Scientific Audiovisual Materials And Linked Open Data: The TIB Perspective

Paloma Marín Arraiza,
German National Library of Science and Technology (TIB), Germany

Abstract
Libraries are starting to use Linked Open Data (LOD) to provide their data (li-brary data) for reuse and to enrich them. However, most initiatives are only available for textual resources, whereas non-textual resources stay aside. Firstly, this paper discusses the potential of library data to be published as LOD. Secondly, it focuses on the library data related to the management of audiovisual scientific materials in the TIB|AV-Portal. The use LOD Standards to support multilingual functionalities and data reuse is outlined. Future developments lead to the building of semantic applications based on LOD Struc-tures

Keywords: scientific videos, linked open data, data publishing, semantic applications, grey literature, library data.

1 Introduction

A large amount of data is generated by global science every day from multiple sources. Libraries' task is to organise and classify all the resources and their respective data, generating new data known as 'library data'.
Data on the web exists in a structured form in databases, and in semi-structured forms in textual and non-textual collections. To deal with both the flood of information as well as the range of hetero-geneous data formats, a new approach was needed for information searching and access. The data should no longer be isolated but connected to other data, and become accessible, explorable and discoverable by both people and machines. Thus, the HTML based web changed into the Web of Data, also referred to as Semantic Web. It derives value from data and creates pathways between datasets and resources (Heath and Bizer, 2011).

The derivation of value from the data is possible due to the structure this data presents. In the web of data, information and knowledge are stored in simple structures known as ´triples´, which consist of three parts: subject, predicate and object. To publish and interlink structured data on the web, RDF (Resource Description Framework) is used. Data linked to other and published under open licenses (such as CC0[1]) are known as Linked Open Data (LOD).

Tim Berners-Lee (2006) outlined the principles structured data need to follow in order to be published as Linked Data:

1. Use URIs (unique resource identifier) as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

The use of semantic technologies and LOD concerns libraries and has a lot of advantages. LOD is better retrievable and easier to interlink since URIs, that define things, are more stable than URLs (uniform resource locator), that define addresses. The interoperability and further use of the data is improved since LOD is based on open web standards and besides the used data model, RDF, is more flexible than other library standards (e.g. MARC) (Pohl and Danowski, 2011). Therefore, W3C Stand-ards become a reliable and flexible alternative.
The aim of this article is to present the use of LOD in libraries and at the TIB, specifically in the TIB|AV-Portal, and how LOD support multilingualism, data reuse and information retrieval applica-tions.
The structure of the paper is as follows. Section 1 introduces the topic and describes the structure of the paper. Section 2 defines library data and specifies the use of LOD in libraries for textual and non-textual materials. Section 3 presents the use of LOD in the TIB|AV-Portal. Section 4 concludes the article.

[1] Creative Commons Zero, public domain.

2 Linked Open Data in Libraries

2.1 Library Data and new types of scientific information

In the Library Data Incubator Group Final Report (2011), library data is defined as 'any type of digital information produced or curated by libraries that describes resources or aids their discovery'. So far, scientific written texts such as books and journal articles were the main resource of academic libraries. Non-textual materials were considered to be inappropriate for academic purposes, and were usually regarded as general-interest publications rather than as 'proper' scientific publications (Löwgren, 2011). However, the scientific landscape changes constantly and academic librarians must be aware of the importance of this multimodal scholarship (Spicer, 2014). This means, they should provide platforms to support these materials, index them properly, and promote their accessibility and reuse.

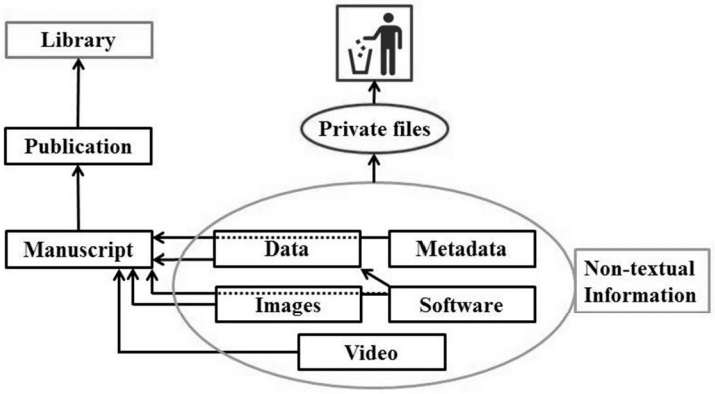

Fig. 1. Flow of scientific information from research to publication (modified after Klump et al., 2006[2]).

