E-041

LinkedMusic Project: Integrating Music Databases

藤永一朗†
Ichiro Fujinaga

1. Introduction

Digital technology has significantly advanced online text searches, enabling users to explore research questions quickly and efficiently. In contrast, online music search is still in its infancy. Despite the increased availability of music on the Internet, large-scale searches of music data remain challenging due to the diversity of musical material, which includes scores in various notations and audio/video recordings in multiple formats. Unlike text documents, where metadata can often be inferred and utilized for search, musical content cannot be easily searched using “Google-style” text queries.

Hundreds of specialized online music databases contain valuable metadata and musical content, but they often use different metadata schemas, making cross-database searches difficult and limiting their usefulness for music research. Music search faces unique challenges: while musical metadata may be in text (e.g., title, composer), the content is usually in diverse formats such as notation or audio recordings, and most musical content does not convey meaning in the same way as text. Therefore, tools effective for text searches are often inadequate for music searches for text searches are inadequate for music searches.

Furthermore, music is complicated. It has been said among librarians that “if you can catalogue music, you can catalogue anything” [21]. A piece of music can be represented by a score, a collection of performing parts, an arrangement (e.g., for a choir), a sound or video recording, a computer program, etc. Each of these representations may have different and multiple contributors, such as composers, performers, recording engineers, and video producers. Traditional library cataloguing standards were mainly designed for books, and music cataloguers have struggled to work around these limitations. Additionally, copyright and licensing issues remain significant, especially with popular music.

Currently, when people, including musicians, scholars, journalists, educators, and the public, look for musical information (musical scores, performer information, etc.), they might need to visit several different databases, each with its own user interface and search methods, and combine the search results to obtain their desired answer. Through this project, users with varying interests and goals will be able to query across multiple databases without having to visit each one (or even knowing that each one exists independently), greatly accelerating and enriching their search experience. This new environment will enable new research methods (e.g., applying big data approaches to a broader range of musical material), change the way music is created and performed, and deepen our understanding of how music relates to the human experience.

Partnering with nine major international institutions at the forefront of music database management systems, our goal is to build the world’s first comprehensive global digital music library by fully incorporating their metadata. Metadata for music may include title, composer, arrangers, lyricist, performers, date, duration, publisher, etc. In any library, metadata is the key to locating items. Our digital music library will include not only music scores but also audio/video recordings, performer biographies, album covers, concerts, playlists, and music reviews.

There are several challenges in creating a global digital music library. Although a significant amount of music data is already available in hundreds of different online databases, it is difficult to search across these databases due to varying query interfaces and metadata schemas. Furthermore, the number of databases and different schemas is rapidly increasing across the Internet. Therefore, there is a pressing need to create a framework that will facilitate the interlinking of databases, allowing for greater dissemination and re-use. This is particularly critical in the context of open educational resources and remote learning, where free access to high-quality online materials is increasingly important. The phenomenon of metadata schema proliferation is widespread and not limited to music [13][20]; thus, successful outcomes of this project can be applied to interlinking databases in many other domains, such as healthcare and manufacturing.

We will achieve our goal by developing tools to convert existing metadata into a flexible and extensible format based on linked data principles and then store the metadata in a data lake. This unique strategy avoids the expensive and unrealistic task of developing a new, monolithic metadata schema that would require adoption by hundreds of database providers. The aggregated metadata will be indexed in an online open-source metasearch engine, we call SESEMMI (Search Engine System for Enhancing Music Metadata Interoperability), where many different musical resources can be searched simultaneously and linked back to the individual databases for the content. The extensibility will be vital to making SESEMMI inclusive and culturally sensitive, as we can incorporate new vocabulary and variables as the project develops, effectively and efficiently including more music. By performing large-scale searches across multiple platforms and in multiple languages, users will be able to track the movement of different genres and traditions (e.g., K-pop, Carnatic music, or rock) around the world; search by demographic information from Wikidata (e.g., place, gender, or race); track and compare performers’ careers; and document traditional and indigenous music cultures and materials.

Our partners include major music libraries, an international music library consortium, and non-profit organizations hosting large, well-known music databases. LinkedMusic Project brings together leading experts from (ethno)musicology, library and information sciences, and music information retrieval, whose combined expertise is necessary to make these critical and timely breakthroughs in music research and to help train a new generation of experts. By interlinking existing online metadata while considering inclusivity and accessibility, our project aims to

support new fields of inquiry and fundamentally change future scholarship, creation, listening, and performance.

