P. Baranowska, T. Basaglia, A. Kohls, M. Moskovic, H. Tzovanakis
DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Scientific Information Service

Project Overview and Identified Needs

Summary of findings during Milestones 1 and 2

13 January 2023

1. Executive summary

Scientific information is the key output of CERN's activities. Consequently, the efficient management, preservation and dissemination of information is a core activity of the Organization. Scientific information undergoes a natural lifecycle: from its creation, through dissemination and management, to finally being discoverable and reusable for further scientific studies. Due to CERN's historically grown, fragmented and non-interoperable landscape of numerous information management tools and services, the creation, management, reuse and discoverability of CERN's scientific information is not possible today without significant human efforts and manual workarounds. Given the importance of these activities for CERN, a project was launched by the Director for Research and Computing in order to investigate the status quo, highlight concrete shortcomings, and develop a proposal for a holistic information management landscape for CERN's research products.

Based on this mandate, the CERN Scientific Information Service launched an extensive study of the as-is situation of CERN's information management. Observations from a series of stakeholder interviews within the CERN research community but also authors of scientific information across all departments, a survey amongst creators and consumers of scientific information, and from its direct observations in the day-to-day management of research outputs were compared with today's best practices in scientific information management in order to identify key gaps and weaknesses at CERN.

The identified weaknesses in the status quo of CERN's suite of scientific information management systems can be mitigated or corrected through a number of recommendations developed by the project team of SIS. All recommendations together will result in the development of a holistic and efficient information management landscape for CERN, including a) a new central workflow management to support the creation, approval and dissemination of research products; b) a new streamlined institutional repository focusing on the preservation and dissemination of scientific artefacts, accompanied by existing repository solutions such as EDMS, HEPData and Zenodo; and eventually c) a Current Research Information System (CRIS) including a federated search functionality for scientific information (if a further analysis confirms a sufficiently strong added value over existing third party solutions). The implementation of such a new landscape needs to be supported by an organisation-wide training and education programme.

The proposed improvements and new developments will not only significantly improve the efficiency of the entire scientific information lifecycle across all CERN departments and experiments, it will also lead to better visibility and discoverability of CERN research, will allow CERN to showcase its impact

on society more prominently, and will ensure the adequate long-term preservation of CERN research products. As such, implementing the proposed measures should be seen as a key strategic initiative of the Organization in the coming years.

2. Project goal and roadmap

2.1. Problem Statement

A key aspect of CERN's mission is to perform world-class research in fundamental physics, a process which involves both the application and production of scientific knowledge. The scientific knowledge acquired at CERN is represented in various research products that are stored in multiple media formats across a range of information systems. CERN is mandated through its founding Convention to make its scientific findings generally available. Any solution to preserve and expose CERN's research results, particularly scientific documents, should therefore serve the needs of both the CERN scientific community as well as the broader public.

Since the beginning of the digital age, several CERN departments have developed various information management systems (see Annex I) to handle their respective scientific and administrative documents, often including dedicated submission and approval workflows. Over time, these systems were expanded to address the changing needs of the Organisation, and today host a wide range of artefacts, including operational procedures, policies, internal guidelines, etc.

The Scientific Information Service (SIS) is one of the shepherds and main information manager of CERN's scientific knowledge, ensuring its global accessibility and applying the FAIR[1] principles. In this role, SIS has identified three problem areas resulting from the historical evolution of CERN's information systems:

1. Due to parallel development over many years, a significant overlap in roles and responsibilities along the production chain has emerged, resulting in enormous complexity and inefficiencies for information managers.
2. The ambiguity in the roles of different systems (CDS, INSPIRE, ArXiv, EDMS, Zenodo etc.) generates confusion or even frustration on the side of authors, resulting in the risk of information not being shared in an adequate form.
3. Not all tools for managing information have been built with long-term document preservation and persistence of document identifiers in mind; as a consequence, important information might be lost or not findable.

The above-described problems are exacerbated by the range of involved stakeholders, adopted software solutions, and storage and distribution systems. This creates an ever-increasing organisational effort to manage CERN's scientific knowledge: redundant information creates duplication of work; unclear information ownership, overlap and ambiguity across the numerous systems make it impossible to implement streamlined processes; and inconsistent, incomplete metadata can prevent the easy dissemination of scientific artefacts and often requires manual

[1] Findable, Accessible, Interoperable, Reusable, see https://www.go-fair.org/fair-principles

curation. As a result, the knowledge produced at CERN is suboptimally exploited, and the cost (both human and material) for information management increases. Ultimately, the Organisation faces the risk of not meeting the expectations of its member states regarding the general availability of its scientific findings.

This project aims to develop a proposal for a coherent information management landscape for CERN's scientific information that reflects state-of-the-art technology, organisational constraints and realities, and is suited to the evolving needs of the research community at CERN.

2.2. Scope of the project

CERN, as a knowledge-based organisation, both consumes and produces vast amounts of information as part of its activities. For the current project, it is useful to distinguish between two broad categories of information:

- Scientific information: any piece of knowledge deriving from or related to scientific research and recorded in some material form by a CERN (associate) member of personnel or by a CERN experimental collaboration or using the CERN facilities.
- Non-scientific information: anything else.

A more extensive discussion of what scientific information means in all its nuances can be found in Annex III. The current project's focus is on the former only and does not attempt to address the needs of the Organisation related to non-scientific information. There are two reasons for this. Firstly, restricting the scope in this way makes the project more manageable and increases the likelihood of success. Secondly, scientific information, as a major output of CERN's activities, presents specific characteristics and needs, in particular for dissemination outside the Organisation, which justifies a dedicated approach.

Note that the scientific/non-scientific distinction is not clearly articulated in the current information landscape. For instance, the CERN Document System (CDS) not only serves as an institutional repository for scientific articles but is also involved in procurement processes and hosts various administrative documents. Consequently, the proposed changes in the scientific information landscape will also affect many non-scientific documents, and a solution will need to be found.

A series of stakeholder interviews were conducted (see next chapter) to create a sound understanding of organisational needs, including questions about publications, data and software research products. Since existing organisational procedures around publications are significantly more advanced than the dissemination of other research products, the amount of feedback related to "traditional" artefacts, namely articles, was considerably higher. However, based on the increased momentum towards holistic open science, supported by the new CERN Open Science Policy, it is expected that active dissemination of all types of research products will increase in the years ahead, and a proposal towards a new scientific information landscape will need to take this trend into account.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

2.3. Methodology and roadmap

In the past, several non-coordinated attempts were made to improve specific aspects of the information landscape (e.g. the introduction of CDS Videos). However, these efforts consisted of ad-hoc technical solutions applied towards addressing specific problems or approached the problem for only a subset of the user community. This approach lacked an overarching strategic viewpoint and exacerbated the existing situation, adding further complexity. Some years ago, a similar initiative was started to streamline the information landscape at CERN. But due to the lack of support from relevant stakeholders, these efforts were unsuccessful. The active engagement with the CERN community in this project will help to mitigate the problem.

To avoid biases towards existing solutions or processes, this project adopts a green-field approach, i.e. developing a new ideal solution architecture that best addresses all user requirements and considers today's general information management's best practices.

As a first step, organisational priorities were defined, and general information management needs through input from scientific information professionals at CERN (primarily within SIS) were collected.

Subsequently, the project team conducted 21 guided interviews with key stakeholders in the creation or dissemination of scientific information as a qualitative research method to document the as-is situation and to identify demands from specific user communities or additional general information management needs. The interviews were equally used to validate (or amend) previously defined requirements. A list of interviewees was created based on an in-depth stakeholder analysis and was expanded based on input received during the first interviews. A list of interviewees and questions asked is attached in Annex II. In addition, the project team had numerous side conversations and spontaneous chats with stakeholders across the Organization to complete the picture.

While sufficient input related to creating CERN scientific information was obtained through these interviews, the consumers of scientific information are even more diverse. To capture this diversity, a survey was conducted amongst the CERN community (196 complete responses).

Based on the aggregate input and following consultations with the CERN IT engagement experts, the project team developed the information landscape vision outlined in this document. Working closely with CERN IT, a target architecture will be developed as a next step, which will subsequently be validated with a subset of stakeholders. Based on the final architecture, the technology experts in CERN IT will be able to identify and assess suitable solutions by applying a strict 'build vs buy' approach. The final outcome of this project will be a concrete recommendation, including an implementation roadmap and an estimation of budgetary impacts for consideration by the Enlarged Directorate.

It is well understood that the actual development and implementation of the proposed future information landscape will be a major and ambitious project involving many stakeholders. Such a project will require thorough planning and project management, including risk management. The implementation roadmap will need to consider such aspects. However, while aiming for a structured

approach, this initial analysis project will not follow formal project management methods as the scale of the analysis does not justify the additional effort.
The analysis approach described above translates into the following project phases:

Process Step	Indicative Timing	Involved Stakeholders
1. Needs & Priorities	Mar 2022 - Apr 2022	SIS
2. Stakeholder Engagement	May - Aug 2022	SIS, EP (+ experiments), IT, TH, ATS, IR
3. Documentation	Sep - Dec 2022	SIS, EP (+ experiments), IT, TH, ATS, IR
4. User survey	Dec 2022	All CERN departments and experiments
5. Principle Architecture	Jan 2023 - Feb 2023	IT, SIS
6. Solution assessment	Mar 2023	IT, SIS, EP (+ experiments), ATS
7. Recommendation	Apr 2023	SIS, IT

Table 1: indicative project phases and timing

3. Scientific information life cycle

3.1. General scientific information life cycle

Scientific information typically passes through five stages, which can be characterised as creation, dissemination, organisation and management, discoverability, and usage[2], as shown in Figure 1:

[2] Similar terminology is used, for example, by the UCLA library.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

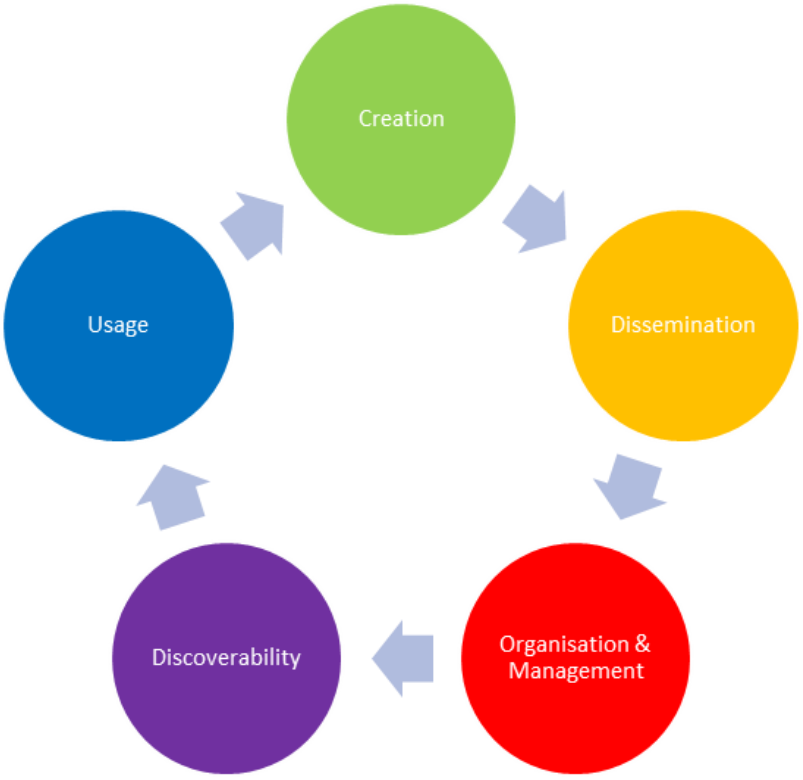

Figure 1: General Scientific Information Lifecycle

Creation is the first stage in the existence of any scientific information product. As an outcome of their research, scientists discover results that lead to creating scientific information products in various forms (articles, datasets, software, etc.). Engineers that are given a certain task (e.g. to develop a new material or technology) report their findings and proposed solutions through the creation of scientific information products.

Dissemination is the primary goal of any scientific information product. A scientific paper, a data set, or other research output is supposed to initiate or contribute to scientific discourse and hence needs to be shared with the widest possible population within a given target audience. In order to allow such a target audience to discover the product and hence participate in the discourse, metadata are attached to the product, which is used by repositories, information aggregators or indexes to interpret the information.

Making scientific information available in a relevant information outlet requires the **organisation and management** of associated metadata that have been disseminated by a publisher or have been imported from a repository. Information managers such as librarians enrich and correct the existing metadata and organise and collate them in a manner that is useful for the scientific community.

Discoverability is arguably the most important step in the described life cycle, as information is useless unless it can be retrieved by others interested in the topic. The broader the dissemination and the better the organisation and management of metadata, the better the overall discoverability of the information. However, a powerful and easy-to-use search engine is necessary to enable users to find the needed information efficiently.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Usage can be interpreted in two ways. Researchers use discovered scientific information as input to their own creation process; hence, the usage in this context closes the life cycle loop. However, information managers or organisational leaders will use scientific information in the form of aggregated summaries – such as the number of publications or citations – in order to measure their scientific output and demonstrate the effectiveness and success of the organisation.

3.2. The scientific information lifecycle at CERN

During the discussions with a wide range of stakeholders across CERN, it became apparent that, despite the organisational and scientific diversity, many processes related to the management and dissemination of scientific results are very similar. While the complexity of review and approval processes vastly varies, the overall workflows are comparable. However, further alignment of workflows should be promoted to minimise technical complexity, as demonstrated in the described findings later in this document.

The schematic presented in figure 2 represents the scientific information lifecycle at CERN (background colours correspond to the previously described lifecycle stages). Key elements of the process that are subject to recommendations from this analysis are highlighted in blue. In chapter 4, we will describe all process steps in more detail and analyse their efficiency as well as possible shortcomings.