Therefore, as mentioned in the Pisa Declaration (2014), 'grey materials' should be persistent identified and linked, as far as possible, with other publications. This generates a big amount of library data that aids the classification and finding of relevant information. Not doing so lead to a loss of scientific information, as shown in figure 1.

2.2 The potential of library data for Linked Open Data

The W3C consider libraries an important focus for LOD because of being important content providers. Being usually structured information, library data present a high potential to be shared as LOD. Peset et al. (2011) highlight libraries projects where LOD play a crucial role: Authorities and vocabularies of the Library of Congress, Linked Data service of the Deutsche Nationalbibliothek (DNB), and Libris, the collaborative Swedish catalog. Moreover, the Web is considered a global data space where libraries should publish their data using RDF and open licenses. Doing so, they promote and support data reuse (Saorín, 2012).

However, most library content is text-based providing few non-textual materials. An example of aggregation environment with textual and non-textual materials is Europeana. Europena is classified by the Library Data Incubator Group under 'archives and heterogeneous data´ since it manages metadata from diverse types of materials and formats (Ríos-Hilario et al. 2012). Some tasks performed by Europeana to promote LOD in aggregation scenarios are the experimentation with sematic search based on RDF within collections, knowledge organization systems alignment, and data aggregation system with LOD export.

In this context, we present the TIB|AV-Portal as future use case of the LOD approach for non-textual materials. Some applications, such as multilingualism, are already implemented in the portal. Further implementation is focused on the weaving the data into LOD to enable metadata enrichment in scientific audiovisual materials.

3 The TIB|AV-Portal: Current practices and perspectives with Linked Open Data

3.1 Multilingualism

Textual metadata from text and speech recognition are linked to entities of the *Gemeinsame Normdatei* (GND, German for: Integrated Authority File), which is mainly managed by the German National Library. The data of the GND are available with the formats MARC 21

[2] Klump et al. (2006) Data Publication in the Open Access Initiative. Data Science Journal, Volume 5, 15 June, 2006, page: 80.

Authority, MARC21-XML and RDF-XML under Creative Commons Zero[3] (CC0). This publishing as Linked Open Data enables the linking between entities in the portal.

From the complete dataset of the GND, the TIB|AV-Portal uses the terms related to the six TIB core subjects. The GND includes synonyms, homonyms, hierarchical relations between terms and cross-references between related terms. These properties are used in the portal to perform a semantic search.

In other to achieve a multilingual and semantic portal, further terms in English were needed, since the GND possessed few English labels. A mapping into other controlled vocabularies was required. For this purpose, labels from the DBpedia[4], Library of Congress Subject Headings[5] (LCSH), results of the project 'Multi Lingual Access to Subjects[6] (MACS) and the WTI thesaurus 'Technology and Management'[7] were selected. These terms were also available as LOD and are now saved in the local RDF store (Strobel, 2014).

The combination of vocabularies and bilingual labels enables a cross-lingual search in the portal, as shown in figure 2.