2. Previous Work

A parallel project in the field of visual art, known as Linked Art [1] involves prominent institutions like the Louvre, Rijksmuseum, Metropolitan Museum of Art, Museum of Modern Art, and Victoria and Albert Museum. The Linked Art Editorial Board includes members from the Canadian Heritage Information Network, Europeana, and Oxford University. We plan to consult their conceptual model and adopt their general principles, [2] including the design goals of Linked Open Usable Data (LOUD).

In the realm of music, past projects such as the Sheet Music Consortium (SMC)[3] and the DOREMUS Project [1] have aimed to interlink various music collections. The SMC successfully connected 36 different collections, benefiting from the relatively homogeneous nature of American sheet music, which primarily consists of a few pages of piano music. In contrast, our project encompasses a diverse array of musical information, including scores, reviews, and recordings. The DOREMUS Project aimed to interlink three major music databases in France by studying and consolidating their metadata schemas.

The challenge of interlinking databases has a long history across various fields, from genomics to large corporations [2][6]. Ideally, every music database would adopt a common metadata schema, similar to the Europeana consortium's strategy, where members convert their metadata to the central Europeana Data Model [7]. This is unrealistic for our project, however, as most organizations hosting music databases have limited resources.

The desire to link online resources originates from the concept of the Semantic Web [4], where machines can interpret data using Linked Data principles [5]. Linked Data uniquely identifies concepts in a machine-readable way, allowing for precise queries. For example, a keyword search for "Schumann" might yield results for both Robert and Clara Schumann but searching by identifier removes this ambiguity. Despite significant progress over the past two decades in deploying Linked Data, the dream of "query the Web as a single global database" [9] remains unfulfilled. One major challenge is the disparate metadata schemas used by different data sources [3].

While general standard metadata schemas, like FOAF (Friend of a Friend) [7] and Dublin Core [22] exist , specialized subdisciplines have created their own schemas to meet specific needs. This proliferation of custom schemas has resulted in databases that cannot communicate effectively with each other, even within a single discipline such as music.

3. Methodology

In this project, we will combine metadata from 14 different music databases, each with its own schema, to enable users to search across all databases from a single site. This will involve developing strategies to integrate different metadata schemas and creating a metasearch engine for music. Given the complexity of musical information, instead of striving to create a perfect metadata schema standard, we are exploring the use of data lakes [17] and "lazy searching."

Our novel approach ensures that hosts of existing databases will not need to modify their schemas. Instead, we will import their metadata and store it in our data lake. This is a powerful concept because modifying the schema of an existing database is a complex and expensive process that has prevented many legacy databases from being integrated into larger or newer systems in both academia and industry [11][23].

Lazy searching is analogous to lazy learning in machine learning, where the model defers most computations until it receives a query to make a prediction. Unlike "eager learning" methods (e.g., artificial neural networks), which build a general model based on the entire training dataset before receiving any queries, lazy learning methods, like the k-nearest neighbour classifier, do not create a general model in advance. Instead, they store the training dataset and only perform computations when a prediction is required. Thus, with lazy searching, minimal effort is exerted during the ingestion of various metadata, avoiding complicated ontology mapping or schema alignment [19]. Instead, more effort is spent at the query stage, analyzing the query using methods such as Large Language Models [16] to deliver useful answers by searching the data lake of linked data.

Our strategy involves having each musical database host provide us with a complete data dump (e.g., database dump or a CSV file). We will convert the data dump to RDF (Resource Description Framework) independently of other databases and deposit it into our data lake. This process will not involve standard metadata mapping across different sources. The issue of disparate schemas is addressed through lazy searching.

Once the metadata is gathered in the data lake, we will create an open-access metasearch engine: SESEMMI. This website will allow metadata to be searched across different databases. Localization and internationalization are critical concepts for this project. SESEMMI will be multilingual, enabling searches in various databases by global users.

Linked Data enables us to define data and relationships independently of human language, allowing users to search and see results in their preferred language, regardless of the language of the database they are searching. During the mapping process, every item in the original metadata will be semi-automatically assigned a unique identifier, called a URI (Uniform Resource Identifier), including names and concepts. Many of these names and concepts already have URIs, for example in Schema.org[4] and Wikidata. [5] This approach minimizes the need to choose a "standard" vocabulary or spelling for an entry and allows us to describe relationships between concepts in a machine-readable way. For example, a URI can unite variant spellings of Tchaikovsky in any script, all the modern and historical names of

[1] https://linked.art
[2] https://linked.art/model/
[3] https://digital.library.ucla.edu/sheetmusic/
[4] https://schema.org
[5] https://www.wikidata.org

a particular region, and even all the different names for a specific musical instrument.

