2017

The Mystery of the Schubert Song: The Linked Data Promise

Kimmy Szeto
Baruch College

Mystery of the Schubert Song: The Linked Data Promise

Mystery of the Schubert Song

"I think the German group really needs one more song." A music reference situation unfolded as the voice teacher discussed a recital program with her student. "I remember that Austrian soprano...was she Austrian? What's her name? Strada? Estrada? The last song on her album is a Schubert song…it's upbeat; it starts on a high G. What's that song called? It's one word…I think it ends with '*-lein*'..." The teacher thumbed through her volumes of the complete Schubert songs, then started running her finger down the index. In the meantime, the student picked up her mobile phone and pulled up "List of Songs by Franz Schubert" on Wikipedia. She moved on when she realized the songs were listed by opus and catalogue numbers. Then she pulled up "List of compositions by Franz Schubert by Genre," and scrolled to the section "Lieder with piano accompaniment," first the eleven cycles and sets, and then three dozens or so by voice type. At that point, she looked overwhelmed by the sight of the remaining list of 500 or so entries.

I stayed silent as the accompanist should, estimating a maximum of 30 seconds before they would both give up. But my librarian persona leapt into action. I pulled out my mobile phone, searched for "soprano obituary strada" (without the quotes). I realized the name was misspelled (thanks to Google's "Did

you mean…" suggestion).[1] The actual name was Stader and she was Swiss, not Austrian. "Maria Stader?" I asked. "Right, Maria," the teacher said, still scanning the index. I then searched for "maria stader schubert LP." Among the top results were two entries on Discogs, an online marketplace for music collectors. "Was it a live recording of a concert?" The teacher did not think so. So I chose the Discogs entry for the 1958 Deutsche Grammophon studio album *Liederabend.*[2] The only song title that ended with "*-lein*" was not the last song on either side of the LP. It had more than one word, and it was not even by Schubert ("Das bescheidene Wünschlein" by Othmar Schoeck). However, I spotted another track and took a leap of faith: "*Seligkeit*?"

"How did you know?"

I did not. But I knew how to look, I knew when to ask follow-up questions, and I knew when to guess. While retrieving *Seligkeit* on IMSLP, I remarked that to train as a singer nowadays was to train as a librarian. It turned out the high G-sharp was not at the start of the song, but it was the start of the last phrase of the song. Nonetheless, the song was exactly the one the teacher was looking for, and it was perfect for her student's recital program.

[1] Google's "Did you mean…" feature uses multiple probabilistic and machine learning algorithms that are based, in part, on the user's search history. So, this search is not meant to be replicable. It was with an element of chance that I hit upon a good suggestion.

[2] "Maria Stader, Franz Schubert, Felix Mendelssohn-Bartholdy, Othmar Schoeck – *Liederabend*," *Discogs*, accessed November 30, 2016, https://www.discogs.com/release/7821855.

Catalog Searching and Data Connectedness

While singers might benefit from information literacy skills, they should not need librarianship training. But in so many situations like this one, absent a reference librarian, our online services fall short.

Our current bibliographic systems can respond to a search for an LP as the material format, Schubert as the author, and solo songs with piano as the subject.[3] This search yields over 1,000 results. However, adding "strada" or "estrada" to the search yields zero result.

At this point, an experienced searcher would focus on revising the singer's name. Remarkably, adding the correct name "stader" to the search reduces the results to 5 and includes the appropriate answer. The problem is getting to the correct name. Neither the WorldCat public interface nor the Library of Congress offers a name search by nationality. One could attempt to add more keywords: adding "soprano" would yield over 400 results, but adding "Austria" or even the correct country "Switzerland" would not yield bibliographic records of any of Stader's recordings, because such a search would look for keywords in bibliographic records, and nationalities are recorded in a separate authority file.

[3] A WorldCat search with the command "mt=lps au:schubert su:songs with piano"; in the rest of the article, the commands "pn: " and "kw: " are used for WorldCat searches for names and keywords, respectively.

A persistent searcher might at this point use the fact that the recording was old and comb through all the results in chronological order. Going down this path would require examining a minimum of 72 bibliographic records before reaching the one for an album titled *A Maria Stader recital* which includes the same tracks as *Liederabend*.[4] Even then, there is no guarantee the searcher would recognize the spelling discrepancy in the singer's name to select the record for further evaluation.

