Undefined 0 (2014) 1–8
IOS Press

Crowd-sourced Digital Humanities linked data contributing to library datasets: the case of the Listening Experience Database

Alessandro Adamou [a] and Mathieu d'Aquin [a]
[a] *Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom*
E-mail: {alessandro.adamou,mathieu.daquin}@open.ac.uk

Abstract. In this paper, we present a linked dataset for 'early access' to information crowd-sourced as part of the Listening Experience Database project. We call it early access, consistently with the practice in modern software development, as its main aim at this stage is to collect feedback and initial use cases that can support the evolution of the dataset. The Listening Experience Database is a Digital Humanities project aimed at gathering structured and documented evidence of how music is perceived throughout history. The content is largely represented in terms of widespread ontologies for the domains of music and literature. Reuse from external datasets such as DBpedia and the British National Bibliography is guaranteed by the data entry workflow, and reused entities are re-published with data that improve upon the original datasets, for instance by modelling portions of published written works. The dataset is updated daily by both a community of enthusiasts and a team of experts, the latter also being in charge of approving and curating data.

Keywords: Linked Data, dataset, Digital Humanities, data sparseness

1. Introduction

Listening Experience Database (short, **LED**) is the name of both a Digital Humanities project aimed at gathering documented evidence of personal experiences of listening to music throughout history, and the structured database where such evidence is stored. This database is originally available through a Web portal[1] that provides exploration capabilities and the tools for contributing new data [2].

LED as a project aims at providing curated data with special regard to the documentation that is source to the recorded evidence, to which end it implements a governance policy of supervised crowd-sourcing. Any volunteer can contribute new data or additions to existing data, but their contributions only go public upon approval by a restricted team of moderators, to ensure essential scholarly conventions are respected. To aid this data curation phase, it is important to incorporate reuse of data from digital libraries and other external providers in the lifecycle of LED content. Therefore, the LED portal was implemented as a semantic content management system, and contributed data are stored natively in RDF. The lifecycle of the content, from data entry to validation and publication, is handled through different RDF graphs representing overlapping datasets (datasets of contributions from specific users, the 'validation queue', the published data, etc.). In this paper, we focus on the dataset of approved contributions, which is made available on the LED portal, and exposed as a Linked Open Data set through the Linked Data platform of the Open University[2].

As a result of including reuse in the data lifecycle starting from the entry phase, linkage with exter-

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

[2] OU Linked Data, `http://data.open.ac.uk`

0000-0000/14/$00.00 © 2014 – IOS Press and the authors. All rights reserved

nal data is provided right from the beginning. Datasets that we link against include the *British National Bibliography*[3] and *DBpedia*[4]. The ontology by which LED data are represented integrates widespread vocabularies from the bibliographical and musical domains, such as Bibo[5] and the Music Ontology[6].

The LED linked data are provided as an 'early-access' release, in that the available data are a daily snapshot of the content of the original database, which completely supersedes the one from the day before. Versioned releases will be published every time changes are made to the ontology and propagated to the data modelled thereafter.

The first public unveiling of the LED linked open dataset was on September 15, 2014, nine months into the crowd-sourcing campaign of the project. While the current instalment of the project runs until the end of 2015, hosting and administration of the crowd-sourcing portal and the *data.open.ac.uk* platform are scheduled to run and be maintained indefinitely, and so is the automatic update process of the linked dataset.

2. Synopsis

LED linked data are one of the five-star[7] datasets hosted at *data.open.ac.uk*, the Linked Data platform of the Open University. Although most datasets served by this platform are inherent to the assets of the OU in the educational domain (e.g. staff, courses and study material), some datasets contain the body of resources gathered by specific research projects. This can have repercussions on the way data are modelled or named (see e.g. Section 5.2 for naming conventions in LED).

The LED dataset comes as a single named graph that is entirely standalone, in that SPARQL queries restricted to this graph (e.g. using a FROM clause) will return the same data as those that can be browsed on the website[8]. The resources made available with the dataset, and their physical locations (and logical names where applicable), are summarised in Table 1.

