Polifonia

Polifonia: a digital harmoniser for musical heritage knowledge, H2020

D7.1 First Data Management Plan

Deliverable information	
WP	WP 7 Project Coordination and Management
Document dissemination level	PU Public
Deliverable type	ORDP Open Research Data Pilot
Lead beneficiary	KNAW
Contributors	All partners
Document status	Final
Document version	V1.0
Date	30/06/2021
Authors	Femmy Admiraal (KNAW) Task lead Andrea Scharnhorst (KNAW) Task lead Enrico Daga (OU) Marilena Daquino (UNIBO) Elena Musumeci (MiC) Peter van Kranenburg (KNAW) Christophe Guillotel-Nothmann (CNRS-IReMus) Marco Gurrieri (CNRS-IReMus) Valentina Presutti (UNIBO) Marta Clementi (UNIBO) Albert Meroño Peñuela (KCL) Monica Turci (UNIBO) Eleonora Marzi (UNIBO) Antonio Puglisi (DP) Raphaël Fournier-S'niehotta (CNAM)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746

INTENTIONALLY BLANK PAGE

Project information

Project Start Tate: 1st of January 2021

Project Duration: 40 months

Project website: http://www.polifonia-project.eu

Project contacts

Project Coordinator

Valentina Presutti

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
Department of Language, Literature and Modern Cultures (LILEC)

E-mail: valentina.presutti@unibo.it

Project Manager

Marta Clementi

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
Research division

E-mail: marta.clementi3@unibo.it

POLIFONIA consortium

No.	Short name	Institution name	Country
1	UNIBO	ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA	Italy
2	OU	THE OPEN UNIVERSITY	United Kingdom
3	KCL	KING'S COLLEGE LONDON	United Kingdom
4	NUI GALWAY	NATIONAL UNIVERSITY OF IRELAND GALWAY	Ireland
5	MiC	MINISTERO DELLA CULTURA	Italy
6	CNRS	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS	France
	SORBONNE	SORBONNE UNIVERSITE (LinkedTP)	France
7	CNAM	CONSERVATOIRE NATIONAL DES ARTS ET METIERS	France
8	NISV	STICHTING NEDERLANDS INSTITUUT VOORBEELD EN GELUID	Netherlands
9	KNAW	KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN	Netherlands
10	DP	DIGITAL PATHS SRL	Italy

Project Summary

European musical heritage is a dynamic historical flow of experiences, leaving heterogeneous traces that are difficult to capture, connect, access, interpret, and valorise. Computing technologies have the potential to shed a light on this wealth of resources by extracting, materialising and linking new knowledge from heterogeneous sources, hence revealing facts and experiences from hidden voices of the past. Polifonia makes this happen by building novel ways of inspecting, representing, and interacting with digital content. Memory institutions, scholars, and citizens will be able to navigate, explore, and discover multiple perspectives and stories about European Musical Heritage.

Polifonia focuses on European Musical Heritage, intended as musical contents and artefacts - or music objects - (tunes, scores, melodies, notations, etc.) along with relevant knowledge about them such as: their links to tangible objects (theatres, conservatoires, churches, etc.), their cultural and historical contexts, opinions and stories told by people having diverse social and artistic roles (scholars, writers, students, intellectuals, musicians, politicians, journalists, etc), and facts expressed in different styles and disciplines (memoire, reportage, news, biographies, reviews), different languages (English, Italian, French, Spanish, and German), and across centuries.

The overall goal of the project is to realise an ecosystem of computational methods and tools supporting discovery, extraction, encoding, interlinking, classification, exploration of, and access to, musical heritage knowledge on the Web. An equally important objective is to demonstrate that these tools improve the state of the art of Social Science and Humanities (SSH) methodologies. Hence their development is guided by, and continuously intertwined with, experiments and validations performed in real-world settings, identified by musical heritage stakeholders (both belonging to the Consortium and external supporters) such as cultural institutes and collection owners, historians of music, anthropologists and ethnomusicologists, linguists, etc.

Executive summary

This first version of the Data Management Plan (DMP) follows in its core the template structure for an Open Research Data Pilot (ORDP) (section 3). The ORDP launched for all H2020 project has two functions: developing a DMP and providing open access to research data, this way to foster openness and re-use. Polifonia is a project operating between various humanities fields and content providers, relying on knowledge engineering and semantic web technologies as its primary mode of operation. From the specific features of Polifonia, specific requirements[1] for the Data Management follow. To address this specificity, we have extended the formal template of the DMP with two sections. Section 1, the introduction, summarises the specific approach of Polifonia in the creation of a knowledge graph which incorporates various aspects of musical heritage (introducing the Polifonia Pilots). Section 2 reflects how the DMP is used to coordinate the collaboration inside of Polifonia. As we show, the process towards the DMP complements other cross-Work package activities, such as working on the Roadmap towards the Polifonia Portal (WP1) and the activities of the Technical Board. The interdisciplinary character of Polifonia and its envisioned mode of working (as described in the Description of Actions) has influenced what we eventually choose to document in this first version of the DMP.

The core of the deliverable is section 3, the ORDP, which zooms into the Pilots as primary sources for data, even though not all of them are organized around a specific set of data. In other words, this DMP version maps out the legacy from which Polifonia starts. In later iterations of the Data Management Plan the focus will be shifted to components inherently use and produced in Polifonia. How this shift will be accompanied by the next two iterations of the Data Management Planning we discuss in the conclusion.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

Document history

Version	Release date	Summary of changes	Author(s) -Institution
V0.1	07/05/2021	First draft released	KNAW
V0.2	18/05/2021	Revised draft with contributions from pilot leader on data stories	KNAW + Pilot leaders
V0.3	20/05/2021	Revised draft, inclusion of sections 4-6	KNAW
V0.5	21/05/2021-27/05/2021	Revision of all sections, addition of Annex	KNAW
V0.6	31/05/2021	Revision after meeting with Advisory Board member and reviewers (May 28, 2021)	KNAW
V0.7	18/06/2021	Final revisions, added elaborated Appendix I	KNAW
V0.8	24/06/2021	Incorporation of final comments	KNAW; Technical Board
V1.0	25/06/2021	First version submitted to EU	UNIBO

Table of contents

Project information 3
Project contacts 3
POLIFONIA consortium 3
Executive summary 5
Document history 6
1. Introduction 9
1.1. Pilots at the heart of Polifonia 9
1.2. Sources and knowledge graphs 10
1.3. Methods behind this DMP 12
1.4 Structure of this deliverable 13
2. Reflection on the role of DMP's in research project and the special case of Polifonia 14
2.1 The form and function of a DMP 14
2.1.1 The DMP and its role as a formal document in research projects 15
2.1.2 The DMP as management tool inside of Polifonia's organisation 16
2.1.3 Distributed, web-based, cross-domain – the Polifonia as a network of processes and products 19
3. Polifonia DMP 21
3.1. Data summary 21
3.1.1 BELLS 25
3.1.2. ORGANS 26
3.1.3 CHILD 26
3.1.4 MUSICBO 26
3.1.5 TUNES 27
3.1.6 TONALITIES 27
3.1.7 Summary – types of data formats used 28
3.2 FAIR data 28
3.2.1 Making data findable, including provisions for metadata 32
3.2.2 Making data openly accessible 34
3.2.3 Making data interoperable 35
3.2.4 Increase data re-use (through clarifying licences) 36
3.3. Allocation of resources 37

3.4 Data security 38
3.5 Ethical aspects 38
3.5.1 Standards and guidelines 38
3.5.2 Research involving humans 39
3.5.3 Informed consent 39
4. Conclusions 40
References 43
Appendix I: Overview of the main sources used in the pilots 45

1. Introduction

1.1. Pilots at the heart of Polifonia

In this version of the DMP[2] we document the **initial conditions** for data management inside of the project and related necessary documentation work. Using the ORDP template for H2020 projects, we zoom into the *Pilots* which form the corner stones for the organisation of work inside of Polifonia. As visible in the DoA, they are grouped around certain topics.

Preserving musical heritage through knowledge graphs			**https://polifonia-project.eu**
1	BELLS	Preservation of Historical Bell Heritage: dependencies between tangible and intangible	
2	ORGANS	A Knowledge Graph on History of Pipe Organs	
Managing musical heritage collections through knowledge graphs			
3	FACETS	Exploration of music scores collections through statistical features	
4	INTERLINK	Interlinking of collections in digital music libraries and audiovisual archives	
Studying musical heritage through (interlinked) knowledge graphs			
5	CHILD	Exploration of musical heritage for scholarly enquiry: a case study on Music and Childhood.	

[2] We are very grateful for feedback from the Advisory Board member (Elena Giglia) and the reviewers (Erzsébet Tóth-Czifra and René van Horik) on earlier versions of this document.

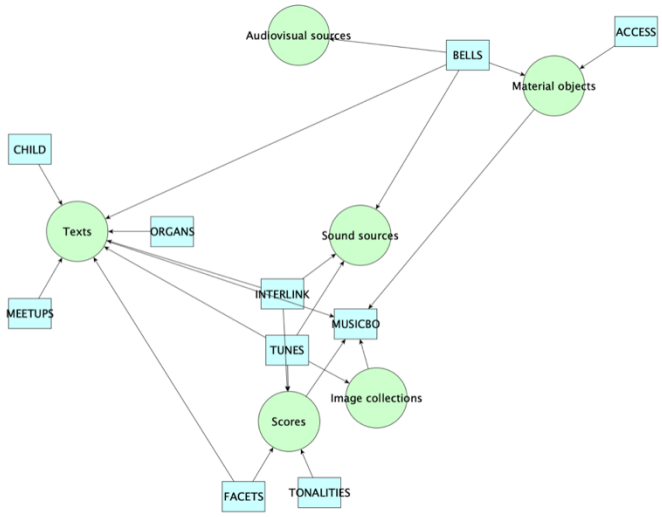
6	MUSICBO	Knowledge graph of Bologna Musical Heritage.	
7	TUNES	Tunes analysis and classification	
8	TONALITIES	Modal and tonal classification of Western notated music from the Renaissance to the 20th century	
Interacting with musical heritage knowledge graphs			
9	ACCESS	Making musical performances accessible to people who are Deaf or hearing impaired	
10	MEETUPS	Musical Meetups: the European musicianship flow	

Table 1: Polifonia's *10 Pilots:* Label and description, Icon and Grouping

1.2. Sources and knowledge graphs

Polifonia is dedicated to the musical heritage of Europe, and represents an all-encompassing approach to it. At its heart is the construction of a distributed knowledge graph (i.e., the project applies semantic web technology). Semantic web knowledge engineering fosters access to sources and enables users to put sources in context, by interlinking them. Knowledge graphs act as threads in a new web of (musical) knowledge which is the main envisioned final data output of Polifonia. The parts of this new emerging new web of knowledge are diverse: they entail existing data collections as well as newly produced one; existing Knowledge Organisation Systems (KOS) (mainly ontologies) as well as newly created one; and existing analytical frameworks and methods as well new one. The 'users' of this new web are equally diverse: from knowledge engineers pushing technological innovation, to humanities scholars of different background, to wider audiences - all interacting in different ways with the traces of musical heritage. At the end of Polifonia, this new web of knowledge is expected to materialise in form of a web portal, which guides users to sources, conceptual spaces, methodological workflows, and provides new ways of experiencing music.