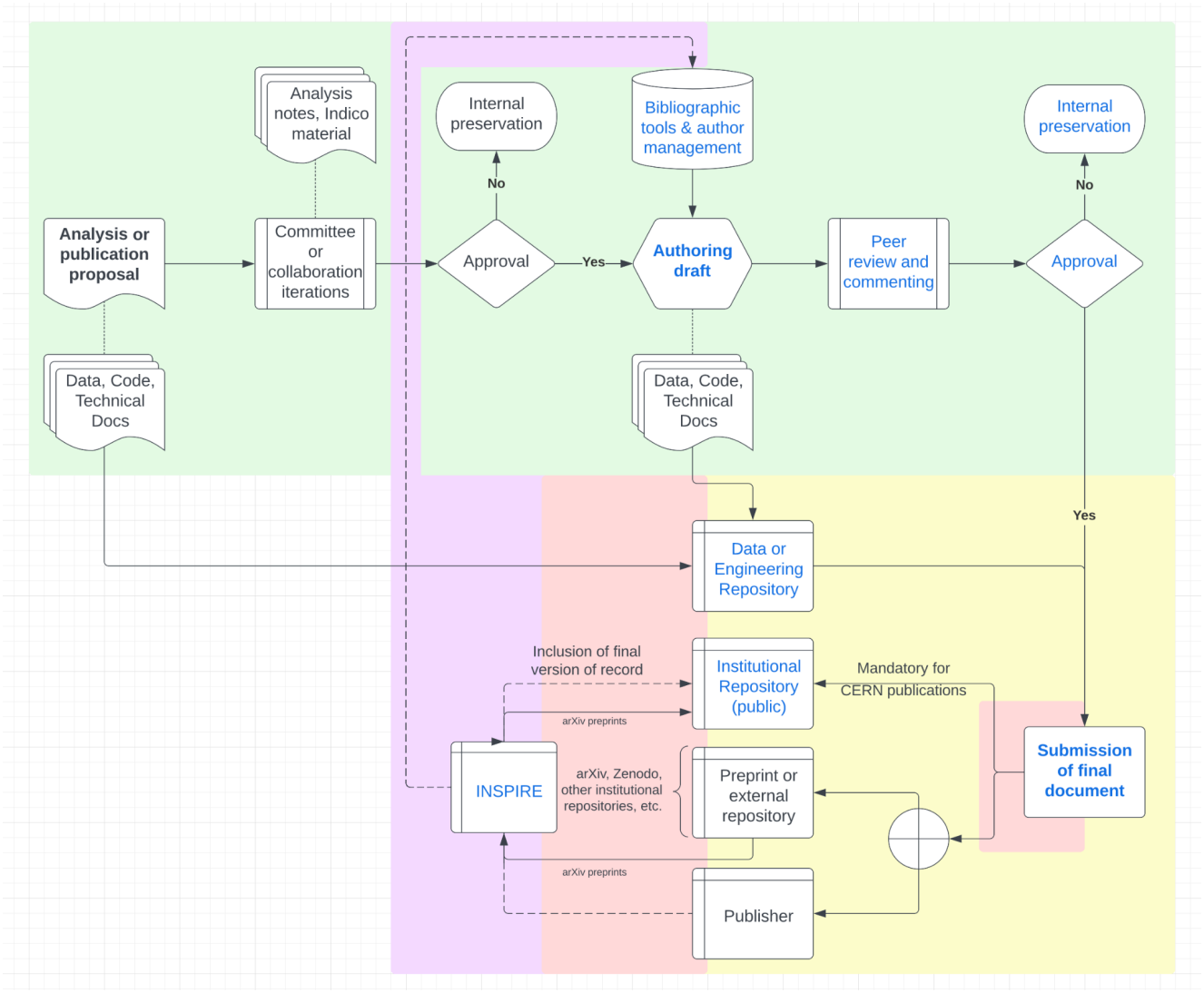

Figure 2: CERN Scientific Information Process

4. Concrete findings related to the CERN scientific information lifecycle

This chapter analyses the five main stages described in chapter 3 individually and identifies the concrete actors, processes, tools and services at CERN. Key findings from the stakeholder interviews associated with each stage are reported.

4.1. Creation

In the CERN context, the **Creation** step of the generic scientific information lifecycle described in Figure 1 can be broken down into two sub-processes: 1) **Authoring,** which includes the actual

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

preparation of an analysis and the actual (possibly collaborative) authoring process, i.e. preparing a first version of a manuscript or dataset; and 2) **Review, Commenting and Approval** where, depending on the specific needs of the experiment or department, several peers, as well as the hierarchy of the author(s), might be invited to review and improve the article, and subsequently authorise its publication. This chapter considers both sub-processes separately as actors, tools and recommendations differ.

4.1.1. Authoring

4.1.1.1. Introduction

The creation or authoring of research products, such as papers, datasets or software, can involve one or many authors. At CERN, an article originating from experimental collaborations can formally contain several thousand authors (members of the collaboration), although typically, only a small subset of these are actively involved in its creation. Whenever several authors are involved in creating an information product, collaborative tools are used to work jointly and in parallel on the product.

In the context of experimental collaborations, the creation process often starts with the proposal of a specific analysis. The documentation of the proposal and its discussion within the collaboration can already be considered a part of the scientific discourse; therefore, these internal collaboration documents should be included in the scope of the overall scientific information management.

4.1.1.2. Actors

Scientific information at CERN is generated through various roles across the Organization. Researchers (across all scientific disciplines), engineers, science educators and even personnel in the administration author scientific articles, collect research data or develop software and algorithms to contribute to the scientific discourse. However, most scientific information artefacts originate from the Research & Computing and Accelerator sectors. The aggregate research output of the CERN community is as diverse as its creators. The overview of CERN Publications by subject (see Figure 3) gives an insight into this diversity. As the meeting point of scientific excellence and the host lab for the most advanced accelerators and detectors in the field, CERN should support information creators as much as possible to create further capacity in the scientific discourse.

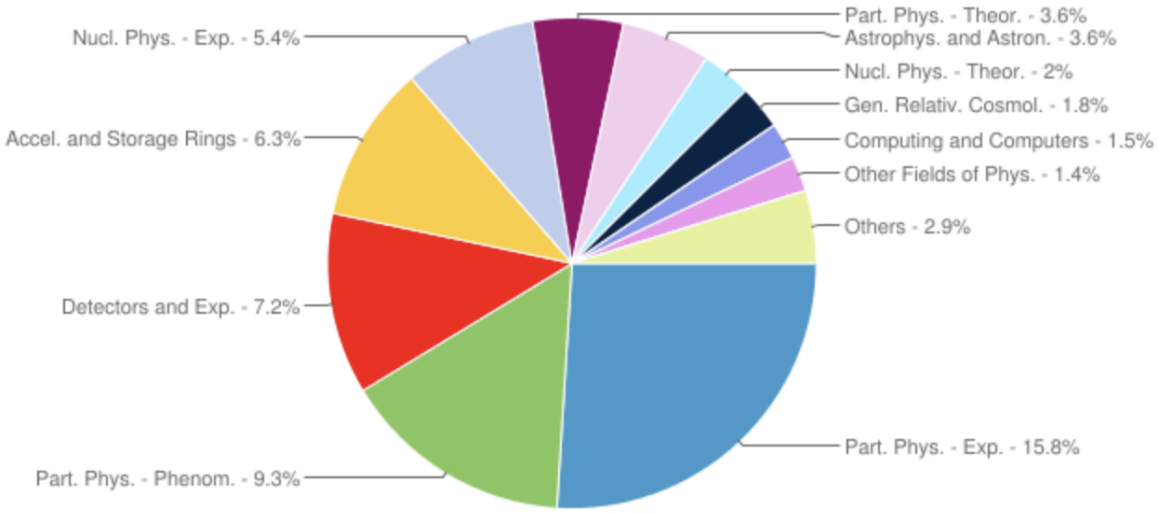

Figure 3: CERN publications 2021 by subject

The following high-level schema in Figure 4 displays the authoring process with actors (in blue) and key tools used (in red).

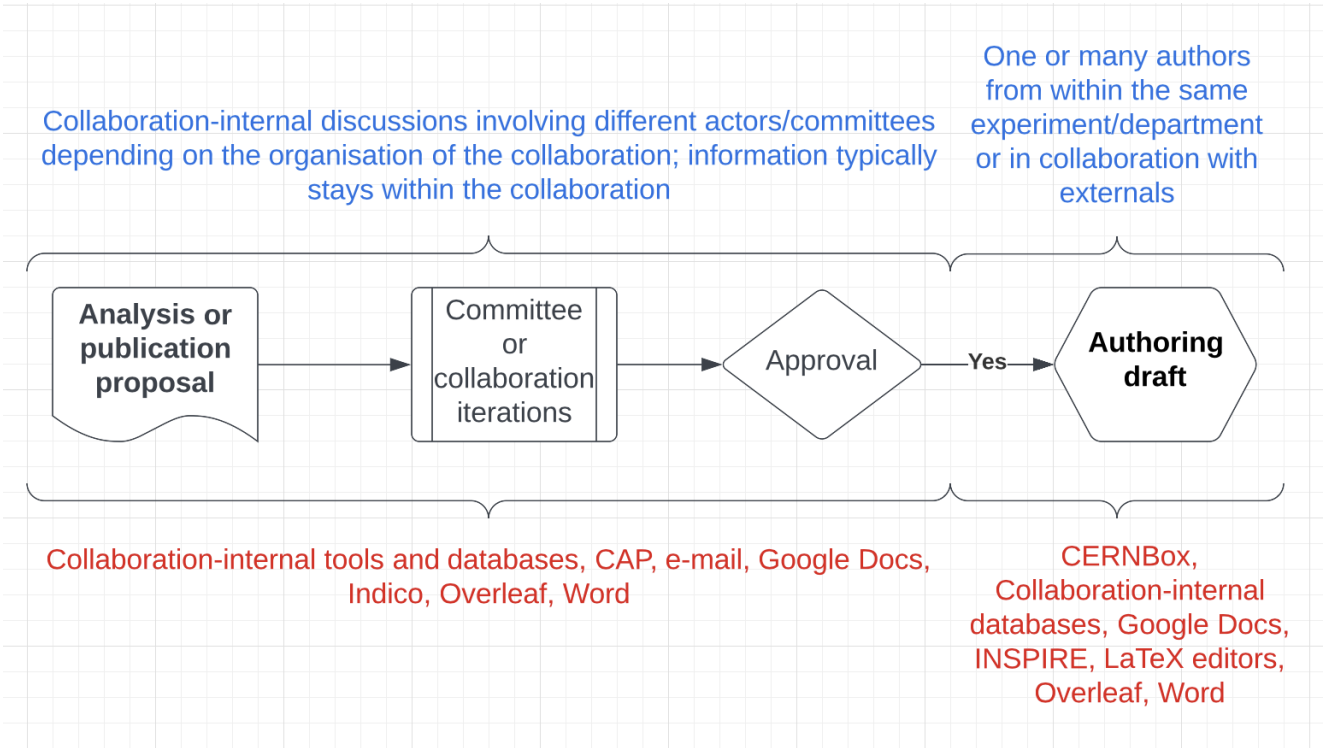

Figure 4: Authoring sub-process, actors and tools used

4.1.1.3. Processes and Tools

Collaboration-internal analysis proposals and approval processes are done today using collaboration databases such as Glance or CADI, standard CERN tools such as Indico, or simply via e-mail. Some experiments recently started introducing CAP as a tool to manage their analyses.

The **format** in which a scientific artefact is created depends on the content as well as the standard practice within the specific scientific sub-community. For example, LaTeX is used more in EP and TH departments, whereas within the ATS sector, word processing software like Microsoft Word is generally the norm.

Collaboration is an important factor when authoring, as typically, several researchers are involved in the creation of a research product. Therefore simultaneous editing of the same document/source file by multiple authors streamlines the process, as otherwise different versions of files are sent back and forth between the collaborators (usually via email) and may create confusion. To facilitate collaborative authoring of articles or reports, Overleaf is increasingly used to manage the editing process. It allows shared access to the source and output files of the in-progress draft, as well as commenting functionality to iterate on the document. However, its use is not yet a universal practice. Some authors resort to local LaTeX editing and a CERNBox and email-based collaboration workflow. Authors regularly use (LaTeX) templates that correspond to the style of their chosen journal or those specified by their collaboration/project.

Concerning the content, authors need to manage bibliographic references, the author list, as well as funding information and acknowledgements. Creating a **bibliography** entails finding relevant records and retrieving their standard citation formats (e.g. BibTeX) from online databases and maintaining an up-to-date local copy. To retrieve bibliographic information, INSPIRE can be used to generate BibTeX citation information. The large LHC experiments separately maintain their own databases with customisations in order to ensure uniformity of style and correct mistakes. While sufficiently functional today, the current practice leads to a significant multiplication of efforts as several almost identical tools are maintained in parallel.

Managing author lists increases in complexity with the number of authors. Persistent identifiers like ORCID and ROR ease the disambiguation of author names and institutional affiliations and, ideally, should be added to any research product. Larger collaborations manage their own author lists (recently with attributed persistent identifiers) in their own databases (for instance, Glance or CADI), but smaller experiments do not have such tools, and hence author lists are typically managed manually, a process that can be cumbersome and error-prone. CERN Databases such as Greybook/Foundation contain some relevant author information, but they are not systematically aligned with Glance or CADI. Author information can also be compiled in an XML file and attached to arXiv and journal submissions, so it can be processed automatically by repositories/aggregators, such as INSPIRE, and the journal workflow, but this practice is usually limited to larger collaborations.

As for **funding information and acknowledgements**, this information is particularly important to measure the scientific output of a specific project (usually indicated by a grant number). This is currently managed by authors on a case-by-case basis, and it is very rarely indicated in the metadata, which complicates retrieval of this kind of information. CDS has metadata fields to indicate

funding/grant information, but it is barely used. Otherwise, no tool currently exists at CERN to handle funding information.

4.1.1.4. Key Discoveries and Recommendations

The **adoption of persistent identifiers** for authors and affiliations should be generalised in the production of author lists, ideally interfacing directly with existing collaboration systems or the CERN staff database. Respective technical interfaces still need to be created. Standardisation should also be introduced in the denomination of document types and in the systematic attribution of a CERN standard artefact identifier. Assigning a **unique CERN identifier** (agnostic of experiments/departments/projects) early in the process could significantly improve matching and linking between different stages of research products (e.g. preprint and peer-reviewed publication or article and underlying dataset; matching on title and authors alone is not sufficient) within CERN's institutional repository.

Authors need to edit **LaTeX documents** collaboratively, which means having shared access to the source and output files of the in-progress draft and commenting functionality to iterate on the document. They need to have easy access to templates that correspond to the style of the chosen journal or the experiment. A centrally managed system is necessary to preserve the full history of the sources of the document for further modifications and archival purposes. Overleaf currently serves this goal well but is not used consistently enough. Similar to (and ideally based on) the workflow and approval system, as described in section 4.1.2., access rights to collaborative documents should be automatically altered based on collaboration rules and the document status.