Linking entities such as individuals to a URI also allows information about individuals (such as sexual orientation or death date) to be edited or updated independently of our interlinked databases. Details such as name, nationality, gender, or profession do not need to be managed locally. In addition to the music databases that will be part of our interlinked resources, we will pull in data from sources like Wikidata. This means queries involving information in Wikidata that might not be in the original music database (e.g., race or gender) will be possible.

For internationalization, challenges extend beyond merely translating metadata entities. Although physical entities, such as a person or a musical instrument, can be represented by a URI with links to various spellings in different languages, concepts or labels like “composer” or “performer” may not translate directly. These concepts can have different meanings depending on the culture. Using Linked Data, we can connect related concepts, showing relationships between similar categories without forcing them into a single framework. In addressing this issue, we draw on current scholarship regarding multilingual and cross-cultural Linked Data [12][18].

4. Current Workflow

We are currently converting each database’s metadata into RDF format while preserving their original metadata schemas. We first use OpenRefine to reconcile each entity with properties and items in Wikidata and other metadata standards (e.g., Schema.org) to a URI and then create RDF statements. Most of these databases are based on standard relational databases, and numerous tools are available to export this type of data into Linked Data [15]. We are storing the results in the OpenLink’s Virtuoso graph database[1] (see Figure 1).

Although the OpenRefine reconciliation process is semi-automatic, we are developing specific conversion scripts for each database we import. This ensures that if we need to re-import the metadata in the future, for example, due to updates to the original database, we can perform the ingestion automatically.

We are also experimenting with ChatGPT[2] to convert natural language queries into SPARQL queries.[3] This promising method would greatly enhance the usability of SESEMMI as users can query using natural languages [14]. We expect other AI technologies in the future to be able to find information in data lakes, beyond using SPARQL queries [10].

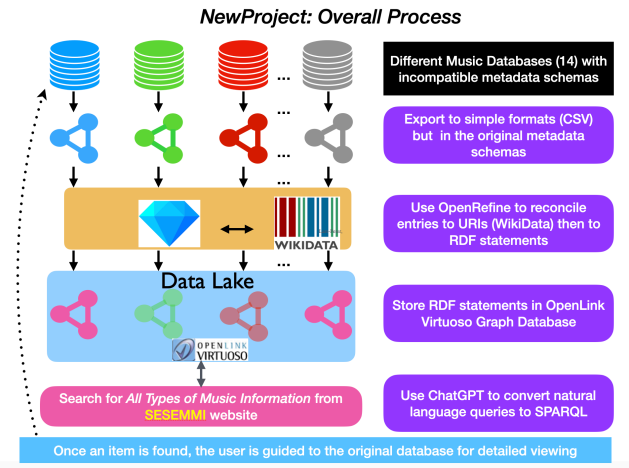

Figure 1: Workflow of integrating various online music databases into our data lake (OpenLink Virtuoso) by converting the metadata of the databases into RDF. We then use ChatGPT for creating SPARQL queries for the search.

5. Conclusions

The results of this project are poised to exert a long-lasting influence on the global music community by providing enhanced access to diverse musical information for scholars, musicians, and the general public. This initiative will transform music research into a more evidence-based discipline through the utilization of extensive data via our networked metadata search. Upon successfully enabling searches across a network of 14 open databases, we plan to expand this network to include additional music databases. A key contribution of this project is our unique framework that integrates multiple metadata schemas, allowing seamless searches across various music databases through our metasearch engine, SESEMMI. Our innovative methodology circumvents the need for data providers to alter their schemas, thereby increasing the visibility of smaller databases and attracting a broader user base.

Acknowledgments

This project is supported in part by the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2022-1004) and the Fonds de Recherche du Québec (FRQSC SE3-303927).

References

[1] Achichi Manel, Lisena Paquale, Todorov, Troncy Raphäel, Delahousse Jean, “DOREMUS: A Graph of Linked Musical Works”, *The Semantic Web – ISWC 2018*, Lecture Notes in Computer Science, Vol. 11137, edited by D. Vrandečić et al., pp. 3–19 (2018).

[2] Antoniou Grigoris, Van Harmelen Fran, *Semantic Web Primer,* 2nd Edition. Cambridge, MA: MIT Press (2008).

[3] Baca Murtha, “Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information”, *Cataloging & Classification Quarterly* Vol. 36, No. 3–4, pp. 47–55 (2003).