The LED ontology is available in machine-readable format in several RDF-based serialisations of OWL (RDF/XML, RDF/JSON, Turtle and N-Triples) and includes `owl:imports` statements on the vocabularies it depends upon. Every term (class, property or individual) defined by this ontology is also available in machine-readable format: note that dereferencing it does not deliver the same RDF as the ontology, rather, the RDF signature of the term itself and of all the terms of the ontology that form incoming links to it.

Currently only one RDF dump is publicly available at one time, that is, a nightly snapshot of the state of the database as contributed by its users and moderated by the project team. No specific version is indicated, so that applications intending to stay up to date can maintain a single, immutable reference. As the datamodel evolves, versioned releases of the dump and the ontology (using the OWL 2 versioning convention) will be scheduled, however, only the nightly snapshot will be mirrored for SPARQL querying and dereferencing.

3. Rationale

The open dataset of LED stems from a data lifecycle that runs continuously. It encompasses curated entries by domain experts (such as professional musicians, musicologists and scholars in English literature) and crowd-sourced entries by connoisseurs and enthusiasts. This means that the dataset grows and evolves every day, but also that the degree of detail in the submissions varies greatly, depending on what information is available to support the evidence and the knowledgeability of contributors on the subject matter. The minimal requirement for a submission to the LED portal is to provide evidence as a quote in natural language text, and a source document indicated by title. However, contributors are given the ability to include much richer information to model the listening experience they are describing, including for example the people involved; the setting of the musical performance (possibly even differing from the listening environment, as with radio broadcasts of live music); and peripheral information of the literary source, such as the collection it was taken from, what specific element of the collection it is (e.g. a letter or diary entry), its original language and who translated it into English.

As part of accommodating this sparseness in the level of detail, there was a need for supporting data multitenancy, i.e. the ability of users to benefit from information entered by other users, and integrate it with their own if possible. For example, a user may provide biographical information on a historical character, like his or her occupation or religion, that another

[3] BNB linked data, `http://bnb.data.bl.uk/`

[4] DBpedia, `http://dbpedia.org`

[5] Bibo, `http://bibliontology.com/`

[6] Music Ontology, `http://musicontology.com/`

[7] as per the five-star open data deployment scheme at `http://www.w3.org/DesignIssues/LinkedData.html`.

[8] not counting linked external entities such as those from DBpedia.

Name	Listening Experience Database	
Graph name	`http://data.open.ac.uk/context/led`	Also VoID description URL
SPARQL endpoint	`http://data.open.ac.uk/sparql`	Graph name can restrict queries
Nightly RDF dump	`http://led.kmi.open.ac.uk/rdf/export/led-SNAPSHOT.nt.bz2`	Compressed N-Triples
Root Ontology	`http://led.kmi.open.ac.uk/ontology`	Includes dependency closure
Documentation	`http://led.kmi.open.ac.uk/linkeddata/`	
# Triples	54,709	As of September 25, 2014
License	CC BY 3.0, `http://creativecommons.org/licenses/by/3.0/`	Inherited from *data.open.ac.uk*

Table 1
Resources of the LED dataset

user only knew by name. Likewise, users should be able to support, or challenge, information provided by other users. Extending this paradigm to Web data, LED users should be able to benefit from existing structured data from open datasets external to LED, especially in the bibliographical and musical domains, but also contribute information that is not covered by said datasets.

The expected long-term effect of this design is that, as the dataset grows, overlaps in the data of submitted entries begin to arise. This should, in turn, prompt the community - comprised of contributors and moderators alike - to resolve conflicts and redundancies deriving from these overlaps, thus "healing" the dataset semantically. Moderators of the LED platform have reported coming across the first occurrences of this phenomenon, where multiple users provided distinct listening experiences and data on two musical performances that turned out to be the same. As a facility of the LED portal, moderators are given data reconciliation tools that merge and/or align matching entities that have been created separately, possibly because their contributors did not have enough information in their possession to relate them at the time.

All the above points – data spareness with support for anytime incremental integration, reuse and reconciliation – fall within the constituent principles of Linked Data. Plus, the British Library provides extensive coverage of structured data of published literature in the United Kingdom[9], which is part of the main focus of the LED project. This has provided motivation for implementing LED natively as Linked Data. Basically, the underlying data store of listening experiences, both under review and approved, is an RDF store. The data entry form comes with autocompletion features that suggest RDF entites from external Linked Data and LED itself as the user types, and automatically fill in information that can still be amended by the user [10]. We refer to our previous publications for details of the data entry and curation mechanisms implemented in the LED portal[1,2]. Implementing the database in RDF and integrating reuse from the early phases of data lifecycle provides a five-star dataset that does not require a post-processing phase prior to publishing it. For the sake of long-term sustainability, the publication step is carried out on the consolidated linked data platform of the Open University.

