British Journal of Educational Technology (2015)
doi:10.1111/bjet.12283

Exploring TED talks as linked data for education

Davide Taibi, Saniya Chawla, Stefan Dietze, Ivana Marenzi, Besnik Fetahu

Davide Taibi graduated at the University of Palermo, currently is a researcher of the Institute for Educational Technologies at the National Research Council of Italy. His research activities are mainly focused on the application of innovative technologies in the e-learning field, with particular emphasis on Mobile Learning, Semantic Web and Linked Data for e-learning, standards for educational processes design, Open Educational Resources. In the last few years, his research has been addressed toward the Learning Analytics field. He has worked as a contract professor at the University of Palermo. Saniya Chawla, (B. Tech-Computer Science) is currently an MBA student at the Indian Institute of Management-Ahmedabad. She has worked with new and emerging technologies like semantic web and linked data. Her current area of research is inclusive of these areas which falls under the umbrella of web science. Along with this, exploring new strategic and financial prospects tied up with digital media and technology also adds up to her research interests. Stefan Dietze is a research group leader at the L3S Research Center (Germany). His research interests are in the areas of Linked Data, Web Science and Artificial Intelligence, and their use in actual application domains. Stefan currently is co-ordinator of several European R&D projects and he is co-chair of several working groups in the Semantic Web area. His work has been published in numerous major conferences and journals, he is member of many organization and program committees and editorial boards and a frequent invited speaker. Ivana Marenzi, PhD, is senior researcher at the L3S Research Center in Hannover. Throughout her career she has specialised in the relationship between technology and communication; her main area of research in Technology Enhanced Learning includes the support of collaborative and lifelong learning. Besnik Fetahu is a PhD student at the L3S Research Center, Leibniz University Hannover. Previously he was part of the Graduate School of Computer Science at Saarland University. He holds a M.Sc. in Computer Science from the University of Sofia and a B.Sc. degree in Computer Science from the University of Prishtina. His main research interests lie in the field of Information Retrieval and Extraction, with the main focus on bridging the gap between unstructured and structured data on the Web. Address for correspondence: Dr Davide Taibi, Institute for Educational Technologies, Research Council of Italy, Via Ugo La Malfa, 153, 90146 Palermo, Italy. Email: davide.taibi@itd.cnr.it

Abstract

In this paper, we present the TED Talks dataset which exposes all metadata and the actual transcripts of available TED talks as structured Linked Data. The TED talks collection is composed of more than 1800 talks, along with 35 000 transcripts in over 30 languages, related to a wide range of topics. In this regard, TED talks metadata available in structured, multilingual and HTTP-accessible form constitute a valuable resource, for instance, for schoolteachers, to explore controversial contemporary topics with their students in order to stimulate awareness and critical thinking or as a means for language learning. Moreover, being compliant with state-of-the-art Linked Data principles, our dataset facilitates the computation of links with related data and resources. The TED dataset is used by a number of educational applications, and it is included in the LinkedUp Data Catalog.

The TED talks dataset

- Location:

Dataset described at: http://datahub.io/dataset/ted-talks
SPARQL endpoint: http://data.linkededucation.org/linkedup/ted/sparql
Dump: http://data-observatory.org/ted_talks/tedtalksdump.nt.gz

- Creator: Saniya Chawla, Besnik Fetahu, Stefan Dietze, Davide Taibi
- Date: June 9, 2014

- Format: application/rdf + xml
- Restrictions to use: Creative Commons—BY license (licensees may copy, distribute, display and perform the work and make derivative works based on it only if they give the author or licensor the credits in the manner specified by these)

Introduction

TED[1] is a series of global conferences spanning over all the topics from business to science to entertainment. Since 2006, TED talks have been made available on the TED web site. Nowadays, more than 1800 talks are publicly available along with a rich collection of 35 000 transcripts in over 30 languages at the time of writing, and the number is constantly growing. TED talks are translated by more than 15 000 volunteers within the Open Translation Project.[2] In order to ensure good quality, transcripts are reviewed by language coordinators before publication. The videos are released under a Creative Commons BY-NC-ND licence so that they can be freely shared and reposted.

