RePIM in LOD: semantic technologies to manage, preserve, and disseminate knowledge about Italian secular music and lyric poetry from the 16th-17th centuries

Paolo Bonora

Dipartimento di Filologia Classica ed Italianistica, Università di Bologna, Bologna, Italia
paolo.bonora@unibo.it

Angelo Pompilio

Dipartimento di Beni Culturali, Università di Bologna, Bologna, Italia
angelo.pompilio@unibo.it

Abstract

The *RePIM in LOD* project aimed to publish the "*Repertorio della Poesia Italiana in Musica, 1500-1700*" (RePIM) as Linked Open Data (LOD) dataset. For the extent and detail of its contents, RePIM is a reference archive for research on Italian secular music from the 16th-17th centuries. In recent years, scholars have been able to access it through a public web-based application. Due to the obsolescence of its information technology platform, the RePIM repository was set to be taken offline. To preserve this precious source, the project migrated its contents into a knowledge base adopting semantic technologies and designing an up-to-date end-user application. The article illustrates the challenges of managing information about madrigal tradition and the digital knowledge preservation of bibliographic and philological information in the field of Italian secular music and lyric poetry of the 16th-17th centuries.

Il progetto *RePIM in LOD* si è posto l'obiettivo di pubblicare il "Repertorio della Poesia Italiana in Musica, 1500-1700" (RePIM) come dataset Linked Open Data (LOD). Per l'ampiezza e il dettaglio dei suoi contenuti, il RePIM è un archivio di riferimento per la ricerca sulla musica profana italiana dei secoli XVI-XVII. Negli ultimi anni, gli studiosi hanno potuto accedervi attraverso un'applicazione web realizzata nei primi anni Duemila. A causa dell'obsolescenza della sua piattaforma informatica, l'archivio RePIM era destinato a essere messo offline. Per preservare questa preziosa fonte, il progetto ha migrato i suoi contenuti in una base di conoscenza adottando

tecnologie semantiche e progettando un'applicazione aggiornata per l'utente finale. L'articolo illustra le problematiche poste dalla gestione delle informazioni sulla tradizione madrigalistica e della conservazione digitale delle informazioni bibliografiche e filologiche nel campo della musica profana italiana e della poesia lirica dei secoli XVI-XVII.

Keywords

Introduction[1]

The "Repertorio della Poesia Italiana in Musica, 1500-1700" (RePIM), is the digital version of the incipit index of Italian poetry in music, a project conceived at the end of the 70s by Lorenzo Bianconi in collaboration with Angelo Pompilio and Antonio Vassalli ([20]). Since 1990, materials collected on paper cards during the 70s-80s, and bibliographic descriptions of poetic and musical sources, have been revised and widely integrated. As a rule, musical sources of secular music do not declare authorship of the intoned literary text. To identify the poet, it is necessary to collate the musical sources with coeval poetry books. The RePIM incipit index was designed to respond to this basic scientific need: to identify the greatest possible number of authors of poetic texts set to music in secular and spiritual compositions of the 16th-17th centuries through the systematic examination of poetry books. A minor subset of the archive was published online in 2006 but has never been fully published in any analogic or digital form. Unfortunately, the technological platform has become obsolete since then and now must be discontinued. The *RePIM in LOD* project identified the publication as a Linked Open Data (LOD) dataset to preserve it as a free and unlimited resource for the community of present and future scholars. Contents have been migrated in Resource Description Format (RDF) adopting FRBR Object Oriented (now Library Reference Model) and CIDOC Conceptual Reference Model as reference ontologies. The result is a Knowledge Base (KB) relying on a set of open and well-established formal ontologies and then fully interoperable within the Linked Data ecosystem. Finally, to develop a new web-based application, the ResearchSpace platform ([19]) has been selected as an open-source, cultural heritage-oriented, and scholar-friendly tool which also aims to address digital knowledge preservation and maintainability over time.

---

1   While both authors contributed to the research being illustrated here, P. Bonora is responsible for the Introduction and section 2 and 3; A. Pompilio is responsible for section 1; both authors contributed to section 4.

1. The challenge of telling the tradition of madrigal texts

The production of secular music on Italian language lyrics occupies approximately half of the entire musical production between the early 16th century and the first three decades of the following century. It is a vast repertoire with more than 53,000 compositions, preserved in about 3,500 printed books[2].

In a prince's court hall, an aristocratic family, or an academy, ladies, nobles, officials, poets, and musicians would linger to listen to four, five, or six singers who gathered around a table, and sang lyrical poems of a few lines, usually between four and fourteen: madrigals, canzonettas, stanzas of canzoni, octaves, sonnets. The prince, the court entourage, the nobles, and academics gathering in these private rooms are the promoters of this musical practice. The composers obtain the texts directly from the hands of the poets or through intermediaries who frequent these cenacles, from the many books of poems published in those years, or from other musical compositions of the same genre that appeared in other music books, they intone them and entrust their performance to *cantori* and *dame cantatrici*[3]. In the musical intonation, the composer arranges the poetic text very freely: iterating one or more words, inverting their position, taking up and re-proposing entire portions of the text or modifying the dictation by replacing individual words or larger portions, sometimes deleting certain verses or adding new ones. The poetic text printed under the notes may therefore be identical to the original dictation but may also differ significantly.

After composing a certain number of them, roughly 20 to 30, the composer assembles these in a collection homogeneous in terms of genre (madrigals or canzonettas) and the number of voices, to be published. Identifying the poetic texts sung and comparing the versions of the texts found in literary and musical sources can therefore provide valuable information on the circulation of this poetic-musical production throughout the Italian peninsula (but also beyond the Alps) and the textual tradition of the individual texts.