4. Data

Data in the LED dataset are represented in accordance to an OWL2-DL ontology identified as, and located at, `http://led.kmi.open.ac.uk/ontology`. As this is the only public release of the ontology since the launch of the LED dataset at the time of writing, the only public version IRI is the same as the ontology IRI [5]. The LED ontology is largely built upon existing standards for representing knowledge on bibliographical, musical and personal entities. A summary of the vocabularies reused in LED and the scope of their usage is given in Table 2[11].

On top of these vocabularies is a model for the listening experiences themselves. Such notion has found little to no coverage in existing linked data, presumably due to their scarce availability. Further models are also provided to satisfy special database design requirements, which eventually prompted us to modularise the LED ontology. These ad-hoc modules are:

- a controlled vocabulary of document types for listening experience sources;

[9] The British Library, `http://www.bl.uk`

[10] For performance reasons, these queries are not performed live on external SPARQL endpoints, but on local search indices that rely on the stability of external URIs, also known as their "coolness" [6].

[11] Usage of Schema.org and the DBpedia ontology is largely provisional and scheduled for removal in future versions, in favour of terms from other dependencies, or custom terms that subsume them.

Name	Namespace	Scope
Bibliographic ontology (Bibo)	`http://purl.org/ontology/bibo/`	Sources of listening experiences
DBpedia ontology (for release 3.9)	`http://dbpedia.org/ontology/`	Specific features of participants in listening experiences (e.g. occupation, religion)
Dublin Core metadata terms	`http://purl.org/dc/terms/`	Employed as a dependency of Bibo
Event ontology	`http://purl.org/NET/c4dm/event.owl#`	Base model of Listening Experience items
Friend-Of-A-Friend (FOAF)	`http://xmlns.com/foaf/0.1/`	Participants of listening experiences
The Music Ontology	`http://purl.org/ontology/mo/`	Subjects of listening experiences: musical works (e.g. songs) and performances thereof
OWL-Time	`http://www.w3.org/2006/time#`	Base model for underspecified temporal indicators (cf. Section 4.3)
Schema.org	`http://schema.org/`	Basic creative work authorship; personal information not covered by FOAF

Table 2
External vocabularies reused in the LED ontology.

– a taxonomy of occupations and social status (cf. Section 4.2);
– an extension of the EDTF specification for fuzzy date/time representation (cf. Section 4.3);
– a metadata vocabulary.

The data published by the LED dataset can be domain-wise classified as follows:

Bibliographical. Published and unpublished literature in English, if some evidence of listening experience was found in it, is represented in LED. Instances include published books, collections, scholarly publications, newspapers, diaries or chronicles, but also portions of them (e.g. a diary entry, a letter in a compiled collection, or a newspaper article). Data are largely modelled using a combination of Bibo and Dublin Core vocabularies, plus elements of Schema.org. Reused external entities include instances of `dc:BibliographicResource` in the **British National Bibliography (BNB)** and of `dbpedia-owl:WrittenWork` in **DBpedia**, within the range that we shall discuss in Section 4.1.

Musical. The subjects of listening experiences are musical works or performances thereof (e.g. live or on the radio). The choice of either class depends on whether the contributor can only identify the music being heard by its title or mnenomic identifier, or can also provide additional information concerning how, where and by whom it was performed in a specific occurrence. These data are represented using the Music Ontology plus additional (provisional) terms from the DBpedia ontology. Musical genres are a combination of user-defined entities and objects from `dbpedia-owl:genre` assertions in DBpedia. Instruments include the MusicBrainz instrument taxonomy in SKOS[12], but also everything that is an object of `dbpedia-owl:instrument` assertions in DBpedia. The latter, in the present DBpedia revison, include broad categories (e.g. 'chordophones'), classes (e.g. 'electric guitar') and even their realisations (e.g. 'Gibson Les Paul'). We are currently allowing for all of these, though, depending on the usage trend of database contributors, we are investigating the possibility of restricting to the MusicBrainz taxonomy only. Reused entities from the musical domain are currently imported from DBpedia, with the inclusion of datasets from the musical domain currently under investigation.