Maintaining several **bibliographic databases** in parallel is an unnecessary duplication of efforts and should be replaced by one central database. While existing services such as INSPIRE can serve as bibliographic databases due to their complete coverage of the field (beyond CERN), current search capabilities are insufficient to support the use cases of the experiments (for instance, effectively searching for specific particles). In collaboration with the four LHC experiments, either a new central database should be developed, or INSPIRE capabilities need to be enhanced.

Managing author lists and funding information is a recurring and tedious yet crucial task for experimental collaborations. While it has to remain the responsibility of the actual collaboration, CERN could significantly improve the efficiency of the process by introducing a central tool maintained by CERN, which would eliminate the parallel maintenance of individual tools by the different collaborations. Such a **central author and funding information management system** can automatically assign persistent identifiers for authors (ORCID), affiliations (ROR), and funders/grants (DOIs) to further reduce manual curation efforts during the information lifecycle. Through interoperability with other important CERN databases, such as the Greybook or Foundation, the consistency of information can be ensured. Furthermore, by centrally managing access rights, data privacy requirements (OC11) can be systematically ensured. Such a tool should be available also for smaller experiments and flexible enough to support the varying rules around authorship in the different collaborations.

4.1.2. Review, Commenting and Approval

4.1.2.1. Introduction

Before a CERN paper or other scientific information product is published (i.e. made available to external readers), it typically undergoes an **internal peer-review and approval process**, either within an experimental collaboration or department or including some wider approval process (e.g. EP preprint approval). Approvals might be undertaken for the scientific correctness and relevance of the product or for other means (e.g. publishing open access, licensing terms, etc.). Such processes may be characterised by varying levels of complexity across different collaborations/groups/projects, depending on their respective needs and practices.

4.1.2.2. Actors

The review and approval procedure of a scientific artefact is typically managed internally within experimental collaborations or CERN departments/projects. While, in principle, all research products undergo internal review and approval procedures, the rigour, complexity and number of involved reviewers or approvers vary significantly. Scientific publications from CERN Experiments typically also include an EP Department approval procedure subsequent to the collaboration-internal process.

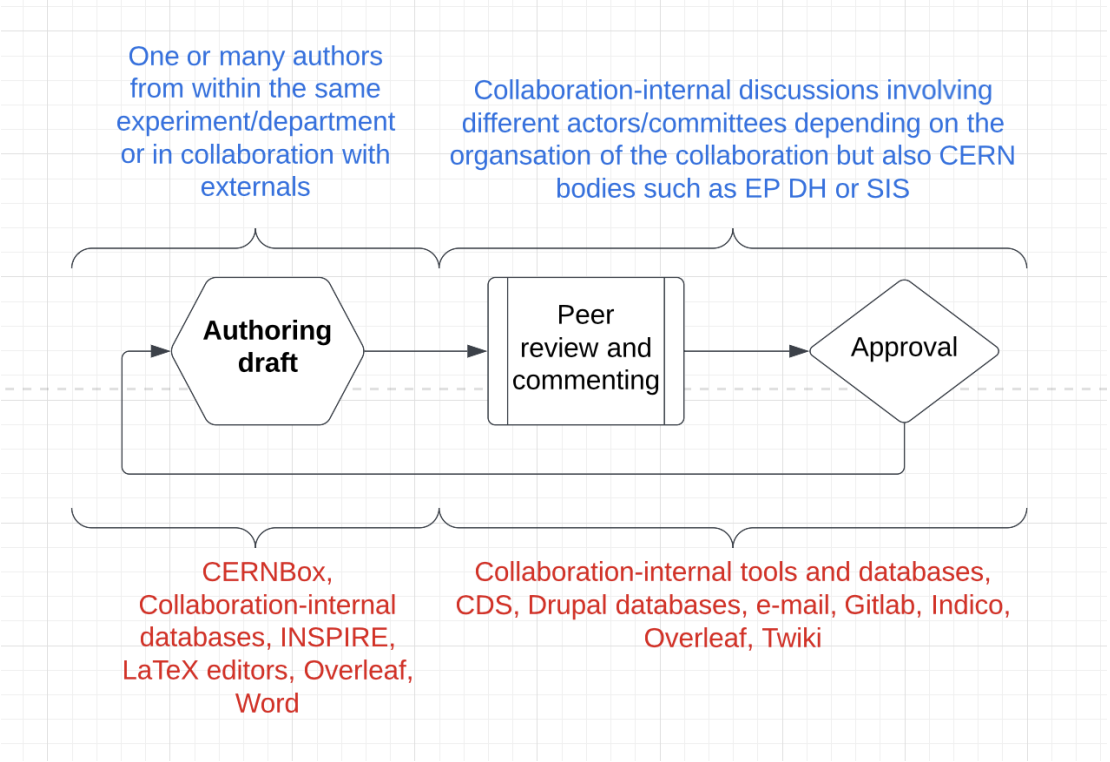

Figure 5: Approval sub-process, actors and tools used

The high-level schema in Figure 5 displays the authoring process with actors (in blue) and key tools used (in red). The review, commenting and approval process, by its own nature, might require updates and iterations from the artefact authors and hence part of the authoring process will be re-executed as part of the overall approval procedure.

4.1.2.3. Processes and Tools

The different phases of today's peer-review and approval processes (see figure 5) are carried out in **multiple systems in parallel**, which are not typically interoperable. Examples of systems used for reviewing and commenting include Glance or CADI (which have been created and are maintained directly by the different LHC collaborations), CDS, Google Docs or spreadsheets, Overleaf, Twiki, Gitlab, Drupal websites, notes in Indico, as well as e-mail or personal discussions. Documents originating from the Accelerator and Technology Sector (ATS) are also regularly reviewed and approved through EDMS.

CDS is often used for **commenting** on physics papers. However, due to limitations in editing and/or formatting, comments are often created on Twiki, MS Word, and Google Docs and then copied to CDS. As in some cases, collaboration databases are used to track progress, it happens that subsequently, the same comment is copied again back into the collaboration database.

Approvals typically happen both before the actual authoring of an artefact (internal analysis approval), as well as after its creation. Today, this is partly done through collaboration-internal systems, partly via CDS (for physics papers) or EDMS (for engineering papers). While a collaboration-internal approval of data releases to the CERN Open Data Portal is also done prior to dissemination, there is currently no systematic and technically documented approval process for such releases. Equally, formal approvals of public releases of analysis software in Github or Gitlab are not systematically documented or do not exist at all.

4.1.2.4. Key Discoveries and Recommendations

The diversity of tools and process inconsistencies, or rather the lack of a consistent and uniform approval workflow, often requires high (and duplicated) maintenance efforts and creates risks of information loss (e.g. of specific comments or approval conditions) or even insufficient approvals during the process. Some of the tools that have been created directly by the collaborations, like CADI, which is used by CMS, are becoming increasingly difficult to maintain and urgently need replacement. However, a more modern tool cannot be developed due to a lack of resources. Therefore, if a uniform **tool for commenting, peer review and approval** could be offered across CERN, it would allow for more effective collaboration and enforcement of policies. During our interviews, all experimental collaborations expressed interest in such a tool, provided it is flexible enough to address their specific needs.

For physics papers, CDS is used today for commenting, but due to limitations in its editor, comments are often created using other tools and then copied to CDS. CDS does not allow the linking of comments to specific versions of papers or a specific place in a document, nor does it offer the functionality of responding directly to individual comments. This often leads again to parallel processes (e.g. adding comments in parallel to collaboration databases). The lack of consistency and the need for manual workarounds in the commenting process was perceived by the interviewed stakeholders as key weakness. Providing a **transparent and flexible commenting functionality** (Google Doc style) with the ability to link comments to a specific area in a specific version of a document could certainly streamline processes. LaTeX should be supported in such comments. Other

stakeholders require to be able to respond to specific comments. Authors/owners of a document need to be able to resolve comments, but the commenting history needs to be preserved.

Today, the approval procedures are partly done through collaboration-internal systems or via e-mail, partly via CDS. The current mix of roles of CDS, i.e. workflow management tool and repository, should be streamlined by a new **central workflow management layer** on top of the actual repository layer, which includes a commenting functionality beyond the actual approval process. The link between the two layers (repository and workflow) should be transparent, and the metadata format should comply with the one adopted for the information management system.

Additionally, approval and reporting needs (beyond the original research product release) do not seem to be connected or aligned. For example, open access-related aspects (i.e. which journal to publish in; approval by SIS) are not integrated with the EP preprint approval process. The EU office's approval of milestone reports is also not linked with actual publication processes. In general, the **lack of feedback on approval processes** in CDS to other systems leads to manual copying/pasting back to other tools used in parallel. There is also no mechanism in place to reject a preprint which has been submitted to CDS and directly inform the submitter that it should be submitted somewhere else, e.g., as a CERN-OPEN document that is linked to a CERN activity but does not fit in a specific group's submission or collection.

Approval process rules need to be easily maintainable by the experimental collaborations or departments (internal rules) or by respective central management teams (e.g. EP, SIS). The workflow and approval system needs to be seamlessly integrated with a plurality of commonly used authoring tools (e.g. Overleaf) and should use the CERN institutional repository as an underlying database layer to store document versions during the approval workflow, i.e. all research output (already in draft) should be captured systematically in the repository for preservation. Depending on the document's status (e.g. in peer-review, approved, etc.), the workflow management system needs to alter the **access rights to the respective document** depending on predefined rules (e.g. not approved documents only visible to the experiment or department). Equally, access rules to comments need to be flexible depending on the rules and practices of the respective experiment/department. All **comments and approvals need to be preserved** and become part of the actual document history, together with the respective versioning of the document itself.

Any effort towards standardisation of review, commenting, and approval workflows should consider **rules and practices of the individual groups/projects/collaborations** aimed at streamlining more formalised workflows and improved structuring of more informal ones. The registration and discussion of a planned analysis, conference contribution, or thesis, the internal peer-review and commenting process, and any publication-related approval processes (incl. open access and other standardised approvals) should be managed in one **holistic workflow management system** across all CERN departments and experiments. While a certain standardisation of the approval procedures is desirable to be aligned with the requirements stated in CERN Operational Circular 6, such a workflow system needs to allow for extensive customisation to support the varying level of complexity originating from collaboration internal requirements. The registration should include a process to assign **CERN-standardised keywords and experiment/project affiliations** which will be used throughout

the approval process as well as in the metadata of the final repository record and downstream dissemination outlets to simplify search/reporting.

The envisioned workflow management system needs to **actively communicate with approvers** and authors, i.e. send notifications of pending approvals or received comments, including reminders if there is no reaction to a required workflow step within a defined number of days.

4.2. Dissemination

The dissemination of research products can generally be categorised in two ways: dissemination through formal (peer-reviewed) publication or dissemination via depositing one or several research artefacts in repositories. Both routes are followed either in parallel or the publication step follows the deposit to the repository. This chapter considers both routes of dissemination separately, but the recommendations for both are closely linked.

4.2.1. Dissemination through repositories

4.2.1.1. Introduction

CERN authors are interested in disseminating their research products as widely as possible. By reaching a large audience, results can contribute towards the scientific discourse, and authors receive more credit for their work (e.g. through citations). To achieve that goal, research products need to be easily findable and accessible to other researchers. A range of dedicated repositories is available for different use cases, such as arXiv.org, HEPData, or Zenodo. In addition, all CERN publications must be preserved through CERN's institutional repository (CDS).

4.2.1.2. Actors

The dissemination of scientific artefacts is, in principle, the key responsibility of its authors and should only take place following the successful completion of the approval described in chapter 4.1.2. In larger experimental collaborations, dissemination is often centralised in the collaboration secretariats or publication committees, while in smaller experimental collaborations or in publications originating from CERN departments, one of the authors typically submits the research output to the respective outlets. The Scientific Information Service can assist authors with advice concerning outlets, correct metadata definitions, and can even facilitate submission on behalf of authors.

The following high-level schema in Figure 6 displays the submission process with actors (in blue) and key tools used (in red). Today the process basically consists of the decision by the authors of where to deposit their research output, as well as a subsequent manual submission to the chosen outlets.

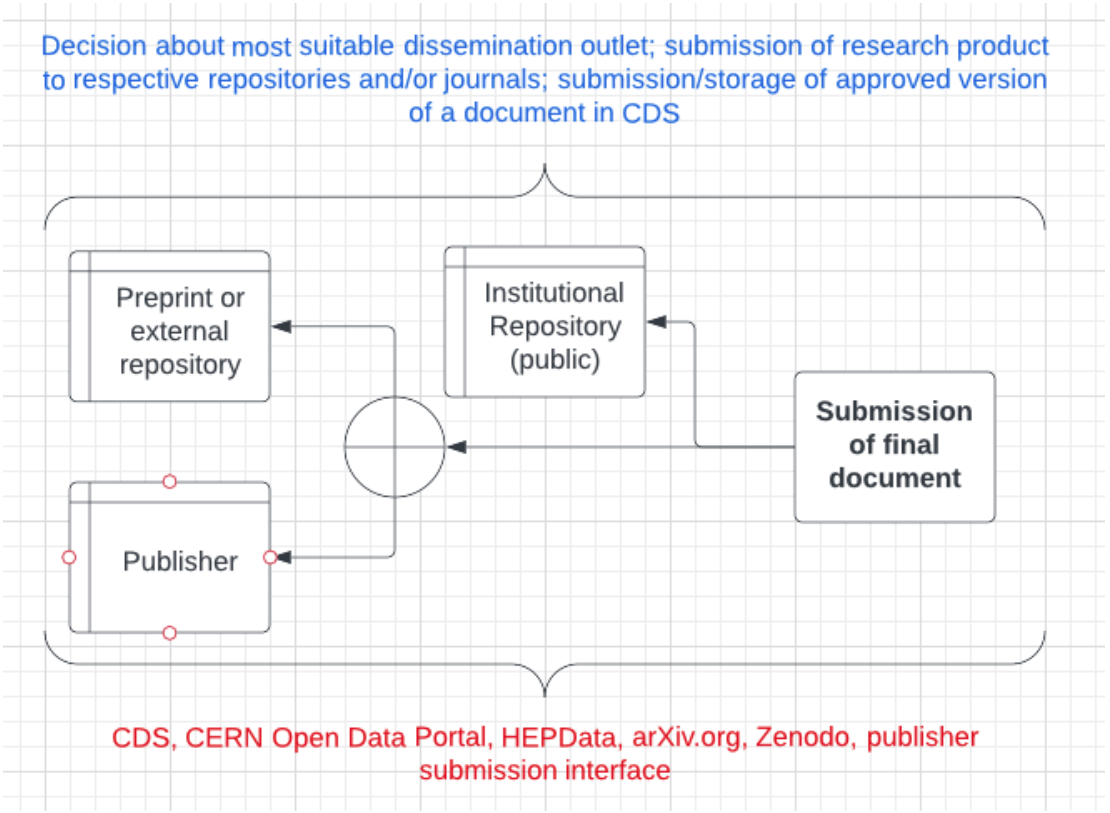

Figure 6: Dissemination sub-process, actors and tools used

4.2.1.3. Processes and Tools

The **core component of CERN's information landscape is CDS**, CERN's institutional repository. It hosts all research outputs related to approved analysis proposals. The hosting can either be exclusive (in the case of non-approved publications that remain only accessible through the institutional repository) or can happen in addition to dissemination through external outlets.