The ingredients for this emerging web have been called *sources*, and this is no arbitrary terminological choice (see D1.1 Roadmap). The term *source* has a wider meaning, it can be a concrete database, a classification scheme or an ontology, but also a certain conceptual framework – an epistemic reference point. The survey conducted in WP 1 with all pilots had been "conceived to offer mutual support in defining general goals, roles and expectations of the ten leading pilots of Polifonia." (D1.1). But, in the meetings about its design, it turned into a much more fundamental inventory of the 'knowledge needs' of the Polifonia community. This process led a conceptual mapping of types available sources one the one side and envisioned 'knowledge units' as elements of research enquiry on the other side (see Figure 1 and Figure 2 in D1.1., Figure 1. copied here).



Example 1: Survey – nature of the sources – Conceptual map of Polifonia's knowledge sources (Part 1).

Figure 1: Reproduced from D1.1. (WP1)

Pilots are ordered originally according to four different types of practices around musical heritage (see Table 1 above): **Preserving, Managing, Studying, and Interacting.** Each of these practices comes with its own requirements to data management, and foci shift between enhancing existing sources to creating new sources to creating new ways for research and wider audience to connect to sources. During the interaction with the Pilots we realised that some pilots are more **'data-driven'** (composed around one specific data base, or in quest for bringing to life a new collection); others are more **'method-tool-driven'** (towards developing new methods and new tools or implementations of methods), and where others are more **'audience-driven'** (oriented to serve end-users, which can be envisioned researchers, wider audience or specific audiences).

From what has been said so far, it is also clear, that Polifonia does not operate on one set of data, nor is it focussed on one type of data. As visualised in Figure 1 from D1.1. data types include: **audio-visual**

sources, scores, sounds, texts, images and material objects. They come from different sources and are used across Pilots (Figure 1). Hence, it would be too short cut, to think about each of the Pilots as a 'container' for which a DMP 'subplan' could be defined and iterated during the project. This would of course also defeat the whole integrative idea behind the Polifonia project design.

As a consequence of the complexities described above: there is no easy or straightforward way of documentation. Not all aspects of the DMP template (the Open Research Data Pilot) are of the same relevance - if at all – to all Pilots. Therefore, next to the factual documentation (which has been executed, and which in part also reveals what knowledge is still missing) this DMP version is piece of reflection – complementing the Roadmap (WP1), but also the discussions in the Technical Board, and in WP2.

1.3. Methods behind this DMP

As mentioned above, the DMP also took the *Pilots* as starting point. Consequently, questions relevant in the context of a DMP (such as size of the data, used KOS to access and annotate them, institutional location, technological readiness level etc.) have been integrated into the survey (see D1.1. Appendix, Survey template). The answers of the survey formed the starting point for interview sessions and mail exchange with the Pilot task leaders conducted in WP7 and eventually led to what we now called *data stories* as part of this DMP.

Excerpt of the Survey (executed in WP1) - highlighted parts are DMP driven

1. DOMAIN SPECIFIC PART.
    1.1. Identification & characterization of the sources, and available digital corpora or datasets.
        1.1.1. Availability
            1.1.1.1. Are the sources digitized?
            1.1.1.2. If so, are the sources born digital?
            1.1.1.3. Can you provide links to the sources in a (partial) list?
            1.1.1.4. Where is the data stored?
            1.1.1.5. How scattered is musical knowledge data over different sources?
        1.1.2. Source characterization. What is the nature of the sources?
            1.1.2.1. Score
            1.1.2.2. Sound source
            1.1.2.3. Image collection (iconographic items, diagrammatic items, etc,)
            1.1.2.4. Texts (poetry, librettos, writings about music, correspondence, theoretical sources, documents about gestuality, technical documents, etc.)
            1.1.2.5. Audiovisual sources
            1.1.2.6. Material objects (bells, instruments, theatre equipment, sources listed above explored from the perspective of their materiality)
        1.1.3. Legal issues
            1.1.3.1. Are there any copyright or licensing issues?

1.1.3.2. Who has the rights to the used datasets?

1.2. Form, type, extent of research outcomes

…..

2. **TECHNICAL PART**. The technical part specifies the actual implementation of the conceptual model:
   2.1. Characterization of existing datasets (more specific technical questions will be addressed in the data management plan).
      2.1.1. Metadata
         2.1.1.1. What formats are used?
         2.1.1.2. What standards are used?
      2.1.2. Datasets
         2.1.2.1. What formats are used?
         2.1.2.2. What standards are used?
   2.2. Knowledge organization systems and interoperability
      2.2.1. Please list the controlled vocabularies, taxonomies, thesaurus, ontologies, etc. that you will use in your pilot
      2.2.2. Please indicate if you plan to expand existing knowledge organization systems or if you plan to create a specific one
      2.2.3. Do you plan to support interoperability towards third parties?
      2.2.4. Do you have API's on your resources?
   2.3. Are there planned overlaps (already defined or envisioned) between the pilot's resources and other resources used within *Polifonia*?

……

3. **SOCIO-PEDAGOGICAL PART.** The socio-pedagogical part aims to identify the target groups and their conditions

Conceptually, we equally rely on information science (documentation) and science and technology studies or science of science. Methodologically, we applied *participatory observation* (in meetings of WP1, WP2, and the Manipasti sessions) next to interview appointments and mail exchange with Pilot task leaders in the actual writing process of this report. We also held one meeting with coordinator in WP7 (early May).

1.4 Structure of this deliverable

As was described above, this deliverable is structured around the different pilots. However, we do take the DMP template, in which the aspects of FAIR are addressed by mean of open questions, as the guiding overall structure of the present document. This results in a DMP that presents the **status quo of the data management process at the first stages of the project**. In later iterations of the DMP, the focus

shifts to the data management practices that correspond to that particular phase of the project (M18 and M36) (see also conclusions).

Section 2 presents a wider more general reflection on the status of DMP in research projects and the special complexity for DMP making in the case of the Polifonia project. It also details the methods applied to collect the information for this DMP.

Section 3 represents the DMP, which according to the ORDP consists of 5 sections:

- Data Summary
- FAIR
- Allocation of resources
- Data security
- Ethical issues

The section **Data Summary** contains tables with overview on the sources and is complemented with a data story for each data driven Pilot. The stories have been built up, wherever possible, based on the questions mentioned under 'Data Summary' in the DMP template. The **FAIR section** starts with an overview of metadata schemes and standards handled in the project and the envisioned ontologies to be used. Also here, we return to the Pilots wherever possible, and describe as well how the pilot in question makes sure its data is handled in a FAIR way. For many Pilots, though, at this stage of the project, not all template questions can be answered to the detail yet. The last four sections of the DMP are addressed on the level of whole Polifonia, with examples from the pilots where appropriate.

The *Data stories* (and this first version of the DMP) define the points of departure including points of uncertainty (missing information, research plans still adapting). In the course of the collaboration inside of Polifonia, the Polifonia community will develop her own sources and knowledge units. How to preserve them in the mid and in the long-term is the ultimate goal of the contribution of Data Management as a process. In the conclusions, we develop of vision how to design the second version of the DMP (M18), and the third DMP version (M36).

2. Reflection on the role of DMP's in research project and the special case of Polifonia

2.1 The form and function of a DMP

In any research project, a Data Management Plan is a **formal document**, hence a product, often formalized as an obligatory part of the application process, or, as is for example the case in Open Research Data Pilots (ORDP), as one or more deliverables. However, in the course of a research project the management of data itself is a **process**. This document addresses both aspects.

2.1.1 The DMP and its role as a formal document in research projects

As a formal document, the DMP documents different phases in the Data Life Cycle which accompanies, and as such is interwoven with, the Research Cycle. Both are depicted schematically in Figure 2.

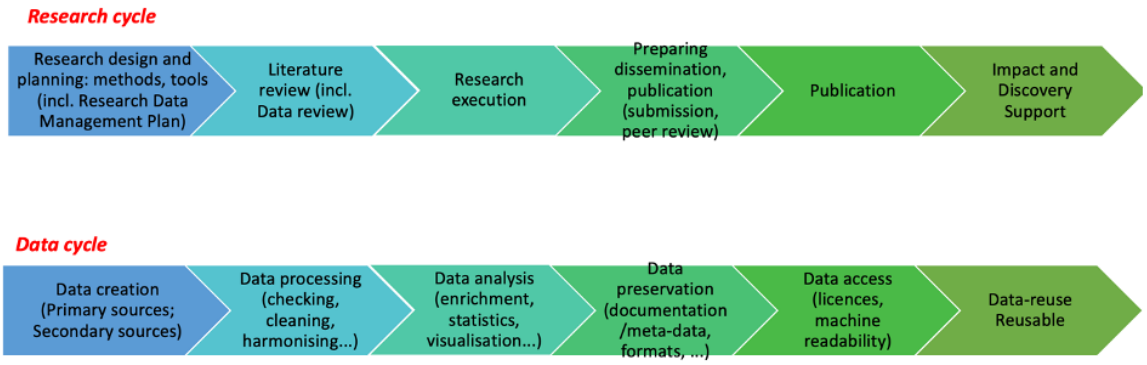


Figure 2: The interrelationship between the research cycle and the data life cycle.[3]

The model of the data cycle is reflected in the outline of many of the available templates for Data Management Planning. For this H2020 project, we use the Open Research Data Pilot (ORDP) template, which emphasizes specifically the aspects of FAIRness and sustainability.