Choosing another path, a savvy searcher might hone in on the name and nationality of the singer using the Virtual International Authority File,[5] but would still come up empty: a search for "soprano" yields too many results but narrowing down by "Austrian," "Strada," or "Estrada" does not provide any further clues. The VIAF record for Maria Stader does turn up if the search phrase includes "Swiss." In other words, this is a dead end where searching with the wrong nationality could not correct the name, and searching with the wrong name could not correct the nationality.

[4] "A Maria Stader recital," *OCLC WorldCat*, http://www.worldcat.org/oclc/2764096. The WorldCat record for the same *Liederabend* album appears 18 records later: "*Liederabend* Maria Stader," *OCLC WorldCat*, http://www.worldcat.org/oclc/ 30023258.

[5] http://www.viaf.org. Even though Maria Stader's nationality is recorded in the Library of Congress Name Authority File, searching the public interface on id.loc.gov yields no results. To perform a search for the field that includes nationality requires a tool such as the Connexion software search with an OCLC authorization credential.

The searcher is eventually turned back to square one: examining an index, be it the chronological list of sound recordings, the list in a Wikipedia article, or the title index in the *Complete Songs*.

Can bibliographic systems do better than this? One possible technological solution is linked open data.

Search Strategies and Data Strategies

Linked open data is a set of design principles for making data freely available on the Internet in a structure that allows machine processing to understand, connect, and enrich the content represented in the data.[6] This web of machine-parsable data enables the creation of new knowledge as machines make inferences based on integrating existing data sets[7] from disparate sources. Could linked open data enable machines to solve the mystery of the Schubert song? Very likely, had data from WorldCat, VIAF and Discogs been available as linked data for machines to make inferences beyond the known, and somewhat incorrect, information.

In the search for *Seligkeit*, the teacher and the student both tried to browse a title index, based on two pieces of data (Schubert, song). Had they been in a

[6] In this paper, I use the term "machines" to refer to computers, as well as all other computing devices, learning machines, and neural networks.
[7] In this paper, "data" refers to individual pieces of data and "data set" refers to pieces of data grouped together into a machine-readable structure.

library, a reference librarian might, at first, try the catalog searches discussed above, based on several more pieces of data (Schubert, song, singer's name, singer's nationality, format of recording, date of recording). This is not to say that the index browse and the catalog search could not have led to the answer. But six hundred songs are not easily browsed, and the catalog search was only able to reduce that figure by a fraction. My strategy, given only a mobile phone and 30 seconds, involved looking, in a particular order, for three pieces of data: Who was this singer (name)? Which LP recording was it (Schubert)? Which track in the recording was it (title ending with "-*lein*")? The reason for this particular order was to narrow down answers as quickly as possible, so that I could take a guess before time was up. This two-searches-and-a-guess strategy was neither unusual nor unique, but could machines have come up with it? How can we make more use of machines as an analytical tool? Machines are only as good as the programs we run and the data we supply, and there is much the library community can do about the data. After all, creating and managing data is one of our areas of expertise.

Computers are machines designed to perform arithmetic and logical instructions on data. Through a process called decomposition, humans translate complex problems into sequences of simple machine instructions and break down data into machine-parsable sets. The simple and repetitive nature of computing works well when we supply data sets with a uniform structure in which what the

data represent is unambiguous and atomic, that is, already in the lowest level of detail. In our search for *Seligkeit*, the voice teacher offered several pieces of information: a soprano (with a possible name and a possible nationality), an LP with a Schubert song (with a possible portion of the title). Figure 1 shows one possible way to decompose the data based on my search strategy.[8]

*** FIGURE 1 ***

The problem, and the challenge, is to start with these pieces of data and somehow end up with the song title *Seligkeit*. Web searching, for the most part, means to enter the data as text strings and look for where they appear on Web pages. Catalog searching finds records that have these text strings in particular fields. While field searching in a library catalog is more precise, the results are limited to bibliographic records in library systems. Linked open data, on the other hand, not only offer a global web of data for field searching, but also allow computer programs to evaluate and return additional data that ordinarily would fall outside the scope of web and catalog searching.