Biographical. Any agent, namely person or group, that has a role in a listening experience, such as listener, performer, composer of the original music or editor/compiler of the evidence, is represented using a combination of FOAF, Schema.org and the DBpedia ontology. For people, we include basic biographical data, such as places and dates of birth or death, gender and alternate names, but also some basic information to frame that person in a social and economic profile (cf. Section 4.2). Reused entities can belong to DBpedia, the BNB, or the **Virtual International Authority File (VIAF)**[13], which bridges the two other datasets.

Contextual. On top of the above, there are the listening experiences themselves. These are custom-modelled as a subclass of personal events and as such inherit from the event model of the Event Ontology,

[12] MusicBrainz instrument tree, `http://purl.oclc.org/net/MusicInstruments`
[13] Virtual International Authority File, `http://viaf.org`

with the addition of specific context properties introduced to satisfy design requirements, such as indicators of whether these personal events occurred indoors or outdoors, live or in playback, or in a different environment than the one where the music is performed. It should be noted that, as the data workflow of the LED project is ListeningExperience-centered, so is its dataset. In other words, users contribute to the database only by submitting new listening experiences, and in doing so they provide peripheral information on the other entities involved, eventually integrating the data contributed earlier. Likewise, for every RDF resource in the dataset there is a path from an instance of `<http://led.kmi.open.ac.uk/term/ListeningExperience>`, and the entire graph can be walked starting with that class.

4.1. Contributions to existing datasets

Reuse of data from external providers occurs in LED since the data entry phase, by means of forms that support autocompletion of the text fields where the user types, and automatic filling of the corresponding form based on what data are present on the external dataset about the selected entity. The user is still able to amend auto-filled forms, though the governance policy of the dataset ensures that these proposed amendments will propagate to the open dataset only upon approval by moderators in the project consortium. This way, the LED dataset becomes an additional node in the Linked Data cloud that provides complementary knowledge about entities managed by other nodes in the cloud.

This is especially true for British National Bibliography data from the British Library, which are extensively imported in LED and are subject to design choices that limit the scope of the description of a published work. Although the implicit datamodel that emerges from the entry forms in LED is mostly aligned with the typical signatures of BNB entities, there is some unique knowledge that LED provides for them, that lies outside of the scope of BNB. These include:

1. Co-authorship and editing: the BNB typically lists a single author for a published work, with co-authors listed as `dc:contributor` alongside editors and compilers. LED distinguishes authors from editors using the corresponding Bibo properties.
2. Translators from the original languages: these are not present at all in BNB data.
3. Relationships between manuscripts and their published counterparts, as manuscripts are out of the scope of BNB.
4. Volume/issue numbers where applicable.
5. Portions of bibliographic resources, such as a letter in a collection of correspondence, a diary entry or a newspaper article; BNB does not model parthood in published works to this degree.

The model of biographical data in LED also integrates information regarding the social and cultural status of participants, that is only sparsely found in DBpedia and not at all in BNB, including their occupations and religion. Also note that, for data on musical works, whether generated internally or imported from DBpedia, LED becomes a unique provider of Linked Data entities for their performances. With the bulk of data on modern music events expected to grow in LED, this opens up for several data integration opportunities, such as events on *Last.fm*[14].

Another issue is that of ontological alignment between entities across datasets, such as strict equality using `owl:sameAs`. In general, there is no direct alignment between entities in DBpedia and the BNB, nor does the *sameAs* interlinking service[15] provide a path. On occasion, a bridging aligmnent can be found by traversing the corresponding VIAF authority file entry that can be reached from both datasets. However, LED employs a reconciliation tool that site moderators can use to declare that multiple entities (e.g. one generated internally and others imported from BNB and DBpedia) match. This results in `owl:sameAs` links from the designated primary entity in the LED dataset.