Figure 3: The main elements of the ORDP-DMP template.

The framework of FAIR has been first introduced in 2014 and were two years later detailed as the FAIR principles (Wilkinson et al. 2016). These principles operationalise aspects of Findability, Accessibility, Interoperability and Reusability, in a total of 15 guidelines that help researchers as well as data supporters to increase the re-usability of research data. Currently, an intense discourse and research is taken place to further sharpen guidelines, and standards to be applied to guarantee the FAIRness of data (Hugo et al. 2020).

[3] https://library.ucf.edu/about/departments/scholarly-communication/overview-research-lifecycle/

2.1.2 The DMP as management tool inside of Polifonia's organisation

Schemas as those depicted in Figure 2 present an ideal-typical categorization of processes which in the real world are usually less tidy - a network of elements rather than a clear sequence, and temporarily executed simultaneously, or in a different order than expected, and often in an iterative way (cf. Beaulieu et al. 2013, see also Reichmann et al. 2021).

Aware of the complexity of research processes, Polifonia purposefully chose an iterative approach to the DMP, that aligns with the evolution of the project and the data management as a process, as outlined above. In a first step, the DMP documents and evaluates the state of the *data* envisioned to be used. Here, we zoom in into the pilots, which form the corner stones of Polifonia work, by presenting *data stories* about some of the pilots.

In the overall Polifonia structure, there are various cross-Work package activities. Ultimately, a web portal will be designed to allow access to Polifonia resources (WP1). This design process has been informed by a **survey** which covered both envisioned 'sources' (collections, datasets, metadata schemes, ontologies), as well as envisioned operations around those sources (D1.1., 2021, DL1.3. 2021) This led to a mapping of the methodological and conceptual space of Polifonia. The mapping process has been further informed by the introduction of ideal-typical personas (Polifonia **stories**), representing future users of the Polifonia portal (D1.1 2021). In working sessions called **Maninpasta** a number of high-level research trajectories were defined in a collaborative and co-creative way (DL1.1, 2021). In parallel, the Technical Board keeps an **inventory** of machine-readable data sources, and available ontologies. GitHub is used as a 'hold-all' of the technological achievements during the process. While GitHub is used for data collection and management during the project, given the nature of GitHub (as commercial platform) it is not suitable for long-term preservation. The different versions of the DMP documentation zoom into selected elements of the sources Polifonia uses. This way the DMP complements by a *bibliographical view* the other perspectives (research and wider audiences; and knowledge engineering) present in the project (Figure 4).

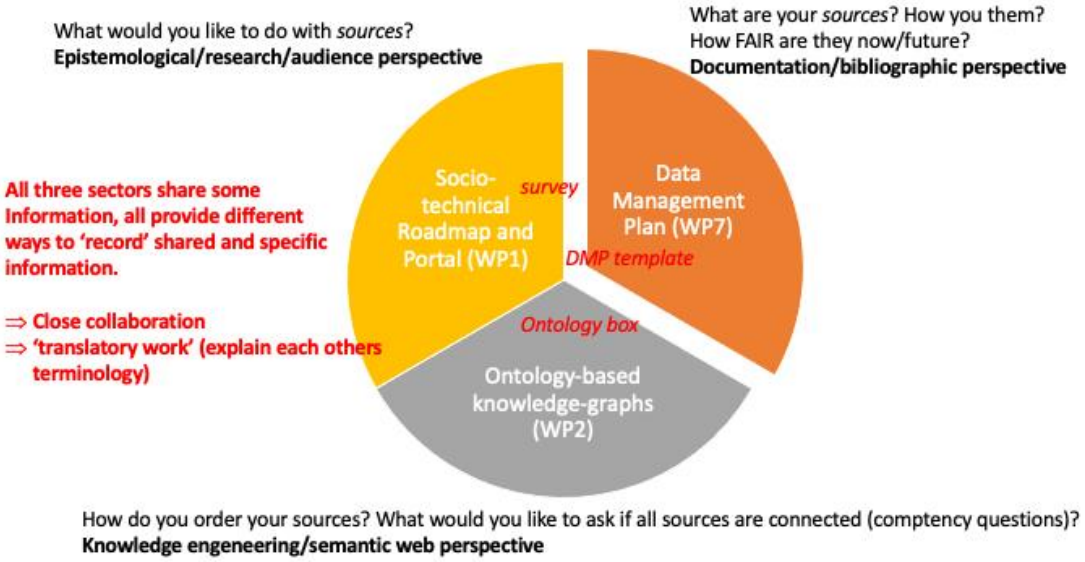


Figure 4: The DMP as integrated part of the Polifonia project.

Three versions of the DMP are planned at M6, M18, and M36. These versions are accompanied by two webinars providing expert knowledge on FAIR requirements and preservation strategies of FAIR Digital Objects to the Polifonia Consortium.

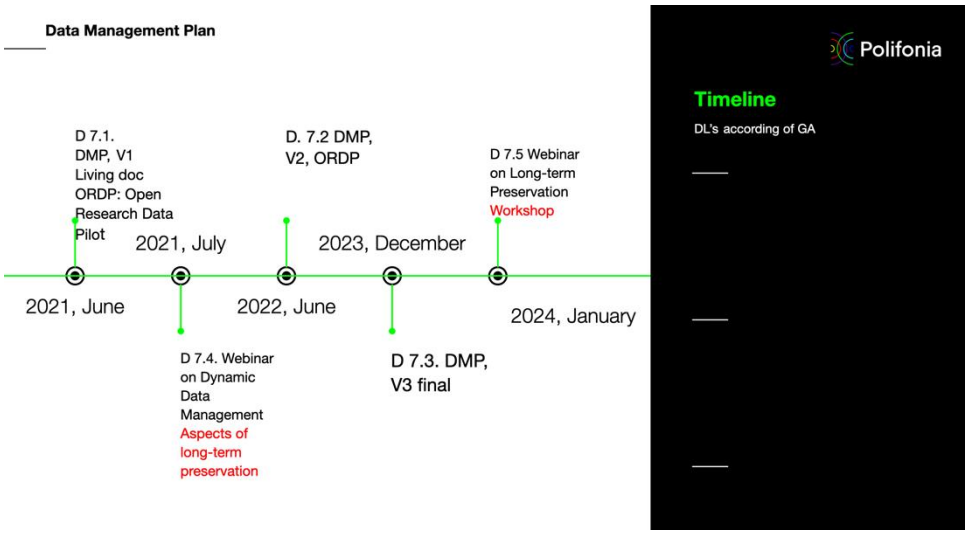


Figure 5: Timeline for the making of the DMP

The first deliverable of the Technical Board (D1.3 - Pilots development – collaborative methodology and tools, 2021) lists the following components of the Polifonia Ecosystem:

Registries– indexes of resources of interest to Musical Cultural Heritage. A preliminary example is the MusoW catalogue of Musical Resources on the Web. Other registries can be developed to fit specific needs (for example, the catalogue of resources useful to the CHILD pilot)

Ontologies– produced in the context of the Polifonia project to support pilots and use cases, ontologies specify domain knowledge and are used as means for developing a shared understanding of the domain and as software artefacts applied in published datasets•

Datasets– structured data offered following best practices in (Linked) Open Data publishing. If data was not produced as Linked Data, when possible, the resource is also publishedin its original format. Multiple ontologies can be applied and alternative Linked Data versions of the same source data are possible, to fit the needs of different use cases.

Repositories and corpora– collections of digital assets relevant to Polifonia use cases.

Knowledge Graph– a distributed but unifying view of musical cultural heritage knowledge, is a virtual composition of Linked Data resources to be reused for large scale integration, for example, to support unified indexes for exploration and discovery.

Services– Web APIs that expose reasoning and data processing capabilities. Services are run by Polifonia consortium members and instantiate specific components to the OpenWeb. Among those there are Linked Data services such as SPARQL endpoints – live data services publishing the above components for querying with SPARQL.

Software libraries– reusable code produced by the project to support pilot activities. Software libraries are used by programmers in their own applications

CLI tools– ready-made tools to be used by developers in scripting data manipulation pipelines

User interfaces– targeting domain experts, citizens, developed to support specific activities in the context of the Polifonia Pilots, user interfaces can be reused across similar applications targeting different datasets and scenarios

Experiments– code and dataset of scientific experiments used during the research activity of the project. Experiments are meant to be documented, reproducible, and linked to re-search outputs. However, experiments are not expected to produce code and data which is directly reusable. The developers may produce derived assets as independent components of the ecosystem

Applications– applications targeting specific use cases, possibly as direct outputs of the Pilots. Applications may reuse ecosystem components and function as demonstrators in tutorials and referenced by the documentation of the ecosystem.

Containers– which wrap applications or services ready to be deployed in a computing infrastructure (a developer's laptop or a cloud service).

Stories– requirements from the world of Musical Cultural Heritage preservation, exploitation, and scholarship; stories are the starting point of the collaborative methodology and the sense-making layer of the Polifonia Ecosystem, giving context and purpose to the components. Stories and scenarios are linked to relevant components of the ecosystem.

Tutorials– a showcase of the Polifonia Ecosystem through end-to-end tutorials, inspired from the Pilots and displaying the capabilities of the components in concrete applications. Tutorials are also an excellent starting point for developers.

Web Portal (a KG view on Polifonia Linked Data resources) – the aggregator of the Polifonia Knowledge, exploiting the Knowledge Graph as underlying integration method.

Documentation– documentation associated to each component. Developed autonomously, to fit the requirements and needs of the specific type of component.

Polifonia Ecosystem Website (GitHub) – The resource for accessing the Polifonia Ecosystem, browsing the components and associated the documentation, accessing the resources for developers, and joining the project team in building the next generation of tools for Musical Cultural Heritage."

From this list, it is quite clear that Polifonia uses the term 'data' in a very broad meaning (including tools, workflows, services). In the realm of the Data Management Plan we will discuss which type of the above listed components and which concrete examples are a target for a more detailed preservation strategy.

It is clear also from this overview from the Technical Board, that Polifonia will rely on existing sources, enhancing them, creating new sources from a recombination of what exists, and from its own research endeavours.

