://linkededucation.org
[7] LinkedUp - Linking Web Data for Education Project: http://linkedup-project.eu
[8] http://lod-cloud.net
[9] http://thedatahub.io
[10] LearnWeb, also known as LearnWeb2.0, http://learnweb.l3s.uni-hannover.de/lw

Thus, this historical data of LearnWeb provides useful information to estimate the educational relevance of particular Web data sources. Our primary research question addresses the perceived educational relevance of Web resources and the diversity of sources, as an indicator for the educational relevance of social media and Web data in general.

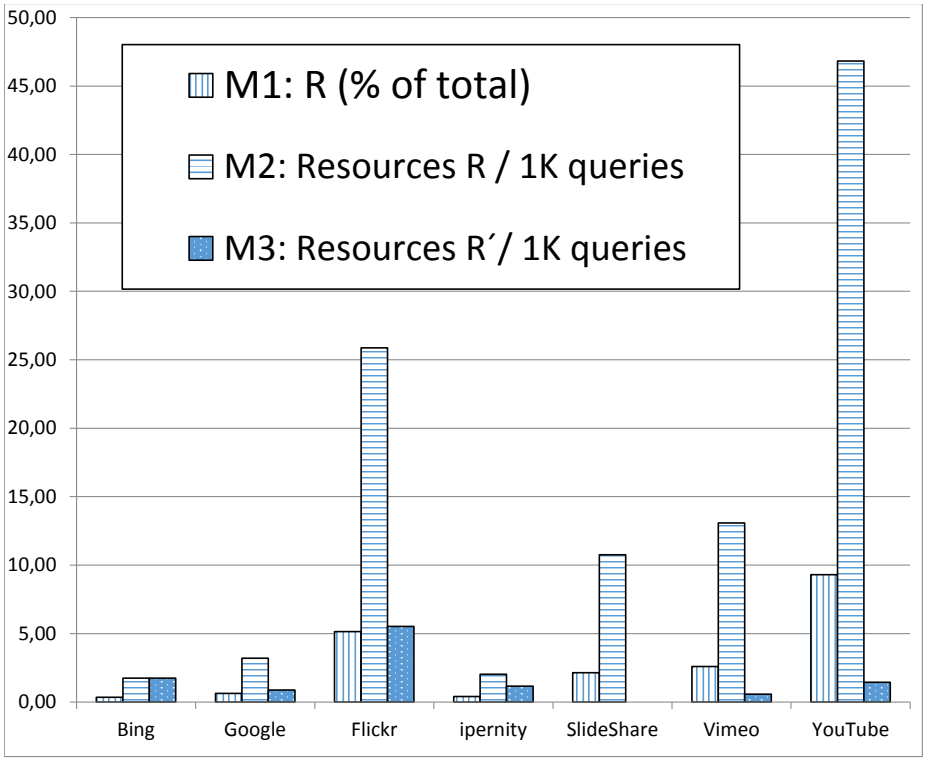

Fig. 1. Resources qualified by Learn Web users as educationally relevant - M1 (left), M2 (middle), M3 (right)

For this purpose, for each Web source Si we investigate 3 measures as indicator for educational relevance. (M1) is the proportion (%) of the overall repurposed resources R retrieved from Si, (M2) the ratio of *repurposed resources* R' from Si per 1K submitted queries, and (M3) the ratio of high-rated, *repurposed resources* R' from Si per 1K queries. With respect to M3, we consider a resource as high-rated if its average user-assigned score is equal or above the threshold T (here T=2,5; rating scale integers 1..5). Results are depicted in Figure 1 and Table 1. Note, though typical queries, e. g. "astronomia aurora boreale", were defined with the intention of finding educational material for a certain topic, mostly these appear to not contain any indicator of the learning intention.

Table 1. Resources qualified by LearnWeb users as educationally relevant - total R, total R', M1, M2, M3 measures per source

	#R	# R'	M1	M2	M3
Bing	6.00	6.00	0.35	1.74	1.74
Google	11.00	3.00	0.64	3.20	0.87
Flickr	89.00	19.00	5.14	25.88	5.52
ipernity	7.00	4.00	0.40	2.04	1.16
SlideShare	37.00	0.00	2.14	10.76	0.00
Vimeo	45.00	2.00	2.60	13.09	0.58
YouTube	161.00	5.00	9.31	46.82	1.45
Upload	296.00	35.00	17.11	n/a	n/a
Links	1082	45.00	62.54	n/a	n/a

Additionally, users were able to manually add links to useful informal learning resources (category "Links" in Table 1). A breakdown of the most frequently provided domains confirmed results from Table 1 by revealing YouTube as most frequent source (416), while the remaining domains varied heavily (Figure 2), including a range of museums, media broadcasters or educational sources. Although the investigated samples are rather small, the results of our preliminary investigation suggest that, in particular social media sources such as YouTube, SlideShare or Flickr contain a high proportion of material of use for knowledge acquisition and skill development, while a widespread variety of highly diverse sources provides additional resources on more specific topics or domains. We are currently conducting a more elaborate study based on a wide range of structured datasets retrieved from the DataHub[9]. While the DataHub contains 57 datasets which are explicitly tagged with the term "education", our current studies consider not explicitly labeled educationally relevant datasets, such as Europeana[11], BBC Programmes [12] or scholarly publications. Datasets are investigated for their educational relevance by computing their semantic similarity to learning-related WordNet[12] synsets and user studies. Additionally, coverage of domains and resource types are explored through mappings of datasets and schemas to DBpedia[13]. While results will be presented as part of future work, initial assessments show a correlation with the above results.