The dataset described in this paper makes available all metadata and the actual transcripts of TED talks as structured Linked Data (Bizer, Heath & Berners-Lee, 2009), an increasingly common practice in educational settings (Dietze *et al*, 2013). Following established and state-of-the-art Linked Data principles, the goal is to enable reuse and take-up of TED talks particularly in educational scenarios and to facilitate interoperability and interlinking of TED talks with related resources of educational relevance in the Web of data (Dietze *et al*, 2012).

TED talks are considered valuable resources in educational settings from two different perspectives: (1) as knowledge resources in their own right containing valuable and accessible content and insights for learners and (2) as educational material dedicated to language learning, particularly facilitated through the wide range of multilingual transcripts.

Recently, the dataset has been included in the LinkedUp[3] Data Catalogue (d'Aquin, Adamou & Dietze, 2013). Moreover, the dataset has been integrated in the LearnWeb-OER platform[4] in order to improve search results related to teaching and learning scenarios. LearnWeb-OER is a collaborative system, extensively used by teachers and students of various universities and schools, which empowers its users to collect and share resources from various Web sources, such as YouTube, Bing, Flickr, within a single environment.

Creating linked data-compliant metadata of TED

In order to make the TED data available as Linked Data, the TED portal was crawled to extract information about the TED talks. The process of converting extracted metadata into RDF was implemented by means of a four-step pipeline (Figure 1): (1) crawling TED website for videos and relevant attributes, (2) creating an appropriate RDF vocabulary (schema) for the collected attributes, (3) lifting the collected data into RDF in accordance with the defined schema from the previous step and (4) uploading the RDF to an actual triple store.

In the initial step, the pages in the TED website were crawled and parsed in order to extract the required information from the HTML pages. The extracted metadata is listed in table 2 together with the corresponding properties used to represent the data in the RDF dataset. Given the well-structured nature of the TED HTML pages, extraction was straightforward, and most effort was dedicated to lifting data into a suitable RDF vocabulary. Following established Linked Data principles which are geared towards wider reuse and interoperability of data, we have reused

1 http://www.ted.com/.
2 http://www.ted.com/about/programs-initiatives/ted-open-translation-project.
3 LinkedUp is an EU-funded project with the aim to push forward the exploitation of the vast amounts of public, open data available on the Web, in educational sector.
4 http://learnweb.l3s.uni-hannover.de/lw/ & https://www.l3s.de/projects/internal/LearnWeb-OER.

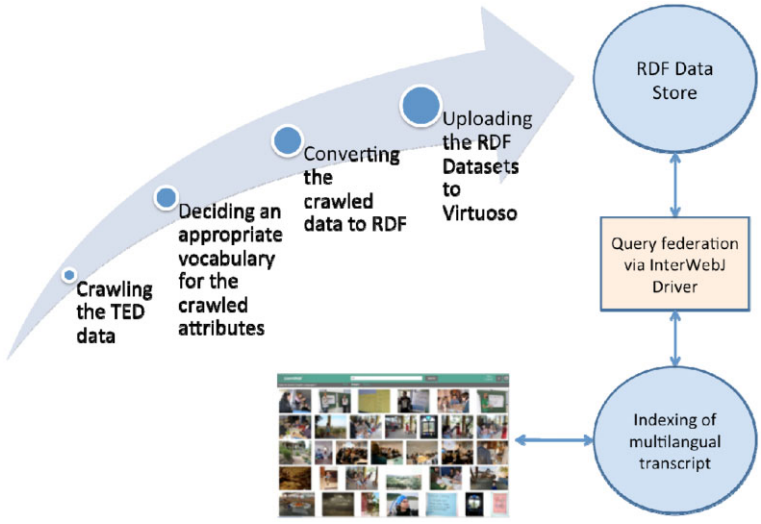

Figure 1: Four-step pipeline approach

concepts and predicates from already published vocabularies. The main RDF vocabularies are the DC-Terms Vocabulary,[5] the W3C Ontology for Media Resource,[6] the BIBO ontology,[7] Dublin Core[8] and the Schema.org[9] vocabularies. A complete list of concepts and predicates is listed in tables 1 and 2. Data were then uploaded and stored in a RDF storage based on Open Link Virtuoso,[10] which provides a public SPARQL endpoint for querying and dereferencing features.