2.1.3 Distributed, web-based, cross-domain – the Polifonia as a network of processes and products

Polifonia is centred around music. Its ambition is to embrace the 'phenomenon' of music in an all-encompassing way (see https://polifonia-project.eu for the mission of the project). Music is both a practice as well as an object of study (musicology). Musicology ranges from the study of the transmitted notations of music (scores, recorded sound), to the instruments used to produce music, to the actors creating and performing music. The study of the perception of music – which is also part of Polifonia – is highly interdisciplinary in itself. Research in Polifonia even extends into buildings used to perform sound (bell towers) and hence touches into architecture and art. Each of these academic fields comes with its own practices around the documentation and use of sources, including formal ways to annotate them (specific knowledge organization systems).

Polifonia's approach to this complexity is practice driven. Using the collective wisdom of the project partners, sources are collected and tested for (re-)usability. In an iterative circle of versions of knowledge graphs, it is expected that some sources become more 'popular' due to availability and suitability for the research goals of Polifonia. Those responsible for the DMP will closely watch and interact with these dynamics to be able to propose for the second version of the DMP a selection of sources which represent the legacy of the project itself.

Semantic web technologies are at the heart of Polifonia, and knowledge graphs the engine of knowledge production. For the making of the DMP it is important to realize that the project deals with a network of sources (data), accessed and processed over the web. Different from the ideal typical picture of 'a dataset' which is object for data management, Polifonia deals with a network of sources. Even more importantly, not for all sources Polifonia is in charge. The project researchers deal with other institutions which are content provider, or KOS service provider. Some of them are part of the consortium, others are not.

As the project unfolds, the data management process also evolves and hence the DMP matures as well. For the second draft of the DMP we expect to be challenged with a selection of sources to be documented which are core to Polifonia (see above), and in the third draft of the DMP we present sustainability scenarios depending on the 'ownership' of sources.

For the next phases of documentation, existing guidelines for semantic web technologies, and for FAIRification of semantic artefacts (FAIR Digital Objects in general) will be leading. The final version of the DMP document serves as a best-practice documentation in itself.

3. Polifonia DMP

3.1. Data summary

Polifonia's knowledge management

Polifonia starts from a number of web-based resources (Pilots), but also has defined other web-based spaces on which to exchange and temporarily store information.

Resources	Pointer	Main use/Openness	Preservation strategy
Main sources used in the pilots	See Appendix I for the full list: BELLS (5); ORGANS (1); FACETS (2); INTERLINK (9); CHILD (3); MUSICBO (10); TUNES (11); TONALITIES (4); ACCESS (1); MEETUPS (5).	Input for Polifonia consortium (Public)	In the responsibility of the original content providers
Polifonia Website	https://polifonia-project.eu Build on Wordpress, with back-up's via the InternetServiceProvider chosen by DP	Dissemination about the project (Public)	Submitted to webarchive
Polifonia Github instance	https://github.com/polifonia-project	Organisation space used as the primary infrastructure for supporting collaborative development of the pilots, including collection of requirements, software development, testing, and documentation. A key element of the approach includes an open-by-default policy, where contributions to the development is encouraged also outside	Not yet determined

		the Polifonia consortium.	
Certain Polifonia Github repositories	To be determined	Deposited as releases with Zenodo	Following Zenodo
Polifonia Sharepoint space	https://liveunibo.sharepoint.com/sites/polifonia	Private sharepoint group hosted at UNIBO; used for project documentation (meeting notes, DL drafts, presentation, work documents of WP's and other groups)	Not yet determined
Polifonia Portal	In preparation, webspace hosted at UNIBO Requirements in discussion Server: OS: Ubuntu Linux (no less than version 20) already installed RAM: >4GB CPU: 1-2 AMD Epyc or Intel Xeon, 8 core (no less than 4 core) Storage: no less than 50GB IT Support: Less than 2 working days response Backup	Access control: Allow access to partners belonging to different institutions (including: UNIBO, OU) SSH authentication; Sudo users with HTTP / HTTPS / SSH access	To be determined in detail in DMPv3

Table 2: Web-bases spaces in use in Polifonia.

Polifonia's data stories

Terminologically, the DMP uses the term 'data'. However, what data are is a complex question (Borgman 2015). What is seen as data strongly depends on the context: the scientific field or knowledge domain

you operate in, and the norms, values, methods, standards applied by it. More recently, in the context of the EOSC even broader definitions have been used which use 'data' as a placeholder for all varieties of Digital Objects (such as images, software, workflows, ontologies, etc.) (Corcho et al 2021). In Polifonia, during first sessions of WP1 we coined the term *sources*, which is broader, and encompasses data as well as formal metadata, as well as other contextual information.

As we start from the Pilots, we try to classify them according to at seems to be relevant under a preservation perspective. We used four different types, that each roughly correspond to one or two WPs.

- **Data driven:** Pilots clearly centered around an existing collection or aiming to build a collection (WP2)
- **Method driven:** Pilots centered around pushing for an innovation of methods with the aim to answer a concrete research question (WP3 and WP4)
- **Tool driven:** Pilots centred around new technological solutions, which emphasis of genericity of these solutions (WP3 and WP4)
- **Audience driven:** Pilots centered around enabling new ways to experience music (WP5)

Next to this 'preservation typology' we rely on the identification of types of sources as provided in D1.1, which orders sources according to the type of data in them.

Table 2 presents an overview of the 10 pilots of the Polifonia project. The pilots that rely most heavily on specific Musical Heritage resources are described in the data stories. The pilots that aim specifically at providing technological solutions (e.g. tools) for better accessing to more resources are used as examples in the sections that address the various FAIR criteria. The pilot ACCESS aims to test technology that can be used to make live musical performances accessible to audience members who are Deaf or hearing impaired. As this is a technology-driven pilot and does not rely on external data sources, it is not included in the data stories.

The DMP aspects change depending on what the focus in terms of sources and goals is. Some Pilots start from a database, and aim to produce a 'derived', enriched version of (part) the database. Other Pilots focus on developing a new ontology (partly re-using existing ontologies, partly extending ontologies). In both cases, what to preserve, in which form will be different. This is why we use Pilots at different places of the overall ORDP structure as examples to illustrate certain aspects of making data FAIR.

Preserving musical heritage through knowledge graphs			**Nature of sources** (D1.1.)	**Typology** (focus of a preservation strategy)
1	BELLS	Data story in section 2.1	Audiovisual sources Texts Material objects	Method-Driven Data-Driven

2	ORGANS	Data story in section 2.2	Texts	Data-Driven
Managing musical heritage collections through knowledge graphs				
3	FACETS	Example of OA software in section 3.2	Scores Texts	Tool-driven
4	INTERLINK	Example of Interoperability in section 3.3	Sound Sources Scores Texts	Tool-driven
Studying musical heritage through (interlinked) knowledge graphs				
5	CHILD	Data story in section 2.3	Texts	Method-driven
6	MUSICBO	Data story in section 2.4	Material objects Texts Image collections Scores	Data-driven
7	TUNES	Data story in section 2.5	Image collections Scores	Data-driven Tool-driven
8	TONALITIES	Data story in section 2.6	Scores	Method-driven
Interacting with musical heritage knowledge graphs				
9	ACCESS	Technology-driven	Material objects	Audience-driven
10	MEETUPS	Example of reusability in section 3.2	Texts	Method-driven

Table 3: Overview about the Pilots and where they appear in the DMP.

The data stories focus on the sources that are used in each pilot. They aim to answer to the questions about the Data summary as they are formulated in the DMP template and which are repeated in the text box below. These data stories are complemented with an overview in Appendix I, that lists all sources

identified so far with more detailed information on file size, data formats, metadata standards, access rights, licencing and copyright, as far as this information is available at this point.

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

What types and formats of data will the project generate/collect? Will you re-use any existing data and how?

What is the origin of the data?

What is the expected size of the data?

To whom might it be useful ('data utility')?

3.1.1 BELLS

The BELLS pilot investigates the possibility to build a conceptual model concerning bell sound in its tangible aspects (bell types, connection with bell towers structures, connection with places, diffusion of sound in space), as well as its intangible aspects (occasion of sounds, tradition and practice transmission, connection with religious time and civil time scan). In the Polifonia web portal, the geo-referenced knowledge graphs will be available and linked to the collected sources (audio, audio-visual, textual documentation). This enables historians, ethnomusicologists, linguists and other academics to investigate bells as a complex system, starting from the sound recording, but connected to written and oral sources to data about typologies of bells, socio cultural contexts, performing practices, transmission practices. In addition, the result will be of incredible value to cultural heritage institutions as well as to institutions involved in curating bells and their surroundings.

The existing sources on which this pilot builds have been partially digitized, but partially they are only available still in analogue form (e.g., tapes). The sources are stored in local archives, such as the Historical Diocesan Archives, the State Archives (for textual sources), the archives of Bell Ringers associations (for sound and audiovisual sources, that will be digitized and imported in the Institute's digital repository), as well as the General Catalogue of Cultural Assets. The sources include audio-visual sources, images, architectural descriptions, texts about bells (poetry, librettos, technical documents, etc.). in a wide range of formats (e.g. jpeg/tiff, xml-rdf, etc.).

3.1.2. ORGANS

The aim of this pilot project is to improve the access to the detailed descriptions of historical pipe organs in the Netherlands. In order to do so, the full text of *Het Historische Orgel in Nederland* (1997-2010), a 15 volume, 4,500+ pages encyclopaedia containing histories and images of almost 2,000 Dutch organs, published by Nationaal Instituut voor de Orgelkunst (NIvO) will be made accessible in the form of a knowledge graph. Though originally a non-digital source, currently, there is a digitized version of the full text available at the NiVO to which the Polifonia pilot leader has access. At the moment, the sources are available in MS Office formats only (Word, WordPerfect).

In the pilot project, this vast amount of knowledge will be fully made accessible in the form of a knowledge graph. This will enable detailed study of the history of organs and organ building in the Netherlands and abroad. Such a knowledge base will be highly valuable for music historians, but also for organ builders and organ advisors who are involved in restoration, maintenance, or reconstruction projects.

Via the Polifonia portal, this vast amount of knowledge will not only be better accessible, it will also be linked to the other sources of the Polifonia project. This is of great interest to musicians and academics of various disciplines (musicology, organology), as well as curators and organ building companies.

