Documenting the research process. Opportunities and challenges for Bibliometrics and Information Retrieval

Daga, Enrico ; Daquino, Marilena; Fournier-S'niehotta, Raphaël; Guillotel-Nothmann, Christophe; Scharnhorst, Andrea

published in
Proceedings of the 13th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 45th European Conference on Information Retrieval (ECIR 2023)

2023

DOI (link to publisher)
urn:nbn:de:0074-3617-4

document version
Publisher's PDF, also known as Version of record

Documenting the research process. Opportunities and challenges for Bibliometrics and Information Retrieval

Enrico Daga[1,†], Marilena Daquino[2,†], Raphaël Fournier-S'niehotta[3,†], Christophe Guillotel-Nothmann[4,†] and Andrea Scharnhorst[5,*,†]

[1]*Knowledge Media Institute, The Open University; Walton Hall, Milton Keynes, MK7 6AA, United Kingdom*
[2]*University of Bologna; Dipartimento di Filologia Classica e Italianistica Via Zamboni 32, Bologna, Italy*
[3]*CEDRIC Laboratory, CNAM Paris, 75003 Paris, France*
[1]*CNRS Délégation Paris B: Paris, Île-de-France, France*
[5]*Royal Netherlands Academy of Arts and Sciences, Data Archiving and Networked Services*

Abstract

This paper reports about knowledge management experiences in the EC funded project **Polifonia** (Research and Innovation Action funding scheme). Polifonia is a challenging project which aims at developing a methodological framework for musical heritage information. The project encompasses sources from text, sound, scores, settings (buildings), and experiences. It is organized around 10 Pilots which cover various actions such as preserving, studying, managing and interacting with musical heritage. Its advantage is that it uses semantic web technologies (ontologies and resulting knowledge graphs) as *lingua franca* binding the different Pilot data together. More specifically and additionally to the common use of GitHub repositories in research projects, Polifonia adds an additional organizational structure, what we call a *Research Ecosystem*. The Polifonia Research Ecosystem documents project outputs and their mutual dependencies as semantic artifacts, developing annotations for both (output and dependencies). This paper details the design and implementation of such a Research Ecosystem as a specific approach to effectively coordinate collaboration and related software production. Using the case of the Polifonia project, the paper reflects on the opportunities and challenges arising when it comes to formalize best practices to execute innovative research processes. Finally, we discuss the potential impact that such developments could have on future bibliometrics and information retrieval practices.

Keywords

ontology engineering, knowledge organisation, interdisciplinary collaboration, project-based research, information retrieval

BIR 2023: 13th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2023, April 2, 2023
*Corresponding author.
†These authors contributed equally.
andrea.scharnhorst@dans.knaw.nl (A. Scharnhorst)
http://www.enridaga.net/about/#human (E. Daga); http://raphael.fournier-sniehotta.fr (R. Fournier-S'niehotta)
0000-0002-3184-5407 (E. Daga); 0000-0002-1113-7550 (M. Daquino); 0000-0002-9137-8011 (R. Fournier-S'niehotta); 0000-0002-4817-3686 (C. Guillotel-Nothmann); 0000-0001-8879-8798 (A. Scharnhorst)

1. Introduction

From the earlier days of measuring the sciences (here meant as all academia) there has been a need to be able to document not only the output of research (as done in formal scholarly communication) and the boundary conditions under which it operates, but also the whole research cycle. As knowledge production is so fundamental for knowledge-based societies this fascination cannot not surprise [1]. The fact that most research is funded by public expenditures [2, 3] adds other aspects to such a curiosity, namely: the need to access and manage research funds, and the duty to be accountable and open to society (see e.g., [4, 5]). The paradigm of Open Science encourages all actors involved in research to seek ways to open up the black-box of research itself. In turn, the World Wide Web provides the technology to execute such dissections on an unprecedented scale.

Knowledge organisation, and with it knowledge organisation systems, have been at the heart of any knowledge production. We cannot think without organising, and each scientific community, specialty, or domain, develops its own way to settle knowledge. This aspect is intrinsically linked to the production of 'new ideas', for which 'new ways to think about problems' are needed [6]. But with domain-dependent ways of thinking, also come specific ways to determine what the valuable assets are, and how to properly document them in order to enable scholarly communication. An archetypal example is the long lasting difference in formal scholarly output venues between humanities fields (using books), natural sciences (using journal articles) and computer sciences (using conference proceedings). Likewise, *data* undergo similar diverse processes when being documented, which leads to a diversity of meta-data schemes, formats, and data management procedures, making re-use across domains harder. Structured data are the fuel of automatic processes, and data re-use is the mantra of FAIR research [7]. Indeed, documenting the life-cycle of data artifacts is of paramount importance in many fields and has contributed to the recent resurgence of provenance research, for example, with the notion of *Data Journey*, which aims at describing processes from the point of view of the data objects and their reuse [8, 9]. In short, documentation is first meant to serve domain-dependant requirements, which may clash with requirements from other domains. It has been claimed that 'what is data for one is noise for another' [10]. In fact, the growing specialisation in the science system on one side and the need for interdisciplinary work on the other side, makes the alignment of domain-specific descriptions of research assets a challenge for many, often hampering a broad re-use of assets and procedures across projects.

Information retrieval which is key to reuse relies on various forms of documentation. In this article, we report about a way to conceptualise and formally document research assets in interdisciplinary research. We leverage common ground practices on how to best organise cross-,inter-, transdisciplinary work in project based funded research. In the implementation of our approach, we embrace the decades-long standardisation methods in research documentation, starting with bibliographical documentation of formal scholarly communication (based on standards such as Dublin core, CERIF, FBRB) [1], up to recently proposed standards to ensure even more fine-grained documentation of the 'elementary particles' of research (RO-CRATE [11], and

[1] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/,https://eurocris.org/services/main-features-cerif,http://www.sparontologies.net/ontologies/frbr

others). Our work is situated at the crossroad between knowledge engineering and management (inspired by Open Software development), science and technology studies (to understand interdisciplinary collaboration) and scientific documentation (developing FAIR standards to enable organisation of and navigation through the growing body of scientific knowledge).