CERN's engineering & equipment documents and data are mainly stored in EDMS, CERN's official Product Lifecycle Management (PLM) tool. PLM data and documents are generally factual descriptions and specifications of installations, procedures, etc., and are usually not meant to be shared outside the respective team in charge. These differ from scientific information, usually intended to be made publicly available, for example, as conference contributions, preprints, and/or published articles. However, some scientific papers refer to documents stored in EDMS, which might not be externally accessible and subject to further iteration. Therefore, it should become scientific practice to enable public access to such documents if they are used as references.

Zenodo is regularly used to disseminate and preserve reports and articles that are not captured via CDS (e.g. EU project outputs without CERN authors).

Documents related to conference presentations are disseminated differently based on their form. In case proceedings are published, the corresponding contributions are, in principle, available on CDS.

It is increasingly common for conferences not to publish formal proceedings, in which case only slides are available on indico.

For research data collected at CERN, four levels[3] of complexity have been identified by the Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group, for which varying dissemination processes and tools are utilised. Level 1 data (related to scientific publications) are often deposited to HEPData. Analysis software or other code relevant to publications is often made publicly available through GitHub. Some level 3 data collected by the LHC (or other CERN accelerators) together with required analysis software are made available through the CERN Open Data Portal, in addition to being preserved through the CERN and/or collaboration-internal original research data preservation routines. Other datasets might also be disseminated using Zenodo, in particular in the context of EU projects. However, there is currently **no systematic release of datasets** happening **outside LHC experiments and EU projects**. Given that CERN's recently announced Open Science Policy includes a commitment to making research data openly available, such a systematic release needs to be established.

Due to the **manual depositing of datasets or software** related to a scientific publication, there are often errors and inconsistencies in metadata and citation practices. As a consequence, datasets or software associated with a publication might not be appropriately identified and consequently might not allow the intended reproducibility of findings, despite efforts to follow good open science practices.

A CERN publication could end up in a multitude of dissemination outlets in different stages of its life cycle in parallel: a restricted or publicly available record in CDS, a pre-print submitted to arXiv.org, a peer-reviewed journal article on the journal's website, the peer-reviewed article deposited to a repository or on arXiv.org, a version exposed on the experiment's website, etc. A typical scientific paper will be disseminated in parallel through several, if not all, of the above avenues. In addition, related datasets or software products are also disseminated through multiple outlets. However, all routes are followed manually without any meaningful way to ensure aligned metadata. This not only causes duplicated efforts on the side of the author or experiment secretariat, but these records will also co-exist without a clear reference that they refer to the same publication. This causes significant manual curation efforts, e.g. in the Scientific Information Service, to identify and subsequently combine such duplications in tools such as INSPIRE or CDS, and to ensure consistent representation of the same scientific record in CDS and external repositories.

While the submission to CDS should happen systematically for all CERN outputs before or in parallel to the external dissemination, in many cases, it only happens indirectly after its submission via external dissemination routes (such as arXiv.org or publishing in a scientific journal). The Scientific Information Services runs processes to identify CERN articles not yet in CDS in order to ingest them (typically through INSPIRE). This requires regular active manual matching/merging of records, including the alignment of metadata. This issue is further described in Chapter 4.3.1.

[3] See CERN Open Data Policy for LHC experiments: https://cds.cern.ch/record/2745133/files/CERN-OPEN-2020-013.pdf.

4.2.1.4. Key Discoveries and Recommendations

Several repositories are used today for internal dissemination and preservation (CDS, EDMS, Zenodo). Due to an observed **lack of awareness of the different services** and their original purpose (confirmed in the personal interviews and through survey results), authors at CERN seem to select services based on personal experience or the habits of their teams. In particular, within the ATS sector, EDMS has become the main tool for information storage, including in part for scientific articles. Similarly, experimental collaborations might use CDS to store any project-related documentation that could instead be stored in EDMS. In a future landscape, project or PLM-related information should not be stored in the institutional repository, and documents intended for public dissemination should not remain exclusively stored in EDMS but should be transferred to and preserved in CDS. In cases where related documents are (as intended) stored both in CDS and EDMS, both systems need to link documents properly and ideally directly exchange related metadata to improve the cite-ability of EDMS documents that might have access rights restrictions. As part of the efforts towards a future Scientific Information Landscape, clear guidelines and respective training for CERN authors are required to ensure the consistent and intended use of future tools.

Focusing on its intended core functionality, a new **version of CERN's institutional repository should exclusively collect, serve and preserve the CERN scientific output**. Clearly distinct from other tools such as EDMS (see above), it should store CERN articles related to approved analyses. As part of the preservation layer of the future CERN Information Landscape, all CERN output should continue to be preserved and safely stored for the long term (incl. preservation in potential disaster scenarios) and be accessible at any moment. This is valid not only for traditional outputs such as published articles but also for internal notes produced by experiments or less-conventional documents such as slides, which are typically only stored in Indico currently. Consequently, such documents should be systematically transferred for preservation to CDS.

The **systematic release of research datasets**, as foreseen in the CERN Open Science Policy[4], needs to be supported. Today, there are already adequate tools available for such a systematic approach, but they need to be clearly presented with their distinct use cases, and the CERN community needs to be educated as part of the CERN Open Science Policy implementation efforts. HEPData has established itself as a standard for releasing Level 1 data (plots and tables related to scientific publications). The CERN Open Data Portal architecture is dedicated to large and complex datasets that typically require dedicated analysis software. The portal is already regularly used by the four large LHC experiments but will further expand its application to experiments with similarly complex datasets. For all remaining research data, Zenodo could serve as the standard outlet for dissemination. It seems advisable to create a **dedicated CERN Open Data community in Zenodo** to simplify the discoverability of CERN research data, allow systematic indexing of such datasets in INSPIRE, and support the respective open science efforts at CERN through a dedicated dissemination avenue.

CDS (and possibly other elements of the Scientific Information Landscape preservation layer) should offer the storage and preservation of research artefacts in standardised, non-proprietary,

[4] See https://openscience.cern or directly access to the complete policy text: https://cds.cern.ch/record/2835057/files/CERN-OPEN-2022-013.pdf

well-documented and unencrypted formats (for instance, text-based files, PDF/A or TIFF for articles). As the central repository for CERN scientific articles, it should offer versioning and needs to ensure the unique and consistent naming of records. File names should be unique, consistent, informative and have the ability to be sorted/updated easily. File names may contain information such as project acronym, title, location, year(s), data type, version number, and file type. By **consistently assigning DOIs** for each record, CDS will allow a high level of interoperability with other external repositories, aggregators and dissemination services.

CDS should also interact with collaborations' internal databases that are in charge of preserving non-approved analysis proposals (either through proprietary collaboration tools or CERN Analysis Preservation) to ensure the ability of a complete review of CERN's physics analysis proposals.

The parallel dissemination of research products into a multitude of outlets is important for the scientific discourse and supports CERN's mission to share information. Nevertheless, it leads to duplication of manual submission efforts and introduces a lack of consistent metadata. The latter creates subsequent manual curation efforts or leads to insufficient discoverability and interlinking of research products. Therefore, a **single entry point** for the submission of research outputs to several downstream publication outlets such as repositories (e.g. arXiv, Zenodo, subject repositories, HEPdata) as well as scientific journals (assuming standardised automatic interfaces that allow interoperability) is proposed in addition to preserving the research products in CERN's repositories such as CDS, CERN Open Data Portal or Zenodo. Many CERN experiments maintain external websites to showcase their scientific publications. Such collections should also be automatically populated through this single entry point. The single entry point for submissions should be part of the **central workflow management layer** of the future Scientific Information Landscape.

Such a one-stop submission interface should provide a clear definition of the different functions of the connected (internal and external) platforms, the types of documents that are supposed to be hosted in the different landscape components or that can be submitted to publication outlets, the licences that should be applied for dissemination through the institutional repository or submission to external outlets, etc.

During the unified one-stop submission process, authors will define **standardised and detailed metadata,** which will be partly pre-populated from the central bibliographic database as well as the central author and funding database introduced in chapter 4.1.1. The use of persistent identifiers will be actively recommended or even made mandatory. The metadata will be consistently and systematically disseminated to all relevant CERN repositories and the downstream outlets (e.g. external repositories) to eliminate today's duplication of manual metadata population and to ensure consistent metadata across platforms. For papers deriving from EC-funded projects, metadata should be also propagated to the relevant EC portal in the required format.

During this single entry point submission process, authors also have the ability to correctly link related research products such as articles, datasets and software. By integrating the single entry point mechanism in the workflow layer, its proposed business-intelligence functionality should produce powerful reports and statistics around the dissemination outlets and avenues chosen by the CERN community.

4.2.2. Dissemination by a publisher

4.2.2.1. Introduction

According to CERN's Open Access Policy for CERN publications, "CERN authors are required to publish all of their peer-reviewed primary research articles open access." While dissemination through repositories is typically managed and controlled by the authors, dissemination by a publisher is under the control of a third party. This dissemination route is typically applicable for scientific articles only (but some publishers also offer the publication of software or datasets). The process includes a peer-review process, and the final publication is typically available through a range of discovery services. While publication by a publisher takes significantly more time compared to using repositories and typically involves a high cost for OA publishing, its peer-review and systematic dissemination represent a significant advantage over only depositing in repositories.

4.2.2.2. Actors

The dissemination of scientific artefacts is, in principle, the key responsibility of its authors and should only take place after the successful completion of the approval described in chapter 4.1.2. In larger experimental collaborations, the dissemination is often centralised in the collaboration secretariats, while in smaller collaborations or in publications originating from CERN departments, one of the authors submits the research output to the respective outlets. The Scientific Information Service regularly assists authors with advice concerning outlets, open access conditions, licences to be used or correct metadata definitions. Third-party publishing companies play a key role in the dissemination by publishers as receivers and distributors of scientific information.

Please refer to figure 6 in the previous chapter for a schematic display of the actors and processes.

4.2.2.3. Processes and Tools

Publishers typically run proprietary submission and production systems. Article authors (or their proxies) have to submit their manuscripts through these systems, often including manual input of respective metadata such as author names and affiliations.
There are different ways in which the peer review process is organised, but it always includes regular communication between the publisher/editor and the author(s) about feedback from reviewers and respective updates/additions to manuscripts.

Publishers offer services for disseminating content in the form of feeds of tables of contents and alerts about new publications.

4.2.2.4. Key Discoveries and Recommendations

Due to today's separation of submissions to the different outlets, discrepancies in the descriptive metadata between the preprint and postprint (published) versions of articles are inevitable. To match respective records and to ensure a coherent corpus of the literature, either manual interventions from curators are required, or complex matching algorithms for automated processing need to be developed.

The proposed introduction of a **single entry point** for the submission of research outputs (see 4.2.1) will effectively address this problem, as the same information (terminology, granularity, etc.) will be

provided in a consistent way to all downstream publication avenues. This will obviously require interaction with 3rd party systems of publishers or preprint repositories such as arXiv.org. Such interaction could either be through direct APIs (if the receiving side supports that) or by generating standard metadata exports that can be used for the respective submission processes.

However, such a technical solution should be complemented by **efficient collaboration between the Scientific Information Service, other libraries in the field and metadata providers**, namely publishers. Such collaboration needs to aim for a discipline-wide standardisation of metadata, i.e. bibliographic control. The ultimate goal of such efforts is to establish Universal Bibliographic Control[5], which should lead to sharing the effort of artefact description, to eliminate ambiguity and redundancy through sharing and re-using records.

4.3. Organisation and Management

4.3.1. Organisation by information managers

4.3.1.1. Introduction

Dissemination and organisation are very closely interlinked: Disseminated information has to be carefully curated and further disseminated to any other important information outlets. Therefore, these processes often happen in parallel. Information managers not only monitor the (often automated) ingestion of scientific artefacts into appropriate databases but also verify and enrich metadata, as well as organise existing information, for example, by collating artefacts into a collection. Based on well-maintained metadata, reports and statistics to monitor organisational output can be generated.

4.3.1.2. Actors

Information managers can be split into two main categories, namely content curators and the technical product owners/maintainers of the repositories or databases. At CERN, the tools in use (CDS, Zenodo, Indico) are maintained by the IT department, or in the case of INSPIRE, by SIS. EDMS, on the other hand, is managed within the EN department. An external repository that plays an important role in metadata management at CERN is arXiv, which is managed by Cornell University (its role will be further described in the next chapter 4.3.1.3). Relevant tools for the organisation and management of research data are: CAP, which is maintained by SIS; HEPdata, operated in collaboration between Durham University and SIS; as well as the CERN Open Data Portal and Zenodo, both maintained by the CERN IT department.

The main content curators at CERN are within SIS, who ensure that all CERN-relevant publications end up in CERN's institutional repository CDS. There are various ways a document can end up in CDS, "travelling" through multiple repositories and databases (see chapter 4.3.1.3). This means that a record could potentially be curated by multiple people throughout its life cycle. A paper submitted to arXiv will go through a first check by the arXiv content managers before they are harvested by INSPIRE and curated and enriched with metadata. This record is then exported to CDS, where it is once more verified by dedicated cataloguers. Therefore any curators within arXiv, INSPIRE and CDS are important actors when it comes to information management and organisation.

[5] See https://en.wikipedia.org/wiki/Universal_Bibliographic_Control

Not every tool that is used to store information at CERN has a dedicated content manager. Indico or Zenodo, for example, give users full control over which information they want to share and in which way they want to organise it. How content creators and owners manage and organise information is further described in chapter 4.3.2.

4.3.1.3. Processes and Tools

As active information management and organisation is currently focused on scientific articles, the following section has a similar focus. However, for a future information landscape, similar considerations need to happen for all other types of research products.