Data Linking on the Internet

The World Wide Web connects hypertext documents via hyperlinks, and has grown from handful of pages when first implemented in 1990 to over 1 billion

[8] Although they are legitimate clues, I did not consider the high G or the upbeat nature of the song, because I knew that this information was unlikely to turn up in a Web-based search or recorded in library data.

Web sites 36 years later.[9] Now imagine a similar scale of connected data sets! The concept behind achieving a vast amount of data interconnectedness is surprisingly simple. Basically, it requires a critical mass of data sets to appear on the Internet following four design principles. They are listed in Figure 2,[10] along with current technologies[11] that satisfy their purposes. Their ramifications are elaborated below.

**** FIGURE 2 ***

Today, we have already seen versions of these design principles in practice. The World Wide Web is a familiar example. Documents on the Web use the Universal Resource Locator (URL) as identifier; they are addressed by the prefix http:// (Hypertext Transfer Protocol); and they are marked up in a structured language HTML (Hypertext Markup Language), which provides a method (the <a> tag with the "href" attribute) to link to another document. While

[9] A Web site counter with references to the counting algorithm can be found on "Total Number of Websites," *Internet Live Stats*, http://www.internetlivestats.com/total-number-of-websites.

[10] Table adapted from Tim Berners-Lee, "Linked Data," *Design Issues*, last modified June 18, 2009, https://www.w3.org/DesignIssues/LinkedData.html. Tim Berners-Lee describes these four characteristics are "expectations of behavior" that are often erroneously understood as rules or requirements. URI, HTTP, RDF and SPARQL are listed not as requirements but as technologies of choice for their already widespread use on the Internet. He explains these brief design notes more fully in his presentation "Tim Berners-Lee: The Next Web," *TED*, February 2009, https://www.ted.com/talks/tim_berners_lee_on_the_next_web.

[11] In this paper, "technology" refers to any application of science for practical purposes, which include computing hardware, software, as well as standards and specifications for communication protocols, data models, markup and query languages, etc.

documents are linked on the Web, the use of the URL and HTML constrains machines from taking advantage of the ability to make inferences across data sets. As the identifier, each URL refers to the entire document, but not any content within it. Support for encoding machine-parsable data is also limited in HTML.[12] In other words, data that reside within a Web page are not well identified as data. As a result, a typical Web search is actually reading an enormous index of the text appearing on pages on the Web.

The web of data, on the other hand, will enable machines to understand what the data are about, so that, rather than just looking through indexes, machines will be able to perform reasoning and analysis.[13] The full potential of linked open data, therefore, depends on the way we make data available, or the way that the data can be identified and connected with other data via discoverable links that express an array of meaningful relationships.

[12] Some metadata about the document itself can be recorded in the document header; new tags and attributes have appeared in HTML5, the latest revision of the language, which added the ability to embed custom data and designating meaning for certain types of text. But the specification document acknowledges the issue of machine processing is not adequately addressed by the language. See World Wide Web Consortium, "HTML5: A Vocabulary and Associated APIs for HTML and XHTML," last modified October 28, 2014, https://www.w3.org/TR/html5/introduction.html#introduction.

[13] For example, when provided with the statements: "A soprano is a singer" and "singers are people," the machine will be able to draw the conclusion: "A soprano is a person." Taking this example a step further, given data on names and ages of sopranos, and, from a separate data set, the gender of the names, the machine will, without explicit human input, be able to generate additional understanding, such as, "A soprano is a female person – typically; a soprano is a young male person – seldom."

In the simplest terms, providing a link between data is doing exactly that: when constructing a data set, arrange the data so that each piece of data can be connected by a link to another piece of data. The structure and method that have emerged for this purpose are the data model Resource Description Framework (RDF), and its companion query language Protocol and RDF Query Language (SPARQL). While the model is simple, the actual technical specifications are more involved, and the Web community has been developing and maintaining standards and documentation.[14]

In recent years, using RDF for constructing data sets has gained substantial traction in the library community.[15] We will delve into the details of the model after a short background discussion on this technology and its relationship with library practice.

[14] The suite of RDF standards is one of the many Web standards being developed and maintained by the international membership body World Wide Web Consortium (W3C).