We use the term *Research Ecosystem* to describe the actual research process and its living documentation. Ecosystem means decomposing scholarly activities and elements into assets, and to define the links among them in a machine-readable way. These links concern the different functions and mutual dependencies of those assets . By such a decomposition, an Ecosystem approach supports the retrieval of more granular elements (assets such as data, software, etc...) and facilitates their (re-)use. But, the Ecosystem approach is not exclusively about defining the assets. The development of standardized documentation of assets of the research process is a necessary ingredient, but not sufficient to construct an Ecosystem perspective. Next to data and software (currently in the focus of open access) other important assets are requirements, methods, and even epistemic frameworks. The latter determine what the knowledge units are. In an Ecosystem the 'atomic parts' are bound together in functional units: i.e. data records into collections, software code into applications, user needs into interfaces and so on. This is why we later talk about *components* rather than assets. The emphasis is on the inter–dependencies of components – their role in the overall function of the research process – and on how to determine the right level of deconstructed metadata to describe components and their functions. Our work focuses on an approach to extend documentation towards ingredients and choices, usually invisible, part of a 'research culture' and only embedded in the tacit knowledge of those researchers which collaborate together. In other words, our ambition is to describe the process of interdisciplinary work in a formalised way, to enable a better exchange of best practices of knowledge production in interdisciplinary collaboration.

We demonstrate the design and implementation of a Research Ecosystem for the case of a concrete EC funded project. Walking the reader through our experiences in this specific case, we try to highlight those parts of the Ecosystem design which are generic, and reproducible. We believe that each new research collaboration will need to create its own, unique, Ecosystem. Nonetheless, we also believe that types of components and inter-dependencies are similar for each research collaboration. By presenting an example in which the Ecosystem concept is connected to a formal (ontological) and tangible implementation, we hope to deliver a template on how to develop a Research Ecosystem. The documentation of research processes opens new perspectives also for bibliometrics, and has the potential to contribute to better understanding the pathways towards creating an impact, impact in the project itself, in science beyond one specific project, and maybe even in society.

The remainder of the article is the following. The main section is devoted to the Research Ecosystem. We briefly introduce the Polifonia project, the idea of a Research Ecosystem, its implementation in GitHub, and the ongoing evolution of the Ecosystem. We close this paper with lessons learned and future work.

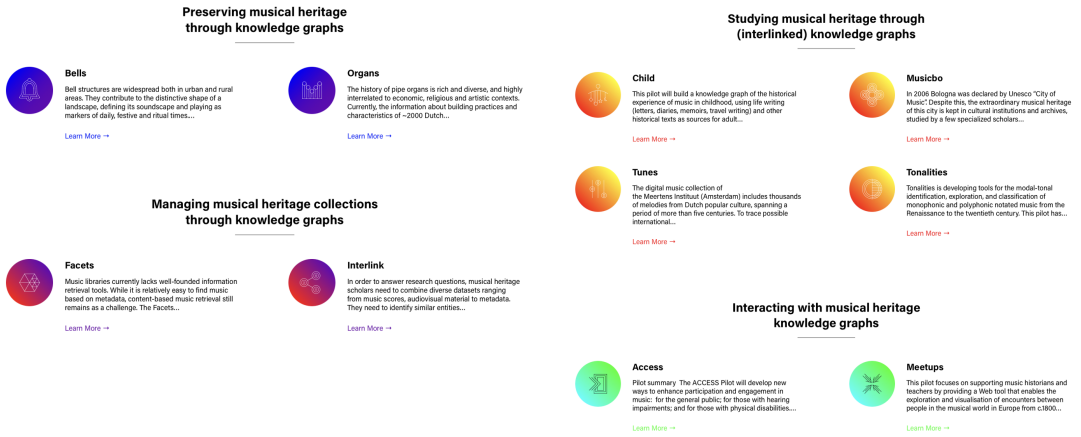

Figure 1: A summary of the Polifonia Pilots

2. The Polifonia Research Ecosystem

2.1. The Polifonia Project - an example of interdisciplinary collaboration

Polifonia is a *EC Research and Innovation Action* funded project [2]. Its goal is to highlight the evolution of European musical heritage in space and time. Themes addressed are as diverse as the transition of music genres across borders, the experiences of music in childhood, and the tracing of musical ideas through musicians' encounters in history. Polifonia is organised around ten Pilots (see Fig. 1), which are very diverse in types of source material used, including texts, scores, sound, images, video and material objects (see Fig. 2). The project consortium brings together anthropologists and ethnomusicologists, historians of music, linguists, musical heritage archivists, cataloguers and administrators, and creative professionals [3]. However, computer science, and more specifically, ontology engineering and semantic web applications, are responsible for the technological backbone of the project, and their methods form the *lingua franca* used to enable an integration of knowledge. This integration covers the Pilots, enriching - in turn - their source material and generally enhancing the access to European musical heritage. But, it also entails the organisation of the knowledge production in the project itself.

From the start, Polifonia took great care to build into the project means to navigate the complexity of assets one can notice in the diversity of Pilots scopes and goals, as well as the different epistemic frameworks from which domain specialists contribute to the project (see Fig. 3). Special care was taken to enable clear and effective collaboration and communication. This effort spans from a specific graphic language (e.g., Pilots are color coded according to their function in the knowledge production life-cycle - see the look and feel of the Polifonia website, and Fig. fig:pilots), to a detailed description of the collaborative methodology based on three methods: Stories, collaborative sessions called *Maninpasta*, and Surveys. Those three methods

[2] running from January 2021 to April 2024
[3] https://zenodo.org/communities/polifonia/about/