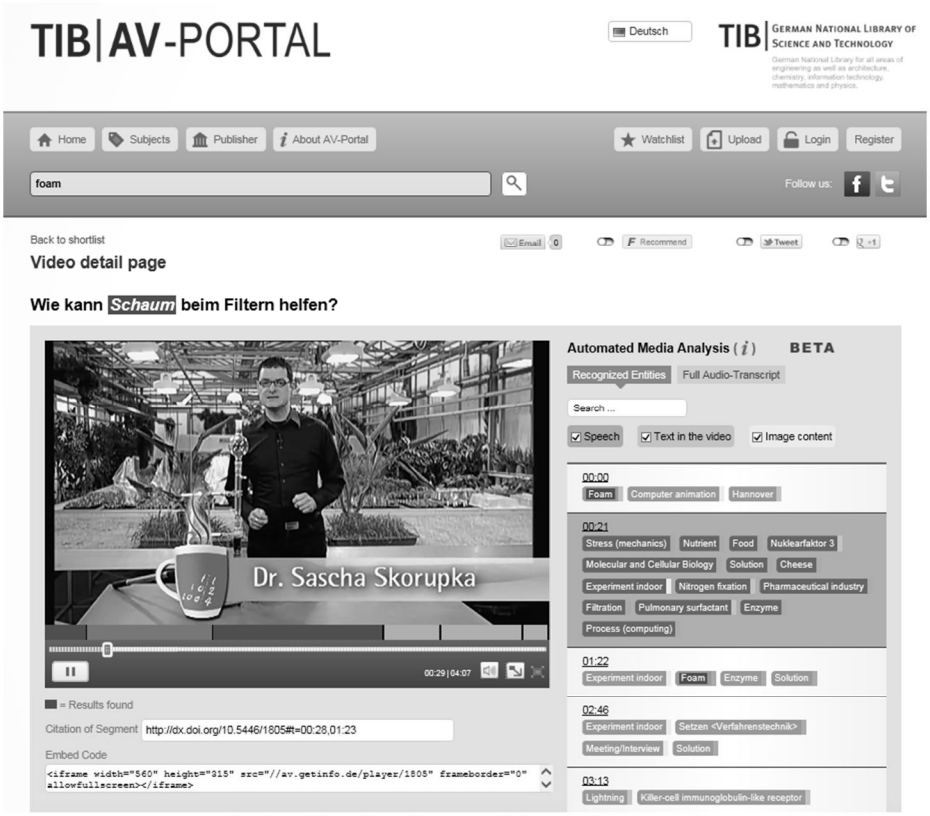

Fig. 2. Video detail page of the TIB|AV-Portal.
Results of the cross-lingual retrieval (foam-Schaum) are shown.

This is likely to increase the completeness of relevant search results and improve video indexing based on LOD properties. The semantic search based on linked data is an interesting research topic in continuous development (Waitelonis and Sack, 2011; 2014).

3.2 Data publishing and semantic applications

A fundamental question is how to publish metadata as LOD. According to Zuiderwijk et al. (2012) there is no standardised way to do it and one finds different approaches. Heath and Bizer (2011) also support this idea and describe possible Linked Data publication patterns. The type of input data (queryable structured data, static structure data or text documents) determines the appropriate publishing pattern. Static structured data coming from XML files must be converted into RDF files directly into an RDF store (Health and Bizer, 2011). The data from a RDF store (also known as triple store) can be directly published as Linked Data if they fulfill the Linked Data principles.

[3] https://creativecommons.org/publicdomain/zero/1.0
[4] http://de.dbpedia.org
[5] http://id.loc.gov/authorities/subjects.html
[6] http://www.dnb.de/DE/Wir/Kooperation/MACS/macs_node.html
[7] http://www.wti-frankfurt.de/index.php/produkte-thesaurus

Due to the structured nature of the data of the TIB|AV-Portal and the existence of an RDF store, the weaving into Linked Data requires the fulfillment of the principles. To do so, parts of the internal ontology and URI scheme need to become dereferenceable. Therefore, we are mapping and merging our internal ontology with existing LOD vocabularies[8]. Following best practices for library data, Dublin Core Terms (dcterms), Schema.org, Bibliographic Framework Initiative (Bibframe) and The Bibliographic Ontology (BIBO) were identified as best vocabularies for our purposes. Dealing with videos and their automatic annotation made also important the use of the Ontology for Media Resources and the Open Annotation Model as part of the mapping process.

The datasets will be exposed as dumps, as already done by other libraries such as The European Library[9]. Dumps are an efficient and fast way to share data with consumers. However, it requires a content and synchronization planning. As data provider, one have to decide which data are part of the dump, it format and actualization frequency. Following other libraries actions, we decide to provide 3 dumps (one with the IWF[10] collection, one with the TIB core subjects collection and one with the whole collection). The chosen formats are RDF application/rdf+xml and RDF in text/turtle, and we will offer a quarterly actualization frequency.

Further, the metadata can be embedded into the portal in form of RDFa. The last specification of RDFa 1.1[11] allows the use of RDF in HTML. This provides more control on data access and may be built on top of an RDF store. RDFa defines attributes for the semantic markup such as *about*, *property* or *vocab* to identify the subject, relations between subject and object, and the vocabularies that are used (Strobel and Marín-Arraiza, 2015). We consider RDFa an appropriate form to provide more relevant information –for instance, about an author– and to retrieve information available elsewhere in our portal or on the web –for instance, articles written by the same author and retrieve through the property *schema:author*–. This should establish a net between our content and related content of other providers, supporting reuse and content promotion.