[15] An extensive report on the adoption of linked data by the library community can be found in Mitchell, Erik T., "Library Linked Data: Research and Adoption," *Library Technology Reports* 50, no. 5 (2013), as well as in his "Library Linked Data: Early Activity and Development," *Library Technology Reports* 52, no. 1 (2016). For an example of RDF use in a library linked data project, see the Linked Jazz Project (http://linkedjazz.org), developed at the Pratt Institute School of Library Information Science. A fuller technical exposition of the linked open data set built for this project can be found in Cristina Pattuelli, Alexandra Provo, and Hilary Thorsen, "Ontology Building for Linked Open Data: A Pragmatic Perspective," *Journal of Library Metadata* 15 (2015): 265-294, doi:10.1080/19386389.2015.1099979.

Linked Data: Technology vs. Philosophy

Even though the design principles—identifier, dereferencing, data structure and query language—are essential, this particular combination of technologies—URI, HTTP, RDF, SPARQL—is not required for building a web of data. Just as HTTP and HTML are not required to build a web of documents, other parallel "webs" based on other technologies exist today.[16] Essentially, the Internet provides the undergirding for multiple network technologies. No matter which "web," any Internet transmission, from the file to the software, through the computer's network cable to the modem into the Internet, triggers a cascade of interconnected and interlocking technologies that share interoperable specifications in spite of different computers, operating systems, or software applications.[17]

[16] For example, today over 140 servers with nearly 5 million files have been connected in "Gopherspace," a linked data environment of computer files communicated over the Internet via the Gopher protocol and a text menu structure since 1991. (The current size of Gopherspace can be found in real time by making a query in the Gopher search engine *Veronica-2*, http://gopher.floodgap.com/gopher/gw?gopher/0/v2/vstat.) Another linked data environment that has been in service on the Internet since the 1980s runs on Z39.50, a communication protocol that is heavily used in the library community for its ability to perform complex, structured searches simultaneously on multiple systems. (The Library of Congress maintains the Z39.50 standard, as well as the "Z39.50 Register of Implementors," last modified September 2016, https://www.loc.gov/z3950/agency/register/entries.html.)

[17] Using the Internet requires adhering to standards involving a broad range of transmission protocols, data formats, markup languages, and query languages, as well as hardware, including modems, switches, routers, and data cables.

Because the size and reach of the Internet provides a positive feedback, new Internet-related technologies, products and services will be developed to be compatible. Initially, the popularity of HTTP and HTML made them the de facto standards for the Web. Then Cascading Style Sheets (CSS) became a ubiquitous language for Web page design and layout when major Web browsers began to support it. Similarly, in the near future, we expect technology standards for the web of data to develop and coalesce,[18] with URI, HTTP, RDF and SPARQL as the basis for this new web architecture.

By employing the Internet, we also subscribe to the philosophy behind Internet architecture that is open, interoperable, evolvable, and network-accessible. MARC, an architecture of library systems and operations since the 1970s, is at odds with this philosophy. As the Internet grew and matured, the library community long recognized the divergence between MARC, the closed architecture of library catalogs and the open architecture of the Internet.[19] Even though MARC stands for MAchine Readable Cataloging, the central purpose of machine processing was to print database records on catalog cards and on

[18] See Berners-Lee, Tim, "Web Architecture from 50,000 Feet," Design Issues, last modified August 27, 2009, https://www.w3.org/DesignIssues/Architecture.html.

[19] With over 11,000 data elements, MARC is a closed data format that making it interoperate on the Internet requires complex procedural workarounds. For an experimental study on MARC authority data, see Papadakis, Ionnas, Konstantinos Kyprianos, and Michalis Stefanidakis, "Linked Data URIs and Libraries: The Story So Far" *D-Lib Magazine* 21, no.5/6 (2015), doi: 10.1045/may2015-papadakis.

computer screens. Since then, we continued to design databases, interfaces, and discovery systems modeled on the catalog card, and contents follow a highly controlled syntax in individually demarcated records. This design allows the library community to create quality-controlled data in robust systems that communicate with each other, but not with the open Internet. By contrast, linked data design is open and dynamic: there are no fixed records, and, at any time, any Internet user, human or machine, can supply data and create links between data. Linked data is as much a state of mind as it is technology.[20]