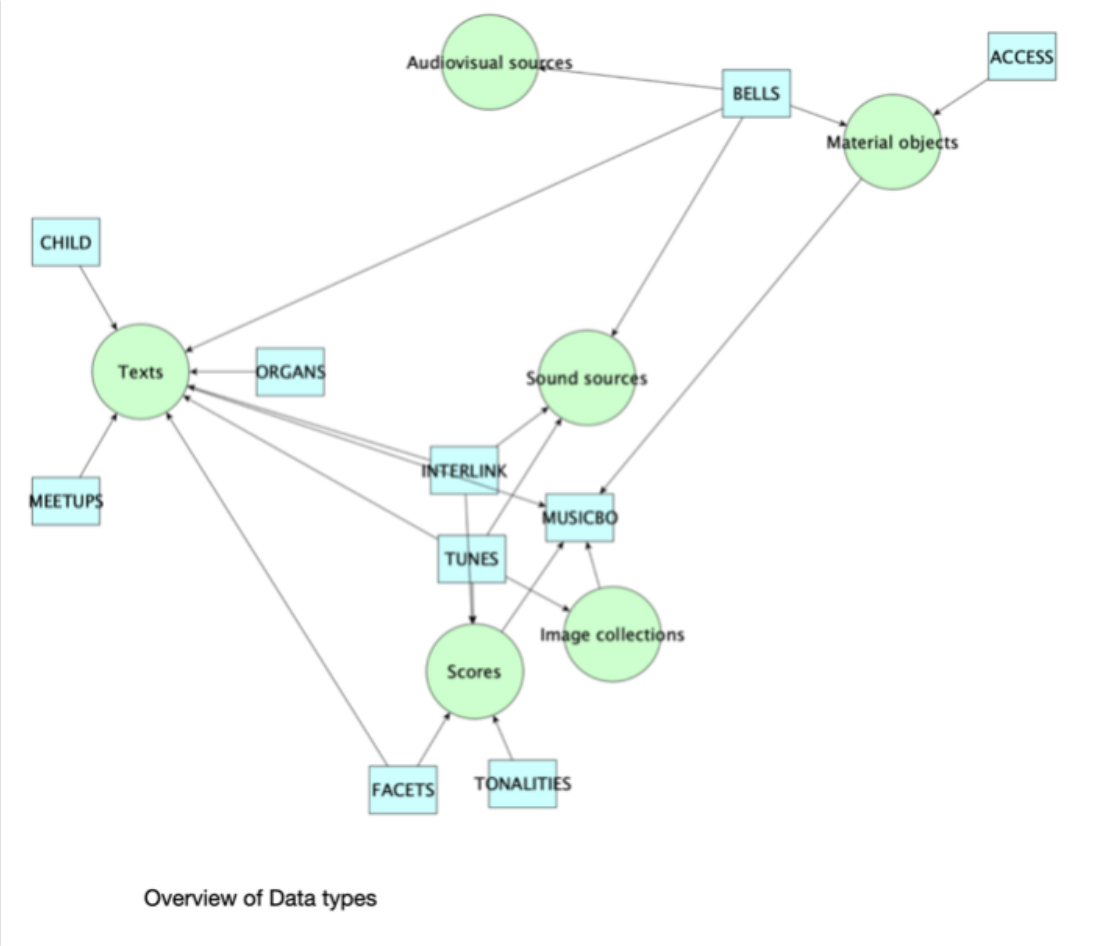

Figure 2: Data types and sources used by Polifonia Pilots - reproduced from D1.1

(also abbreviated with *SMS*) allow us to describe requirements of archetypal stakeholders (Stories), address them in an agile way via hackathons and working groups (*Maninpasta*), accompanied by an iteration of internal surveys addressing conceptual (epistemic) and practical (data, tools) questions[12, 13]. One outcome of the survey were two network visualisation (called *Polifonia Atlases*) which represent the types of data sources used by the Pilots in the project (Fig. 2), and the variety of epistemic dimensions depicted as *knowledge units* to which the project wants to answer (Fig. 3).

Additionally, the project uses the Data Management Plan (DMP) as an element for coordination of knowledge flows inside the project. Three iterations or versions of the DMP are planned: at the beginning, in the middle, and in the last quarter of the project [14, 15].

To summarise, knowledge management in Polifonia is organised along three axes or perspectives: documentation and bibliography, perspective of audiences (inside and outside of the project) and their information needs, and knowledge engineering. In Polifonia, the first axis is mostly addressed by the DMP, the second by the design of an all-in-one Web Portal (which

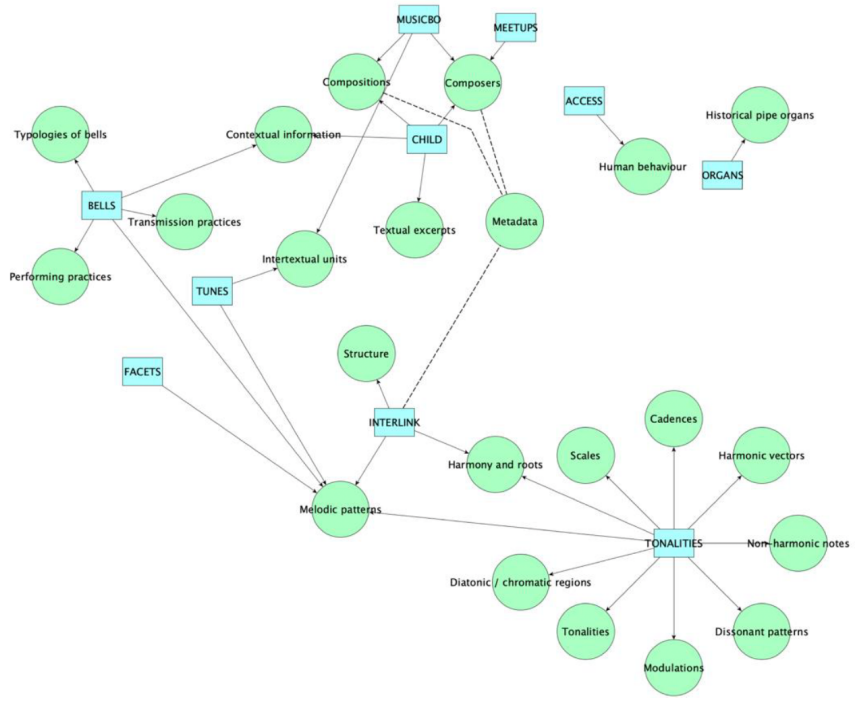

Diagram 2: Survey – Knowledge units investigated by the pilots

Figure 3: Knowledge units representing the different epistemic perspectives of the Pilots - reproduced from D1.1

is still in the making), and the third by the creation of a Polifonia Ontology Network and the Polifonia Knowledge Graphs. The *Polifonia Research Ecosystem* binds all those perspectives, methods and results of knowledge management together in a digital environment where assets and procedures are meaningfully inter-linked and evident to future adopters of research outputs [16].

2.2. The Research Ecosystem design

The Research Ecosystem has been introduced already in the project proposal narrative as an important means to ensure the collaborative work inside of Polifonia. The Description of Action states "Polifonia implements a digital ecosystem for European Musical Heritage: music objects along with relevant related knowledge about their cultural and historical context, expressed in different languages and styles, and across centuries. The ecosystem will include methods, tools, guidelines, experiences, and creative designs, openly shared according to F.A.I.R. principles."[4].

[4]personal communication, DoA document has not been made public

An early deliverable ("Roadmap and pilot requirements 1st version") elaborated further on the design of the Ecosystem: "[the aim 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" [12].

The Polifonia Research Ecosystem can be seen as a set of technologies which effectiveness is shown through the Pilots. "The added value of the Ecosystem – compared to delivering the sum of the pilot applications – will be demonstrated through reuse and interoperability (of software and data) among the different pilots." [12]. In other words, paradigms like FAIRness [7], Open Science [17] and Reproducible Science [18] for the further motive behind the Ecosystem.