3.1.3 CHILD

The CHILD pilot centres around the question what role music has had in children's live, be that through education, play, or family and community practice. In order to study this, children's writings, such as letters and diaries, will be identified in the primary sources. The pilot will
reuse the Listening Experience Database (https://led.kmi.open.ac.uk) both as a direct source
of evidence and as source
of references to relevant books, memoirs, and letters that can contribute to the scholarly enquiry. Relevant sources include public domain repositories such as the Internet Archive and Project Gutenberg.

Although these letters and diaries were originally non-digital, the sources mentioned are open access repositories, containing (parts of) digitized books. In the web portal, those accounts will be linked and related to musical scores and relevant assets in the knowledge graph. The output of this pilot mainly targets scholars, in various fields such as Cultural Studies on Music and Musicology.

3.1.4 MUSICBO

The aim of this pilot is to identify evidence about the role of Bologna as a creative city for music of the centuries. It draws on the Lessico Beni Culturali, the BUB archives, as well as other open access libraries. They contain testimonies of scholars, journalists, travellers, writers etc in a great variety of formats. By means of the knowledge graph, it will be made possible to extract knowledge from the various sources.

This will be of interest not only to scholars (e.g. musicologists, historians), but also to cultural institutions as well as the tourism industry.

At this point in time, the team dedicated to this pilot is planning the corpus. The aim is to create a balanced corpus, representing the variety in genres. Once the corpus is fully compiled, it is expected to comprise around 1 million words.

3.1.5 TUNES

Aiming to investigate to what extent Dutch melodies from the 17th and 18th century are connected to melodic repertoires from elsewhere in Europe, the main source for this pilot is the Dutch Song Database, as well as the corresponding melodic data, both made available by de Meertens instituut. By means of a knowledge graph, this database will be connected to other musical databases (e.g. the Irish Traditional Music Archives, RISM incipits, the Essen Folk Song Collection, the tune index from The Colonial Institute, themes from the Barlow and Morgenstern Collection, and various collections in ABC notation) in the Polifonia portal. Once connected, musicians as well as music historians and others interested, will be able to search for overlap in the melodic patterns. Similarities found between the Dutch melodies and the repertoires from the other sources, in turn, will be annotated in the Duch Song Database.

The content of the Dutch Song Database is the result of a long-term collection process. Originally, most of the content was non-digital, but over the years it all has been digitized. As heterogenous as its origins, the formats vary greatly among the sources (ASCII files in csv format, kern, midi, pdf, png, jpg, txt, lilypond source, mp3, etc.).

3.1.6 TONALITIES

The main aim of this pilot is to develop tools for the modal-tonal identification, exploration and classification of monophonic and polyphonic notated music from the Renaissance to the 20th century. Its starting point is the knowledge, tools and corpuses already available at CNRS is taken as a starting point and the pilot builds up from there. The main resources that it will use are the NEUMA score library and the Meertens Tune collections. In addition, it will strive to incorporate as well as many of the resources of other pilots made accessible by the technology developed in WP2.

The pilot envisions to provide metadata according to the MEI scheme for the score corpora, and CIDOC-CRM and LRMoo for metadata and annotations. As an early estimate, the final dataset is expected to correspond to roughly 1 GB including the MEI corpora and RDF annotations.

The pilot is expected to have an impact on research in the field of theory and evolution of musical language as well as support the understanding collections by students, performers and others interested. As such, the data resulting from this pilot will be relevant for academics such as musicologists, historians of music and music theory, and analysts, as well as for composers, cultural institutes, music industry.

3.1.7 Summary – types of data formats used

Type of source	Pilot	Used formats
Sound sources	BELLS; INTERLINK; TUNES	tar gzip, N-triples, RDF, **kern, midi, pdf, png, jpg, txt, lilypond source, mp3
Scores	FACETS; INTERLINK; TUNES; TONALITIES	MEI, MusicXML, Midi, tar gzip, N-triples, RDF; XML
Audio-visual	BELLS	unspecified
Material Objects	BELLS; ACCESS	N/A
Texts	BELLS; ORGANS; FACETS; INTERLINK; CHILD; MUSICBO; TUNES; MEETUPS	pdf, html, doc, wp5, txt
Image collections	BELLS; MUSICBO; TUNES	jpeg, tiff

Table 4: Overview of types and data formats used (for existing SPARQL Endpoints see Table 4.1. in D1.3. 2021).

3.2 FAIR data

FAIRness is an essential element of the DMP. FAIR stands for Findable-Assessable-Interoperable-Reusable. Since the FAIR principles have been introduced (Wilkinson et al. 2016), their definition, but also how to best implement them, has been further developed.

An overview about FAIR is given in Figure 6. The FAIRaware tool hosted by DANS KNAW complements the factual overview with an interactive guide through what is means to operate in a FAIR manner (see https://fairaware.dans.knaw.nl).

Figure 6: Schematic overview about the FAIR principles (Courtesy of https://ccafs.cgiar.org/open-access-and-fair-principles and inspired by EHRI3_D12.1).

Various elements are crucial for FAIRness: the use of (certified) data repositories, the use of persistent and resolvable identifiers for 'data', the use of standard models when it comes to metadata description (enriched with information on access/licences, provenance), the use of controlled vocabularies inside of metadata models, the preference for machine readable metadata, the use of open formats when it comes to the archiving of data (see the Preferred formats guidelines published by DANS[4]).

Due to the heterogeneous nature of the pilots, as well as of the institutions and researchers involved in the Polifonia project, different levels of FAIRness, and of FAIRawareness, have been noted. In the webinar on dynamic data management that KNAW-DANS is organizing in M9, we aim to identify in detail where our expertise is most wanted/needed, and we'll start working more closely with the pilots on the FAIR issues at stake in their particular situation.

In the section above, the data stories focused on the sources *used* in the projects. Moving from the pilot level of the project to the overall Polifonia level, this section mainly describes the provisions made or envisioned to make the data that is *produced* in the project as FAIR as possible. In addition, here and there examples are added from the pilots to give more concrete examples of how FAIR is interwoven on all levels of the project. The subsections below follow the questions corresponding to FAIR data in the DMP template.

[4] https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats?set_language=en

FAIR and semantic web

The FAIR principles come as a natural to semantic web technology, which by definition operate on the web. Polifonia is dedicated to the Linked-Open-Data philosophy (see https://5stardata.info/en/). But, as recommendations of the W3C group show, in implementing Linked Data technologies a variety of decisions are involved which all have implications both for FAIR and for FAIR in the long run (long-term sustainability). Those include the definition of suitable and durable namespaces for the URI's, decisions on versioning strategies, issues of resolvability, and sustaining data and tooling. (W3C 2017; Siebes et al. 2019)

To ensure FAIRness of the actual research process the consortium applies a combination of using GitHub (specific Github repositories) and the ability to synchronize official releases of a Github repository with a DOI referenceable Zenodo entry. This will make results from the ecosystem directly referenceable also to be used in project reporting.

As for persistent URIs, Polifonia will create a w3id domain with a clear convention for data we produce. Concerning the use of data provided by others, the owners of these datasets can create their own URI's if they don't have one already.

Ontologies

Knowledge engineering as core of Polifonia brings naturally with it, attention to (semantic) interoperability (Poveda-Villalon et al. 2020). Ontologies (and ontology re-use) are at the core of Polifonia. Moreover, experts from the Polifonia consortium are part of groups which contribute to the making of standards, which will be used in the project. An example is the work of the MEI (Music Encoding Initiative). Here, members of the consortium participate in the MEI Linked Data Interest Group (https://music-encoding.org/community/interest-groups.html) which works on standards towards MEI using RDF.

Currently, WP 2 collects a list of ontologies to be used. These ontologies operate on the level of metadata indexing of sources, over general ontologies used in Cultural Heritage and Music, down to very specific ontologies, which address elementary knowledge units (RelatedWork-Ontologies, 2021).

Examples listed in this living document are:

General ontologies (for resources and their descriptions)

- Dublin Core[5]
- Schema.org[6]
- Web Annotation Data Model[7]

[5] https://dublincore.org/specifications/dublin-core/dcmi-terms/
[6] https://schema.org
[7] https://www.w3.org/TR/annotation-model/

General ontologies on music and Cultural Heritage:
CH:

- CIDOC CRM[8]
- Europeana Data Model[9]
- FRBRoo[10]
- LRMoo[11] (FRBRoo is deprecated, it's LRMoo now)
- VIAF[12]
- ArCo[13]
- Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT[14]), Music Terms[15] within LCGFT

Music:

- DOREMUS[16]
- Music Ontology[17]
- Music Theory Ontology[18]
- MusicOWL[19]
- Music Notation Ontology[20]
- LD version of MEI (work in progress[21])
- Performed Music Ontology[22] (partly)
- MusicXML Ontology[23]
- MusoW[24]
- Music Vocabulary[25]

[8] http://www.cidoc-crm.org
[9] https://pro.europeana.eu/page/edm-documentation
[10] http://www.cidoc-crm.org/frbroo/ModelVersion/frbroo-v.-3.0
[11] http://www.cidoc-crm.org/frbroo/sites/default/files/LRM-FRBRoo_V0.6%28PR%29%20-%20Rev.pdf
[12] http://viaf.org and https://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Virtual_International_Authority_File_(VIAF)
[13] http://wit.istc.cnr.it/arco
[14] https://id.loc.gov/authorities/genreForms.html
[15] https://knoworg.org/lcgft-music-terms/
[16] https://www.doremus.org
[17] http://musicontology.com
[18] https://ora.ox.ac.uk/objects/uuid:2301db2f-da56-4a25-a715-7771fcdaa607/download_file?file_format=pdf&safe_filename=music-theory-ontology.pdf&type_of_work=Conference+item
[19] https://dl.acm.org/doi/10.1145/3106426.3110325
[20] https://cedric.cnam.fr/isid/ontologies/files/MusicNote.html
[21] https://docs.google.com/document/d/1fFhF3Bpoj1r1VpJukonG80FzhokNKgcBAX1VMyGfxjQ/edit
[22] https://bioportal.bioontology.org/ontologies/PMO
[23] http://www.ontologydesignpatterns.org/ont/musicml/musicml.owl
[24] https://musow.kmi.open.ac.uk/resources-type#Ontology-div
[25] http://www.kanzaki.com/ns/music

Polifonia Webportal

Central to the data production in the Polifonia is the web portal that is developed in WP1 It will ultimately include:

1) a reference registry of Musical Heritage (MH) resources, including but not limited to all collections used and produced in the project

2) methods for continuous indexing of MH resources

3) methods for automatic and semiautomatic generation of metadata according to the Polifonia ontologies

4) methods for searching, querying, browsing MH resources.