Recognizing RDF Linked Data

Because of linked data's open design, a flexible data model such as RDF has emerged as the standard for the Web of data. The basic structure of RDF is the triple. The RDF triple enables assertions by linking two pieces of data with a one-way relationship between the two. This model appears in various guises in various disciplines, for example: Node-Arc-Node (mathematics/graph theory), Subject-Predicate-Object (linguistics), Object-Attribute-Value (programming); Entity-Relationship-Value (software engineering), Record-Field-Data (relational database), Resource-Property-Value (information science). The RDF model can

[20] For an in-depth discussion on conceptualizing library data models, see Alemu, Getaneh, Brett Stevens, Penny Ross, and Jane Chandler, "Linked Data for Libraries: Benefits of a Conceptual Shift from Library-Specific Record Structures to RDF-based Data Models," *78th IFLA General Conference and Assembly (2012)*, http://www.ifla.org/past-wlic/2012/92-alemu-en.pdf.

also be implemented in various forms. A good way to understand RDF is to recognize RDF in familiar places.

RDF Reading of a Spreadsheet. Because each serves a distinct function, rows and columns of a spreadsheet are not interchangeable. For example, in a spreadsheet for instrumentation of musical pieces such as Figure 3,[21] each row is a record about a piece of music and each column represents the Deutsch number and an instrument used in the piece. The header of each row holds the title of the piece, and the header of each column designates what the information is about in the cells below. In an RDF reading of this table, title is the resource, Deutsch number and instrument/voice are the properties, and each cell contains the value. In other words, to construe a spreadsheet as RDF triples, the row header is the resource, the column header is the property, and the row-column intersection is the value, or: Row-Column-Cell, as shown in Figure 4. In Figure 5, I re-wrote the spreadsheet as a set of RDF triples.

*** FIGURE 3 *** *** FIGURE 4 *** *** FIGURE 5 ***

In essence, the structure of this particular spreadsheet can be configured as shown in Figure 6 and 7. Note that the rows and columns with repeated headers only need to appear once in RDF, because RDF imposes no limits on the number

[21] For the purpose of illustrating contrasting data, I chose two other Schubert songs for this and subsequent examples.

of properties, including repeated ones, a single resources can have. Figure 8 shows these relationships graphically.

*** FIGURE 6 *** *** FIGURE 7 *** *** FIGURE 8 ***

While this spreadsheet can be construed as a set of RDF triples, the structure of this spreadsheet creates several constraints that limit the machine's ability to understand the data fully. The spreadsheet limits the number of entries for instrument/voice to three.[22] The three-column design compels data to be modified in certain situations. For *Auf den Sieg der Deutschen*, we enter "two violins" because entering "voice," "violin," "violin," "cello" requires four columns, so, to fit the data into three columns, the two appearances of "violin" are combined into a single entry "two violins." Allowing the use of the word "two," the meaning of the column is no longer unambiguous: because "two" is a number, not an instrument/voice. Moreover, the data is no longer atomic: because "two" and "violin" are two distinct pieces of data. For *Brüder, schrecklich brennt die Thräne*, there are not enough columns to list all the instruments of the orchestra, so we enter "small orchestra." In this case, the meaning of the column is, again, no longer unambiguous, because "orchestra" is an ensemble, not an

[22] It might be easy to add another column in a spreadsheet application, but if this were a table as a part of a larger relational database, adding columns could be laborious task. Altering the design of a relational database, such as adding a column to, usually requires creating a development copy of the database and testing all existing functionalities against it.

instrument/voice, and "small" is a qualifier of the orchestra, not itself an instrument/voice.

Could we not change the column, then, to "instrument/voice or ensemble and the number thereof" so that we could capture as much information as possible in the limited space? While this appeals to human sensibility, machines would either be confused, or led to make inferences that are incorrect. On the other hand, changing the way we understand what instrumentation is about can lead to us structuring the data in a way that machines can understand.

