6.3 Survey on available datasets

To ensure that the ontology design is aligned with the music datasets publicly available on the Web, a series of activities were carried out to: (i) align with the data collections used in Polifonia's pilots; (ii) overview the current datasets in the field of Music Information Retrieval (MIR), considering the high-quality annotations provided by such sources; and (ii) prioritise the music collections contributed by the Polifonia consortium – to ensure that the in-house expert knowledge on these particular repositories is leveraged as a peculiar asset of the project. Each of these activities is described in the following subsections.

6.3.1 Datasets from Polifonia's pilots

The first goal was addressed through a detailed analysis of the dataset survey initially carried out in the WP1 data survey [1], with the aim of verifying whether the musical context set by the pilots was in line with the competency questions used for requirement collection. The outcome of this analysis raised several technical concerns regarding the plausibility of a subset of competency questions in light of the available data and the state of the art in computational music analysis. This allowed to refine some of the competency questions according to the analytical framework described in Section 2.3, thereby achieving a preliminary alignment among WP2 – as technology provider of the ontological ecosystem and for the transformation of data into knowledge graphs, and Polifonia's pilots. Furthermore, to further align with the latest work carried out in the pilots, WP2 will receive a specific data samples taken from the collections these pilots are actually using.

6.3.2 Datasets in Music Information Retrieval

During the last 20 years, the field of MIR has seen the introduction of an unprecedented number of music datasets, enabling researchers to train and evaluate algorithms for several tasks, from chord recognition and beat detection, to source separation and mood detection. Concerning the second line of activities, we conducted an in-depth literature review involving more than 200 datasets that have been extensively used to accommodate a wide variety of MIR tasks. The motivation behind this literature review twofold: first of all it aims at collecting diverse high-quality data and annotations related to the musical content, and secondly, it aims to understand the limitations of the datasets under analysis and how the Polifonia Ontology Network can help to address them.

From a metodological point of view, our survey is organised in such a way as to describe/catalogue music datasets based on their collection metadata – a list of fields that are usually expected to be found, either implicitly (from websites, files, manuscripts, additional material) or explicitly (stored in a single file) from an MIR collection. In this way, taxonomies can be created from our survey based on these fields, which are individually described as follows.

Music media type(s) If the dataset explicitly provides any musical content, this field is used to describe whether such content is either in audio or symbolic format. A collection can also

provide tracks of both formats, which is becoming a prominent trend in the literature [35].

Duration The approximate duration of each tracks (in seconds), in case the collection does not provide full-length audio recordings (the release of audio clips/excerpts is common for audio datasets, as the complete audio material may not be shareable due to copyright). For instance, the popular music with emotional annotations (PMEmo) dataset [61] provides audio excerpts of chorus sections (as this is copyrighted material), whereas full-length tracks are directly available in the Jamendo collection [62].

Audio format The audio format of the musical material, if full or partial recordings are provided, e.g. MP3, WAV, FLAC. More choices are plausible depending on the collection.

Symbolic format The symbolic format of the musical material, if a valid digital representation of musical scores, related to a notated composition or a transcribed performance. Common options include, but are not limited to MIDI, MusicXML, MEI, although the former is generally more popular considering the high availability of musical data in this format [63].

Other media In addition, or in alternative, to the musical material, datasets might also provide additional information and artifacts related to the musical objects. For example, audio features (e.g. Mel-frequency cepstral coefficients, chromagrams) are usually provided when recordings cannot be released [64], although collections can also release complementary information such as the rank of tracks on music charts (e.g. Billboard) [65].

Records The number of music pieces covered by the collections (compositions and/or performances), regardless of the availability of the corresponding musical material. This number can vary from small collections of 25 compositions [35] or 50 performances [66], to large scale datasets providing more than a million tracks [67].

Genres Music datasets have a tendency to specialise on music belonging to a narrow selection of genres and styles, to make them more consistent with the kind of analysis enabled by the data and the annotations they provide (e.g. the adherence to musical form is generally more strict in classical music compared to contemporary music). Therefore, it is fundamental to contextualise music datasets to the specific genres their music pertain to.