During the project, all partners involved use a common GitHub group as well as other means of online collaboration as working environments (e.g. SharePoint, Overleaf). GitHub acts as space where the Polifonia datasets (collections and tools) take shape. The data produced in this space is not all equally suitable for later archiving. In the course of the project, some parts residing in GitHub will be deposited as releases in Zenodo. Towards the end of the project, a selection will be made what to include in a final deposit with a long-term archive. This final deposit is referred to below as the 'Polifonia data collection'. At the end, we expect this dataset to be a collection of various data deposits.

3.2.1 Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

Will search keywords be provided that optimize possibilities for re-use?

Do you provide clear version numbers?

While the project moves forward, the long-term preservation strategy for the different web-based resources that produce data will be articulated in more detail. As detailed in Table 2, some resources will remain under the responsibility of the content providers, whereas others, such as the resources stored on the Polifonia GitHub and Sharepoint instance as well as the Polifonia Portal, will become part of the Polifonia data collection that is to be archived. The repositories that have been identified for archiving are AMSActa (UNIBO), Open Research Data Online (OU), EASY (KNAW-DANS), and the Digital Repository of Ireland (national trusted digital repository for Ireland). While AMSActa mainly focuses on

research output in the form of papers, the other three repositories specialize in archiving research data. The KNAW-DANS repository as well as the Digital Repository of Ireland are certified with the Core Trust Seal. All deposits at the three data repositories receive a doi and are described with according to a standard metadata model (Dublin Core Metadata). In case a dataset needs updating, this is possible according to the repositories' procedures. For example, the DANS archiving system provides versioning on two levels: minor changes (minor corrections in one or more files) and major changes (adding or removing files, changes in file names, etc.).

If applicable, within the pilots, resources will be provided unique and persistent identifiers as well. In addition, the data produced by the pilots may be accessed by SPARQL endpoints, as is illustrated with the example below.

> *"In the pilot TONALITIES, we use a technical and internal identification system, which relies on UUID, to uniquely identify all resources. These IDs are stable, and make each and every resource "addressable" via an unique URI (eg: http://data-iremus.huma-num.fr/id/123e4567-e89b-12d3-a456-426652340000). This system provides an extremely fine level of addressing (for example, each note in a score has its own identifier).*
>
> *All the data will also be available through SPARQL endpoints, and the corresponding semantic graph will be browsable with a Web interface. As the data will be available through a public SPARQL endpoint, users will always benefit from the most up-to-date version of the LOD graph and no further versioning is needed."*

In the GitHub group, for the moment, naming conventions are left up to the creator of a source, as this is a working environment. These will later be harmonised once it is decided which resources will be included as part of the Polifonia data collection. This is in line with the general methodology used throughout the project, which strongly relies on a bottom-up approach. According to this approach the gathering, documentation and maintenance of requirements for such components, are managed as hackathon-like, grassroots initiatives by collaborations between the Pilots and the software development teams and supervised by the technological WPs (see D1.3 for more details on this methodology).

In the pilots, each of the pilot leaders decides at what level it is necessary to adhere to strict naming conventions and how the names are constructed. At this point in time, not all of the pilots have clear policies in place. This will be addressed, among other things, in the workshop on data Management organized in M9 by KNAW-DANS.

3.2.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default?

How will the data be made accessible (e.g. by deposition in a repository)? What methods or software tools are needed to access the data?

Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Where will the data and associated metadata, documentation and code be deposited?

Have you explored appropriate arrangements with the identified repository?

If there are restrictions on use, how will access be provided?

Is there a need for a data access committee?

Are there well described conditions for access (i.e. a machine readable license)? How will the identity of the person accessing the data be ascertained?

As clearly stated in the Description of Actions, the Polifonia projects aims to release all created resources and implemented methods with an open source approach, with licenses that allow commercial exploitation, such as Apache 2 and CC BY. This approach is illustrated by the example below from the pilot FACETS. However, many of the pilots make use of sources provided by one of the partner institutions or external parties. For these data, other licences may apply. Since Polifonia is strongly dedicated to open access practices, we will investigate if we can maximize open access to these resources made available by the content providers.

"The FACETS pilot aims to design a faceted search engine (FSE) for music score collections, supporting explorations and discovery of scores of interest in large collections. This engine will be demonstrated in Neuma, and its code released in open source for librarians, musicologists and score collections managers, after which other score-oriented musical libraries for cultural heritage will then benefit from the code, such as Royaumont and Gallica-BNF."

All the identified data repositories mentioned above have (machine-readable) licensing as well as authentication procedures in place. As KNAW-DANS is in the lead of the data management in the Polifonia project, they will also be in the lead of collaborating with the other repositories on the exact nature of the deposits. The next version of the DMP, due in M18, will include a more detailed long term preservation strategy, as we expect to have a clearer vision on what the Polifonia data collection will ultimately comprise.

3.2.3 Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

At the very heart of the Polifonia lies the creation of the **web portal**, which will access all the sources as linked open data knowledge graphs. Making data interoperable is thus one of the main aims of the project. Throughout the project, all partners strive to use file formats standard to the field, as well as standard metadata schemas. As described in section 1.3 above, a first inventory of which file formats and which metadata formats will be used by the various pilots was made in WP1[26]. In case no standard vocabularies or ontologies exist, we will aim to build on existing ones and expand from there, rather than creating new ones.

On the level of the pilots, it depends on the nature of the particular pilot whether standard vocabularies and ontologies can be used. Some pilots that strongly rely on vocabularies in their methodology, have already started according to this approach, as the example below from the pilot BELLS nicely illustrates. In other cases, such as the case of the pilot TONALITIES, it is unavoidable to generate new ontologies and vocabularies.

> *"In the BELLS pilot, we will be making use of existing ontologies, such as ArCo Ontologies, the ICCD Vocabulary for Musical Instruments and the ICCD Vocabulary for Intangible Heritage. When needed, we plan to expand the ArCo Ontologies as well as the ICCD vocabularies and thesauri."*

> *"The data used in the TONALITIES pilot are very specific to certain music analysis practices for which there is currently no controlled vocabulary. We will therefore have*

[26] https://github.com/polifonia-project/survey

to create them ourselves. Due to the high specificity of our data, it is not certain that we will be able to provide relevant alignments with other vocabularies or ontologies."

3.2.4 Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

How long is it intended that the data remains re-usable? Are data quality assurance processes described?

As was already mentioned above, the Polifonia project aims to make all the data it produces available using open access licences such as CC BY and Apache 2.0. However, the copyright of some of the resources lie with external parties, and they may decide to licence their resources differently, as the example below from the pilot ORGANS illustrates. In other cases, such as in the case of the pilot CHILD, all resources are freely available to begin with. In Appendix I, the sources and their access conditions are listed.

"The main source in the pilot ORGANS is the Orgelencyclopedie (1997-2010), for which the copyright lies with the Dutch National Institute of Organ Art (NIvO). They have the exclusive rights to the dataset, and this resource will probably not be made publicly available within the Polifonia project, though it will be integrated in the knowledge graph."

"The pilot CHILD uses sources that are openly available on the internet, such as the Internet Archive and Project Gutenberg. The source material can thus be accessed and downloaded via APIs."

A selection of the data that is most relevant to the project will be made towards the end of the project. This data will constitute the Polifonia data collection that will be archived immediately after the end of the project.

At this moment, no embargo is envisioned. In case it is necessary for specific parts of the data, this will be decided in consultation with the repositories. In principle, the data will remain available indefinitely, for which quality assurance and migration procedures are in place.

To maximize the re-usability of the Polifonia data, a dedicated task force in WP6 specifically aims to make the data available for other parties, and to maximize its impact by supervising as the deployment of all technologies by third parties, as well as by providing training events.

3.3. Allocation of resources

What are the costs for making data FAIR in your project?

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

The costs for making the Polifonia data FAIR are covered by the grant. The project partner KNAW-DANS is responsible for the overall data management in the project. A total of 11,5 PMs is reserved for dealing with all issues around data management. Apart from overall guidance on data management throughout the project, KNAW-DANS will lead the formulation of the Data Management Plan in an iterative cycle with versions of the DMP due in M6, M18 and M36. In addition, two webinars will be organized, one on dynamic data management (M9) and on long-term preservation (M37). Finally, during the project life time, KNAW-DANS will guide the process that ultimately leads to the deposit of relevant components of Polifonia (datasets, tools). No additional costs for archiving those components are expected. The strategy is to choose public funded certified, trusted repositories (e.g. CoreTrustSeal), which don't incur fees.

Closely related to one or more specific sub-criteria of FAIR, is the work caried out by the Technical Board (WP1) and in WP6 (Dissemination, Communication and Exploitation) and WP8 (Ethics requirements, see section 3.5).

3.4 Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Is the data safely stored in certified repositories for long term preservation and curation?

The resource holders are expected to have their back-up and storage policies in place and handled by their IT departments. During the project all partners producing data, follow their own protocols for data storage (e.g. on institutional servers), back-up procedures, etc. In the webinar on dynamic data management organized in M9, we will investigate whether all partners have their security policies sufficiently in place or whether additional training needs to be offered, or arrangements need to be made. This may especially apply to those pilots working in one way or another with personal data (see section 3.5 on Ethical aspects).

After the project end, the Polifonia data collection will be stored for the long-term in certified repositories, as described in section 3.2.2 above.

3.5 Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

The Ethical aspects of the Polifonia project are described in detail in the DoA (give reference). Since, at this point in time, there are no deviations from the described policy envisioned, the DoA remains leading. In addition, WP8 is specifically dedicated to Ethics, and will provide seven deliverables, e.g. on the selection of participants, informed consent procedures, etc. Therefore, only a brief summary of the Ethics section from the DoA is given below:

3.5.1 Standards and guidelines

All Polifonia partners agree to apply to the ethical standards and guidelines of Horizon 2020, as well as professional and international standards and relevant national, EU and international legislation, such as

the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights. With respect to personal data, all partners will comply with the General Data Protection Regulation (GDPR), that was implemented in 2018.