The essence of the metaphor of an Ecosystem lies in thinking of a complex system build on **components** which interact in specific ways. In nature, "an ecosystem (or ecological system) consists of all the organisms and the abiotic pools (or physical environment) with which they interact. The biotic and abiotic components are linked together through nutrient cycles and energy flows." [5] In a Research Ecosystem, the components can be as small and fine–grained as needed (such as data, or specific tools) and as large as functional (whole stories of use, whole collections, whole workflows). How the Ecosystem materialises - which components and links are defined - will depend on the concrete research quest. But, in general, any classification or compartmentalisation of research assets is determined by the final **functions** they have in the knowledge production. The other advantage of thinking about knowledge organisation in terms of an *Ecosystem* approach is its flexibility. Ecosystems are no machines but organisms, they evolve, adapt, both as systems themselves as well as concerning their inner structure. We show how this 'evolution' happens later on.

In the original design (see Fig. 4) the components are marked as (yellow) boxes, and grouped into three main columns, pointing out both their type and function for the project. One should read Figure4 from the right to the left. The components at the very right column (Collection, Transformation and Linkage) represent (groups of) research assets and main actions executed around them which form building blocks for both research and information retrieval. The middle column (Development and Deployment) focuses on components relevant to research engineer the technological backbone of the project. The left column (Access, Discovery and Experience) contains all those components which are relevant to communicate the project results to its audiences: the domain specific experts (scholars), other research engineers (developers) and the public (citizens). Components in columns in the design schema follow a kind of temporal order. Reading figure 4 from right to left means also that components to the right are more likely to be build earlier in the research process, while some of the components to the left (e.g., interfaces) will be build later in the project. Still, as the many lines of inter-dependencies and actions indicates there are connections among all three columns back and forward. For instance, *Stories* guide the dissemination/interactive interface part and are therefore placed in the very left column. But, how *Stories* were conceived has been informed by the component *Registries and Repositories* on the very right, and the making of *User Interfaces* in the middle is guided by them too.

By compartmentalising the research process itself in this way, the aim is to create a formal yet flexible inner organisation of the project. This flexibility is not a side-product of the usual

[5] https://en.wikipedia.org/wiki/Ecosystem

Figure 4.1: The Polifonia Ecosystem: overview of the components and main relations. Connection points in black represent connections to all the components.

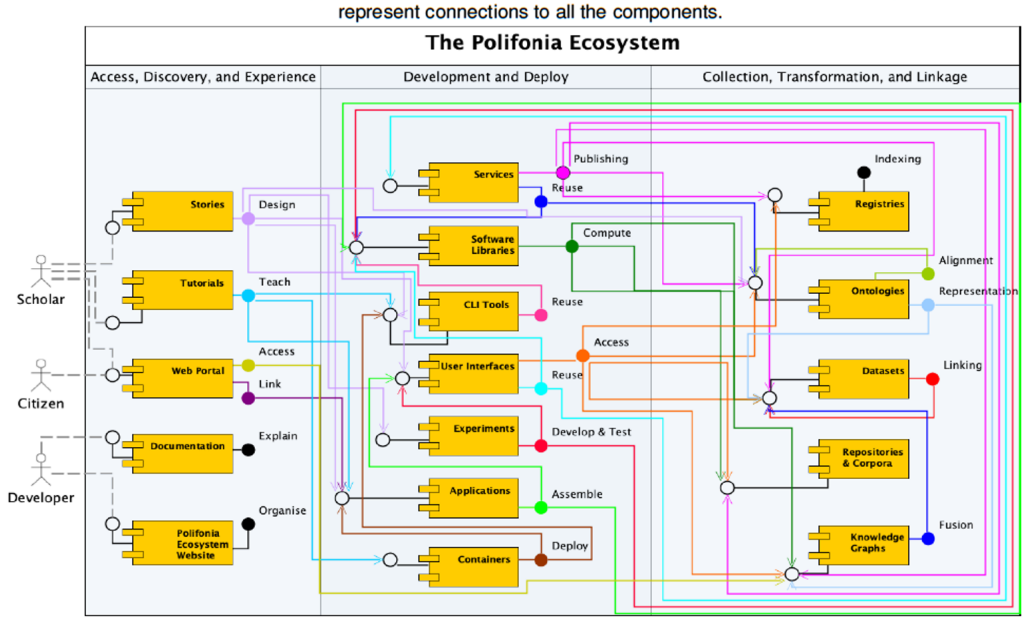

Figure 4: The original design of the Polifonia Ecosystem - reproduced from D1.3[16]

messiness of research [19], it is intentional. From the very beginning the makers of the project foresaw a tension emerging through two competing processes. One the one hand, the Pilots, they proclaimed, must be developed independently, to be useful for the research quest around each specific Pilot. Therefore, they to some extent have separate deployment of applications, Pilot-customised interfaces, services, etc.. On the other hand, and at the same time, an exchange of information and knowledge between Pilots and ultimately also an integration between them need to occur to realise synergy in the outcomes of Polifonia, and to contribute to the overall project goals. To ensure a degree of interoperability, a possible overlap of requirements from the Pilots must be captured and monitored to avoid duplication of effort and to maximise reuse inside the project and beyond. The diversity of the knowledge domains connected to the Pilots and the interdisciplinary nature of the teams involved in the process creates a challenge when it comes to keep a balance between 'research freedom' in the Pilots and research integration in the project. The answer was to create a hub for linking software, data, ontologies and interaction components - realised on a very popular platform - namely GitHub.

2.3. The Implementation of the Research Ecosystem

The implementation of the Polifonia Research Ecosystem has two layers: a governance layer and a technological layer. Development activities in the project are coordinated by a technical board and research outcomes – at any stage of their life – are collected and shared across the