Creating Machine-Parsable Data

Instrumentation, or medium of performance, is a complex concept. Decomposing the data in play reveals four components: part, instrument/voice, player, and ensemble.[23] Illustrating them as RDF properties, these four components are interrelated as shown in the schematic in Figure 9[24]: a piece of music consists of parts; each part calls for instruments/voices; each part also calls

[23] Part, instrument/voice, player and ensemble refer to the abstract concept, rather than the physical printed part, the physical instrument, the actual person, or a specific ensemble.

[24] Earlier versions of this diagram with its technical underpinnings were presented on October 15, 2016 at the chapter meeting of the New York State-Ontario Chapter of the Music Library Association in Toronto, Canada and on July 8, 2016 at the annual congress of the International Association of Music Libraries, Archives and Documentation Centres in Rome, Italy. I would like to thank my international colleagues for their valuable input.

for certain types of players; each player is responsible for one or more parts; and various parts may be grouped into an ensemble.

*** FIGURE 9 ***

This model resolves the atomicity and ambiguity problems we encountered earlier. If the score calls for two violins, as in *Auf den Sieg der Deutschen*, there will simply be two individual links to a violin part. If the score calls for a small orchestra, as in *Brüder, schrecklich brennt die Thräne*, there will be nine individual links to the nine orchestral parts, and then each of the nine parts will link out to a single orchestra.

This model can further resolve problematic situations toward describing medium of performance in current music cataloging practice. For example, instrumental doubling and generic instruments such as “percussion” can be expressed like this: a part is linked to multiple instruments; those instruments are all linked to one player; that player is linked back to the part. This level of specificity is possible because part, instrument/voice, and player are independent properties. Doing so also eliminates the need to enter the number of parts, the number of players, or the number of ensembles, because each of these numbers can be obtained by counting links, a task that machines can accomplish.

Other details of medium of performance can also be captured with more refined properties. For example, the “alternative medium of performance” concept (which is defined with subtle differences in MARC field 382 subfield p and in

UNIMARC field 146 indicator 2 and subfields b to f position 8) can be expressed using properties that signify alternative, and used only for the component in question. This leads to a more precise understanding of what is an alternative to what, in a number of distinct scenarios, including "same piece of music but consisting of different parts" and "same part but calling for different instrument/voice." Expressing alternatives this way not only covers situations where the alternative is explicit, such as a "sonata for clarinet or viola and piano," where the viola part is the alternative to the clarinet part while the piano part is unchanged. It also allows us to see other cataloging concepts in new light; for example, it is possible to express a piano/vocal version of an opera as the piano part being the alternative to all the orchestral parts together while the voice parts remain unchanged. However, we might want to make the distinction between these two types of alternative-ness. With linked open data, we are free to refine the "alternative" property to a "derivative of" property. Or, refining to show various degrees and styles of derivative-ness, such as "part adapted for" (another instrument), "orchestration of," "reduced orchestra version of," "piano reduction of," "re-orchestration of," "adapted for" (a different instrumentation), or even "reconstruction of," "recreation of," "inspired by," "re-styling of." While for a human user using "alternative" will suffice for all these scenarios, more precise properties allow machines to acquire more nuanced understanding, especially for complex concepts and the many degrees of equivalence and similarity.

For machines, it is perfectly acceptable to encounter relations that are not equivalent. Depending on the sophistication of the program, machines can do the job of analyzing the nature of the similarity, evaluating the degrees of similarity, and calculating the likelihood of usefulness when responding to a query, or the machine's version of taking a guess. So, the problem is not that medium of performance concepts are not equivalent between MARC field 382 and UNIMARC field 146, but is the lack of equivalence and similarity relationships defined to bridge the two. The same problem extends outside library data—no equivalence or similarity relationships exist for connecting library medium of performance data with other non-library data sets, such as Discogs. To build a global web of linked open data, providing the means to connect them is key.

Library Data as Linked Data

The library community is fortunate to have quality data created by trained specialists in a uniform, structured database design. The downside is that as cataloging has evolved over time, idiosyncrasies have crept into our practice. Without knowing or realizing the full implication of linked open data technologies that would later emerge, we have inadvertently developed cataloging rules to accommodate data structures rather than atomic data and unambiguous properties, and modified data structures to accommodate conventional human

usage and readability. These developments not only hinder machine-parsability, making it difficult for library data to be processed easily on the open Internet.