Year of release The year when the dataset was first released, without considering when the actual data collection activities started. Revisions to the dataset are also recorded in order to keep track of the major editing activities, and also, to have an approximate measure of how actively maintained the collection is.

Collection metadata Whether the dataset provides metadata at the collection level, trivially including all these fields used for our survey. This is needed because collection metadata can also include additional information that is not covered by the survey, e.g., the name of the project investigator, the university that is in charge of storing the data, etc. Surprisingly, most MIR datasets do not provide collection metadata in a standard, unified and consistent manner, hence this process needs to be done manually for each source (as in our survey).

Content metadata Ideally, dataset should provide a specification of their content – a document containing a list of tuples, where each element provides information specific to a single track, e.g., $(title, artist, release, MusicBrainz\ identifier)$. This last information is funda-

mental to disambiguate among compositions and/or performances, especially if any linking operation allowing the interconnection of different collections (the main goal of INTERLINK – one of Polifonia's pilots) should be operated or considered. Notably, some datasets do not provide this document – where all such information is made explicit for each track, although content metadata may again be implicitly scattered across multiple sources, such as files, websites, and manuscripts [33].

Annotations A list of all the annotations – the actual core of an MIR dataset, as this information is key to enable training, evaluation and testing of computational models for music retrieval and analysis. Annotations are generally contributed by domain experts (musicologists, composition teachers, etc.) when it comes to the detection/attribution of musical features, such as music structure [68], key (tonic and scale) [69], chord progressions [70], at different levels of granularity (hierarchical, flat) and temporal resolutions (global, local). When the annotation task does not require musical expertise, as it involves more subjective and less theoretical interactions with the musical content, annotations are provided by listeners following specific guidelines and frameworks (the annotation methodology); examples include music-induced emotions [71], and listening habits [72].

MIR tasks According to the music annotations provided, a dataset can enable one or more MIR tasks. For instance, the availability of music emotion annotations [73] makes it possible to train and evaluate methods for music emotion recognition [74], whereas the a dataset providing annotations of cadences [75] can be used for pattern extraction and cadence detection [76]. If the source code of these computational methods is provided, their recognition performance/accuracy is often ranked and recorded with respect to each dataset they were tested on (e.g. PapersWithCode[3]). The connection between music datasets and algorithms for music analysis and retrieval is thus a peculiar aspect in MIR – where datasets are more appealing for their computational potential, rather than for the information itself.

Access Music collections can be fully or partially accessible to the public (open), requested for research following a formal procedure for the release (on-demand), or explicitly declared by the authors/curators as unavailable (closed). In the second case, fees, commissions, or cost of licences can be asked by the curators, as done for the RWC dataset [77].

Online Whether the resource can be accessed online or needs to be manually provisioned.

API Whether the database can be accessed through an application programming interface (API), allowing for specific user-defined calls to retrieve musical content of interest. Trivially, this implies that the dataset is online (see the previous field), and is accessible to the user. The automatic access of music datasets is a crucial problem in MIR, as this motivated the development of tools and libraries that can facilitate the process of data acquisition and pre-processing [78, 79].

License/copyright The type of licence and redistribution information, if explicitly provided by the dataset.

References Links to the official website of the dataset or to a web-page describing its content