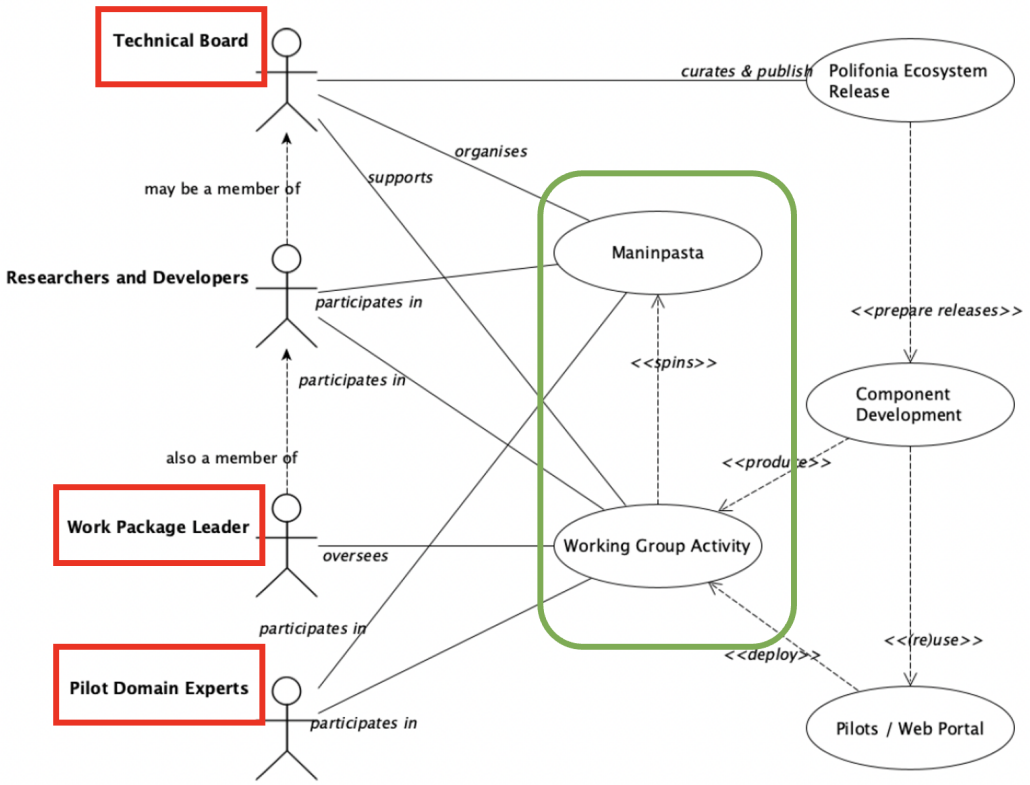

Figure 3.1: Overview of the collaborative methodology as UML use case diagram.

Figure 5: The vision on the role of the Polifonia Research Ecosystem in the overall project management - reproduced from D1.3[16]

consortium via a **GitHub organisation**[6].

The Ecosystem is interlinked with the collaborative instruments already mentioned above. Figure 5 shows (in the middle of the diagram) how agile forms such as temporal working groups and *maninpasta* (which are kind of grass-rooted) are connected to a more traditional project organisation, in terms of WorkPackages and Pilots (to the left of the diagram). The diagram (see Fig. 5) also shows how the Ecosystem acts as mediating level, which is assumed to document and coordinate the collaborative efforts. To do this, the components are not only defined, they are also annotated in a specific way. The annotation scheme covers both attributes of the type of components as well as their inter-dependencies and functions in the overall process of knowledge production. Systematic version control and releases enable re-use of development work (including releases into Zenodo as a repository).

To support such an agile but formal documentation, Git, and the GitHub platform in particular, have been chosen for tracking progresses and sharing code and results inside of the project and

[6]https://github.com/polifonia-project/

with a wide public. GitHub repositories that are part of the project organisation are collectors of research outcomes, spanning from collection requirements to documentation, experiments, and final outcomes. Notably, each repository may include one or more components, i.e. building blocks of the Research Ecosystem.

To maximise the impact of results and ensure a clear path to the preservation of results, preliminary good practices have been setup and shared across data and software providers via the GitHub organisation itself. Those are also documented in the DMP versions. [14, 15] The central element is the so-called **Rulebook**, which is itself developed in the form of a Git repository[7]. The Rulebook includes code of conduct and best practises that Working groups and Pilot teams must follow when sharing their results on the GitHub organisation. In detail, the Rulebook covers (1) the identification of responsible people for each repository (*champions* [8]), (2) specific guidelines for technical deliverable preparation[9], and data or ontology development documentation[10], and (3) the annotation schema[11].

The **annotation schema** includes the metadata set elaborated by the technical board and members of the consortium to describe cataloguing and relational aspects of Polifonia components. The annotation schema is aligned to well-known ontologies. It extends a subset of schema.org[12] predicates with terms belonging to the W3-endorsed PROV Ontology [20] and the SPAR ontologies, i.e. CiTO [21].

The metadata set is formatted as a **yaml heading** to be included in a README file (written in markdown). Each repository including a component must have at least one markdown file with the annotations' header. Repositories can include as many markdown files/yaml headings as the number of components are available in the repository itself. External components, i.e., those useful for developing the Polifonia components but not part of the Ecosystem, may also be referenced, through annotated markdown files. An example of an annotated component is the Sofware CLEF[13] [22]:

```yaml
component-id: clef
name: CLEF, Crowdsourcing Linked Entities via web Form
description: CLEF is a lightweight Linked Open Data native
    cataloguing system tailored to small-medium crowdsourcing
     projects.
type: Application
release-date: 2021-10-27
release-number: latest
work-package: WP1
keywords:
  - crowdsourcing
```

[7] https://github.com/polifonia-project/rulebook
[8] https://github.com/polifonia-project/rulebook/blob/main/CHAMPIONS.md
[9] https://github.com/polifonia-project/rulebook/blob/main/deliverable_guidelines.md
[10] https://github.com/polifonia-project/rulebook/blob/main/ontology-KG-development-documentation-guidelines.md
[11] https://github.com/polifonia-project/rulebook/blob/main/schema.md
[12] http://schema.org
[13] See also the GitHub repository:

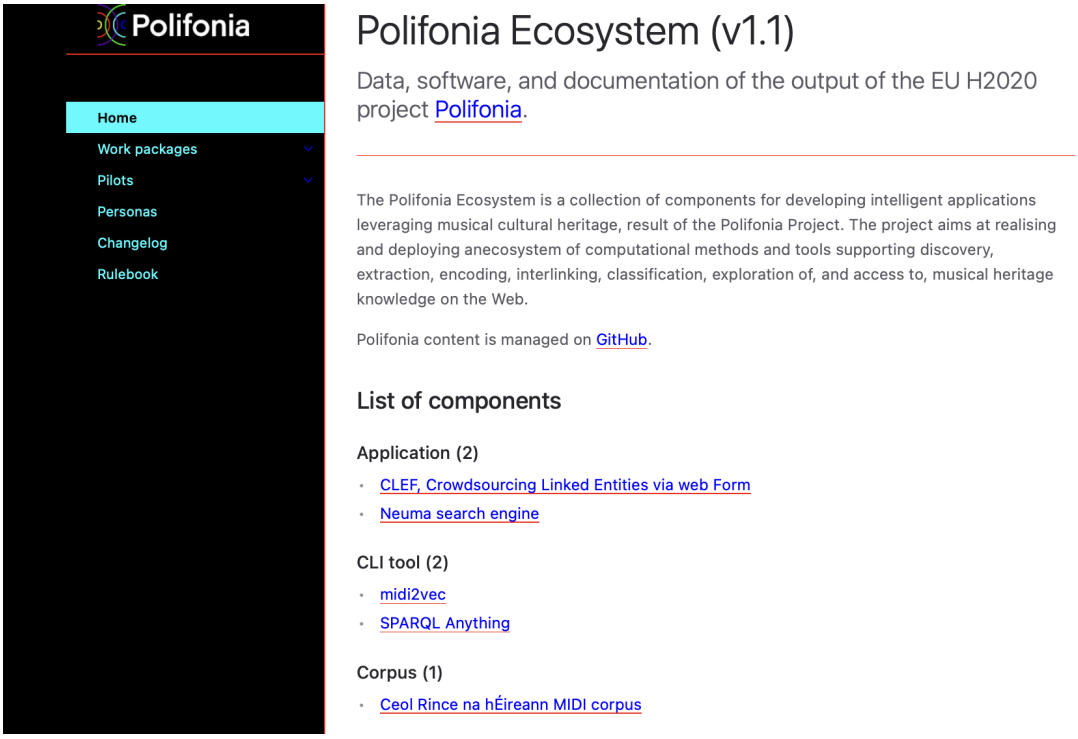

Figure 6: Snapshot of the Polifonia Research Ecosystem Website - access May 12, 2023

```
10    - linked open data
11    - registry
12  licence: ISC
13  release-link: https://github.com/polifonia-project/clef/
      releases/latest
14  demo: https://projects.dharc.unibo.it/musow/
15  credits: Marilena Daquino (UNIBO), Martin Hlosta (FFHS,
      external collaborator), Mari Wigham (NISV), Enrico Daga (
      OU)
```

The yaml headings serve two main purposes, namely: populate the Research ecosystem website and generate a knowledge graph of components in RDF.

To enable a better navigation among the components described in the various GitHub repositories under the umbrella of the Polifonia GitHub organisation[14], a website, populated with components annotations and their links, has been designed.

Fig 6 shows that a selection of component types is used as "first-class citizen" for navigation: most notably the Pilots, and Persona's[15]. But, as the website is automatically created on top of Github, its organisation is flexible too. The website lists 'approved' components under the

[14]https://github.com/polifonia-project/ecosystem
[15]Permalink https://web.archive.org/web/20230512113125/https://polifonia-project.github.io/ecosystem/

various types. To populate the Research Ecosystem website, markdown files are automatically extracted from the repositories members of an organisation and a **validity check** is performed at schema level (i.e. whether mandatory elements are included or not) and instance level (i.e. whether terms of controlled vocabularies are used appropriately). Repositories for which a release is available are preferred outputs to be mapped and displayed on the website.

2.4. The on-going evolution of the Research Ecosystem

So far, we described the concept of an ecosystem as a way to conceptualise and organise research outputs (and related activities) from the *top-down*. However, this is paired with a complementary *bottom-up* approach where elements of the annotation schema as well as the general workflows are both refined and re-evaluated. Crucially, periodic workshops occur where we analyse the current state of affairs of the Polifonia Research Ecosystem and we reflect on possible improvements. Apart from considering this as a natural part of the design process, we observe how the evolution of the schema in time has to be a critical element of the methodology, allowing the approach to incorporate unseen elements (for example, new output types) and to adapt to changes in the landscape of research practices.

Indeed, since the first implementation of the Research Ecosystem on Github, the Ecosystem itself has been in continuous evolution (see Fig. 7. This concerns foremost the subsequent *population* of the Ecosystem, meaning that components of various types are produced, annotated and published. However, due to the character of research as an open search process in a complex landscape of scientific problems [23, 24] also the Polifonia knowledge production is not 'just' an execution of a plan. As written above, in the course of the project, the consortium discusses what types of components are the most central, how they should be described, and how their inter-dependencies develop. To give an example, the attribute 'Licenses' was added to the first annotation scheme due to discussions around the Data Management Plan, the building of first demonstrators [16], and an exploration of what licences and copyright standards are currently used across sources relevant for Polifonia [15]. Working groups work on adapting the annotation scheme, and workshops are organised to implement new versions by collectively annotating selected components. Fig 7 marks some of those changes.

3. Conclusion, lessons learned and future work

The Polifonia Research Ecosystem exemplifies new ways to document the actual research process carried out by scholars from different disciplines. It focuses on the concept of an ecosystem and how this could be implemented in projects with a firm software development component. While semantic web approaches form the methodological ground how to produce and share such a documentation, our approach represents a more in-depth reflection on the process leading the Polifonia Research project. Rather than thoroughly presenting the ontology engineering work, this paper brings together elements from a discourse in fields of science philosophy and science of science how to best understand interdisciplinary work, from traditions

[16] https://polifonia-project.eu/re-watch-polifonia-presentation-at-ai-music-festival-sonar/ Permalink https://web.archive.org/web/20230512185208/https://polifonia-project.eu/re-watch-polifonia-presentation-at-ai-music-festival-sonar/