As demonstrated earlier, a document written by a CERN author can end up in several tools depending on the document type and place of submission. This multitude of dissemination outlets is described in Figure 7 below. Some interoperability and synchronisation is set up for receiving metadata from certain publishers, arXiv and INSPIRE into CDS. However, for information submitted to Indico, Zenodo or any other internal database, there is currently no way to exchange metadata other than manual input. For documents submitted to EDMS, a link to the corresponding record in CDS can be manually added, but this would only create a one-directional reference, while it is impossible to exchange metadata between the two systems.

If an article is submitted to a journal, its metadata might be imported to INSPIRE if the journal falls within its scope. There, the metadata is curated, and if relevant to CERN, it is further exported to CDS. Journals that are outside of INSPIRE's scope have to be captured manually by curators. A CERN-relevant arXiv submission, however, will be automatically imported via INSPIRE into CDS without manual curator intervention. INSPIRE then ensures that records coming from arXiv are matched with the records that come from journals before they are further exported to CDS. For information managers and curators, this **matching is the most challenging task**. If the paper was already submitted to CDS before metadata from INSPIRE arrived, the newly arriving record will either create a duplicate unless it matches certain persistent identifiers. This is why using unique persistent identifiers throughout the entire lifecycle of a document is so important; the matching can be performed automatically if the same identifier is used in every performed submission for the same document. Otherwise, this matching becomes very difficult or even impossible, for example, if the title has changed between the preprint and publication stages.

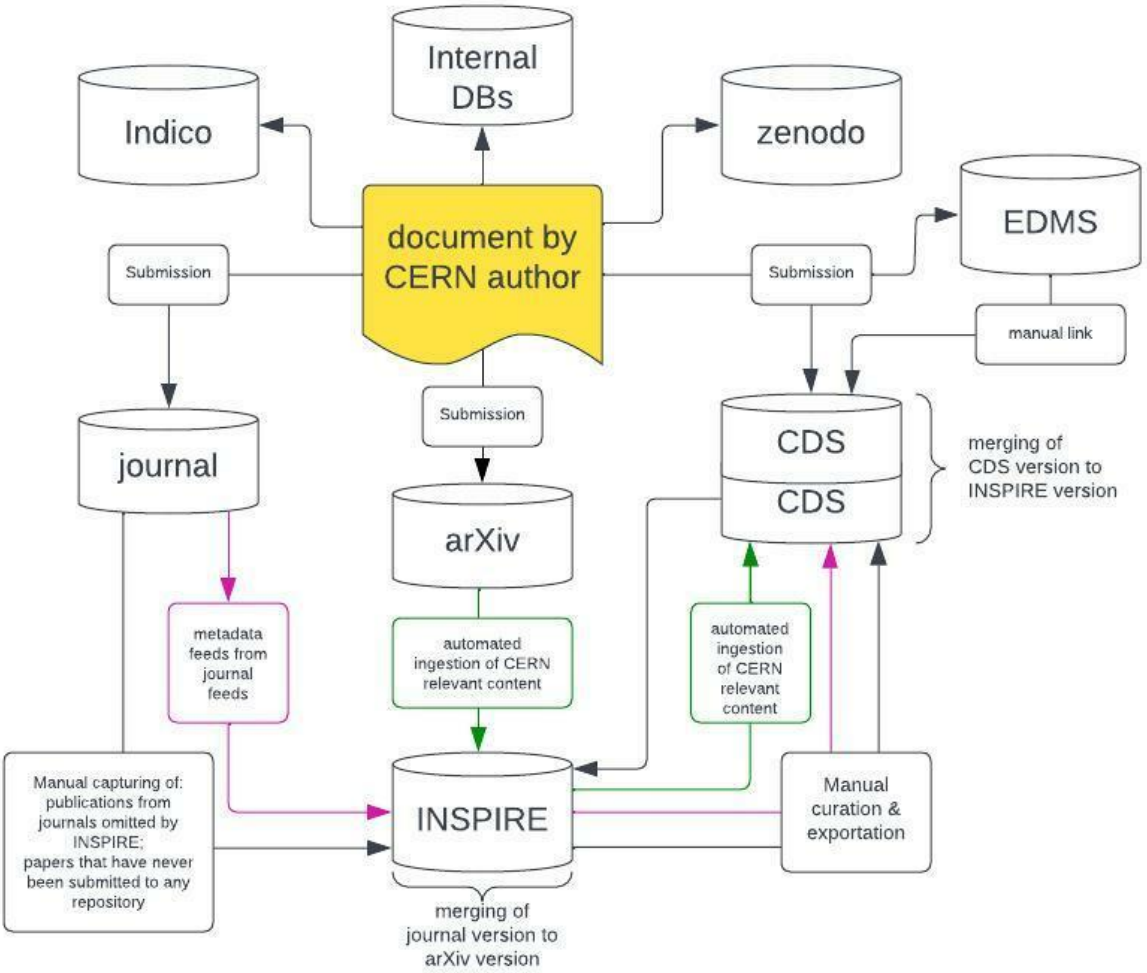

Figure 7: Metadata ingestion & synchronisation workflows at CERN

Once the information is in CDS, it is **generally organised into collections**, depending on the topic or association with a collaboration, experiment, or project. It is then usually further structured according to a wide variety of document types. Collections are set up upon request by the infrastructure managers in IT in conjunction with SIS as content managers. The collections are usually fed by submission forms that are set up on a case-by-case basis, which often involves a significant amount of effort. A simple, standardised submission process would facilitate the organisation of information, as well as create more incentives for users to submit their information.

The content is further organised by a standardised list of broad subjects as well as free text keywords. These keywords, however, are neither used consistently, nor are they normalised in any way, which therefore does not offer much additional value when trying to organise content systematically. This is particularly problematic when searching for specific particles, as these can be written in many forms via TeX formulas and are, therefore, difficult to retrieve in a search. A

maintained ontology of normalised keywords would help to better structure the information and improve discoverability.

In order to further group and collate certain records, metadata tags are assigned for experiments, studies, and projects whenever possible. However, it is not always obvious for curators to identify a paper as being part of a specific experiment without a clear mention in the text; therefore, it is important for authors to add this information already at submission. The submitter can have a big impact on metadata quality and organisation of content, as is described in chapter 4.3.2. below.

4.3.1.4. Key Discoveries and Recommendations

The information landscape at CERN is very broad, and each tool has specific functionalities for different document types. One of the key findings from the stakeholder interviews was that the authors do not fully understand the different use cases of the various tools. This should be taught to every new arrival at CERN on an institutional level, and each tool should have a clear content policy. Authors/submitters are confronted with a choice of where to submit their paper, code or data and usually end up submitting to multiple tools at the same time. A single entry point for submitting would not only be much more efficient for the user but would also allow more **consistent metadata** to be shared across repositories and databases. This would reduce manual curator intervention (as only one single record has to be touched), and reduce the risk of duplication (as mentioned in chapter 4.2.2.4)

As records can be submitted separately to multiple tools at the same time, different repositories gather metadata at a different stage of a document's life cycle. As described in more detail in chapter 4.3.1.3, matching these records and avoiding duplication is an important and difficult task for curators and requires substantial manual intervention. **Assigning a unique persistent identifier** from the very start of the life cycle that follows it throughout would hugely facilitate this matching process.

Furthermore, the metadata received from publishers and journals often differs in quality. Publishers are increasingly enriching their metadata by attaching identifiers like ORCID and ROR, or copyright and licensing information, but these efforts are inconsistent. This again means manual curator intervention is necessary to enrich and correct missing metadata. Using **standardised metadata to be exported across all publishers** would decrease manual curator intervention by a large margin; however, this requires practical changes to occur outside of CERN's circle of influence, i.e. external metadata providers like publishers.

Moreover, the lack of interoperability between Indico, Zenodo and external databases with other tools complicates metadata exchange. This firstly could mean that relevant information that should be stored in CERN's institutional repository is missing or, for example, is only stored in an Indico event and hence not adequately preserved. As already mentioned previously, **relevant documents should be systematically transferred from Indico to CDS** for long-term preservation and consistent discoverability. Furthermore, access rights restrictions could also further hinder the retrievability of crucial information. This is particularly visible when referencing an EDMS document in a paper, as links to EDMS often lead to a restricted document. It should be ensured that referenced documents and data are publicly accessible.

In order to further improve organisation and discoverability, a high-level classification system of research products should be introduced that allows grouping according to various criteria (e.g. experiment, department, type of document, etc.) using a controlled vocabulary maintained by SIS in agreement with other stakeholders. Finally, the ability to explicitly **link related artefacts** together, which allows jumping from one to the other (e.g. paper and data, conference and contributions, book and chapters, author and papers, etc.), is a crucial requirement for efficient organisation and management.

Another functionality that is currently not covered by any of CERN's information management tools is storing information related to publication fees or funding information. As CERN aims to publish 100% of its results in Open Access (OA), the new information landscape should support workflows and **OA monitoring capabilities** for efficient OA budget planning, cost-benefit assessments, determining the eligibility of a paper for the coverage of OA fees, as well as reporting of OA policy compliance (see also chapter 4.5.2). Such OA-related information may include

- Information about the "nature" of authoring collaboration (CERN? EU-funded? CERN recognised experiment?) as a justification for (non-)coverage of OA fees.
- Information about the amount of OA fees or if fees are related to a special issue to determine if a special authorization is needed.
- Information provided by publishers about who paid for OA, in particular, if other than CERN.
- Information about licences to support future reuse in other publications.

EU Projects at CERN are subject to some specific requirements that are not completely met at the moment by any of the existing repositories. This leads to manual workarounds or duplication of efforts.

- Some documents require additional approval layers (e.g. milestone approvals by the CERN EU office) that can be supported in the future landscape by the central workflow management system.
- EU projects need additional specific metadata elements related to the project itself (project, milestone or deliverable identifiers). They need to be captured during the document creation and need to be searchable in the repository.
- Internal documents (i.e. not for public consumption) originating from EU projects might not always have CERN authors. Nevertheless, the project team will need to have a repository that can expose and preserve such documents, even though they are out of scope for a CERN institutional repository.
- All published documents will need to be registered in a central EU monitoring facility. A repository serving EU projects will need to support an automated interface with such a monitoring facility.

4.3.2. Organisation by information owners and creators

4.3.2.1. Introduction

Where it does not require an information manager's intervention, information can be managed by the author or the collaboration itself. This is usually the case for still internal information in order to keep

track of drafts before they are ready to be made publicly available or simply to create information that should remain restricted for other reasons.

4.3.2.2. Actors

At CERN, it is usually the responsibility of the author to submit their documents/data or software. Sometimes this task is taken over by a departmental or experimental secretariat, particularly if there is an approval involved. Where authors can submit documents has been described in Figure 7, from where information managers within the SIS team ensure that the information is always preserved in CERN's institutional repository CDS.

Bigger experiments additionally organise their relevant information like drafts or any other pre-publication-related data in internal restricted tools to keep track of the output of their activities. Usually, there is a publication chair and a scientific committee who make sure that all output is assigned an identifier and the progress can be tracked. However, smaller experiments that do not have the resources to develop their own databases rely on CDS' infrastructure to manage their information. Submission forms and collections, as well as approval processes, have been established to ensure that users can manage and publish their own information on CDS.

For engineering documentation, which is mainly stored in EDMS, super-users called "Local Administrators" mainly manage documentation and data. They can create new folders and manage access rights directly, so self-archiving with some interventions from local administrators is the usual practice.

CERN, as an international organisation, collaborates with many non-CERN partners. Therefore, some authors that are not necessarily CERN-affiliated play an important role within the landscape. This is particularly relevant for EU-funded projects: the output of a project should be publicly available and manageable for authors from different organisations.

4.3.2.3. Processes and Tools

Authors have the option to submit their paper to multiple tools at the same time, for example, to CDS and to arXiv, before they submit a paper meant for publication to a journal. Once submitted, the information managers at SIS ensure that papers end up in the institutional repository, CDS. Sometimes the **submission is delegated to the collaboration or departmental secretariat**. For larger experiments, the papers are internally managed by their own databases, including internal identifiers and keep track of comments and approvals. The process has been further explained in chapter 4.1.2.

In order to create a new collection or submission form in CDS, the CDS support team has to be contacted, who, in conjunction with the information managers at SIS, create the new requested infrastructure. Any changes or additions must go through the support team and cannot be performed by the authors or collaborators themselves.

Authors can also manage their papers in ORCID or INSPIRE on their author profiles, where they can claim authorship of papers and display them in one place and include some statistical data. This requires, of course, that these documents are submitted in the first place.

For collaborations with partners from outside CERN or EU projects, Zenodo is mainly used to ease collating the output of a project where CERN was a collaborator but not the sole contributor. Zenodo communities can be created with access to a specific group of people which makes it an easy-to-use organisational tool for self-archiving, as communities can be created without intervention from IT support.

4.3.2.4. Key Discoveries and Recommendations

When analysing the processes involved in CERN's scientific information landscape, it becomes apparent that there is significant duplication of effort at all stages. A single document can be created in multiple tools at the same time, but only a few of those tools can actually handle merging the different versions or linking/synchronising them. Users are confronted with decisions on where to submit and, in the case of scientific articles, usually have to submit (in parallel or subsequently) to three outlets: CDS, arXiv, and finally, to their selected journal for publication. A single entry point for submission would vastly improve efficiency and facilitate synchronising and merging of different versions (see more detailed explanation in chapter 4.2.1). Additionally, as some tools are not interlinkable, information that is submitted to Indico, for example, is rarely ever also exported to a more suitable tool for preservation and discoverability. Increased **interoperability between the different tools** could vastly improve information availability and organisation.

Interoperability could again be improved by using a **single metadata standard across all tools** (see Annex 5) or, at a minimum, via a single export format so that curation efforts of the same record in multiple tools can be avoided. The current process relies on ad-hoc scripts (e.g. to retrieve, compare and merge information into existing records) instead of exploiting generic, source-agnostic, interoperability solutions.

When it comes to organisation, records can usually be collated by subject or activity, but a **normalised vocabulary** is necessary to make this possible. This is particularly a problem with keywords that are not normalised, which makes it difficult to group documentation together. As for organising and collating information relevant to a single activity, this can be done in Zenodo via communities or in CDS via collections. Zenodo communities can be set up by the collaboration themselves, whereas CDS collections have to be created by CDS support. In order for a user to submit to CDS, a dedicated submission form has to be created, which usually involves significant effort from the CDS support team. Any changes or additions must go through the support team and cannot be performed by the authors or collaborators themselves, which can sometimes lead to unnecessary delays. The proposed unified and customizable workflow management layer of the future information landscape should adequately address this shortcoming (see chapter 4.1.2).