In printed music editions, however, the name of the poet is usually not mentioned. The RePIM project was conceived to resolve, even partially, this information gap and, first and foremost, was aimed to identify the literary authorship of as many Italian-language poetic texts set to music between 1500 and 1700 as possible, becoming an indispensable tool for research in this field of study widely used by scholars in recent years.

A bibliography of printed books of poetry was compiled, also extending to some manuscripts and modern editions from the 18th century to the present, while for the books of music,

---

2   For a general introduction to madrigal see the corresponding entry of the Grove Music Online ([13]). Further readings may include [11], [12], [18], [22].

3   During 16th century, there was not a sharp distinction between professional and amateur singers. Most performers are at the same time singers and audience ([11]).

traditional bibliographic repertories were used[4] and the bibliographic records were verified on the originals. For each music book, the complete description of the contents was verified and the textual incipit, extended to the first two verses, was given for each piece. Based on the incipit of the texts set to music, the poetry books bibliography was consulted to identify the authors of the poetic texts. Data collected within the RePIM and their organization is a substantial innovation to the traditional documentation approach being represented till now, for this repertoire, by the *Nuovo Vogel* back in 1977 ([23]). More recently, some digital repertories have been published: the Répertoire International des Sources Musicales (RISM), a rich collection of musical sources including madrigals without an analytical description of the contained works; the Catalogue de la Chanson Française à la Renaissance a catalogue dedicated to a genre of secular vocal music, in the French language; the Tasso in Musica Project a collection of Torquato Tasso's works set in music; the Printed Sacred Music Database a catalogue, which is close to RePIM in terms of details, but dedicated to sacred music.

RePIM's data allow some preliminary analysis of the lyric poetry of the 16th and early 17th century and its coeval musical intonations and then to grasp some phenomena regarding the circulation of texts in literary and musical sources. The very first consideration is quantitative. Of the approximately 53,000 texts in music, 20,700 have been identified, a little less than 40%[5]; the remaining 32,000, which have not been attributed, thus testify to a numerically substantial portion of lyric poems that, although not published in poetry books, circulated in manuscripts, now dispersed, providing the poetic sources used by musicians for their compositions. This impressive production of lyric poetry is now only attested by music books.

The occurrence of the same poems in musical and literary sources then offers valuable information on the circulation of the texts, their genesis, stylistic features, and lexical peculiarities which are functional to musical intonation. A text in music contains on average 50-60 words, distributed in slightly less or slightly more than 10 lines, usually septenaries and hendecasyllables. When choosing texts to be sung, musicians not only draw on poems with these morphological characteristics but also extend their interest to much larger texts, such as strophic compositions (*canzoni* and sestinas) and epic and pastoral poems (such as *L'Orlando furioso, La Gerusalemme liberata, L'Aminta, Il pastor fido*).

---

4   Emil Vogel, *Bibliothek der gedruckten weltlichen Vocalmusik Italiens aus den Jahren 1500-1700, Hildesheim*, G. Olm, 1962; Emil Vogel - Alfred Einstein - François Lesure - Claudio Sartori, *Bibliografia della musica italiana vocale profana pubblicata dal 1500 al 1700*, Pomezia, Staderini-Minkoff, 1977 (*Il nuovo Vogel*); Knud Jeppesen, *La Frottola*, Aarhus, Universitetsforlaget; København, Einar Munksgaard, 1968-1970; Iain Fenlon - James Haar, *The Italian Madrigal in the Early Sixteenth Century: Sources and Interpretation*, Cambridge, Cambridge University Press 1988; RISM. Répertoire International des Sources Musicales (https://opac.rism.info/).

5   More precise and in-depth research will undoubtedly allow identifying other texts in literary sources, but it is highly unlikely to deviate significantly from the achieved result.

From these poems, musicians freely extract portions of the text, resulting in an autonomous poem, with coherent content and close to the standard length of 50-60 words, in 6-10 verses. For strophic texts, in some cases, the musician intones single stanzas, or portions of stanzas, in others the composer articulates the musical composition in parts, assigning to each part the text of a stanza or a portion thereof.

The sonnet, a kind of text widely used in musical intonations, consists of about 120 words distributed in 14 hendecasyllabic verses, is sometimes sung in full, more often articulated in two parts, corresponding respectively to the two quatrains and the two triplets, about 70 words the first part, 50 the second ([2]).

From the epic poems composed in *ottava rima*, musicians extract single octaves or a succession of octaves, musically articulated in as many parts as the selected octaves.

In the case of the *Pastor fido*, a pastoral drama in five acts of septenary and hendecasyllabic verses, each musician extracts only those verses he considers functional to the musical intonation and employs them in single or multi-part compositions, as in the case of other texts of larger dimension. In extracting and recomposing verses into a coherent literary text, we can see interventions of poetic craftsmanship involving the substitution of certain words, the repositioning of others, and the addition and/or deletion of one or more verses. The RePIM's collection of incipits allows us to trace this complex situation of textual tradition analytically and considers all the variants and textual articulations found in literary and musical sources.