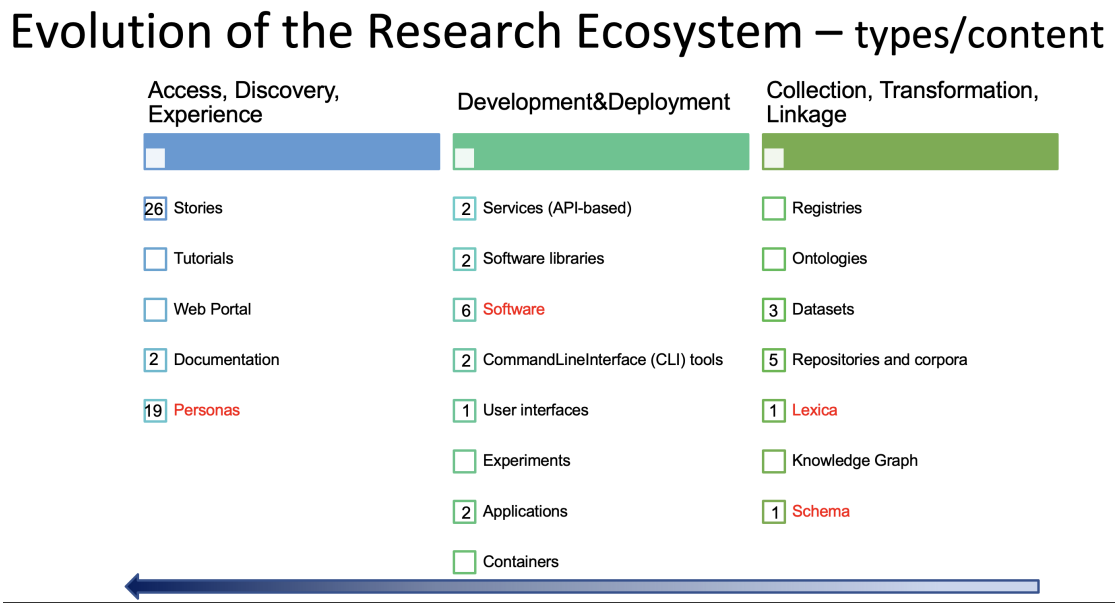

Figure 7: Overview about some of the changes in the Polifonia Research Ecosystem

of knowledge organisation as been developed in information and library sciences, and current semantic web approaches. The motivation to present this paper at the workshop 'Bibliometrics and Information Retrieval' is clear: any formalisation of elements and actions, assets and their relations, products and processes relevant for knowledge production bears in itself the potential to contribute to further standards which can be used for information retrieval. Ontology engineering practices around research assets like data, metadata standards, ontology mappings, and methods to formalise FAIR principles, help us to highlight the possibilities generated by structured information, which can be digested and recombined by machines in unforeseen ways. Rather than proposing new classification schemes of elements/assets/sources of the research process, the goal of the Polifonia Research Ecosystem is different. Its focus lies on the definition and selection of such assets (components and groups of components) which have a specific function, i.e. to answer a specific research question. We would also like to point out that, in general, the documentation of any actual research process is usually more volatile than the documentation of inputs or outputs of this process. With our work we aim at making processes, procedures, requirements, and procedural components first-class citizens of FAIR data management.

Looking at the many ways in which changes in the Ecosystem can occur (see Fig. 8) one might ask oneself if such an approach can ever scale-up and travel from project to project. We believe that it is compelling to seek a formal description the process of knowledge production itself, regardless of this being a one-size-fits-all solution. The formal description (via ontologies) in specific knowledge domains seems to grow in a breathtaking way. The same holds for formal descriptions in newly emerging interdisciplinary domains. A lot of attention is given to the organisation of cross-walks between ontologies across domains. The latter serves the

re-integration of specific domain knowledge into a bigger picture, also with the aim to enable cross-fertilisation again. But, while there is a lot of attention to connect the outcomes of knowledge production across domains, the actual organisation of interdisciplinary work is far less in the spotlight. There are some [25], but not that much attempts to formalise and document reproducible ways of working. On the contrary, each project seems to start again, relying mostly on the tacit knowledge embodied in the researchers in its consortium. This is where the idea, design and implementation of the Polifonia Research Ecosystem aspires to make a difference. It tests what can be formalised, how and how the process of formalisation should be organised.

The main take-away message is to seek for a knowledge organisation whose validation lies in its usefulness to answer certain research questions, and not (only) in the accuracy of the description of its elements. The use of the metaphor of an *Ecosystem* directs the reflection towards seeking a pragmatic compromise. One could also say, we look at the chemistry, not the physics of elementary particles, or even better the living organic adapting substructures needed to been taken care off in a reliable, firm, and reproducible way.

To summarise, the Research Ecosystem concept emphasise the emergence of substructures in the research process (workflows, groups/types of components, ...) on a mid-grain level. It concerns knowledge organisation as intermediary between large, tangible outcomes (such as a portal, a new method documented in a scientific article, or a new ontology) and the level of the elementary (seemingly generic) research objects/assets/pieces of which all research is composed of (e.g. concepts, data, methods/tool). Digitization, automatising, and Web technologies enable new forms of Open Science, reproducible science, FAIR research practices, and citizen engagement. Formalisation (knowledge engineering) also helps to break down and reframe interdisciplinary research. In each research question/each research project, generic elements re-appear at different times of the design and development (e.g. actors, concepts, competences, methods and tacit knowledge, data and documentation types). These are the 'particles of research' and we consider them first-class assets in the Ecosystem.

Finally, Research Ecosystem means deconstructing scholarly activities, and making any asset citable. Ecosystem also means supporting the retrieval of granular elements (data, software, etc...) and their use in real-world settings. Imagining this at scale would open new perspectives for bibliometrics, possibly opening new quantitative or qualitative ways to define and examine pathways to create impact. The Ecosystem idea brings the challenge of finding the right level of abstraction. To find the right level of granularity one needs to look at assets and their dependencies. How much detail is needed and what are the truly important building blocks? What a component or a link is might be different for each new research project, each time, in each domain, with each new research question. The Polifonia Research Ecosystem approach provides us with a conceptual framework (components and relations/functions), together with an implementation (based on GitHub, Zenodo, and Linked Data) and a set of practices and lessons learnt on the way to communicate and to experiment, both at meta-level and in a real-world scenario.

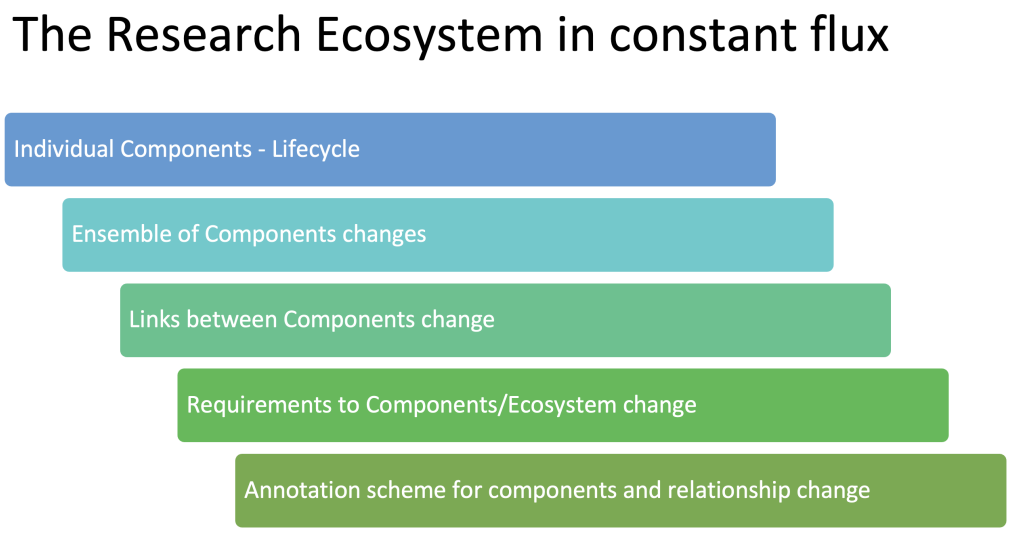

Figure 8: Layers on which changes in the Ecosystem occur

Acknowledgments

This work is supported by a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746 (Polifonia: a digital harmoniser for musical heritage knowledge, H2020-SC6-TRANSFORMATIONS).