3 Facilitating Knowledge Exploration - Towards an educational Knowledge Graph

Having investigated on the educational relevance and diversity of Web resources, we argue that Web-based informal learning calls for support beyond the scope of traditional educational technologies, but through aiding learners in exploring Web data and content.

[11] http://www.europeana.eu/
[12] http://wordnet.princeton.edu/
[13] http://dbpedia.org

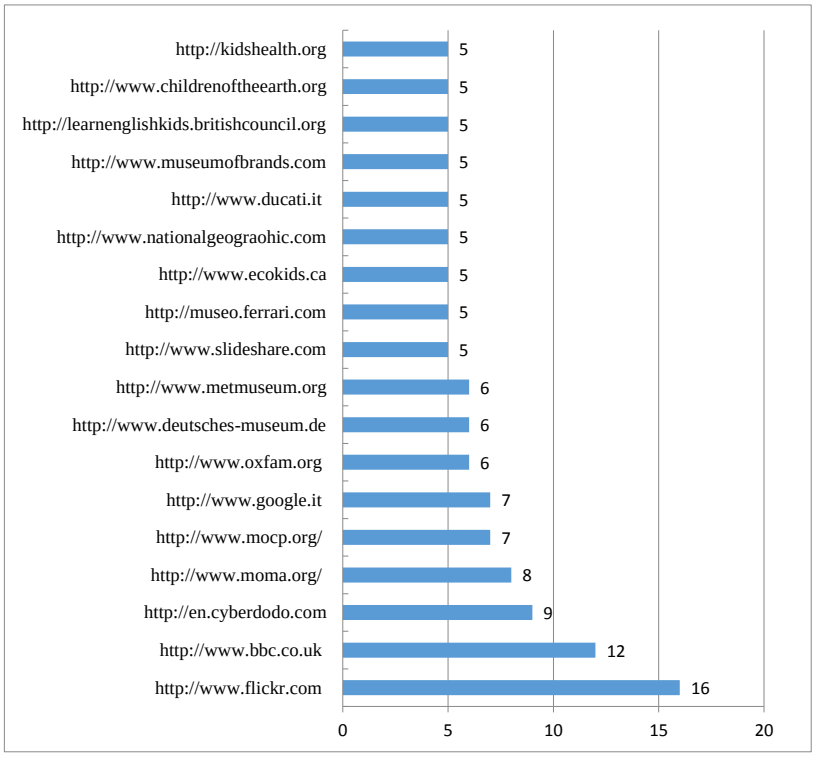

Fig. 2. Frequent domains of resources added manually by LearnWeb users (breakdown of "Links" category in Table 2). Note: the most frequent domain "http://youtube.com" (with 416 occurrences) was excluded for the sake of readability.

This requires taking into account the diversity of knowledge on the Web and to make it accessible and digestible from an educational perspective. While the latter requires the support of learning not by a narrow set of TEL technologies but by adopting a more general view on Web data, these challenges can only be addressed by tackling substantial issues, for instance, with respect to quality, heterogeneity and interoperability of used vocabularies, languages or schemas. In this context, particularly LD techniques - fundamentally aimed at knowledge reuse and sharing on the Web - offer potential to act as facilitator. The opportunities arising from the LD approach are two-fold: (i) LD principles have emerged as defacto-standard for Web data and knowledge sharing in particular also in the field of education [5], (ii), the LD community has produced a vast body of knowledge [7], which in itself constitutes an important resource for educational purposes.

As part of the LinkedUp project[17], we currently conduct a number of data curation activities, aimed at assessing, cataloging and exposing all sorts of Web data of educational relevance (independent of their original intention) where the overall vision entails the creation of an educational knowledge graph which enables learners to explore all forms of suitable Web data and content. This work is two-fold:

- *Community-oriented data cataloging on the Data Hub*: similar to the successful approaches of the Linked Open Data community effort, a dedicated group ("linked-education"[14]) is currently being maintained and populated to gather and tag educationally relevant data
- *Semantic, syntactic and infrastructural alignment*: selected datasets will be annotated with a specific vocabulary which allows data to be exposed in more coherent and accessible ways, aiming to bridge between multiple languages and descriptive approaches.