Typically, poetry texts set to music were identified in poetry books that predate the corresponding music book. This finding suggests that composers usually drew the texts from the poetry editions already available on the market. The same poetic text, however, may have received more than one intonation, from two up to 63 in *Ardo sì ma non t'amo* by Battista Guarini. In such cases, the source the musician relied on may have been either a literary edition or a musical edition. In the absence of literary sources, we can speculate that the source of the text may have been either one of the earlier musical editions or, in the absence of such sources, a manuscript sheet on which the text of the poem was written. The 32,000 works in the RePIM repertoire with no reference to literary sources testifying the existence of a rich production of poems that were perhaps, and at least partially, specifically written for musical intonation. Then this poetic-musical practice was a strong stimulus for poets, who were urged to write texts with morphological, lexical, and stylistic characteristics that were functional to musical intonation.

However, to carry out in-depth investigations into the repertoire of poetry in music from 1500 to 1700 and to study the close relationship between the texts in music and poetry books, the complete texts of both kinds is required for an effective critical comparison. The RePIM project, originally conceived as an incipit index of music texts, has then broadened its information base over the years. It has been enriched by a wider survey of literary sources, not only in print but also in manuscript, and has recently started transcribing complete texts from primary sources or modern editions of music and poetry books: to date, more than 20,000 complete texts from literary and musical sources are available. But this set, although amounting to approximately

one-third of the entire repertoire, is still not enough to effectively carry out the research activities envisaged.

Accessing literary texts of Italian lyric poetry and editions of secular music from the 16th and 17th centuries already available on the web today could help document in a more complete, richer, and more effective way the cultural heritage witnessed by this musical and poetic practice. An example could illustrate the potential of an extended information system. In 1609, Michelangelo Nantermi, organist at the basilica of San Lorenzo Maggiore in Milan, published his *Primo libro di madrigali a 5 voci* in Venice by Ricciardo Amadino.

A four-part madrigal appears in the collection, on a canzonetta text, four stanzas of five verses each. The incipit of the first stanza is *Son io, donne, son io che rimirate*. The poet is not mentioned nor is any information about the given poetic text. From a simple full-text search on Google Books, it was possible to find the text of the canzonetta in the volume *Feste di Milano*, an account of the festivities held in Milan to celebrate the birth of Prince Filippo Domenico Vittorio, compiled by Cesare Parona and published in Milan in 1607, with a dedication to Filippo III. Within the description of the third day of festivities, the parade of the chariot of Venus is described with all the decorations that embellish it, mention is made of the knights parading alongside it and their attire, the sound of trumpets and the dancing that accompanies the parade and precedes the sudden and marvelous appearance of Cupid who, from the top of the theatrical machine, intones the four stanzas of *Son io, donne, son io che rimirate*. The given full text (p. 98) is identical to that sung in Nantermi's four-voice madrigal, but no reference to the poet or author of the music appears. The musical source indicates the author of the music and gives us the musical content of the piece, the textual source gives us precise information about the spectacular context within which that piece of music was performed. Having identified the link between the sources, the partial information from each source contributes to reconstructing the spectacular event from a richer historical perspective and helps us in interpreting this musical composition.

2. Shaping the knowledge: the evolution of the digital container

Just as liquids tend to take the form of a container, knowledge is not immune to the shape of the tool employed to collect, manage, and preserve it. Since the early 90s, the index of incipits and references to sources have been managed only in the digital domain and a printed version has never been published. The increase in contents and the inclusion of new entities were accompanied by the evolution of the information system. Its conceptual model, initially designed to manage the index of incipits, evolved to manage complete bibliographic records and detailed descriptions of the contents of musical and poetic sources[6]. Following the path through different

---

6   Data models and their implementation are evaluated in terms of their effectiveness in serving user requirements. As they evolve, the model must evolve too. When they change in nature, as when the focus moves from the expressiveness of a "research-driven" approach to standardization and

conceptual approaches and their technical implementation is mandatory to fully understand the semantics of the conceptual model shaping data currently available in the repository.

2.2 The origins: the first proprietary model

The first model was designed to implement the basic requirement of taking track of the collation between poetic and musical sources. The resulting model was split into two almost specular parts, each describing a kind of source, that was connected through the relation identified by the index of incipits and the authority file. The description of the poetic source was implemented through a set of bibliographic data, a description of the frontispiece with transcription of names and dedication when available. Each source had a set of references and corresponding collocations. The first two verses of contained poems were extracted to build the poetic incipit index. Incipits with a significant similarity were clustered and assigned a common group identifier, within each group, works were ordered by publication year. This led to the very initial ordering criteria of the index.

Musical sources were modeled similarly: just some metadata required to describe the musical form of the madrigal and its characteristics were added. Both source types referred to a harmonized authority file with a very basic set of metadata required to disambiguate names when required (and possible).

The following diagram (Diagram 1) shows the resulting model, expressed in UML[7]. In yellow, are classes representing core domain entities, and in grey are the corresponding physical table in the relational repository.

---

interoperability, then data shape and detail are affected. A discussion focused on model design and evaluation in the DHs is available in [14].