4 Conclusions

Data organization is a requirement that libraries have to cope with. Nowadays, library data results not just after having managed textual information but also non-textual information.

The TIB|AV-Portal manages scientific videos as part of the non-textual information environment and generates a big amount of data associated with each hosted video. The data is indexed according to international standards and W3C Best Practices. However, it is still possible to go a step further and make the data available to third parties.

In other to execute this action, LOD publishing patterns are being followed. This guarantees that our data are no longer isolated or left behind the information world, which, unfortunately, happens often with non-textual information in science. Moreover, the data can be reused by others, and being part of Linked Data Cloud[12] improves the visibly of the institution and promotes its content.

The TIB|AV-Portal benefits from the use of LOD. LOD enables the multilingual annotation of the content, the semantic search, and will enable the retrieval of new related content thanks to the interlinking.

Libraries such as the German National Library, The European Library or Europeana are currently working with LOD in aggregation scenarios and providing the data to third parties. Big broadcasting companies such as the BBC also use Linked Data to annotate their content. With the TIB|AV-Portal, the TIB aims to do the same and weave into a linked ecosystem.

There is still a long way to go, particularly in the use of LOD for non-textual materials. However, there are already successful use cases[13] generated after the Library Data Incubator.

[8] http://lov.okfn.org/

[9] http://www.theeuropeanlibrary.org/

[10] http://www.filmarchives-online.eu/partners/iwf-knowledge-and-media-1/view?set_language=en

[11] www.w3.org/TR/html-rdfa/

[12] http://lod-cloud.net/

[13] http://www.w3.org/2005/Incubator/lld/wiki/UseCases

References

1. Berners-Lee, Tim: *Linked Data-Design Issues*, 2006.
2. Heath, Tom, Bizer, Christian: *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.
3. Löwgren, Jonas: "The ground was shaking as the vehicle walked pasted me." The need for video in scientific communication. Interactions, vol. XVIII n.1, pp. 22–25, 2011. DOI: 10.1145/1897239.1897246
4. Peset, Fernanda; Ferrer-Sapena, Antonia; Subirats-Coll, Imma: Open data y Linked open data: su impacto en el área de bibliotecas y documentación. El profesional de la información, 2011, vol. 20, n. 2, pp. 164-172.DOI: 10.3145/epi.
5. Pohl, Adrian, Danowski, Patrick: Linked Open Data in der Bibliotherkswelt: Grundlagen und Überblick, 2013.
6. Ríos-Hilario, Ana; Martín-Campo, Diego; Ferreras-Fernández, Tránsito: Linked data y linked open data: su implantación en una biblioteca digital. El caso de *Europeana*. El profesional de la información, 2012, vol. 21, n. 3, pp. 292-297.
7. Sack, Harald, Waitelonis, Jörg: **Linked Data als Grundlage der semantischen Videosuche mit Yovisto**, in T. Pellegrini, H. Sack, and S. Auer (Hrsg.): Linked Enterprise Data – Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien, Berlin: Springer Berlin, 2014, pp. 263-287, ISBN: 978-3-642-30273-2.
8. Saorín, Tomás. Cómo linked open data impactará en las bibliotecas a través de la innovación abierta. Anuario ThinkEPI, 2012, vol. 6, pp. 288-292.
9. Strobel, Sven, Marín-Arraiza, Paloma: Metadata for Scientific Audiovisual Media: Current Practices and Perspectives of the TIB|AV-Portal. E. Garoufallou et al. (Eds.): MTSR 2015, CCIS 544, pp. 159–170, 2015. DOI: 10.1007/978-3-319-24129-6_14
10. Strobel, Sven: Englischsprachige Erweiterung des TIB|AV-Portals. Ein GND/ DBpedia-Mapping zur Gewinnung eines englischen Begriffssystems. In: obib. Das offene Bibliotheksjournal, vol. 1, pp. 197-204, 2014. DOI: 10.5282/o-bib/2014H1S197-204.
11. Waitelonis, Jörg, Sack, Harald: Towards Exploratory Video Search Using Linked Data. In: Multimedia Tools and Applications 53, pp. 1-28, 2011. DOI: 10.1007/s11042-011-0733-1.
12. Zuiderwijk, Anneke, Jeffry, Keith, Janssen, Marijn: The Potential of Metadata for Linked Open Data and its Value for Users and Publishers. In: JeDEM, 2012, vol. 4, n. 2, pp. 222-244.