4.2. Modelling socio-economic profiles

We have not found evidence of consolidated formalisms to represent the social and economic profile of individuals in Linked Data. As it was a requirement that this aspect be modelled in LED data on the basis of factors such as their occupational status, we are currently experimenting with representational models in that respect. The current LED dataset combines objects of `dbpedia-owl:occupation` assertions in DBpedia with occupational classes in the ISCO-08 taxonomy promoted by the International Labour Organization [3]. The latter are being represented in RDF using URNs of the form `urn:x-isco08:{numeric-id}`.

[14] Last.fm in RDF, `http://lastfm.rdfize.com`
[15] `http://sameas.org`

4.3. Representing underspecified time expressions

One issue related to accommodating data sparseness in the population of the LED dataset concerns the representation of time. We have come across frequent instances of listening experiences to be recorded, whose submitters had only partial information in their possession regarding the timeframe where an event, such as the musical performance itself or the birth and death of a participant, occurred. Some synopses of such partial temporal values include *"mid September, 1516"*, *"19 April at 9am"* (as could be found in a diary entry or letter), *"October in the early 18th Century"*, or *"sometime between the 1790's and 1799"*. While it was a requirement that these indications be captured despite their incompleteness, there is no set standard for machine-readable representations of them, let alone for reifying them in RDF. Also, a significant challenge lies in representing them in a way that allows us to relate them to fully qualified date/datetime values where applicable.

With the exception of metadata, LED does not store date and time information as literals. When a temporal entity is fully qualified, either as a date or as a timestamp, the British calendar Linked Data API from *data.gov.uk* is used: for example, the date November 26, 1877 is identified by `<http://reference.data.gov.uk/doc/day/1877-11-26>`[16]. For temporal entities that are not fully qualified, as in the above examples, we are adapting the EDTF specification in RDF. **EDTF (Extended Date-Time Format)** is a proposed standard of the Library of Congress for complex date/time strings that accommodates decades, ranges and unspecified (or uncertain) date elements [4]. Although it has not progressed beyond draft status, it served as the basis for the generation of fully described RDF resources for fuzzy date/time expressions in LED. By request, this specification was integrated with vagueness indicators of the type 'early/mid/late' to denote periods that are willingly underspecified and left to interpretation in the context at hand. So therefore, a natural language expression such as "mid September, 1516" becomes `<http://data.open.ac.uk/time/edtf/1516-09-Mm>`, and "(sometime) between the 1790's and 1799" becomes `<http://data.open.ac.uk/time/interval/179u-uu-uu/to/1799-uu-uu>`, with its start and end indicators linked to it in its RDF description.

Currently, the RDF data of these temporal indicators are materialised and stored in the RDF dataset for only the temporal objects that are linked to some other entity. We are looking into the possibility of moving to a Linked Data API service for stable CoolURIs [6] and on-the-fly RDF generation.

5. Usage

Since its launch at the beginning of 2014, the LED project has attracted more than 1,000 validated contributions from 21 contributors, both experts and enthusiasts. As an open dataset however, the LED linked data have only been available for two weeks at the time of writing. Therefore, usage of this dataset outside the LED portal and the project consortium are still being expected. In this section, we provide an indication of the way data can be discovered, either by SPARQL querying or by exploring URIs.

5.1. Querying

The following SPARQL query retrieves all the people who have written about their own experiences[17]:

```
PREFIX led: <http://led.kmi.open.ac.uk/term/>
SELECT DISTINCT ?person
WHERE {
   [] a led:ListeningExperience
     ; led:is_reported_in/<http://schema.org/author> ?person
     ; <http://purl.org/NET/c4dm/event.owl#agent> ?person
} ORDER BY ?person
```

In the following, we obtain all the listening experiences that occurred in a place that is *not the same* as where the music was performed, either because it was aired on the radio or because the listener was within earshot of a live performance, but elsewhere:

```
PREFIX event: <http://purl.org/NET/c4dm/event.owl#>
PREFIX led: <http://led.kmi.open.ac.uk/term/>
PREFIX mo: <http://purl.org/ontology/mo/>
```

[16] A guide to the URI scheme of the British calendar RDF API can be found at `http://www.epimorphics.com/web/wiki/using-interval-set-uris-statistical-data`.