7   The UML's class diagram notation identifies a bi-directional association relation between classes with a straight line, a generalization relation with an arrowed line, and an implementation relation with the dotted arrowed one.
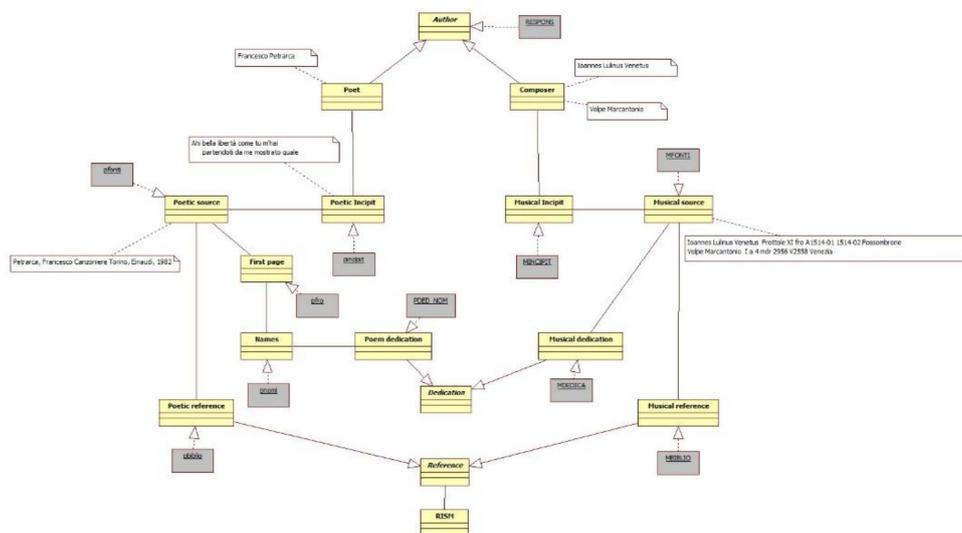
Diagram 1: the RePIM original conceptual model

The alignment between the index of poetic works and the musical ones was performed by comparing the corresponding incipits, their year of publication starting from the most ancient, which is marked as "the archetype", and then proceeding onward through the tradition of the text. While the model effectively supported this basic task, its main flaw was to set a hard separation between the two sides: music and poetry. At the same time, it did not support a complete description of the structure of a work. That is required when just a part or a section of the poetic work is set to music. In this case, only those verses should be added to the index to be matched to the corresponding musical work.

These limitations led to a major redesign of the model in 2005. The main aim was to harmonize the description of works maintaining the separation between sources (carriers) and works (contents) by merging the two sides of the model. At the same time, the description of works should allow a detailed representation of their attributes and structure (i.e., segmentation in stanzas or subsets of verses). Also, the index of incipits should be unified to solve the distinction between music and poetry incipits. A hard-wired partition of incipits coming from poetry or musical works is fictitious as, at this level, verses are just textual matter that may belong to both. For instance, an incipit may be related to the second stanza of the poetic work and be the whole text set to music in the first part of a madrigal. Then, what is important is to have an analytical description of the structure of the work to be able to relate the incipit to the corresponding section of each work, be it poetry or musical. This relation represents the proper context of the origin of the two verses. The model's structural partition implied that any full-text search among different sources must be done over several different tables. This means, for instance, that searching for a book title, the same query must be run on both music and poetry book tables to

obtain consistent results. The same complexity applies to the search for dedicatory information, such as patronage, dedicatory and preliminary texts, which had to be performed separately for poetry and musical ones.

Moreover, the model design led to less-than-ideal support for general querying such as traditional bibliographic searches. To widen the functional range of the user application, the data structure should adopt a more harmonized design. These considerations suggested the adoption of the FRBR model to organize the logical layers: sources being the physical ones, their contents represented as Expressions, while incipits became keys to corresponding complex, individual, or partial Works.

2.3 Towards WEMI: the adoption of the FRBR-ER model

Since the early 2000's the FRBR guidelines and the Group 1 entity set WEMI: Work, Expression, Manifestation, and Item set the reference conceptual articulation of levels required for an accurate, properly structured, and reusable description of bibliographic sources and their contents ([15]). The distinction between Work and Manifestation levels is instrumental in enabling bibliographic, musicological, and literary analytical search. The aim is to allow users to explore literary and musical production from a single perspective which was not fully supported by the previous model. While the distinction between the physical carrier and contents was already present in the original design, the distinction between Expression, the text contained within a specific edition, and the corresponding Work, which may be edited in different sources with some (even minor) variations, was not.

Anyway, we must consider that the object of study of RePIM is the poetic text of the madrigal. This means that the musical counterpart is only described to represent its use in musical sources. That reduces the complexity of representing the nature of music works, and the compromises required when describing music performances with the FRBR model [17]. We need to model the basic distinction between work types, the separation between the work as an "abstract distinct creation" ([15]) and its expressions,[8] and finally to set the references to their sources.

Anyway, this required introducing a separation of the two levels obtained from the "synthetic" partition of the original set of metadata. Those describing the work, such as the title, the lyrical or musical attributes, etc. were related to the Work. Those related to the specific edition, i.e., the page number, the title within the source, and the diplomatic transcription of the title page, became attributes of the Expression. Works, Expressions, and Manifestations are then related to the authority file. Each relation is annotated with specific attributes stating for example the role of the author, the name as stated in the source, etc. Musical and poetry books, corresponding to

---

8   Taking into account just the "textual" portion of the madrigal let us describe the musical work's expression considering only those musical features related to the intonation, being the other metadata describing the poetical ones. This simplified many of the issues in musical performance representation as FRBR expression ([16]).

the Items in the WEMI hierarchy, are related to the corresponding Manifestation representing the specific edition. The relation between Works and index of incipit, a concept that is not present within the WEMI model, is mediated by the entity "Title". The scope of this "association class" is to carry the text that has been adopted to "name" each Work and thus becomes its identifier. The new implementation of the index then merges verses coming from both musical and poetry works. The resulting index is then decoupled from the description of the Work and serves as an abstract entry point for both types.