The SPARQL endpoint is available at the following URL: http://data.linkededucation.org/linkedup/ted/sparql.

An example of the RDF/XML representation of a transcript can be accessed at: http://data.linkededucation.org/resource/ted/talks/alex_wissner_gross_a_new_equation_for_intelligence. Table 1 reports the number of talks and corresponding transcript contained in the dataset.

Application and usage

One of the goals of the present work was to improve data access to the TED talks collection. Sugimoto and Thelwall (2013) raised the issue of the quality of data available to compile a list of TED Talks for analysis; even the TED-endorsed list provided on the official TED blog was not comprehensive. The availability of TED talks metadata as Linked Data and its inclusion into the LinkedUp Data Catalog[11] has promoted the development of applications facilitating the access to TED videos. For instance, HyperTED[12] provides an innovative way to explore TED talks at a finer level of granularity. Based on the approach followed by HyperTED, fragments of the video are connected with their corresponding concepts and topics. In this way, students can focus on important parts of a video.

5 http://dublincore.org/documents/dcmi-terms/.
6 http://www.w3.org/TR/mediaont-10/.
7 http://bibliontology.com.
8 http://dublincore.org/documents/dcmi-terms/.
9 http://schema.org.
10 http://virtuoso.openlinksw.com.
11 http://data.linkededucation.org/linkedup/catalog/.
12 http://linkedtv.eurecom.fr/Hyperted/.

Table 1: List of property used in the TED talks dataset

Domain	*Property*	*Vocabulary*	*#*
bibo:AudioVisualDocument (talks)	Title	ma:title	1753
	Speaker	ma:hasContributor	1753
	Description	ma:description	1753
	Location	ma:location	1659
	Image	schema:image	1753
	Duration	mMa:duration	1659
	Date	dc:date	1753
	Keywords	dcterms:subject	1753
bibo:Document (transcript)	Language of transcript	dc:language	45 982
	Transcript Of	bibo:transcriptOf	45 982

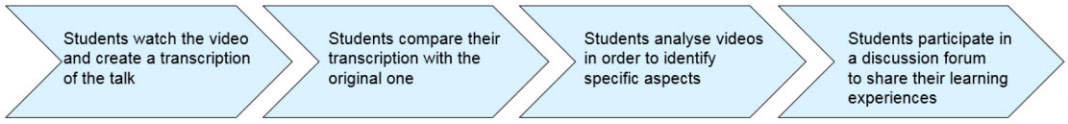

Figure 2: Learning scenario

As stated in the introduction, an important application of the TED talks in education has been by using the LearnWeb-OER system. One of the largest communities using LearnWeb-OER is the YELL/TELL professional online community of English language teachers belonging to the University of Udine, Italy (Bortoluzzi & Marenzi, 2014). The community of teacher–trainers, trainee–teachers and students has started in January 2012 and includes in total 538 users at the time of writing. They use LearnWeb-OER not only to share their resources and teaching experiences, but also to search the Web for additional educational resources for their students. Regarding the data related to TED talks, an indexing phase has been added to improve quality and efficiency of search results. The indexing phase is initially performed on full text in order to enable keyword-based search, including search on transcripts. Moreover, the properties "title," "description" as well as the whole text of the transcripts were indexed to facilitate efficient free-text searches over these fields.