In recent years, however, the cataloging community have incrementally positioned itself to enable linked data implementations. Theoretical work and case studies have been done with the content standard RDA, the underlying conceptual model FRBR, and the future MARC format replacement, BIBFRAME.[25] For music, there are several active linked data initiatives under way.[26] Nevertheless, it is extremely important to recognize that our cataloging practice has been focused on enabling human tasks. We operate on a set of looming assumptions that: there is a thing (physical or electronic); people are intentionally looking for it (or stumble upon it while looking for something else); people want to get it into their

[25] For an explanation on modeling RDA in RDF, see Szeto, Kimmy, "Positioning Library Data for the Semantic Web: Recent Developments in Resource Description," *Journal of Web Librarianship* 7, no.3 (2013): 305-321, doi: 10.1080/19322909.2013.802584; an analysis of FRBR and its applicability to the linked data environment, see Coyle, Karen, "Bibliographic Description and the Semantic Web," *FRBR Before and After: A Look at Our Bibliographic Models*, (Chicago: ALA Editions, 2016), 137-156; a technical paper on modeling FRBR, RDA and BIBFRAME and the tension between closed and open data can be found in Baker, Thomas, Karen Coyle, and Sean Petiya, "Multi-Entity Models of Resource Description in the Semantic Web: A comparison of FRBR, RDA, and BIBFRAME," *Library Hi Tech* 32, no. 4 (2014): 562-582, doi: 10.1108/LHT-08-2014-0081.

[26] For example, Linked Data for Production: Performed Music Ontology (https://wiki.duraspace.org/display/LD4P/Performed+Music+Ontology+Project), DOing REusable MUSical data (www.doremus.org), the Europeana Data Model (http://pro.europeana.eu/page/edm-documentation), and the Music Notation Community Group of the World Wide Web Consortium (https://www.w3.org/community/music-notation).

possession (physically or electronically); and, once in possession, they want to "use" it (to read, to play, to deploy, to somehow consume its content). By contrast, in the web of data, "people" make up a shrinking subset of the users while machine processing is promoted. Programs and algorithms crawl the web of data to build knowledge of their own and to answer human queries. The questions for us today are how to supply data to this web of data[27] and how to harness machines' analytical power for library users.[28]

The Linked Data Promise

As to our original search for the Schubert song, I can safely say linked open data could enable machines to overcome the uncertainties: Misspelled name? Google suggested the correct one. Wrong country? Geographic proximity would lead to singers from Switzerland assigned a higher likelihood. The song title not really ending with "*-lein*"? "*-keit*" would more likely be found as a partial match. Possibly a one-word song title? Short song titles would be given more weight.

[27] A comprehensive overview of the linked data vision can be found in Tim Berners-Lee and Mark Fischetti, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, (San Francisco: Harper Collins, 1999), and Tom Heath and Christian Bizer, *Linked Data: Evolving the Web into a Global Data Space*, (San Rafael, Calif.: Morgan & Claypool, 2011, doi: 10.2200/S00334ED1V01Y201102WBE001.

[28] Philip Schreur discusses how this paradigm shift affects library technical services in his araticle "The Academy Unbound: Linked Data as Revolution," *Library Resources & Technical Services* 56, no. 4 (2012): 227-237. doi: 10.5860/lrts.56n4.227.

Possibly on the last track on the LP? Machines would understand tracks are often shuffled in re-issues, thus giving this criterion less scrutiny. And, what about the high G? Software can now read and notate music with much improved accuracy, and G-sharp is in close proximity. The upbeat nature of the song? Proprietary online music streaming services have been developing algorithms to capture mood in music.

Linked open data invites us to re-orient our approach to creating, managing, and curating data. In return, it lowers the barriers to accessing information and enables knowledge production on a massive scale. The technology is there, and we can, in fact, do better. But first, at least for the library community, we must do a better job working with machines so that machines can work better for us.

Abstract

Linked open data promises global interconnectedness of a vast amount of data. Web technologies promise to lower the barriers to accessing information and to enable knowledge production of massive scale. But can the web of data answer a music reference question? Starting with a seemingly impossible search for a Schubert song, this article describes how linked data technologies could overcome some limitations of catalog searching. However, technical and conceptual challenges are intertwined in the library community's effort to publish linked data. Through an analysis of contrasting data models, this article offers a linked data reading of medium of performance and how the data can be tweaked to improve machine processing. This example leads to a discussion on general strategies towards an open, interoperable, evolvable, machine-actionable network that enables computers to become more effective tools for answering human questions.