[17] This example query, as the ones that follow, could be restricted to the LED dataset by means of a clause such as `FROM <http://data.open.ac.uk/context/led>`, but since there is currently no other use of the class `led:ListeningExperience`, we can safely omit it and benefit from any linkage contributed by other datasets in *data.open.ac.uk*.

```
SELECT DISTINCT ?lexp ?medium
WHERE {
    ?lexp a led:ListeningExperience
    ; led:has_medium ?medium ; event:place ?lplace
    ; <http://purl.org/dc/terms/subject>
        [ a mo:Performance ; event:place ?pplace ]
    FILTER (?lplace != ?pplace)
}
```

The following federated query lists all the countries where a national anthem was performed at least once, according to the listening experiences in LED[18]:

```
PREFIX dbpo: <http://dbpedia.org/ontology/>
PREFIX event: <http://purl.org/NET/c4dm/event.owl#>
PREFIX mo: <http://purl.org/ontology/mo/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?country
    (COUNT(DISTINCT ?performance) as ?performances)
WHERE {
    ?performance event:place ?place ;
    dbpo:genre|mo:genre
        <http://dbpedia.org/resource/National_anthem>
    SERVICE SILENT <http://dbpedia.org/sparql> {
        { ?place dbpo:country ?country }
        UNION { ?place a dbpo:Country
            BIND(?place as ?country) }
    }
} GROUP BY ?country
```

Note that we do not store any data from the signature of a location in DBpedia or any other external dataset that provides geographical information. This is in true in general for any entity that is not curated by users of the LED system (cf. Section 6).

5.2. URI schemes

Having an insight as to how URIs are constructed is part of the interface of a dataset to client applications. The naming convention for generated URIs in LED is `http://data.open.ac.uk/led/{stype}/{uid}`

where `stype` is a short form of the name of one of the types of the entity and `uid` is a unique identifier. Note that the convention of including `led` in the namespace is adopted, as in other project-specific datasets, in slight contrast with the general convention of *data.open.ac.uk*, which follows `http://data.open.ac.uk/{stype}/{uid}`. This is due to the fact that *data.open.ac.uk* mostly hosts data on educational material related to the Open University. Since data from non-contemporary history tend to be an exception, we draw a syntactic line to semiotically separate, for example, the URI of an OU employee from that of a 17th Century musician.

Example values for `stype` are `person`, `lexp` (listening experience), `source` and `performance`. `uid` is usually constructed by concatenating the URL-encoded form of a mnemonic literal (e.g. the `foaf:name` or `dc:title` of the entity) with a long integer, as in `<http://data.open.ac.uk/led/person/Samuel+Pepys/1407246853714>`. Exceptions include listening experiences (whose UIDs are constructed with integers only) and a few classes for which the mnemonic literals are deemed sufficient to uniquely identify the resource, e.g. occupations, which are being assembled in a controlled vocabulary.

6. Maintenance

The open dataset of LED mirrors an RDF graph in the original quad store that backs the LED content management system, and that resides in a separate host environment. This graph, which we call *master* graph, is the only one in the whole quad store that is publicly visible, with the only exceptions of graphs where the LED ontology is stored. The other graphs contain data that have not yet been submitted or approved, and are therefore meant to be only accessible by the respective owners or by moderators. When a (possibly not authenticated) user browses the LED website, the content displayed is an HTML5 rendering of the data in the master graph, with embedded RDFa 1.1 annotations: the quad store and its local SPARQL endpoint are only accessible in loopback by the CMS itself.

The RDF dump mentioned in Section 2 is updated once a day with the content of the master graph, with previous dumps preserved for backup. This dump is also imported verbatim by *data.open.ac.uk*, which rebuilds the corresponding graph from the ground up. On the current scale, the process runs in less than a minute.

Hosting the LED graph on *data.open.ac.uk* allows it to benefit from the features of the platform, such as URI dereferenceability and interlinking with other hosted datasets, as well as its long-term sustainability plan. Whilst the LED platform is set to remain online indefinitely, updates to the dataset beyond 2015 will

[18] This query is slightly simplified. Since the DBpedia SPARQL service does not collapse redirects, property paths should be concatenated with redirection properties, as in `?place dbpo:wikiPageRedirects*/dbpo:country ?country`

be subject to funding a team of moderators, which is currently being actively sought. Migrating to a peer-review governance policy is also being contemplated.