The resulting model is depicted in the following diagram (Diagram 2).[9]
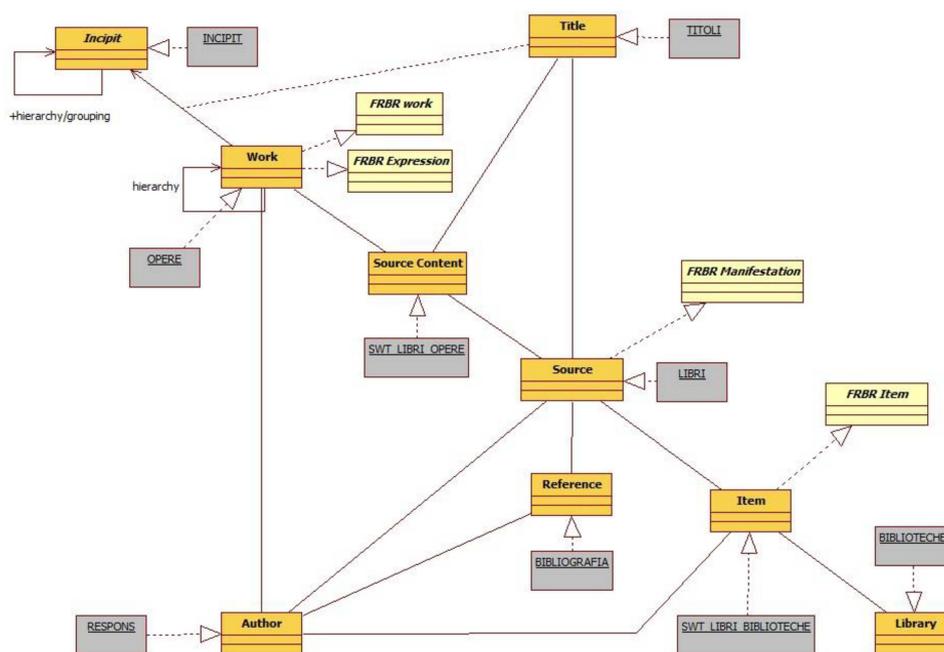
Diagram 2: The FRBR-compliant conceptual model

Obtaining a coherent description of the Work and Expression level starting from the available set of information has not been a straightforward procedure. While the process adopted to extract and make the two entities discrete presented some critical issues, it allowed information that had hitherto been implicit to become explicit and formalized. Discretizing the two levels allowed each entity to be described in more detail through the adoption of specific attributes. For instance, the inclusion of Expression analytically documents the form that each text takes in the

---

9   Core entities are depicted in orange, corresponding physical tables in the RDBMS are in grey, and conceptual FRBR classes are in yellow.

different sources testifying the same Work and records in detail all its accompanying historical information.

The FRBR-compliant RePIM model supports both the incremental collation of works and effective user-oriented navigation. Users can search books by their contents, and their editorial metadata and navigate relations deriving from collations which is a significant advancement from the functional point of view. At the same time, the original partition of musical and poetry sources and works can still be obtained through logical views implemented within the relational database. The adoption of FRBR affected also the cataloguing process. While metadata maintenance is improved, as the descriptive set of data is harmonized for each entity type, the drawback is the effort required to add new entries as any new element must be described at each of the four levels of the WEMI model to be consistently represented.

While the new design complies with the FRBR general model, its physical implementation was specialized to implement domain requirements and had to deal with some constraints originating from the data migration from the previous model. The hierarchical nature of the FRBR model and data normalization, which has been done by adopting domain-related conventions, led to a multi-layered data structure. The result is a relational model consisting of several tables. This required a complete reengineering of the user application. The role of the application is not only to enter and query data but also to produce effective representations of information to the users. This was achieved with the design of logical views as an abstraction layer mediating between the data structure and the user's perspective. From a conceptual point of view, these views provide a domain coherent interpretation of the underlying physical structure of data. In other words, they provide a functional interpretation of the semantics of the data logical structure. Unfortunately, this interpretation, due to the nature of the system architecture that delegates this role to the business logic of the application, is parted from the data itself. That means that to make raw data fully readable by third parties (being a human reader or an external system), even a basic description of the logical model and related views must be provided along. Even the new FRBR-compliant model, being less proprietary and though being more functionally oriented than the original one, would still lack data interoperability without proper documentation. Moreover, the technological platform used for the implementation of the web application in 2006 went out of maintenance in 2020. This stressed the necessity of a third technical evolution of the archive and to address the problem of further maintenance to preserve and keep the knowledge available for the community of users. The idea of publishing the dataset as Open Data along with reengineering the application seemed the most effective solution.

3. Into the LOD ecosystem to disseminate and preserve

Publishing research data as Open Data following current leading guidelines such as FAIR principles requires a significant level of accessibility and interoperability to achieve the aim of reuse.

To make data accessible, domain conventions in information representations and praxis in using and accessing them should be explicated, together with an introduction to decisions taken during the design of the conceptual model. The semantics of the core entities and their attributes, dictionaries, and related conventions and guidelines to arrange a presentation of data must be also provided. The first two would give a guide for interpreting data correctly, while the latter would give a valid, still not the only, functional view for the user. Documenting these three layers could be challenging: both conceptual and technical details should be published. Doing this in natural language would give a reference guide to the human reader but it would not enable direct machine interoperability and, again, it would be parted from the dataset.

In 2021, following previous similar experiences ([5]), we decided to adopt semantic technologies to publish both the conceptual model and the knowledge of the RePIM.