Author bio

Kimmy Szeto is an assistant professor and metadata librarian at Baruch College, City University of New York, where he oversees metadata management for digital resources. His recent research focuses on the technical and conceptual tensions between cataloging practice and the linked data environment. He is an active presenter at Music Library Association meetings and has published in the *Journal of Web Librarianship*, the *Journal of Electronic Resources Librarianship*, and the *Encyclopedia of Information Science and Technology*. Outside the library and academia, Kimmy can be heard as the chamber arranger of symphonic works and as a collaborative pianist in theaters around New York City.

Figure 1. A decomposition of data involved in the mystery of the Schubert song (data in bold are without uncertainty)

a person
- who performs as
  - **a soprano**
- who resides in
  - Austria — *or possibly somewhere in Europe*
- who has the name
  - Strada
  - Estrada — *or possibly something with a similar sound or spelling*

a song
- having been composed by
  - **Franz Schubert**
- has instrumentation/voices consisting of
  - voice (solo)
  - piano
- has the title
  - that has one word — *possibly plus an initial article*
  possibly with more words it's most likely not a long title
  possibly has the word that ends with "-lein"
  or something similar to that

an LP
- was recorded on a date
  - not too recent
- **includes a track that is** — *possibly on the last track on one of the sides*
  - **the song above**
  - **performed by**
  - **the person above**

Figure 2. Design principles for linked open data

Design Principle	Purpose	Current Technologies
Identifier	Identifiers allows data and links to be uniquely identifiable, globally.	URI
Dereferencing	Dereferencing a URI is retrieving a representation of that resource. A global addressing system enables URIs to be accessed and to self-identify.	HTTP URI
Structure and Method	Data can be useful only if queries return data. A common method or language for accessing data in a common structure makes the data globally discoverable.	RDF, SPARQL
Participation	The success of this vision of the linked data environment rests on connecting a vast amount of data across the Internet.	Include links to other URIs

Figure 3. Deutsch number and instrumentation of three Schubert songs

	Deutsch Number	Instrument/ Voice	Instrument/ Voice	Instrument/ Voice
Auf den Sieg der Deutschen	81	voice	two violins	cello
Brüder, schrecklich brennt die Thräne	535	soprano	small orchestra	
Seligkeit	433	voice	piano	

Figure 4. RDF reading of a spreadsheet

	Has Property: Deutsch Number	Has Property: Instrument/Voice	Has Property: Instrument/Voice	Has Property: Instrument/Voice
Resource: *Auf den Sieg der Deutschen*	Value: 81	Value: voice	Value: two violins	Value: cello
Resource: *Brüder, schrecklich brennt die Thräne*	Value: 535	Value: soprano	Value: small orchestra	Value: <empty>
Resource: *Seligkeit*	Value: 433	Value: voice	Value: piano	Value: <empty>

Figure 5. RDF triples of the spreadsheet in Figure 3

Auf den Sieg der Deutschen	→ *has Deutsch Number*	→ 81
Auf den Sieg der Deutschen	→ *has instrument/voice*	→ voice
Auf den Sieg der Deutschen	→ *has instrument/voice*	→ two violins
Auf den Sieg der Deutschen	→ *has instrument/voice*	→ cello
Brüder, schrecklich brennt die Thräne	→ *has Deutsch Number*	→ 535
Brüder, schrecklich brennt die Thräne	→ *has instrument/voice*	→ soprano
Brüder, schrecklich brennt die Thräne	→ *has instrument/voice*	→ small orchestra
Seligkeit	→ *has Deutsch Number*	→ 433
Seligkeit	→ *has instrument/voice*	→ voice
Seligkeit	→ *has instrument/voice*	→ piano

Figure 6. Data structure represented in the spreadsheet in Figure 3

	Deutsch Number	Instrument/Voice
Piece	number	name

Figure 7. RDF reading of the spreadsheet in Figure 6

Piece → *has Deutsch number* → Number
Piece → *has instrument* → Name