LED: curated and crowdsourced Linked Data on Music Listening Experiences

Alessandro Adamou[1], Mathieu d'Aquin[1], Helen Barlow[1] and Simon Brown[2]

[1] The Open University, United Kingdom
{alessandro.adamou, mathieu.daquin, helen.barlow}@open.ac.uk
[2] Royal College of Music, United Kingdom
simon.brown@rcm.ac.uk

Abstract. We present the Listening Experience Database (LED), a structured knowledge base of accounts of listening to music in documented sources. LED aggregates scholarly and crowdsourced contributions and is heavily focused on data reuse. To that end, both the storage system and the governance model are natively implemented as Linked Data. Reuse of data from datasets such as the BNB and DBpedia is integrated with the data lifecycle since the entry phase, and several content management functionalities are implemented using semantic technologies. Imported data are enhanced through curation and specialisation with degrees of granularity not provided by the original datasets.

Keywords: Linked Data, Crowdsourcing, Digital Humanities, Data Workflow

1 Introduction

Most research on listening to music focuses on investigating associated cognitive processes or analysing its reception by critics or commercial indicators such as sales. There is only sporadic research on the cultural and aesthetic position of music among individuals and societies over the course of history. One obstacle to this kind of research is the sparsity of primary source evidence of listening to music. Should such evidence be compiled, we argue that the adoption of explicit structured semantics would help highlight the interactions of listeners with a range of musical repertoires, as well as the settings where music is performed.

With the **Listening Experience Database (LED)**[1], we aim at covering this ground. LED is the product of a Digital Humanities project focused on gathering documented evidence of listening to music across history and musical genres. It accepts contributions from research groups in humanities as well as the crowdsourcing community, however, the data management workflow is supervised to guarantee that minimum scholarly conventions are met. Being conceived with data reuse in mind, LED is natively implemented as Linked Data. All the operations in the data governance model manipulate triples within, and across, named RDF graphs that encode provenance schemes for users of the system.

[1] LED, online at `http://www.open.ac.uk/Arts/LED`

Several content management functionalities available in LED, such as content authoring, review, reconciliation and faceted search, incorporate Linked Data reuse. Reused datasets include DBpedia[2] and the British National Bibliography (BNB)[3], with music-specific datasets currently under investigation. Reused data are also enhanced, as the LED datamodel is fine-grained and allows for describing portions of documents and excerpts, which are not modelled in the datasets at hand. LED therefore also aims at being a node by its own right in the Linked Data Cloud, providing unique content and contributing to existing data too. At the time of writing, the LED dataset stores about 1,000 listening experience records contributed by 25 users, half of whom being volunteers from the crowd.

2 Related work

A similar effort in aggregating structured data in primary evidence was already carried out for reading experiences [1], though the process was not data-driven and the resulting Linked Data were only marginally aligned. We also acknowledge a project being carried out, which gathers direct personal experiences of young users of the system, albeit with a minimal data structure[4]. We also drew inspiration from earlier accounts of using DBpedia for music, such as the *dbrec* recommender [3]. Crowdsourcing is also gaining the attention of the Semantic Web community, with very recent attempts at tackling data quality aspects [4].

3 The Listening Experience Database

We define a **listening experience** (LE) as a documented (i.e. with a quotable and citable source) engagement of an individual in an event where some piece of music is played. In terms of conceptual modelling, a LE is a subjective *event*, and one *document* describing it is the quoted evidence reported in the database.

The lifecycle of data in LED involves the roles of *contributor*, *consumer* and *gatekeeper*, and states called *draft*, *submitted*, *public* and *blacklisted*. Every artifact stored in the system exists in one or more of these states (except blacklisted ones, which exclude all other states), and a state determines if a user with a certain role can "see" an artifact or not. What these artifacts are, depends on the specific phases in the workflow, which are transitions between these states.

Authoring. Contributors populate the knowledge base by entering data on a LE and its associated entities. The entry forms are dynamic and provide suggestions and autocompletion data from LED and external datasets in real time (cf. Figure 1). Artifacts declared during this phase remain in a draft state, only to enter a submitted state once the contributor submits the LE to gatekeepers.

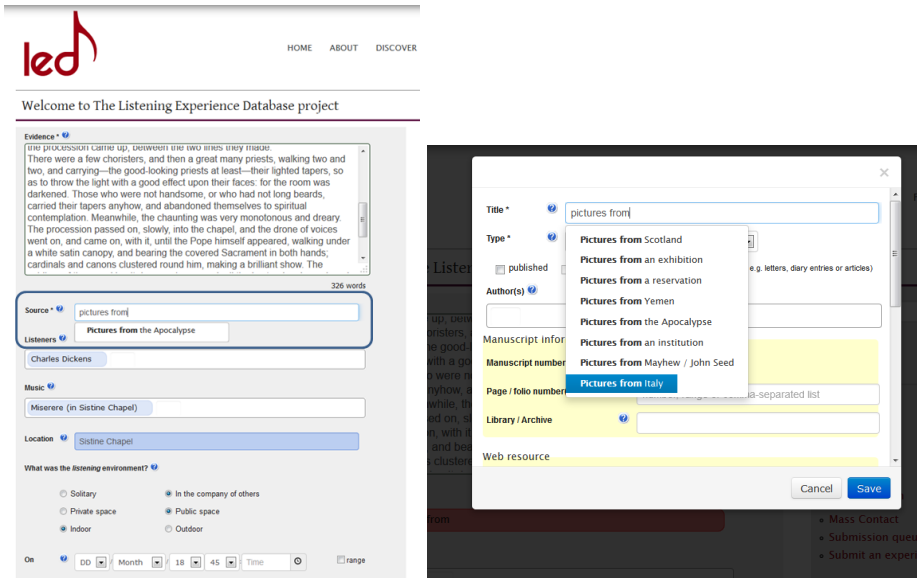
Review. Privileged users with the gatekeeper role review a submitted artifact and either promote it to the public state, or reject it for blacklisting, or demote it