[3] https://paperswithcode.com

Curator	Music media type	Other media	Records	Genre(s)	Accessibility	Online	Year	License	Link
The name of the institution and/or any reference to the person(s) curating the collection.	audio, symbolic, digitalised scores, etc. (more may applky)	visual, etc.	approximate number of records (tracks for ex.)	classical, pop, etc.	open, closed, on-demand	yes, no	when it was released	the copyright license of the dataset (eg. CC BY)	link to the dataset page, or paper
Albert Meroño	Symbolic	Text (embedded in MIDI)	~500K	Classical, videogames, pop	Open	Yes	2017	CC0 1.0	https://zenodo.org/record/579603#.YPAwPhMzbm1
Johan Oomen	audio							CC	https://musopen.org/music/performer/european-archive/
Johan Oomen	audio						from 2014	CC	https://soundcloud.com/beeldengeluid
Johan Oomen	audio							CC	https://www.europeana.eu/nl/collections/topic/62-music?page
Danny Diamond	Symbolic		~7K	Irish traditional	Currently undefined	Yes	2015	No licence spe	http://port.itma.ie/welcome
Danny Diamond	Symbolic		~18k	Irish traditional	open	Yes	2001 (continu	No licence spe	https://thesession.org/tunes/
Danny Diamond	Symbolic		1,224	Irish traditional	open	Yes	1964-1999 (pr	No licence spe	http://www.capeirish.com/webabc/
Peter van Kranenburg	Audio and Symbolic	Images, midi, lyrics	~20K	Dutch Song Culture	Open	Yes	2014-	CC	http://www.liederenbank.nl/mtc

Figure 6.1: Overview of music collections curated within the Polifonia consortium.

in as much detail as possible. This is also complemented with a link to any academic manuscript formally describing the data collection activities and the annotation process.

Besides the limited availability of audio data, the survey also revealed two central issues: (i) MIR datasets are commonly provided as independent and isolated collections, with little or no alignment at the metadata and annotation level; (ii) even when tracks/compositions are coupled with universal identifiers (e.g. MusicBrainz IDs, ISRC), there is no direct way to access and link heterogeneous music-related data from online databases, such as Wikipedia, Genius[4], and Songfacts[5]. The disconnect among music datasets jeopardises their potential integration, and hence their extension and the combination of annotations of different kinds. Simultaneously, the low level of linkage with other databases discourages multi-modal research in the field, where the availability of heterogeneous music-related data (text, images, locations, etc.) is an essential asset. Therefore, the most common method to link multiple music collections is to implement complex data collection and integration pipelines as done by [80].

In sum, MIR datasets are particularly appealing considering the scope of Polifonia, as these collections provide high-quality annotations of musical features – including melodic and rhythmic patterns, chord progressions, musical structures, cadence points, tonalities and so forth – contributed by musicologists and music experts. Although these collections are primarily used for training and testing methods for computational music analysis, their annotations are rarely reused outside the computational domain despite their intrinsic value. If the integration problem is addressed with an ontological ecosystem that can represent and describe MIR datasets within the same infrastructure, their high-quality annotations would be preserved and their connection to the corresponding musical objects would enable the exploration of rich and diverse music-related data, and the automatic discovery and extraction of knowledge. Finally, the work conducted in this survey will serve not only to catalogue the available datasets, but also to make different resources interoperable, which thanks to the Polifonia Ontology Network can be handled as a unique corpus.

[4] https://www.genius.com
[5] https://www.songfacts.com

6.3.3 Internally curated collections

To conclude, a last survey was conducted internally to identify the music collections[6] curated by partners within the Polifonia consortium. For this particular case, a subset of the fields detailed in the previous subsection (collection metadata) was preliminary selected, and all the contributors in the Polifonia consortium were invited to provide the collection metadata of any music dataset that had personally curated. Seven different music collections curated by four Polifonia partners resulted from the survey (c.f. Figure 6.1). Of these datasets, three provide audio recordings, four are based on symbolic music and another include both symbolic and audio tracks.

In addition to the collections mentioned before, it is also worth to remark that the Polifonia consortium can also count on NEUMA[7], thanks to the direct involvement of the Conservatoire national des arts et métiers (CNAM) and Irémus in the project. NUEMA is a large digital library providing rare corpora of music in MEI format, that can be easily accessed, browsed, and searched by users. The library also includes utilities for the annotation of musical scores, thereby realising an online platform that can be used to contribute new material to extend the ecosystem.

[6] In this context, the term “music collection” is intended in a more general sense, to denote music-related data.
[7] http://neuma.huma-num.fr/