3.5.2 Research involving humans

Due to the nature of the project, the great majority of the data that is processed or produced during the project does not involve humans. So far, two concrete cases have been identified in which the research involves human participants:

- The pilot ACCESS includes research with human participants who are deaf, hearing impaired or otherwise disabled. They will be involved in evaluating the interfaces developed in WP1 and WP5.

- In WP6/Task 6.5, the GirlsCodeIdBetter (GCIB) project, Polifonia's partner Officina Futuro Fondazione will develop a 1-day laboratory to be experimented at the premises of the Polifonia partners.

In principle, the data collection is not aiming at collecting sensitive data, however personal data may be collected. This data collection processes will strictly adhere to the GDPR regulation. All personal data will be anonymised. In the case it is unavoidable to collect identifiable data (e.g. contact details for participants taking part in multiple sessions), these will be removed at the earliest opportunity.

Each partner that is subject to the obligation to appoint a DPO according to the GDPR, has appointed a Data Protection Officer (DPO). The contact details of the DPO will be available to all data subjects involved in the research lead by that specific partner (qualifiable as data controller) through the informed consent and the information sheet. In case there is any doubt about the procedures to follow, an independent Ethics Advisor will be consulted prior to the pilot study activities.

3.5.3 Informed consent

As a rule, the Polifonia project will only accept participants for studies who are fully able to give informed consent, or, in case of minors, consent will be gained from the parents, guardians, or legal representatives. All participants will be provided with an information sheet and, after being informed about the project, asked to sign a consent form. The consent form as well as the information sheet will clearly state the purpose of the project and the methodology used. It will also describe how the data will be stored who has access to it. All participants are pointed to their right to withdraw and full contact details of a contact person will be provided.

4. Conclusions

Polifonia operates in a practice oriented and iterative way. In the first 6 months, the consortium created a shared conceptual space (WP1, D1.1) and identified the nature of sources, as well as main methods and approaches (Components (see section 2.1.2)) envisioned to be used during the project (WP1, D1.3.).

Obviously, there is a tension between visions what research and design questions we want to be able to answer and what we can do, based on actual sources available and time available to create new sources. In this interplay, one can expect that in practice some solutions become more widely explored, shared, and developed. In other words, in the course of the iterative research cycle, we expected that certain types of sources and components become more popular than others. Popular means here in terms of frequency of use and spread of use through different pilots and Work packages

Figure 7 depicts schematically the way from the Components to the Portal. In this figure the *popularity* of some components is indicated by a red circle. We also already envision that some components will be direct part of the Portal (Output, Sources), some of them will contribute to the building process of the Portal (Means). For the Portal itself, we expect two types of challenges for preservation to occur:

- How to preserve data, tools and functionalities for actual use? (**living data**)
- How to preserve data, tools and functionality in the mid- and long-term? (**static data**)

We indicate this by the two containers (in Figure 7): sustainability and preservation (here meant as long-term preservation).

Institutional commitments are required to ensure sustainability in a functional, living way. For that a clear governance between all involved institutions is required. Concerning long-term preservation, archives, data repositories, software repositories are primary partners.

For both types of sustainability (mid-term and long-term), two aspects are crucial. First to make a good **selection** (content-wise) for what to be preserved and to **negotiate** (governance-wise) by whom to be preserved. Concerning 'living data' alignment inside of the Polifonia consortium (Technical Board, + WP1, WP2 in particular) is important. For the long-term preservation a discourse among stakeholders in and around the Polifonia consortium will be established.

For both processes, it is important to keep in mind, that Polifonia infrastructurally operates in distributed network way. Hence, some components (e.g., services and applications) are in hands of the Polifonia consortium, for other Polifonia might rely on other content and service providers. So, for different components there are different stakeholder groups.

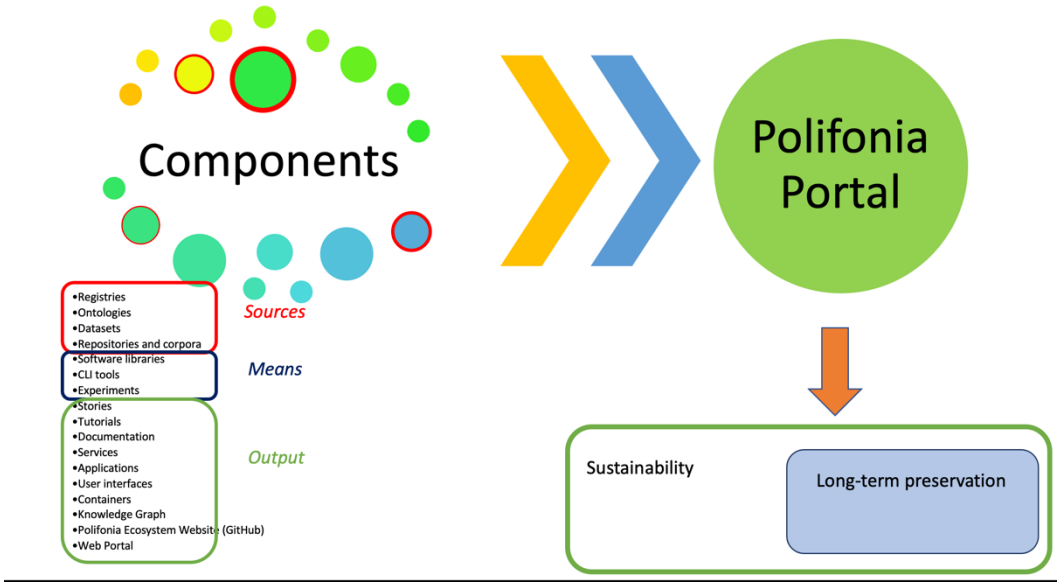


Figure 7: Schematic representation of the pipeline from Components to Portal.

For Polifonia Components we apply a layered strategy of FAIRness and Openess:

- Publications are published as much as possible in Open Access venues. In all cases pre-prints are made available via OA preprint archives; preferable Zenodo, alternatively arxiv.org; or AMSActa (UNIBO).
- Datasets produced by Polifonia and used in publications which themselves represent re-usable units will be deposited with a certified research data archive (either Zenodo, Open Research Data Online (OU), EASY (KNAW-DANS), or the Digital Repository of Ireland)
- All tooling will be published and documented in Github (with release copies in Zenodo).

Similar to the overall Polifonia research design, we will apply an iterative approach also to the making of the DMP. In Figure 8, we align the versions of the DMP to a very generic research model. This Input-Process-Output model belongs to one of the fundamental models to monitor research processes; and has been extensively applied for instance in science of science (Scharnhorst et al. 2021). The green arrows in the 'process phase' indicate the iterative working of the Polifonia. We also would like to emphasise that Polifonia also contributes classical output (publications and presentations). These output types provide relevant context for other 'output' such as dataset, data papers and so one.

As depicted in Fig 8, the first iteration of the DMP looks into the initial conditions of the project (the sources related to the Pilots).

The second iteration will shift the focus towards the Polifonia Ecosystem (the ensemble of sources and all kind of output used and produced by Polifonia). Here, we will exemplify scenarios of possible preservation strategies for selected 'data' (meaning sources and products), so we move from Pilots based data stories to data preservation stories.

The third iteration is envisioned to produce lessons-learned and best practice examples for mid- and long-term preservation again based in on a certain selection. So, second and third iteration of the DMP shift more towards the 'output side'.

The iterative process of DMP making is accompanied by webinars. In which we will update the Polifonia consortium about the current state-of-art of preservation strategies (webinar 1) and report to a wider audience about best es developed by Polifonia (webinar 2).

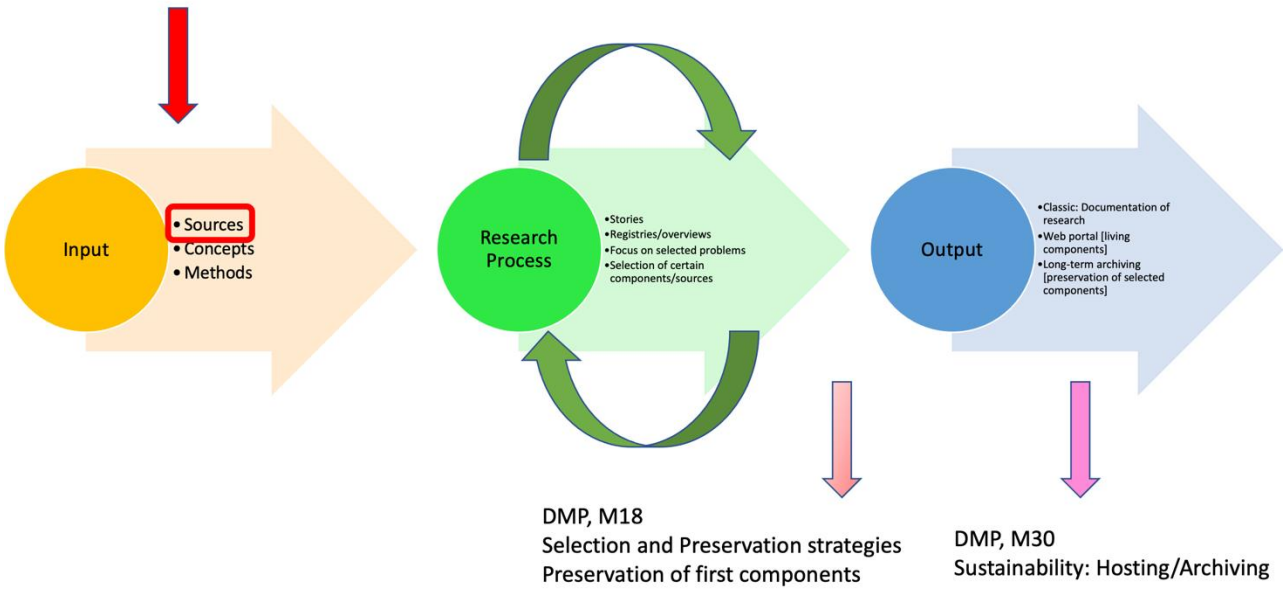


Figure 8: Alignment of the DMP iterations with the overall Polifonia research process.

References

Yasemin Bagci, Yasemin, Jesse de Vos and Nicole Emmenegger. 2021. Polifonia D6.1. Plan for Dissemination, Exploitation and Communication (draft).