TED videos are a valuable source for teaching languages especially because of the presence of multilingual transcripts. In general, TED videos provide a useful source of information to learn a subject or a language or to be updated on the latest news or research. A number of language teachers use the TED videos and data to teach English to their students at school or at university. In particular, the YELL community of teachers, including trainee teachers, primary and secondary school teachers, and researchers, use the TED talks for their lessons in class.

The availability of the TED dataset allows teachers to carry out educational activities with their students, for example, to highlight specific terms in the transcripts, to easily display synonyms, and to find more contextual information about the topic or the subject of a talk from available Linked Data knowledge sources, such as DBpedia, Freebase or Yago.

The TED Talks dataset is currently used at the University of Lecce, Italy, within the course "Interpretazione lingua inglese I" supported by LearnWeb-OER. The implemented learning scenario is described in Figure 2 above. At the third step, students analyse the video in order to identify key concepts, markers of the textual structure, expression indicating the stance of the speakers, expression indicating epistemic mode and expression of deontic mode.

Ethical considerations

Given the nature of the data, covering metadata about educational resources, user-related and personal data have not been collected and exposed in any way. While our dataset is generated by crawling data from the official TED portal,[13] all data complies with the TED terms of use, where TED organizers monitor user-generated contents (eg, comments and discussions related to the talks) in order to avoid illegal content, or content not compliant with the main purposes of the TED conferences. The TED dataset provides all metadata available from the TED portal, but by exploiting Linked Data principles, it allows the reuse through standard HTTP interfaces and, hence, facilitates enrichment and correlation with related data and resources. In LearnWeb-OER, on the other hand, basic user data are collected, where data are stored and exchanged only via secure means and protocols. While LearnWeb-OER provides statistical analysis of activities, all data are anonymised.[14]

Limitations

While the processing pipeline is stable and tailored to re-crawl new data as a daily process in an automated fashion, there is a lag of maximum 24 hours between the emergence of new TED talks on the Web and their appearance within our dataset. In addition, while the Linked Data approach fundamentally relies on creating links with other datasets, this activity is currently left to the data consumer or tools, eg, LearnWeb-OER or HyperTed, which can use the TED talks dataset to interlink it with related Web resources, such as correlated educational materials or factual knowledge, for instance, from DBpedia/Wikipedia or language-related definitions from WordNet.

Acknowledgements

This work has been partially supported by the European Union Seventh Framework Programme (FP7/2007–13) under grant agreement No 317620: LinkedUp project (http://linkedup-project.eu/).

References

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems*, *5*, 3, 1–22. doi: 10.4018/jswis.2009081901.

Bortoluzzi, M. & Marenzi, I. (2014). YELLing for collaborative learning in teacher education: users' voices in the social platform LearnWeb2.0. *International Journal of Social Media and Interactive Learning Environments*, *2*, 2, 182–198. doi: 10.1504/IJSMILE.2014.063402.

d'Aquin, M., Adamou, A. & Dietze, S. (2013). Assessing the educational linked data landscape. In *Proceedings of the 5th Annual ACM Web Science Conference* (WebSci '13) (pp. 43–46). New York, NY, USA: ACM. doi: 10.1145/2464464.2464487.

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N. & Taibi, D. (2012). Linked education: interlinking educational resources and the Web of data. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (SAC '12) (pp. 366–371). New York, NY, USA: ACM. doi: 10.1145/2245276.2245347.

Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H. Q., Giordano, D., Marenzi, I. *et al* (2013). Interlinking educational resources and the web of data: a survey of challenges and approaches. *Emerald Program: Electronic Library and Information Systems*, *47*, 1, 60–91. doi: 10.1108/00330331211296312.

Sugimoto, C. R. & Thelwall, M. (2013). Scholars on soap boxes: science communication and dissemination in TED videos. *Journal of the American Society for Information Science and Technology*, *64*, 4, 663–674. doi: 10.1002/asi.22764.

13 http://www.ted.com/about/our-organization/our-policies-terms/ted-com-terms-of-use.
14 http://learnweb.l3s.uni-hannover.de/lw/statistics.jsf.