The aim is threefold: to publish the RDF dataset as a LOD resource via an Open Science platform like Zenodo,[10] to use the same RDF dataset as the source for a new web application, and, in the future, to support also the cataloguing process. The first tries to foster the resilience of the KB towards the technological obsolescence that affected the archive during the last forty years. We also expect that this should increase the resilience leveraging dissemination of knowledge among different projects and players. Similarly, the second responds to users' requests to keep the access point to the KB online, while the third is looking forward to a complete maturity of semantic systems to replace relational databases as repositories[11]. Unfortunately, the latter requires not only technological advancements but also the development of a cataloguing LOD-based ecosystem in which different actors (i.e., libraries, archives, scholars, etc.) publish their digital resources and establish coordinated procedures for their dissemination, incrementation, and maintenance.

3.1 Going explicit

The use of semantic technologies to transpose data contained in relational databases has been an established practice for more than a decade ([6]). At the same time, a wide range of formal ontologies specifically designed for the cultural heritage sector is now available ([3]). Their very application has been evaluated in several specific fields, from the documentation of ancient books

---

10 DOI: https://doi.org/10.5281/zenodo.6635606

11 An example is the initiative LD4L which aims to create a Scholarly Resource Semantic Information Store and promote its adoption as the main cataloguing system within the participating institutions. See https://wiki.lyrasis.org/pages/viewpage.action?pageId=187176106 (last accessed 27/09/2022).

to the description of performing arts and music history ([9]). We, therefore, had no doubts about the maturity of the tools available to transpose RePIM content into RDF format ([4]). It was simply a matter of identifying the reference ontologies to be used as the foundations of the explicit semantics conceptual model. Starting from the analysis of the coverage obtainable through reference (and possibly already widespread and established, if not even standard) ontologies, the introduction of new classes or properties as extensions to express specific domain concepts would then be evaluated. This approach is aimed at maximizing the interoperability of the dataset while keeping the most accurate definition for domain-related concepts. We expect that the publication of RePIM as a LOD resource through Open Science platforms would lower the barriers to data access, facilitating both the reuse and extension of the KB by a wider community.

Since the adoption of FRBR as a reference for the second data model, the use of an FRBR-compliant ontology was almost natural, then we selected FRBRoo[12] (waiting for LRM to be published) and CIDOC CRM[13] as reference ontologies. These cover almost all the entities within the RePIM domain, starting from the WEMI hierarchy to the authority files of actors (authors and institutions). Only peculiar domain-related relations or concepts and some required to detail the structure of a document we found missing. We also considered the adoption of ontologies dedicated to the musical domain such as Music Ontology ([21]) and DOREMUS ([1]). The first one is focused on the music production process, actors, and products and is built on top of the FRBR model. The Musical Work and Musical Manifestation classes represent the specialization of FRBR's Work and Manifestation entities. The second is designed as an extension of the CIDOC CRM and FRBR to describe semantically catalogs of music works and events. In this second case, specific classes such as Performance Work and Recording Expressions are introduced to better describe music catalog entries. As the RePIM does not include information about musical scores being focused on the poetical production, in both cases, we found that definitions already available in the two reference ontologies were satisfactory and no gain in expressiveness would derive from adopting more specialized ones. Besides, the two specialized ontologies, do not provide application-related concepts such as the "incipit" or document articulation and structure such as those required to describe dedicatory texts: both missing also in the reference ones. Then we decided to adopt the most general models to design ours which, in any case, would result being natively aligned with specialized models sharing the same foundational ontologies.

The RePIM Semantic Model (Repim SM)[14] results from the composition of the set of reference ontologies and a limited set of classes and properties defined as specializations.

---

12 For a brief introduction refer to: https://www.cidoc-crm.org/frbroo/short-intro-frbroo

13 The version adopted is the CIDOC CRM version 6.2.1 in RDFS: https://cidoc-crm.org/Version/version-6.2.1

14 The RePIM SM is available through GitHub: https://github.com/paolobonora/RePIM-LOD

The transposition process assigned a target class from the RePIM SM for each of the core entities in the original ER model. For instance, works instances extracted from the table "OPERE" were assigned to class frbroo:F15_Complex_Work for musical and to class frbroo:F14_Individual_Work for poetry ones. Corresponding Expressions were instantiated from the table SWT_LIBRI_OPERE (see Diag.2) and defined as frbroo:F22_Self-Contained_Expression. Manifestations, instances of frbroo: F3_Manifestation_Product_Type, come from table "LIBRI". Corresponding Items, marked as frbroo:F5_Items, derive from table "SWT_LIBRI_BIBLIOTECHE". Similarly, describing attributes extracted from each table field were assigned a semantically appropriate property from the semantic model. CIDOC CRM helped to transpose entities from the authority file and other references: authors from "RESPONS" table are generically asserted as crm:E39_Actors but also accompanied by their biographical data when available. The more specific class frbroo:F11_Corporate_Body has been instead selected for Libraries (extracted from the table "BIBLIOTECHE"). Even these very basic assignments illustrate the advantage deriving from the adoption of an explicit semantics model. Anyone (a human or a machine) who can interpret the formal definition of the model can understand, and then use, RePIM's data expressed in RDF while even the naming conventions of the original relational model lack readability.