[1] https://virtuoso.openlinksw.com
[2] https://openai.com/index/chatgpt

[3] SPARQL is an RDF query language. See: https://www.w3.org/TR/sparql11-query/

[4] Berners-Lee Tim, Hendler James, Lassila Ora. “The Semantic Web”, *Scientific American* Vol. 284, pp. 29–37 (2001).
[5] Bizer Christian, Heath Tom, Berners-Lee Tim, “Linked Data: The Story so Far”, *International Journal on Semantic Web an Information Systems* Vol. 5, No. 3, pp. 1–22 (2009).
[6] Doan AnHai, and Halevy Alon Y., “Semantic Integration Research in the Database Community: A Brief Survey”, *AI Magazine* Vol. 26, No. 1, pp. 83–94 (2005).
[7] Doerr Martin, Gradmann Stefan, Hennicke Steffen, Isaac Antoine, Meghini Carlo, Van de Sompel Herbert, “The Europeana Data Model (EDM)” *World Library and Information Congress: 76th IFLA General Conference and Assembly* (2010).
[8] Graves Mike, Constabaris Adam, Brickley Dan, “FOAF: Connecting People on the Semantic Web”, *Cataloging & Classification Quarterly* Vol. 43 No. 3–4, pp. 191–202 (2007).
[9] Heath Tom, Bizer Christian, *Linked Data: Evolving the Web into a Global Data Space*, San Rafael, CA: Morgan & Claypool, p. 107 (2011).
[10] Hertling Sven, Paulheim Heiko, “Olala: Ontology Matching with Large Language Models”, *Proceedings of the 12th Knowledge Capture Conference*, pp. 131–139 (2023).
[11] Hsu Cheng, *Enterprise Integration and Modeling: The Metadatabase Approach. Springer Science & Business Media* (2012).
[12] Kontokostas Dmitris, Bratsas Charalampos, Auer Sören, Hellman Sebastian, Antoniou Ioannis, Metakides George, “Internationalization of Linked Data: The Case of the Greek DBpedia Edition”, *Journal of Web Semantics* Vol. 15, pp. 51–61 (2012).
[13] Lois, Chan Mai, Zeng Marcia Lei, “Metadata interoperability and standardization–a study of methodology Part I”, *D-Lib magazine* Vol. 12, No. 6 (2006).
[14] Meyer Lars-Peter, Stadler Claus, Frey Johannes, Radtke Norman, Junghanns Kurt, Meissner Roy, Dziwis Gordian, Bulert Kirill, Martin Michael, “LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT”, *Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow*, pp. 103–115 (2023).
[15] Michel Franck, Djimenou Loïc, Faron-Zucker Catherine, Montagnat Johan, “Translation of Relational and Non-Relational Databases into RDF with xR2RML”, *Proceedings of the International Conference on Web Information Systems and Technologies,* pp. 443–454 (2015).
[16] Minaee Shervin, Mikolov Tomas, Nikzad Narjes, Chenaghlu Meysam, Socher Richard, Amatriain Xavier, Gao Jianfeng, “Large Language Models: A Survey”, *arXiv preprint arXiv:2402.06196*, (2024).
[17] O’Leary Daniel E., “Embedding AI and Crowdsourcing in the Big Data Lake”, *IEEE Intelligent Systems,* Vol. 29, No. 5, pp. 70–73 (2014).
[18] Reid Geneviève, Sieber Renée, “Do Geospatial Ontologies Perpetuate Indigenous Assimilation?”, *Progress in Geography,* Vol. 44, No. 2, pp. 216–234 (2020).
[19] Shvaiko Pavel and Euzenat Jérôme, “Ontology Matching: State of the Art and Future Challenges”, *IEEE Transactions on Knowledge and Data Engineering,* Vol. 25, No. 1, pp. 158–176 (2013).
[20] Patrício Helena Simões, Cordeiro Maria Inês, Ramos Pedro Nogueira, “From the Web of Bibliographic Data to the Web of Bibliographic Meaning: Structuring, Interlinking and Validating Ontologies on the Semantic Web”, *International Journal of Metadata, Semantics and Ontologies,* Vol. 14, No. 2, pp. 124–134, (2020).
[21] Vellucci Sherry L., “Bibliographic relationships and the future of music catalogues”, *Fontes Artis Musicae,* Vol. 45, No. 3–4, pp. 213–226, (1998).
[22] Weibel Stuart L., “The Dublin Core: A Simple Content Description Model for Electronic Resources”, *Bulletin of the American Society for Information Science,* Vol. 24, No. 1, pp. 9–11 (1997).
[23] Xu Boyi, Xu Ke, Fu LiuLiu, Li Ling, Xin Weiwei, Cai Hongming, “Healthcare Data Analytics: Using a Metadata Annotation Approach for Integrating Electronic Hospital Records”, *Journal of Management Analytics,* Vol. 3, No. 2, pp. 136–511. (2016).