To support their author communities with the required information (bibliographies, author lists, grant information), larger collaborations currently maintain their own internal databases, resulting again in duplication of efforts and inconsistent data. As outlined in chapter 4.1.1, the results of this analysis strongly suggest the development of a centrally maintained database that could replace individual (collaboration-specific) solutions as long as a certain degree of flexibility is available.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Organisational information like funding and grant information is also not systematically stored in the metadata but rather in the acknowledgements, which leads to the necessity of full-text search functionality to enable searching for specific grant numbers. A research information management system should store this information within the record.

Interoperability across all CERN information management systems (including those that are not entirely meant for scientific information, e.g. EDMS and Indico) and with key external services (repositories and databases) should be ensured by metadata format and taxonomy standardisation, extensive adoption of persistent identifiers, and by workflows exploiting actionable metadata, therefore minimising manual input. Process simplification across the entire scientific information value chain, for instance, by establishing a **single point of metadata submission**, should pave the way to the creation of a "container record" describing all interrelated manifestations of research output and, as such, facilitate the easy adoption of open science practices. A particularly important example is the interoperability with and linking to external databases for code management (such as GitHub) or datasets (such as HEPData).

4.4. Discoverability

Discoverability of scientific artefacts needs to be considered from two perspectives. Firstly, for the preparation of their next research product, researchers need to be able to discover relevant information authored by scientists from all over the world, i.e. not limited to CERN research products (see also Usage in chapter 4.5). Secondly, CERN's research output itself should be discoverable for the rest of the scientific community in order to allow the widest possible reuse and application of the research. In the following section, both aspects are considered separately; we consider these two elements by discussing how to allow CERN researchers to find scientific information and how to best make CERN's research output discoverable for the rest of the world.

4.4.1. Discoverability for the CERN Community

4.4.1.1. Introduction

Discoverability in this context refers to enabling the CERN community to find all relevant information for ongoing research within a CERN department or experiment. Researchers need to be able to find a comprehensive list of research artefacts relevant to their respective work in a way that meets their regular work patterns and can support the unspecific (direct) exchange within the research team, i.e. the general exploration of ideas and theories. Research artefacts include not only scientific articles but also datasets, software, etc., that are relevant to the research or need to be considered to allow a comprehensive conclusion on a specific research question.

4.4.1.2. Actors

CERN researchers need to be able to easily find relevant articles/datasets for the purpose of including their original findings in new research and, respectively, creating a compilation of cited references. Typically, authors need to extend their search beyond the domain of research originating from CERN. Information Managers should enable researchers to perform such extended searches and need to provide tools and services related to this task.

4.4.1.3. Processes and Tools

There are plenty of tools available to the CERN community. These include generic search engines such as Google; tools that focus on academic content (Google Scholar, ResearchGate); or discipline-specific search engines or aggregators such as INSPIRE or ADS. In addition, fee-based bibliographic databases (Web of Science, Scopus) allow retrieval of information about publications in selected journals or conference proceedings and complement the services of open databases. However, CERN today has no active subscription to such databases.

A survey amongst the CERN community (196 complete responses) has investigated the typical tools used for discovering scientific content. The results show some differences between tools used to discover traditional literature (articles, book chapters, etc.) and those to find other research products (data sets, software, etc.).

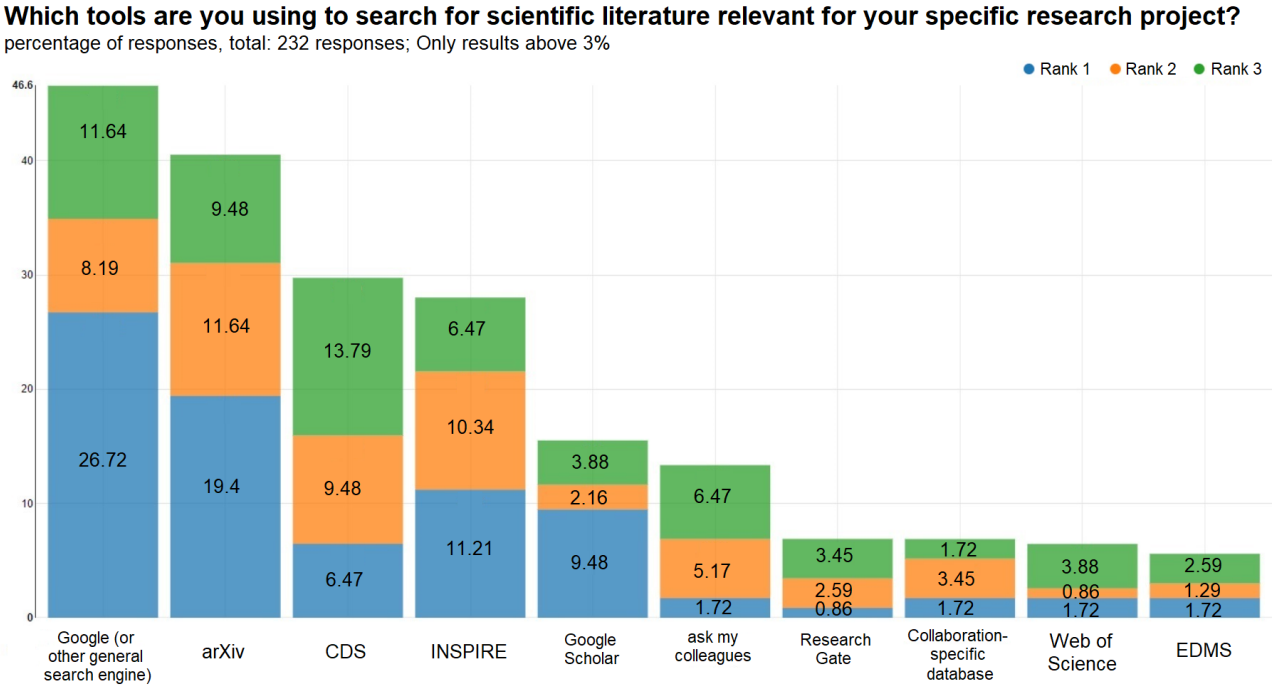

Figure 8: Survey results: Which tools are researchers using to search for literature?

To find scientific literature, the first choice for a majority of researchers at CERN is Google (or other general search engines) or Google Scholar, followed by a search for pre-prints in arXiv and discovering literature in INSPIRE. CDS still plays a role in finding literature, but typically only as a second or third choice, i.e. if the search in other sources was not successful.

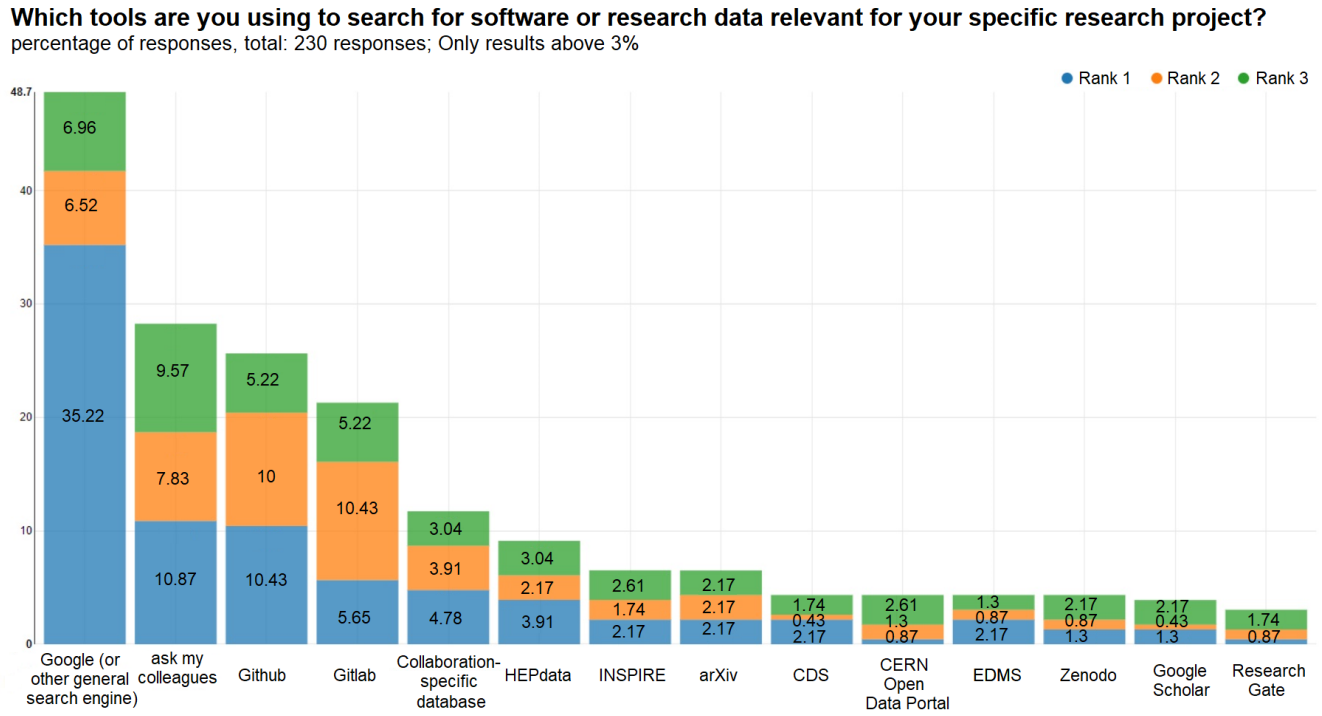

Figure 9: Survey results: Which tools are researchers using to search for software or data?

Similar to the results for literature, Google (or other general search engines) is the main choice for CERN researchers to find Software or research data. The dominance is even more prevailing due to the lack of comparable alternatives such as arXiv or INSPIRE for literature. It should be noted that the latter will introduce a search for software and datasets in 2023. Asking colleagues directly is a frequent practice for software and data, which underlines the lack of a proper entry point to search for such artefacts.

4.4.1.4. Key Discoveries and Recommendations

As the survey results have demonstrated, the **CERN community regularly uses a wide range of tools in order to find relevant research products**. This is not only the result of the habits of a diverse research community, but it is also a sheer necessity due to the fragmentation of the relevant information. Many users direct their efforts first to general search engines such as Google to find relevant information. Subject-specific aggregators such as INSPIRE are equally a useful starting point to allow researchers to look for information across a widespread scientific corpus. However, such a tool can never deliver a 100% complete list of relevant information as researchers will often also need input from institution-internal engineering documents or datasets, which are not typically widely disseminated. Such information is stored across several CERN databases as they are also meant for different audiences or represent information from different stages of the research process.

To meet the crucial need of a researcher to find a comprehensive list of information related to a specific topic, it could be a solution to create a **federated search service** that combines results from all internal databases with results from subject-aggregators such as INSPIRE or general repositories such as arXiv.org. To further improve the results of such a federated search, it will be useful to develop a **CERN-specific set of keywords** that are then systematically assigned in the underlying

proprietary databases and mapped to global standards, such as the INSPIRE keywords assigned by DESY. However, given the dominance of Google or similar third-party tools as the first entry point for researchers, and considering that researchers seem to be generally happy with the results obtained (see Figure 10), **efforts to develop an advanced search functionality for researchers should be carefully assessed against their actual added value** for the CERN community. It might be only useful if integrated into a CRIS functionality (see chapter 4.5.2), which provides further capabilities for information managers or even researchers to find and interlink different scientific information, for instance, by using dedicated bibliometric indicators for refining their search for scientific information.

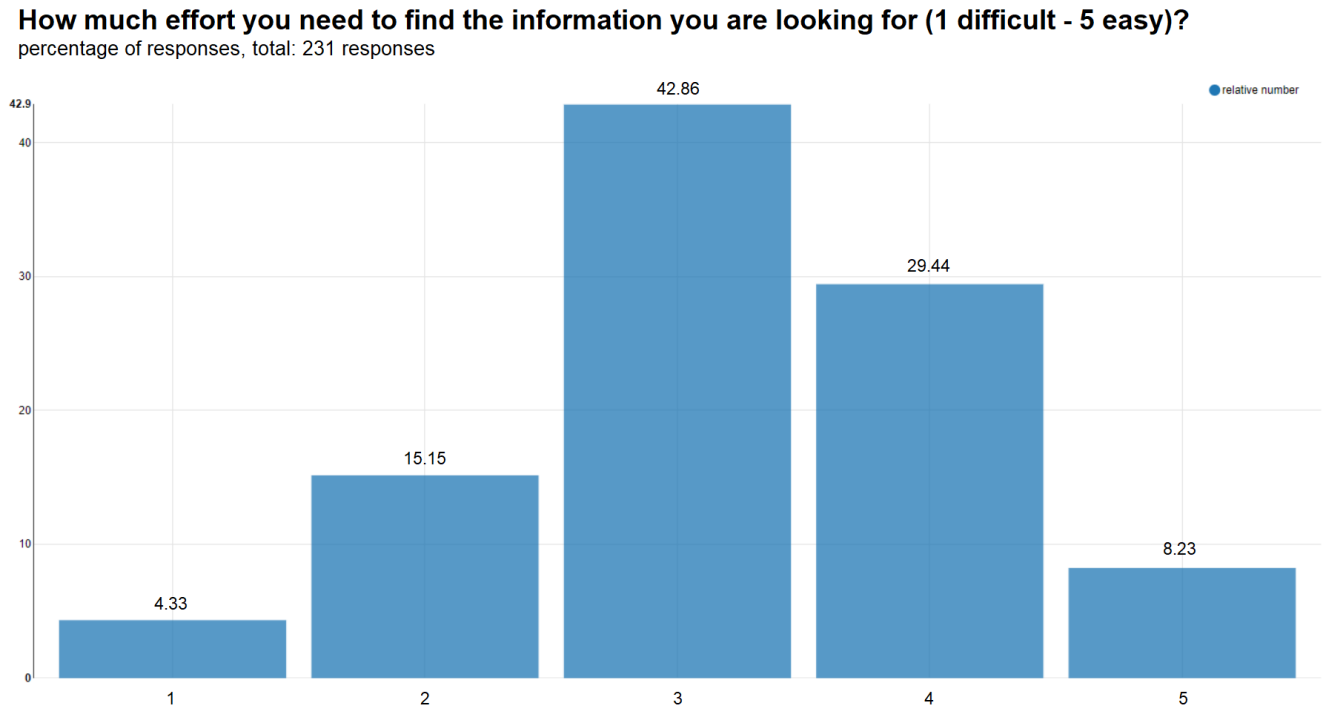

Figure 10: Survey results: How much effort does it need to find information?