The RePIM collects information from several sources from which a task-oriented subset of information is extracted. Concepts such as the Incipit as intended by design are very specific and do have not a direct equivalent within the two reference ontologies. Then we decided to define the repim:Incipit class as a subclass of crm:E33_Linguistic_Object and crm:E42_Identifier to assert its twofold nature of text used to identify works. Some other classes have been introduced as a specialization of the crm:E55_Type class to represent instances of the musical or poetic form of a madrigal vocabulary (i.e., repim: RC4_Musical_Form and repim: RC5_Poetic_Form) as long as the genre (repim:RC7_Genre). Some ancillary object properties are dedicated to tracing relations between linguistic objects such as the repim:RP1_group_member_of which allows incipits clustering and the repim:RP3_has_literary_source indicating links between musical and poetic ones. As a rule, specialization has been adopted to always give the capability to read data through the corresponding hypernymy hierarchy when a more "general" interpretation is required. This also allows third parties to reuse already available alignments with CIDOC CRM to read assertions based on the RePIM SM which, again, we expect to introduce more flexibility fostering interoperability and then reuse of data.

The resulting conceptual model consists of 29 classes (12 from CIDOC CRM, 10 from FRBRoo, and 3 from both DOCO and Repim SM) and 34 properties (16 from CIDOC CRM, 5 from FRBRoo, 8 from Repim SM, and others from DoCO[15], BiRO[16], DCTerms and RDF). Data were extracted from the legacy database management system with a Relational-to-RDF mapping tool following the semantic alignment between the FRBR-ER inspired proprietary

---

15 DoCO: Document Components Ontology [7], http://www.sparontologies.net/ontologies/doco.

16 BiRO: Bibliographic Reference Ontology [8], http://www.sparontologies.net/ontologies/biro.

relational model and the Repim SM. The resulting knowledge graph contains 745.258 class instances and 2.477.159 property assertions (more than 3.2 million triples in total).

Diagram 3: Part of the graph representing the work "Al verde e vago April de' bei vostri anni" in "Rime" by Paolo Gottardo Pontio, 1572

The transposition in RDF is the outcome of a process of evolution of the descriptive model that has gone through three distinct phases in more than two decades. From a proprietary model focused on collecting the information required to carry out the collation of madrigals with their poetic source, we moved on to a model that is still proprietary but designed to respect the functional approach of FRBR. Therefore, from a purpose-oriented data container, we moved to an organization of information that responds to more general requirements while remaining domain-specific. The introduction of explicit semantics, by significantly lowering the cost of access to raw data should further extend the possibility of their reuse.

Both the model and application design aimed at the development of a research tool for scholars. They try to answer very basic research questions expressed by the community of users such as the simple search for a text in the sources; the relationship between poetic and musical works; which texts and authors are chosen by composers for their works; what the poetic forms or the places of editions are. At this stage, we have not performed a more comprehensive evaluation of the conceptual model rather than verifying the effectiveness of the information representation obtained through the application by team members. A wider evaluation will be done with a panel of domain experts. Besides, only time and the statistics of the downloads of the dataset will tell whether the expected dissemination of the dataset will occur.

3.2 Search and collate: the end-user application

To let the wider community of users that may not be interested in raw data reuse, we developed a new web-based application that uses the RDF dataset as the data source. The ResearchSpace platform has been selected to implement basic functionalities plus a set of requirements expressed over the years by the community but never fully implemented. The design of the new application followed a functional approach trying to achieve an effective representation of knowledge. This

meant reducing the complexity of the conceptual model with views that increase the readability of RDF data. We designed dedicated views for the three core entities: the work, poetic and musical, and the book. Each view gives the user a synthetic but complete representation of attributes and relations of the resource. One of the main objectives is to let the user navigate the network of relations without getting lost. At the same time, we want to hide del complexity of the analytical representation of the graph resulting from the adoption of fine-grained ontologies such as the CIDOC CRM.

We introduced an experimental tool to help the user to identify new relations between works (depicted in Figure 1).

Figure 1: The comparison tool of the new RDF-based application

Starting from a specific intonation of an incipit, the user opens a web page presenting the work's record on the left side. Then one of the "siblings" works belonging to the same incipit group can be selected to be shown on the right half of the page giving the possibility to compare the two. Then, navigating the underlying graph of relations between incipits, the user can explore it in

search of new possible collations between works. The tool leverages two features of the semantic-based conceptual model: the capability to generalize the nature of entities and the possibility to navigate the graph of different types of relations seamlessly. The first lets the application treat musical and poetic works as generic while searching and differentiating during analytical representation is required. The second allows the exploration of the relations network following user-defined criteria rather than model boundaries. The composition of the two enables a more functional-oriented approach to application development. At the same time, the new application leverages some ResearchSpace's features such as full-text search; Part of Speech and lemma-based searching capabilities; graph-based browsing of relations between entities (works, authors, sources, etc.); Named Entity Recognition of names, toponyms, etc.; faceted browsing of topics. Yet in an early phase of development[17], the application demonstrates advantages deriving from the adoption of a semantic conceptual model to represent RePIM data and RDF as a data source. The ResearchSpace's templating mechanism allows for the arrangement of an effective resource representation as user's views, although it does not support their formalization within the model as the introduction of a dedicated semantic layer would allow ([5]). Besides the framework supports a basic mechanism for expert-led refinements and expansions of the RePIM knowledge graph. This will be leveraged in the future to introduce editing and annotation features into the web application and would lead toward distributed cataloguing and shared responsibilities among a wider community of scholars. At the same time, the LOD ecosystem would provide the technological infrastructure to include external sources to fill gaps and reach other documentation, such as Parona's *Feste di Milano*, carrying additional information about the performative context.