References

[1] L. Leydesdorff, Knowledge-based innovations and social coordination, in: L. Leydesdorff (Ed.), The Evolutionary Dynamics of Discursive Knowledge: Communication-Theoretical Perspectives on an Empirical Philosophy of Science, Springer International Publishing, 2021, pp. 1–35. doi:10.1007/978-3-030-59951-5_1.

[2] Y. Yin, Y. Dong, K. Wang, D. Wang, B. F. Jones, Public use and public funding of science, Nature Human Behaviour 6 (2022) 1344–1350. URL: https://doi.org/10.1038/s41562-022-01397-5. doi:10.1038/s41562-022-01397-5.

[3] K. Aagaard, P. Mongeon, I. Ramos-Vielba, D. A. Thomas, Getting to the bottom of research funding: Acknowledging the complexity of funding dynamics, PLOS ONE 16 (2021) e0251488. doi:10.1371/journal.pone.0251488.

[4] R. von Schomberg, A vision of responsible research and innovation, in: M. H. Richard Owen, John Bessant (Ed.), Responsible Innovation, John Wiley & Sons, Ltd, 2013, pp. 51–74. doi:10.1002/9781118551424.ch3.

[5] L. Bezuidenhout, The relational responsibilities of scientists: (re) considering science as a

practice, Research Ethics 13 (2017) 65–83. URL: https://doi.org/10.1177/1747016117695368. doi:10.1177/1747016117695368.

[6] A. Scharnhorst, R. P. Smiraglia, The need for knowledge organization. introduction to the book, in: R. Smiraglia, A. Scharnhorst (Eds.), Linking Knowledge, Nomos Ergon, 2021, pp. 1–23. doi:https://doi.org/10.5771/9783956506611-1.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018.

[8] S. Leonelli, N. Tempini (Eds.), Data journeys in the sciences, Springer Cham, 2020. doi:https://doi.org/10.1007/978-3-030-37177-7.

[9] E. Daga, P. Groth, Data journeys: explaining ai workflows through abstraction, Semantic Web (2023) Early–Access.

[10] C. Borgman, Big Data, Little Data, No Data, MIT Press, 2017.

[11] S. Soiland-Reyes, P. Sefton, M. Crosas, L. J. Castro, F. Coppens, J. M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, et al., Packaging research artefacts with ro-crate, Data Science 5 (2022) 97–138.

[12] C. Guillotel-Nothmann, T. Bottini, V. Carriero, J. Carvalho, P. Cathé, F. Ciroku, E. Daga, M. Daquino, A. Davy-Rigaux, M. Gurrieri, P. Van Kemenade, E. Marzi, V. Presutti, A. Scharnhorst, D1.1 Roadmap and pilot requirements 1st version, Technical Report, 2021. doi:10.5281/zenodo.7124253.

[13] C. Guillotel-Nothmann, J. De Berardinis, T. Bottini, P. Cathé, E. Daga, M. Daquino, A. Davy-Rigaux, M. Gurrieri, P. Van Kranenburg, E. Marzi, J. McDermott, A. Meroño Peñuela, P. Mulholland, A. Scharnhorst, R. Tripodi, D1.2 Roadmap and pilot requirements 2nd version, Technical Report, 2022. URL: https://doi.org/10.5281/zenodo.7116561. doi:10.5281/zenodo.7116561.

[14] F. Admiraal, A. Scharnhorst, E. Daga, M. Daquino, E. Musumeci, P. Kranenburg, C. Guillotel-Nothmann, M. Gurrieri, V. Presutti, M. Clementi, A. Meroño Peñuela, M. Turci, E. Marzi, A. Puglisi, R. Fournier-S'niehotta, D7.1 First Data Management Plan, Technical Report, 2021. doi:10.5281/zenodo.6963585.

[15] A. Scharnhorst, R. Van Horik, E. Daga, M. Daquino, E. Musumeci, P. Van Kranenburg, C. Guillotel-Nothmann, M. Gurrieri, V. Presutti, M. Clementi, A. Meroño Peñuela, M. Turci, E. Marzi, A. Puglisi, R. Fournier-S'niehotta, D7.2 Data Management Plan (Second Version), Technical Report, 2023. doi:10.5281/zenodo.7660299.

[16] E. Daga, A. Meroño Peñuela, M. Daquino, F. Ciroku, E. Musumeci, M. Gurrieri, A. Scharnhorst, F. Admiraal, R. Fournier-S'niehotta, D1.3: Pilots development – collaborative methodology and tools (V1.0), Technical Report, 2021. doi:10.5281/zenodo.7712909.

[17] P. Mirowski, The future(s) of open science, Social Studies of Science 48 (2018) 171–203. URL: https://doi.org/10.1177/0306312718772086. doi:10.1177/0306312718772086.

arXiv:https://doi.org/10.1177/0306312718772086.
[18] R. D. Peng, Reproducible research in computational science, Science 334 (2011) 1226–1227. URL: https://www.science.org/doi/abs/10.1126/science.1213847. doi:10.1126/science.1213847. arXiv:https://www.science.org/doi/pdf/10.1126/science.1213847.
[19] A. Beaulieu, M. Ratto, A. Scharnhorst, Learning in a landscape: simulation-building as reflexive intervention, Mind & Society 12 (2013) 91–112. doi:10.1007/s11299-013-0117-5.
[20] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, Prov-o: The prov ontology, W3C recommendation 30 (2013).
[21] S. Peroni, D. Shotton, The spar ontologies, in: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17, Springer, 2018, pp. 119–136.
[22] M. Daquino, M. Wigham, E. Daga, L. Giagnolini, F. Tomasi, Clef. a linked open data native system for crowdsourcing, Preprint arXiv:2206.08259 (2022).
[23] A. Scharnhorst, Citation - networks, science landscapes and evolutionary strategies, Scientometrics 43 (1998) 95–106.
[24] W. Ebeling, Karmeshu, A. Scharnhorst, Dynamics of economic and technological search processes in complex adaptive landscapes, Adv. Complex Syst. 4 (2001) 71–88.
[25] A. Oelen, M. Y. Jaradeh, M. Stocker, S. Auer, Organizing scholarly knowledge leveraging crowdsourcing, expert curation and automated techniques, in: A. S. Richard Smiraglia (Ed.), Linking Knowledge, Nomos-Ergon, Baden-Baden, 2021, pp. 181–198. doi:10.5771/9783956506611-181.