4.4.2. External Discoverability of CERN Output

4.4.2.1. Introduction

While it is important for the research community at CERN to find all relevant information, it is equally crucial for the success of a research product that the scientific community outside CERN can easily discover it. Only by wide discoverability of CERN results can they optimize their impact on society. As such, external discoverability considerations need to consider both the availability of key elements to allow wide exploitation of CERN results and the ability to measure the extent to which such exploitation occurred. The latter is strongly connected to CRIS functionalities which are further discussed in chapter 4.5.2.

4.4.2.2. Actors

The CERN community aims to maximise the dissemination and the findability of its output. Dissemination takes place through harvesting by external services and should be optimised through the definition of the most relevant results to be shared with the external world. Findability for external users should be the result of the optimisation of subject keyword indexing and of relevance ranking criteria, as well as search engine optimization.

4.4.2.3. Processes and Tools

Dissemination is an inherently passive process, though it should be enhanced through proactive dissemination to selected, high-visibility channels.
Interoperability with external indexing services and repositories should be optimised through the adoption of common metadata format standards, adaptable to metadata producers' APIs.

4.4.2.4. Key Discoveries and Recommendations

A high degree of granularity of **descriptive metadata** can ensure good findability if the **research products are indexed in key search engines**, both scientific-oriented services and general search providers such as Google. Indexation could be also achieved indirectly through subject aggregators such as INSPIRE. Though useful in some specific cases, full-text searchability cannot compensate for the absence of structured, highly granular metadata. For example, funding and grant information is usually not in metadata. Therefore, users must rely on full-text search (analogy to chapter 4.3.2.4).

4.5. Usage

4.5.1. Use for new research products

4.5.1.1. Introduction

The use of scientific information for new research products closes the loop of the lifecycle introduced in chapter 3.1. In the typical scientific discourse, new scientific information created by authors uses and re-interprets other scientific information in the form of scientific articles, research data, software, etc. Therefore, findings and recommendations for the usage and respective authoring steps will be critical as their positive (or negative) effects are relevant over and over again in the scientific process.

4.5.1.2. Actors

As the usage of scientific information for new research products is, in fact, identical to the originally described authoring step (see chapter 4.1.1.), the actors are identical and have been described already earlier in this document.

4.5.1.3. Processes and Tools

As the usage of scientific information for new research products is, in fact, identical to the originally described authoring step (see chapter 4.1.1.), the processes and tools are identical and have been described already earlier in this document.

4.5.1.4. Key Discoveries

Today, many experimental collaborations maintain their own proprietary bibliographic databases to allow targeted search for relevant research products. This causes a significant duplication of efforts as these databases are maintained in parallel serving mostly the same content (partly in slightly differing formats). A **centralised bibliographic database for all CERN experiments** would eliminate this duplication and ensure a common quality standard for bibliographic data. Such a central service could be either built on top of existing discipline-specific services such as INSPIRE or can be established as a stand-alone service within the new Information Landscape. However, it seems crucial to overcome the fragmentation and duplication of these efforts.

Once respective information is found and re-used in a scientific article or other research product, good academic practices require citing the original source. However, due to the fact that many different citing formats exist and a list of references could be very extensive, the proposed federated search service (see chapter 4.4.1.) should also provide a **standard format export of references** that can be easily imported into the new research product. The INSPIRE BibTeX export can be used as a best-practice example for such functionality.

4.5.2. Use for statistics

4.5.2.1. Introduction

As a research lab, scientific information is the key output of CERN. Therefore, statistics around the number of research products disseminated (e.g. the number of CERN articles published) are important for measuring the success and efficiency of the Organization as a whole (for instance, as part of the CERN Annual Report) or of individual projects or experiments. In addition, CERN Member States or associate Member States regularly measure the effects of their CERN participation through such statistics. Beyond the sheer number of disseminated objects, their impact is also of high relevance, e.g. measured through the number of citations. Finally, also other external stakeholders, such as funding agencies, are regularly expecting metrics around the scientific output of supported projects or experiments. Many such reports and statistics are generated centrally by the CERN Scientific Information Service (SIS), but there are also other actors, as outlined below.

4.5.2.2. Actors

SIS is the principal shepherd of CERN's scientific output and is responsible for all central monitoring of research products. Information specialists in SIS need to regularly prepare different views and reports related to CERN's scientific information. While reporting on scientific articles was traditionally in focus, statistics related to other research products have become increasingly important.

The (a)MS relations team at CERN International Relations is expected to respond to regular enquiries from governments. Due to the lack of any directly accessible statistics, the IR team typically requests SIS to create respective reports on an ad-hoc basis.

CERN Experiments are typically independent collaborations with their own governance and funding mechanisms. Therefore, they need to regularly report to their members and funding organisations on the effective outcomes of their research activities. Experimental secretariats regularly generate reports and statistics, and larger experiments showcase their publications on dedicated websites.

Many projects at CERN are funded through the European Commission. Such funding typically comes with specific reporting requirements on the number and type of generated research products. The actual project team, SIS, and the CERN EU Office are involved in the generation of respective reports, particularly for regular milestone reports or project reviews.

4.5.2.3. Processes and Tools

SIS currently generates two **regular annual statistics related to CERN publications**. For the CERN Annual Report, the statistics cover the total number of publications, their respective split by subject and journal/publisher. In addition, an annual report on the state of Open Access is produced for the Scientific Information Policy Board and the SIS Annual Report. Both statistics are primarily generated using CDS but require tedious data cleaning and manual mapping.

Other reports, such as statistics for CERN (a)MS, are generated ad-hoc again using CDS but also other sources such as INSPIRE or Web of Science. Due to the high manual effort, reports are only created upon request and need to be planned well in advance to ensure the availability of resources.

In addition to the centralised reporting, experimental collaborations and EU project teams (or the EU project office) generate statistics related to the respective output of experiments or projects. Such reports are typically done manually (in Excel etc.) or using some collaboration internal tools. For the **central EU projects reporting**, the CERN EU Office typically uses Zenodo, as not all project outputs consistently have CERN authors and, as such, are not necessarily captured in CDS.

4.5.2.4. Key Discoveries and Recommendations

Given the importance of scientific information as a key product of CERN's activities, the ability to generate statistics and additional knowledge using metainformation around the organization's scientific information corpus is extremely limited. **Reports are cumbersome to generate and typically involve manual efforts** to correct inconsistent metadata. Available reporting capabilities are very much limited to traditional statistics related to the actual number of published articles or created internal records. Further quantitative or qualitative assessments, such as impact analyses, are not available at all, as this would require the consideration of linked research products (e.g. secondary literature using CERN output as a basis). **Statistics covering other research products** (such as datasets or software) **are not yet generated**, partially because there was not yet sufficient demand to establish respective processes; partially because such products are not preserved/stored in a consistent manner and hence there is no suitable data source for central monitoring. This generates a significant disadvantage for CERN when competing for future research funding and creates significant manual efforts to meet minimum reporting requirements from funders.

The distributed and diverse nature of publication and approval processes (see chapter 4.1) does not yet allow for systematic measurement of process efficiencies, and hence possible improvements are difficult to identify/quantify. It is also not yet possible to systematically measure adherence to CERN's related policies, such as the Open Access or Open Data policies.

The proposed new central workflow management layer introduced in chapter 4.2 should not only simplify the dissemination of all scientific products and improve the consistency and quality of

metadata, but it also represents the opportunity to implement a **business-intelligence component to produce reports and statistics** and run analyses around the approval processes (e.g. time to approval, the average paid OA fee, publications by department/experiment, dissemination outlets, etc.). In addition to reports and statistics out of the workflow management layer, the future CERN institutional repository should equally enable systematic and consistent reporting on CERN's scientific output. This could be established through a **CRIS (Current Research Information System)**[6], which could be directly included in the roadmap towards a new holistic scientific information landscape, or (if not immediately in scope) could also be developed later, in which case the original repository architecture should be designed considering such a foreseen future evolution into a CRIS. A CRIS will support the CERN Directorate's strategic priority to assess the impact of its scientific output on the scientific community and on society at large. Its implementation is therefore strongly recommended and should follow established best practices[7].

As opposed to today's manual, ad-hoc reporting, the systematic availability of key metrics will allow more proactive use of such data in the relationship management with all CERN stakeholders and enable experimental and departmental leadership to revisit the efficiency and effectiveness of scientific dissemination processes. In the context of the recently announced new annual CERN Open Science report, the requirements around scientific information-related KPIs will further increase, and hence systematic reporting will become mandatory.

5. Summary of Recommendations

Need for central workflow management for scientific information

CERN is a large and complex research environment with a hugely diverse research programme. However, from the series of user interviews, we conclude that the requirements and practices around scientific information management are surprisingly similar. While there are natural differences in process complexity and the number of involved researchers (depending on the size and organisation of the experiment or department), the overall process management needs are comparable. Today, all experiments have set up individual processes and supporting databases, partly with custom software solutions and partly using basic tools such as e-mail. From the discussions, one of the core assumptions of this project was confirmed: it is not only feasible but will dramatically increase the overall efficiency amongst the CERN research community if CERN, as the host lab, would provide a **central scientific information management workflow service, including a centrally maintained database for author and funding information management and necessary bibliographic data.** Such a workflow service will set certain standards but needs to remain customisable to the needs of the experiments. Ideally, such a service will support the systematic generation of statistics and reports, and be interoperable with downstream publishing outlets such as preprint servers (arXiv.org) or submission systems of publishers, to minimise the duplication of manual handling of metadata.

[6] A CRIS is a system to store, manage and exchange metadata for the research activity conducted at a research-performing organisation. It includes a service layer on the top of the repository for the assessment of impact and cost-benefit analysis of the organisation. An exemplary detailed case study can be found here: https://doi.org/10.1016/j.procs.2019.01.090

[7] Examples: CRIStin in Norway, ACRIS in Finland, PT-CRIS in Portugal, OMEGA-PSIR in Poland

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Redefined role of an institutional repository

Research in particle physics is global. Researchers at CERN perpetually collaborate with peers at CERN and other labs. For the research process (in all domains), it is essential to have access to all relevant scientific information independent of its origin. Therefore, the role of an institutional repository needs to be revisited. Due to its natural limitation to CERN-originating content, in many cases, it is not useful for researchers to find scientific information. Instead, HEP researchers use tools like Google, Google Scholar, INSPIRE or arXiv.org. This reality should be taken into consideration when designing a new institutional repository. Built-in search functionality does not need to focus on the needs of researchers but rather on other internal users such as experimental/departmental administrators, Member State relations or SIS. The core functionality for the future CERN institutional repository is the **preservation of all scientific output of the Organization, accompanied by additional CRIS functionality to measure CERN's impact on society at large**. To meet the demands of researchers to find scientific content across different (internal and external) databases, **a federated search** across all CERN repositories and key external sources such as INSPIRE and arXiv.org **could be established if clear added value** over third-party services such as Google or Google Scholar can be demonstrated.

Lack of awareness

As stipulated in the problem statement of this project, the current landscape is complex and contains numerous duplications of functionality and considerable overlap in the coverage of scientific literature across the different tools. None of the interviewed stakeholders was fully aware of all available tools and services and their respective original roles. Equally, from the interviews, it became obvious that there are still significant knowledge gaps around open science practices, licensing and copyright requirements for publications, etc. A new landscape should not only be simplified by providing easier interfaces and aligned processes. Its implementation will need to go along with a **dedicated training and education programme for the CERN community** to avoid a similar future situation where documents and processes are widely mixed across tools, depending on personal preferences or driven by a lack of awareness. Current tools that remain in use in the future landscape should have a clear definition of the type of information to be stored and foreseen use cases.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Annex I: Current CERN tools and services involved

AI.1) Systems managed at CERN

CADI (CMS Analysis Database Interface) is the CMS collaboration database to manage collaboration memberships, author lists, workflows for preparing CMS papers and notes, as well as other collaboration-internal management tasks.

CERNbox is a cloud synchronisation service for end-users: it allows syncing and sharing files on all major mobile and desktop platforms (Linux, Windows, MacOSX, Android, iOS), aiming to provide offline availability to any data stored in the CERN EOS infrastructure.

CERN Greybook is a database that stores data on the Organization's research programme, including all the experiments and projects, their collaborating institutes and their participants.

CDS, the CERN Document Server, is CERN's institutional repository. In this role, it aims to aggregate, preserve and expose CERN's scientific output. Currently, it hosts almost 550,000 records, ranging from scientific publications to internal notes, administrative documents, policies, minutes of governance meetings, videos of internal CERN events or lectures, a database of photos, etc. This immense complexity and diversity of content is creating inefficiencies in the information management of scientific content. Beyond its core functionality as a repository (storing, preserving and archiving documents), CDS also acts as a workflow management system for a wide range of publication workflows. This is managed through a large number of unique submission forms that all serve different use cases in terms of covered document types and involved stakeholders. CDS also hosts CERN's e-tendering process, a workflow to manage competitive procurement processes.

CDS Videos is a sister service of the main CDS. In an attempt to better serve the needs of video content, the CDS team developed CDS Videos and started to migrate some content to the new platform, but a large number of videos also remain on the core CDS platform. CDS Videos hosts 5,900 recordings today, while more than 25,000 videos (primarily recordings of talks and seminars) are still hosted on CDS.

CERN Analysis Preservation (CAP) is a repository for the description and preservation of experimental analysis assets. Developed in close consultation with LHC experiments, CAP enables researchers to preserve and document the various components of their physics analyses, e.g. datasets, software, documentation, so that they are reusable and understandable in the future. Reusing existing collaboration tools and a flexible data model, CAP aims to be the end-to-end solution for scientific preservation that can be easily integrated into researchers' workflows, supporting open science best practices.

CERN Library Catalogue includes the collections available at the CERN Library. It includes more than 115,000 books, 23,000 conference proceedings, standards, electronic journals. In order to access documents, one can search within the catalogue, and either access the document online or borrow

the physical copy from the Library. Respective circulation management and workflow modules are part of the system.

CERN Open Data Portal is the access point to a range of data produced through the research primarily performed at CERN. It disseminates the preserved output from various research activities and includes accompanying software and documentation needed to understand and analyse the data. Currently, it hosts some 13,000 records that are shared under open licences and are issued with a Digital Object Identifier (DOI) to make them citable objects.