4. Further developments

The current implementation of the application aims to support basic search and browsing functionalities of the current knowledge expressed within RePIM's KB. The next phase should aim to let users annotate further findings achieved through the exploitation of already available sources or newly added ones by the community. The ResearchSpace platform supports the implementation of data editing procedures as well as the creation of semantic narratives and visual representations of resources from the knowledge graph. This would lead to a collaborative approach to knowledge extension and maintenance. This will require fine-grained information provenance tracing and sharp profiling of contributors. Besides, the outcome would be a shared effort of knowledge preservation and extension among the wider community of users through the years to come. A further step is to look at the LOD ecosystem as a source to widen the perimeter of the current KB to better fulfill its scope by integrating a wider set of documentation.

17　The application is available at: https://repim.itatti.harvard.edu/login

For now, we believe that publishing the RePIM archive as LOD will preserve it and offers, both music scholars and those interested in Renaissance and Baroque poetry, a useful tool and a valuable source of information, resulting from over 40 years of research, which otherwise would be lost in the digital mist.

Acknowledgements

References

[1] Achichi, Manel, Pasquale Lisena, Konstantin Todorov, Raphaël Troncy, Jean Delahousse. 2018. 'DOREMUS: A Graph of Linked Musical Works'. The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II, edited by Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, Elena Simperl, 11137:3–19. Lecture Notes in Computer Science. Springer.

[2] Bianconi, Lorenzo. 1986. Parole e musica: il Cinquecento e il Seicento. in *Letteratura italiana*, v. 6: *Teatro, musica, tradizione dei classici*, 319–63. Torino: Einaudi.

[3] Bikakis, Antonis et al. 2021. "Editorial: Special Issue on Semantic Web for Cultural Heritage". 1 Jan. 2021. 163 – 167.

[4] Bizer, Christian and Andy Seaborne. 2004. "D2RQ-treating non-RDF databases as virtual RDF graphs". In *Proceedings of the 3rd international semantic web conference (ISWC2004)*. Vol. 2004.

[5] Bonora, Paolo, Angelo Pompilio. 2021. "Corago in LOD. The debut of an Opera repository into the Linked Data arena". *JLIS* 12 (2): 54–72. https://doi.org/10.4403/jlis.it-12699

[6] Bonora, Paolo, Charlotte Ossicini, e Giuseppe Raffa. 2006. "From Relational Metadata Standards to CRM Ontology: a Case Study in Performing Arts". In *CIDOC2006 Proceedings*. Gothenburg.

[7] Constantin, Alexandru, Silvio Peroni, Steve Pettifer, David Shotton, e Fabio Vitali. s.d. "The Document Components Ontology (DoCO)", 13.

[8] Di Iorio, Angelo, Andrea Giovanni Nuzzolese, Silvio Peroni, David Shotton, e Fabio Vitali. s.d. "Describing Bibliographic References in RDF", 12.

[9] Doerr, Martin. 2009. "Ontologies for Cultural Heritage". In *Handbook on Ontologies*, S. Staab, and R. Studer (eds). Berlin: Springer. https://doi.org/10.1007/978-3-540-92673-3_21

[10] Durante, Elio, e Anna Martellotti. 1989. *Cronistoria del Concerto delle Dame principalissime di Margherita Gonzaga d'Este*. Firenze: S.P.E.S.

[11] Einstein, Alfred. 1971. *The Italian Madrigal*, 3 v. Princeton University Press.

[12] Fenlon, Iain and James Haar. 1988. *The Italian Madrigal in the Early 16th Century: Sources and Interpretation*. Cambridge: Cambridge Univ. Press.

[13] Fischer, Kurt von, Gianluca D'Agostino, James Haar, Anthony Newcomb, Massimo Ossi, Nigel Fortune, Joseph Kerman, and Jerome Roche. 2001. "Madrigal." Grove Music Online. Oxford University Press.

[14] Flanders, Julia, and Jannidis Fotis. 2015. *Knowledge organization and data modeling in the humanities*. https://d-nb.info/1102830453/34

[15] IFLA Study Group. 1998. *Functional Requirements for Bibliographic Records: Final report*. München: K. G. Saur.

[16] Iseminger, Damian. 2012. "Works and expressions in RDA: problems and solutions". In Directions in music cataloging, ed. by P.H. Lisius and R. Griscom. Middleton: A-R Editions.

[17] Miller, David and Patrick Le Boeuf. 2005. "Such Stuff as Dreams Are Made On: How Does FRBR Fit Performing Arts?". Cataloging & Classification Quarterly. 39:3-4, 151-178.

[18] Newcomb, Anthony. 1980. *The madrigal at Ferrara: 1579-1597*. Princeton University Press.

[19] Oldman, Dominic and Diana Tanase. 2018. "Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace". In *The Semantic Web – ISWC 2018*, edited by Denny Vrandečić, et al., 325–40. Cham: Springer International Publishing.

[20] Pompilio, Angelo. 2006. "Il Repertorio della Poesia Italiana in Musica, 1500-1700 (RePIM): un aggiornamento". In *Petrarca in musica*, a cura di Andrea Chegai e Cecilia Luzzi, 391-396. Lucca: LIM.

[21] Raimond, Yves, Samer A. Abdallah, Mark B. Sandler, Frederick Giasson. 2007. "The Music Ontology". In Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007, επιμέλεια Simon Dixon, David Bainbridge, Rainer Typke, 417–22. Austrian Computer Society.

[22] Roche, Jerome. 1990. *The Madrigal*. Oxford University Press.

[23] Vogel, Emil, Alfred Einstein, François Lesure, Claudio Sartori. 1977. *Bibliografia della musica italiana vocale profana pubblicata dal 1500 al 1700*. Pomezia: Staderini-Minkoff (*Il nuovo Vogel*).