As the LED dataset is released in 'early-access', the quality of data is strongly tied to the continued reviewing work of the team of moderators. Though their everyday task is to revise, amend and approve submitted entries piecemeal, other data curation tasks, such as the reconciliation of redundant entities, are performed off-schedule as issues are detected. A ticketing system is being introduced to keep track of these issues. Any data transformation that is a consequence of these tasks is propagated to the linked dataset the next day.

Another issue is the management of (outgoing) external linkage. The general policy for data of entities coming from other datasets such as BNB and DBpedia is to only store them if they have been confirmed by the contributor and moderators, provided there is a choice to edit them. For instance, if a user on the bibliographical source entry form selects a book from BNB, any relevant data about it will be imported from BNB and auto-filled in the form. The user then has a choice to amend an auto-filled value, or accept it by leaving it as it is. In the latter case, the confirmed value is stored anyway. No data are ever stored, which have not been at least visually presented to the user. That is the case of places and organisations, for instance, which by policy do not have a corresponding form in LED: in those cases, only the URI reference is stored, and whenever it needs to be presented, external data will be imported on the fly using local indices and federated queries.

Alongside maintaining a nightly snapshot of the data in early-access, stable releases of the LED dataset are being scheduled: one in November 2014 to coincide with the Listening Experience symposium[19], and one in towards the end of the first crowd-sourcing campaign in late 2015. Each release will be accompanied by a persistent version of the corresponding ontology.

7. Future work

As part of the release plans for the LED dataset, several improvements are planned for the short to mid term. The current priority is to introduce, in the form of local search indices, at least one dataset for the musical domain, with special attention to the recent third-party Linked Data export of *MusicBrainz*[20], but we are also looking into integrating the aforementioned RDF version of *Last.fm*. This would also allow us to raise the bar of supported knowledge on musical items: at present, whereas the bibliographical component is much more articulated, the entry form for musical items allows for minimal detail only, in order not to discourage non-expert contributors from submitting their data. The submission workflow will evolve to accommodate greater detail, but if we can benefit from linking against long-running musical knowledge bases, this detail can mostly be optional.

As part of the ontological evolution of the dataset, we are planning to phase out a set of custom terms provisionally developed within LED, as well as limiting its usage of the DBpedia ontology, in favour of more intensive usage of Bibo and the Music Ontology. We are also investigating on extending the breadth of ISCO-08 usage in LED (cf. Section 4.2) to encompass its entire occupational taxonomy, as well as releasing the EDTF time data from Section 4.3 as a Web Service.

Other avenues for evolving LED include mining social media for listening experiences, along with setting minimal quality criteria they should satisfy in order to be part of the dataset. Once these quality metrics are set, they will also be implemented as an automated quality control phase in the data curation workflow for experiences submitted via crowd-sourcing.

[19] Listening Experience Symposium, `http://www.rcm.ac.uk/events/listings/details/?id=431576`

[20] LinkedBrainz, `http://linkedbrainz.org`

References

[1] A. Adamou, M. d'Aquin, H. Barlow, and S. Brown. LED: curated and crowdsourced linked data on music listening experiences. In *The 13th International Semantic Web Conference (ISWC 2014) demos*. CEUR-WS, 2014 (to be published).

[2] S. Brown, A. Adamou, H. Barlow, and M. d'Aquin. Building listening experience linked data through crowd-sourcing and reuse of library data. In *1st International Digital Libraries for Musicology workshop*. ACM, September 2014.

[3] International Labour Organization. International standard classification of occupations (ISCO-08). ILO resolution, International Labour Organization, March 2008. http://www.ilo.org/public/english/bureau/stat/isco/isco08/.

[4] Library of Congress. Extended date/time format (EDTF) 1.0. Draft submission 13 January 2012, Library of Congress, 2012. http://www.loc.gov/standards/datetime/pre-submission.html.

[5] B. Motik, P. F. Patel-Schneider, and B. Parsia. OWL 2 Web Ontology Language: Structural specification and functional-style syntax (second edition). W3C recommendation, W3C, December 2012. http://www.w3.org/TR/owl2-syntax/.

[6] L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web. W3C interest group note, W3C, December 2008. http://www.w3.org/TR/cooluris/.