[2] DBpedia, `http://dbpedia.org`
[3] British National Bibliography, `http://bnb.data.bl.uk`
[4] Experiencing Music, `http://experiencingmusic.com`

to draft again, which they can do by either taking over the artifact and amending its data themselves, or sending it back to the original contributor.

Reconciliation. Gatekeepers can align and merge duplicate artifacts that are found to match. They can compare candidate duplicates with other artifacts in LED and third-party data. This operation does not modify their state.

Faceted search. Consumers can navigate LE's by filtering keyword search results by bespoke criteria which are not necessarily stored in LED, but also reused from third-party datasets. Only public artifacts contibute to searches.



(a) Listening experience submission. (b) Autocompletion from the BNB dataset.

Fig. 1: Example of data entry for "Pictures from Italy" by Charles Dickens.

With a native Linked Data implementation, we can immediately integrate reuse with every stage of the data lifecycle starting with data entry, and eliminate *a posteriori* revision and extraction phases from the workflow, thereby reducing the time-to-publish of our data and having them linked right from the beginning. Also, the named graph model of quad-stores can encode provenance information with the granularity of atomic statements [2], thus lending itself to fine-grained and complex trust management models.

To encode the above workflow entirely in RDF, we used the named graph paradigm in order to represent states and artifacts. Deciding on the scale of the latter was an issue: while we intended to give gatekeepers control on single RDF triples (or quads, from the named graph perspective), and to contributors a way to support the truth or falsehood of a triple, this can be complex and time-consuming. Therefore, artifacts are encapsulated into LE's, musical works,

literary sources, agents (e.g. people, groups or organisations) and places: these are, for instance, the classes of artifacts that gatekeepers may want to review or reconcile. However, LE's remain the core artifacts of the sytem: only by creating or editing them can their associated artifacts be generated.

The LED knowledge base is partitioned into *data spaces*, each belonging to a user or role. Every contributor owns two RDF graphs, one for draft artifacts and one for submitted ones. Thus, we can keep track of which contributors support a fact by reusing it (e.g. `<Messiah_(oratorio) composer Georg_Frideric_Handel>`). There is a single graph for public artifacts, and one for blacklisted ones. Contributors have access to the graphs they own plus the public graph; gatekeepers can access every user's submitted graph and the public and blacklist graphs. State transitions are realised by parametric SPARQL queries that selectively move RDF triples across these graphs. Along with these data spaces there are rules that determine the visibility of triples to each user, depending on the content of their private graphs. In general, these rules assume contributors have greater confidence in the facts in their possession, and when missing, they should trust those provided by the community or other datasets.

4 Demonstration

The audience will be given a live demonstration of the LED system, but from the point of view of users with the privileged roles of *contributor* and *gatekeeper*. We will show the benefits of reusing data from indexed datasets during the entry phase, as well as the implementation of our governance model in Linked Data and its effects on the representation of a resource as seen by the general public or a specific user. Data reuse and enhancement will be demonstrated through a LE entry form to be auto-populated in real time and open to input by audience members. To demonstrate the governance model, we will run two distinct entries with shared data through the whole draft-submission-gatekeeping lifecycle. We will then show how differently the shared data and their RDF representations appear to each user, based on the trust and provenance policies in place.

References

1. Matthew Bradley. The Reading Experience Database. *Journal of Victorian Culture*, 15(1):151–153, 2010.
2. Jeremy J. Carroll, Christian Bizer, Patrick J. Hayes, and Patrick Stickler. Named graphs, provenance and trust. In Allan Ellis and Tatsuya Hagino, editors, *WWW*, pages 613–622. ACM, 2005.
3. Alexandre Passant. dbrec - music recommendations using DBpedia. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *International Semantic Web Conference (2)*, volume 6497 of *Lecture Notes in Computer Science*, pages 209–224. Springer, 2010.
4. Elena Simperl, Maribel Acosta, and Barry Norton. A semantically enabled architecture for crowdsourced linked data management. In Ricardo A. Baeza-Yates, Stefano Ceri, Piero Fraternali, and Fausto Giunchiglia, editors, *CrowdSearch*, volume 842 of *CEUR Workshop Proceedings*, pages 9–14. CEUR-WS.org, 2012.