As part of the latter a general vocabulary ("linked-education" vocabulary[15], adopting VoID[16]) for the description, cataloging and alignment of educationally relevant datasets is under development. An initial classification of datasets into educationally relevant categories (such as "educational resource", "scholarly paper" "video lecture") provides suitability indicators for particular tasks. While we intend to apply the schema in combination with data interlinking techniques to provide a learning-oriented view on Web data in general, it is currently being applied initially to an exemplary dataset[17] containing 5,953,623 distinct resources (around 60 million RDF statements) extracted from different sources (Table 2 depicts the sources along with the number of avalable resources). Sources were selected to reflect the diversity documented above and include Europeana, BBC Programmes, mEducator Linked Educational Resources [5], ACM Digital Library[18], DBLP[19] and Linked Universities [3], a collection of university video lectures from YouTube. Schema mapping and data interlinking techniques are applied to facilitate queries across data independent of its origin, for instance, to retrieve YouTube videos and BBC programs together with academic publications about astronomy topics. Enrichment of data with references to joint LD reference vocabularies (such as DBpedia) together with basic named entity recognition (NER) and clustering techniques proved as efficient means to provide a unified and interlinked data graph, resolving fundamental problems arising from distinct vocabularies and languages. Hence, from a learning and knowledge exploration perspective, these techniques have shown significant potential for offering learners a more integrated view on diverse Web data.

[14] http://datahub.io/group/linked-education
[15] http://data.linkededucation.org/ns/linked-education.rdf
[16] http://vocab.deri.ie/void
[17] SPARQL endpoint:
http://data.linked.education.org/openrdf-sesame/repositories/linked-learning
[18] http://acm.rkbexplorer.com/
[19] http://dblp.l3s.de/

Table 2. Resources per origin dataset in experimental Linked Education dataset

Dataset	**# Resources**
Europeana	2251749
DBLP	2988891
ACM	1576988
LinkedUniversities	17900
BBC	15966
mEducator	2219

4 Conclusions

In this paper, we have motivated the need for a broader view on Web-based education and in particular educational technologies and provided insights on how Web data in general can facilitate this paradigm shift. In particular Linked Data, offering a body of structured knowledge and set of techniques, together with traditional collaborative search and data engineering methods can act as a means to integrate learning-related knowledge into a coherent educational graph and support enrichment of the corresponding resources with useful metadata. More recent Microformats and efforts like schema.org[20] can further bridge between the structured Web of data and materials on the Web such as web pages, images and videos by providing a common vocabulary enabling search engines to discover unstructured resources.

Our investigation points out promising directions for larger substantial studies to address issues with regard to the diversity of data quality, domain coverage and varying levels of trust, particularly when dealing with distributed data. These represent fundamental obstacles for data consumers in general and apply even more in learning scenarios. Additional and so far under-acknowledged issues arise from the diversity of licensing schemes and policies adopted by various distributed data providers. Hence, highly multidisciplinary communities are required and currently assembled, e. g., by the LinkedUp project, to advance the redefinition of learning on the Web through exploitation of existing Web knowledge and data.

[20] http://schema.org: a joint effort of, for instance, Bing, Yahoo, Google or Yandex to describe common Website markup vocabularies using microdata

References

1. Dietze, S., Gugliotta, A., Domingue, J.: A semantic web service oriented framework for adaptive learning environments. In Franconi, E., Kifer, M., May, W., eds.: The Semantic Web: Research and Applications. Number 4519 in Lecture Notes in Computer Science, Berlin, Springer (2007) 701–715 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, 2007. Proceedings.
2. Dolog, P., Henze, N., Nejdl, W., Sintek, M.: Personalization in distributed e-learning environments. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters. WWW Alt. '04, New York, NY, USA, ACM (2004) 170–179
3. Drachsler, H., Hummel, H.G.K., Koper, R.: Personal recommender systems for learners in lifelong learning networks the requirements, techniques and model. Int. J. Learn. Technol. **3**(4) (July 2008) 404–423
4. Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H., Giordano, D., Marenzi, I., Nunes, B.P.: Interlinking educational resources and the web of data - a survey of challenges and approaches. Emerald Program: electronic Library and Information Systems **47**(1) (2013)
5. Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D.: Linked education: interlinking educational resources and the web of data. In: The 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications. (March 2012)
6. Schugurensky, D.: The forms of informal learning: Towards a conceptualization of the field. (2000)
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3) (2009) 1–22
8. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., Weitzner, D.: Creating a science of the web. Science **313**(5788) (August 2006) 769–771
9. Siemens, G.: Learning analytics: Envisioning a research discipline and a domain of practice. In: Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge. LAK '12, New York, NY, USA, ACM (2012) 4–8
10. Yu, H.Q., Pedrinaci, C., Dietze, S., Domingue, J.: Using linked data to annotate and search educational video resources for supporting distance learning. Journal of IEEE Transactions on Learning Technologies **5**(2) (April/Spring 2012) 130–142
11. Marenzi, I., Zerr, S.: Multiliteracies and active learning in clil. the development of learnweb2.0. IEEE Trans. Learn. Technol. **5**(4) (January 2012) 336–348
12. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media meets semantic web — how the bbc uses dbpedia and linked data to make connections. In: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications. ESWC 2009 Heraklion, Berlin, Heidelberg, Springer-Verlag (2009) 723–737