Beaulieu, A., Ratto, M., & Scharnhorst, A. 2013. Learning in a landscape: simulation-building as reflexive intervention. Mind & Society, 12(1), 91–112. doi:10.1007/s11299-013-0117-5 (Preprint available at http://arxiv.org/abs/1108.5502)

Borgman, C. 2015. Big data, little data, no data. Scholarship in the networked world. MIT Press.

Corcho, O., Eriksson, M., Kurowski, K., Choirat, C., Van de Sanden, M., & Coppens, F. 2021. *EOSC Interoperability Framework. Report from the EOSC Executive Board Working Group (WG) FAIR and Architecture*. Brussels. https://doi.org/10.2777/620649

Guillotel-Nothmann, Christophe et al. 2021. Polifonia D1.1. Roadmap and pilot requirements 1st version (draft).

Daga, Enrico et al. 2021. Polifonia D1.3. Pilots development – collaborative methodology and tools (draft).

Hugo, W, Le Franc, Y, Coen, G, Parland-von Essen, J & Bonino, L 2020, D2.5 FAIR Semantics Recommendations Second Iteration. Zenodo. https://doi.org/10.5281/zenodo.4314321

RelatedWorkOntologies (2021), Working document of Polifonia WP2. Internal (https://liveunibo.sharepoint.com/:w:/r/sites/polifonia/_layouts/15/Doc.aspx?sourcedoc=%7B47E9B027-21C1-40E1-8128-C664139CDF2D%7D&file=RelatedWork-Ontologies.docx&action=default&mobileredirect=true)

Reichmann, S., Klebel, T., Hasani-Mavriqi, I., & Ross-Hellauer, T. (2021). Between administration and research: Understanding data management practices in an institutional context. *Journal of the Association for Information Science and Technology*, asi.24492. https://doi.org/10.1002/asi.24492

Scharnhorst, Andrea, Katy Börner, and Peter Besselaar (Eds.). 2012. Models of Science Dynamics. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23068-4.

Siebes, Ronald, Coen, Gerard, Gregory, Kathleen, & Scharnhorst, Andrea. (2019, October 3). Linked Open Data - 10 Things to reach the realm of LOD (part of TOP 10 FAIR DATA & SOFTWARE THINGS). Zenodo. http://doi.org/10.5281/zenodo.3471806

Poveda-Villalón, María, Paola Espinoza-Arias, Daniel Garijo, and Oscar Corcho. "Coming to Terms with FAIR Ontologies." In International Conference on Knowledge Engineering and Knowledge Management, pp. 255-270. Springer, Cham, 2020.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

W3C Recommendation 31 January 2017: Data on the Web Best Practices;
https://www.w3.org/TR/dwbp/

Appendix I: Overview of the main sources used in the pilots

	Pilot	Source	Link/reference	File size	Metadata model/format[27]	File format(s)	Access rights	Licence/ copyright	API
1	BELLS	General Catalogue of Cultural Assets	http://www.iccd.beniculturali.it				open	CC BY-SA 4.0 (data and content)	
		Patrimonio Culturale Immateriale (catalogue sheets)	http://paci.iccd.beniculturali.it/iccd/cards/ricercaPaci?data%5Bkeyword%5D=tecnica+campanaria&data%5BricercaLiberaInventario%5D=Cerca&data%5Bdoaction%5D=cerca				open	CC BY-SA 4.0 (data and content)	
		Campanologia.it	https://www.campanologia.it/			html	open	copyright	
		Campane.org	http://www.campane.org/			html	open	copyright	
		Linee guida per la tutela del Patrimonio Campanario	http://www.archeobologna.beniculturali.it/pdf/Protocollo%20di%20Int	8,2 MB	N/A	pdf	open	copyright author	N/A

[27] While we recognize the difference between a metadata model and the file format in which it comes, we decided to conflate this information in the table. For many sources, it is still unknown which metadata models are adhered to (if any), but for some, we already do know in which file format the metadata is stored.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746

			esa_Patrimonio%20Campanario%20Storico.pdf						
2	ORGANS	Het Historische Orgel in Nederland(1997-2010)	https://vu.on.worldcat.org/oclc/782207831	60 MB	N/A	doc, wp5	restricted	copyright NiVO	N/A
3	FACETS	Neuma digital library	http://neuma.huma-num.fr/			MusicXML, MEI, ly, png,pdf	registered users/open		yes
		Irish Traditional Music Archive (Port collections)	http://port.itma.ie/welcome						
4	INTERLINK	Dutch Song Database	www.liederenbank.nl	630 MB	Filemaker Database	Filemaker Database	open	https://www.meertens.knaw.nl/cms/nl/over-het-meertens-instituut/disclaimer	no
		MTC-FS-INST-2.0	www.liederenbank.nl/mtc	1.3 GB	csv	**kern, midi, pdf, txt, lilypond source	open	CC BY-NC-SA 3.0	no
		MTC-ANN-2.0.1	www.liederenbank.nl/mtc	84 MB	csv	**kern, midi, pdf, txt, lilypond source	open	CC BY-NC-SA 3.0	no
		Melodic Features for Meertens Tune Collections and ESSEN Folksong Collection	https://zenodo.org/record/3551003	81.6 MB	N/A	json	open	CC BY-NC-SA 3.0	no

		Stichting Omroep Muziek catalogue	http://muziekschatten.nl						
		Musopen	https://musopen.org/						
		CLARIAH MediaSuite	https://mediasuite.clariah.nl/						
		Common Thesaurus for Audiovisual Archives	https://old.datahub.io/dataset/gemeenschappelijke-thesaurus-audiovisuele-archieven						
		Neuma digital library	http://neuma.huma-num.fr/			MusicXML MEI. ly, png,pdf	registered users		yes
		Irish Traditional Music Archive (Port collections)	http://port.itma.ie/welcome						
		Irish Traditional Music Archive (LITMUS)	https://www.itma.ie/litmus						
		RISM Catalog of Musical Sources	https://rism.info	6.4 GB	XML	MARC-XML, RDF	open	CC-BY	no
5	CHILD	Listening Experience Database	https://led.kmi.open.ac.uk						
		Internet Archive	https://archive.org						

		Project Gutenberg	https://www.gutenberg.org						
6	MUSICBO	Lessico Beni Culturali	http://corpora.lessicobeniculturali.net/					CC BY-NC-SA 4.0	
		Grove Online Dictionary	https://www.oxfordmusiconline.com/				registered users	copyright Oxford University Press	no
		Retrospective Index to Music Periodicals	https://www.ripm.org/?page=ROAoverview						
		Internet Archive Open Library	https://openlibrary.org/						
		Gallica	https://gallica.bnf.fr						
		Biblioteca Universitaria di Bologna	https://bub.unibo.it/it/bub-digitale						
		Biblioteca della Musica	http://www.bibliotecamusica.it/cmbm/scripts/gaspari/src_com.asp						
		Biblioteca Conservatorio P. Martini	http://www.consbo.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/23						
		Bibliothèque Nationale	https://www.bnf.fr/fr/acces-aux-catalogues-numerises						
		Biblioteca Nacional España	http://www.bne.es/es/Catalogos/BibliotecaDigit						

			alHispanica/Inicio/index.html						
7	TUNES	Dutch Song Database	www.liederenbank.nl	630 MB	Filemaker Database	Filemaker Database	open	https://www.meertens.knaw.nl/cms/nl/over-het-meertens-instituut/disclaimer	no
		MTC-FS-INST-2.0	www.liederenbank.nl/mtc	1.3 GB	csv	**kern, midi, pdf, txt, lilypond source	open	CC BY-NC-SA 3.0	no
		MTC-ANN-2.0.1	www.liederenbank.nl/mtc	84 MB	csv	**kern, midi, pdf, txt, lilypond source	open	CC BY-NC-SA 3.0	no
		Melodic Features for Meertens Tune Collections and ESSEN Folksong Collection	https://zenodo.org/record/3551003	81.6 MB	N/A	json	open	CC BY-NC-SA 3.0	no
		Neuma digital library	http://neuma.huma-num.fr/			MusicXML, MEI, ly, png,pdf	registered users		yes
		Irish Traditional Music Archive (Port collections)	http://port.itma.ie/welcome						
		RISM Catalog of Musical Sources	https://rism.info	6.4 GB	XML	MARC-XML, RDF	open	CC-BY	no

		Essen Folk Song Collection	http://www.esac-data.org/		txt	txt (ascii) : Essen associative code	open		no
		Tune index from The Colonial Institute	https://www.cdss.org/elibrary/Easmes/Index.htm	11 MB	HTML	txt (ascii) : HTML	open		no
		Barlow and Morgenstern Collection	https://onesearch.library.uwa.edu.au/permalink/61UWA_INST/1iju3hj/alma9998597202101	51 MB	N/A	txt (ascii) : **kern	open		no
		ABC notation	http://abcnotation.com/	unknown	txt	txt (ascii) : abc encoding	open		no
8	TONALITIES	NEUMA	http://neuma.huma-num.fr/			MusicXML, MEI, ly, png,pdf	registered users		yes
		Dutch Song Database	www.liederenbank.nl	630 MB	Filemaker Database	Filemaker Database	open	https://www.meertens.knaw.nl/cms/nl/over-het-meertens-instituut/disclaimer	no
		MTC-FS-INST-2.0	www.liederenbank.nl/mtc	1.3 GB	csv	**kern, midi, pdf, txt, lilypond source	open	CC BY-NC-SA 3.0	no
		MTC-ANN-2.0.1	www.liederenbank.nl/mtc	84 MB	csv	**kern, midi, pdf, txt, lilypond source	open	CC BY-NC-SA 3.0	no

		Melodic Features for Meertens Tune Collections and ESSEN Folksong Collection	https://zenodo.org/record/3551003	81.6 MB	N/A	json	open	CC BY-NC-SA 3.0	no
9	ACCESS	Live and recorded performances at The Stables Theatre							
10	MEETUPS	Listening Experience Database: https://led.kmi.open.ac.uk							
		Internet Archive	https://archive.org						
		Project Gutenberg	https://www.gutenberg.org						
		Biblioteca Italiana	http://www.bibliotecaitaliana.it						
		BNF France	https://www.bnf.fr/en/bibliotheque-nationale-de-france-catalogue-general						