EDMS is CERN's Engineering & Equipment Data Management Service, already established in the 1990s. The tool was built to allow safe and structured storage and management of documents, but it also allows a collaborative workflow during the lifecycle of a document incl. clear version control and a highly sophisticated access rights concept. EDMS hosts 1.7 M documents, used by more than 4,000 users, mainly in the engineering and technical domain but also including financial and administrative documents. While in many cases EDMS documents are cited in scientific documents, the links are not always externally accessible and when they are, there is no guarantee they will stay so.

Foundation is the access point for referential data for the organisation. Foundation provides across all services that need access to persons, CERN's hierarchical structure of units, addresses, and roles.

Gitlab is an open source code repository and collaborative software development tools for small and large scale projects. CERN's Gitlab's instance manages hundreds of projects with more than 7,000 active contributors.

Glance was initially developed by ATLAS, with ALICE and LHCb joining the effort recently. It includes several systems (collaboration membership, institutes, author lists, conferences...) but it is also used for the management of publication workflows. ATLAS uses it to manage the first phase of some of its publications (draft circulation and approval) before those publications are then submitted to CDS.

Indico is an open-source event management software. In the realm of scientific information, it contains material submitted to conferences or similar events. No efficient and systematic mechanisms for importing metadata and the full text of documents are in place, but documents are typically submitted manually by conference participants/speakers. While in some cases, Indico documents are cited in scientific documents, the system doesn't guarantee long-term preservation or persistence of links. Several projects are ongoing or have been discussed for Indico to expand its role in scholarly communication, for instance, by introducing a robust proceedings publishing module.

INSPIRE is developed and operated as a collaboration of CERN, DESY, Fermilab, IHEP, IN2P3, and SLAC and serves researchers globally. INSPIRE hosts at the moment almost 1.5 million preprints, published articles, conference proceedings, research data from various sources in the scientific disciplines served, i.e. high-energy physics, particle physics, quantum computing, and accelerator physics. INSPIRE serves about 200,000 searches/day launched by 50,000 active users. Following the DORA principles, the service computes citations in HEP across all types of publications to enable institutions all over the world to fairly assess the impact of researchers. INSPIRE also serves as a

hub to post jobs in HEP as well as relevant seminars. It also hosts a journal, conference, and author database.

Twiki is an open source wiki platform for team collaboration and document management. There are numerous Twiki instances installed at CERN.

Zenodo is a large-scale open repository launched in 2013 as part of the EU-funded OpenAIRE project. Zenodo is operated by CERN but serves the worldwide scientific community. It hosts more than 2.2 million publications, datasets, software, images, videos and more across all scientific disciplines. It promotes the use (and direct minting) of persistent identifiers (DOI's) as well as versioning for its content. Zenodo is a trusted repository according to the EC grant conditions and is actively promoted by the European Commission as a central open repository for European research. Content can be organised in communities, and during the COVID-19 pandemic crises, Zenodo was actively used to accelerate the scientific exchange. The majority of worldwide software DOI's are minted through Zenodo.

AI.2) External systems relevant for CERN's scientific information

arXiv.org is a pre-print repository operated by Cornell University. Initially launched in 1991, it has become the standard platform to share research results in particle physics before publication. Today, nearly 100% of HEP articles, later published in journals, are shared as a preprint via arXiv.org. arXiv hosts more than 2 million publications, proceedings, design reports, etc. Thanks to its high level of adoption in the field and its dissemination speed, many physicists at CERN would consider arXiv.org as their primary source for scientific information. It has been selected as the recommended repository for publications originating from ERC funded research in the domain of physical sciences and engineering.

GitHub is a website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code. Github has more than 67 million users worldwide and over 352 million repositories.

HEPData, the Durham High-Energy Physics Database, has been built up over the past four decades as a unique open-access repository for scattering data from experimental particle physics. It currently comprises the data points from plots and tables related to several thousand publications, including those from the Large Hadron Collider (LHC). HEPData is funded by a grant from the UK STFC and is primarily managed by the IPPP at Durham University, while CERN supports the technical operation.

JACoW, the Joint Accelerator Conferences Website, is a platform, operated by an international collaboration, that publishes the proceedings of the main accelerator conferences. The operation started in the mid-nineteen-nineties. To allow for speedy and efficient publication of the proceedings, administrative tools, such as the Scientific Programme Management System (SPMS), have been integrated with the system. JACoW content is available open access, and recent proceedings have PIDs assigned on the paper level. Several CERN colleagues have official roles within the JACoW Collaboration.

Overleaf is an online collaborative authoring tool, allowing researchers all over the world to jointly author scientific documents using LaTeX. It integrates with many publisher submission systems and hence provides the final output in the correct format for further processing by journals. CERN introduced Overleaf several years ago and since its inception, it has enjoyed wide acceptance at CERN: Every month, about 3,500 CERN users are actively using the system on average for 10 hours per month.

OJS, the Open Journal System, is a platform to support the end-to-end publication process for scientific articles. At CERN OJS is primarily used for the publication of the CERN Yellow Report series but also includes the CERN Environmental Report, Annual Reports, and more.

SCOAP3 repository is a platform used for validating the compliance of the SCOAP3 funded articles, monitoring and reporting. It is also used for further distribution of information, thanks to its API. All articles funded by SCOAP3 appear in the SCOAP3 repository upon publication, alongside the publishers' own platforms. Several formats are available, including PDF, PDF/A and XML. Articles are published under a CC-BY licence and can be freely downloaded and further disseminated. The scope of the SCOAP3 repository is neither to duplicate arXiv nor the publisher platforms, it doesn't offer value-added services for the end users such as those of INSPIRE.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Annex II: Stakeholder engagement

AII.1 Questions asked during interviews

The following list of questions represented the starting point for the semi-structured interviews. They were adapted individually to the respective conversations and the specifics of the experiment/department.

Guiding questions as information producer:

- What type of research products do you usually produce (i.e. conference papers, journal articles, slides, reports, code, data etc.)
- How do you manage these?
    - How do you manage and store textual documents & papers (i.e. locally, in a specific database? Which papers are submitted to repositories like arXiv, CDS?
    - How do you manage and store data / How do you manage and store code?
- Describe the current life cycle of a scientific paper from conception to publication:
    - How do you handle commenting/approving in versions prior to submission?
    - Do you claim any existing papers in CDS, and if so how?
    - Do you use any (TeX) templates/standards?
    - Who approves the paper? Is it the same group of people for every paper?
    - Do you use any persistent identifiers?
    - Once a preprint is submitted, what is the further process?
    - How do you decide which journal to publish in?
    - Do you use arXiv for submitting preprints, and/or do you update the records with publication information? Who submits to arXiv (authors or pubcomm?)
    - How do you handle author lists? Do you use any PIDs (ORCID, ROR)?
    - How do you link to HEPdata, when in the process? Do you link to code or other resources to supplement the publication (where)?
    - How satisfied are you with the current process of handling the scientific papers?
    - What are its strong points / What are its shortcomings?
    - Are you missing specific functionalities?
    - Is the current process meeting your needs?
    - What should be changed, what should an ideal life cycle look like?

Additional information as part of the research process

- Do you store funding information, and if so how and where?
- Do you provide such information in your papers?
- Do you track citations of published papers, data, code etc. and how?
- When you apply for funding, do you use ORCIDs?
- Do you check the publication and citation record of colleagues or (other) experiments? If so, where and what is the use case in those cases?

Guiding questions as information consumer:

- When you start working on a new subject or you intend to deepen your knowledge about it, how do you go about collecting information? What tools and sources are you using? How do you identify further relevant literature?
- Are you usually satisfied with the results of your searches? In your opinion, what are the characteristics of an effective search tool or of a reliable/high-quality information source?

- Where would you be looking for code or data relevant to your research? Please describe some use cases when this is relevant to you and how you do it (e.g. ask my colleague, search on HEPdata)?

Other Questions

- Is there anything else you'd like to discuss regarding your handling of scientific information?
- Is there anybody else in your collaboration/department we should interview who might have complementary insights to yours?

AII.2 List of interviewees

AII.2.1 Experiments

ALICE	Javier Castillo Castellanos, Francesco Prino
ATLAS	George Redlinger
CMS	Greg Landsberg
LHCb	John Walsh
AEGIS	Michael Doser
FASER	Jamie Boyd
ISOLDE	Karl Johnston
MoEDAL	Albert De Roeck
NA62	Brigitte Bloch-Devaux

AII.2.2 CERN Departments

BE	Yannis Papaphilippou
EN	Stephan Petit, David Widegren, Stefano Sgobba, Alessandro Bertarelli
EP	Richard Hawkings
IR	Sascha Schmeling
IT	Stefan Roiser
SY	Ray Veness, Yves Thurel, Vasilios Vlachoudis
TH	Michelangelo Mangano, Elena Gianolio

AII.2.3 Projects and others

arXiv.org (Cornell Tech)	(various conversations outside the typical semi-structured interview)
ATS DO	Alexia Augier, Cecile Noels, Frank Tecker, Irene Garcia Obrero
EU Project Office	Sabrina El Yacoubi & Cloe Levointurier-Vajda, Svetlomir Stavrev
JACoW	Ronny Billen

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Annex III: What is Scientific Information?

AIII.1 Definition

For this project, scientific information is to be understood as any piece of scientific knowledge recorded in some material form, the research artefact by a CERN (associate) member of personnel or by a CERN experimental collaboration or using the CERN facilities.

AIII.2 Dimensions

Scientific information can be classified among several dimensions. For the project to be successful, a future information landscape needs to adequately address the different needs of the different artefacts (or research products), audiences, authors, etc. with regard to storage, dissemination, metadata management, long-term preservation and workflow management.

AIII.2.1 Artefact type

- Literature: anything where the output consists mainly of text meant to be read by a human. This could be a formal write-up (which until recently was the full scope of scientific information) like an article in a journal, a monograph, a technical report, a document discussed at a scientific committee, but also more informally, a conference poster, slides, internal note, report etc.
- Research data: data produced as part of the research at CERN. This is primarily data recorded by the CERN experiments (at various levels of processing).
- Software: scientific software produced to deal with various parts of the research workflow. This is mainly used to process/simulate research data in one way or another. For example, event generators, detector simulation software, data analysis tools, theory computation tools, etc.
- Multimedia content: photos, videos or audio files describing a scientific discourse or documenting an experimental proposal. Examples could include recordings of scientific talks, etc.

While these are the most common currently produced artefact types, this list is not exhaustive. There are hybrid artefacts that defy categorisation (e.g. scientific notebooks or analysis workflows that bridge software and data) or entirely new categories that are becoming relevant but not firmly established yet (e.g. open hardware designs). More will certainly appear after completing the project. The proposed solution should take that diversity into account.

AIII.2.2 Lifecycle stage

- Draft: scientific output that is in the process of being shaped for public dissemination by the research group. Examples: paper being drafted by authors and commented on by internal reviewers, dataset being curated, etc.
- Public: scientific output that is publicly available. Examples: paper in preprint stage, conference poster or slides, a dataset on CERN Open Data Portal, released software, etc.

- Published: scientific output that is published in a standard outlet, and this is mainly relevant for literature. Examples: paper in a journal, published book, etc.

AIII.2.3 Audience

- Internal: scientific output that is private and only meant for the benefit of the research group itself. Examples: internal notes, minutes of scientific committees, raw data, custom software, etc.
- Experts: scientific output that is generally highly technical meant for the benefit of the global research community. Examples: a scientific article, research software, Level 3 data, etc.
- General public: scientific output that is popularised and targets a broader public. Examples: Popular article for outreach

AIII.2.4 Origin

- CERN: scientific output from CERN activities, that is produced totally or in part by CERN personnel or users. It's worth distinguishing two subcategories here based on the authors' departments.
    - Research departments: departments whose primary purpose is to conduct research. Examples: EP and TH.
    - Non-research departments: departments responsible for different parts of the CERN laboratory and operations might produce scientific research as part of their activities even if it's not their main goal. Examples: ECO, HSE, EN, IT, etc.
- Non-CERN: scientific output not originating from CERN activities but still relevant for the holistic management of CERN's scientific information.
    - Scientific output generated outside CERN using CERN scientific information in a significant way and hence relevant to measure CERN's impact.
    - Other scientific content not associated with CERN but containing valuable input for CERN activities; most of the CERN library catalogue falls into this class.

AIII.2.5 Authorship

- Large experimental collaboration: a large research collaboration associated with an experiment that typically has highly structured processes in place, particularly for managing its scientific output.
- Other formal research groups: a long-lived research group of a relatively large size that requires significant coordination but is often less organised as far as the management of its scientific output is concerned. This could be, for example, a project funded by an EC grant.
- Informal research group: a smaller group of researchers collaborating on a specific project with a limited scope. This could be for example a small group of theorists or a subset of a large experiment communicating about a more specific topic.
- Individual: a single researcher. This is typically the case for theses or conference papers.
- Scientific committees: Minutes or proposal documents discussed at scientific committees regularly form part of the scientific discourse and as such should be preserved as scientific information.

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

AIII.2.6 Subject

As a large research laboratory, CERN produces scientific output on a vast range of subjects. Most notably:

- High-energy physics (experiment and theory)
- Nuclear physics
- Instrumentation
- Accelerators
- Applied physics
- Computing
- Engineering & technology
- Teaching & outreach

Although those are the primary research subjects, they are by no means exhaustive.

AIII.3 Notable omissions

Anything not in the previous list is considered as out of scope for the CERN scientific information landscape. Most notably, this excludes the following artefacts, some of which have historically been handled on the same platforms as scientific information (e.g. on CDS).

- Non-scientific reports: any document reporting on CERN activities as a whole or in part and does not contain scientific output. Examples: CERN activity report, CERN environment report, group activity reports, etc.
- Policy documents: documents detailing various CERN policies or regulations. Examples: Operational Circulars, COVID measures, HR procedures, etc.
- Operational artefacts: any artefact produced in support of CERN operations, but not of scientific nature. Examples: procurement process, engineering documents, software for accelerator control or HR, etc.
- Public communication: documents produced as part of the communication activities of CERN, both internally and externally, when they are not scientific. Examples: Annual reports, CERN bulletin, stand-alone brochures/pamphlets, CERN website, etc.
- Historical content: documents relating daily or exceptional activities at CERN that might be of historical interest but don't have a scientific character. Examples: Photo collection, CERN archives, non-scientific content, etc.

Annex IV: CERN Policy Constraints

Operation Circular 3

Operational Circular 3

Unpublished CERN reports, technical notes and specifications are considered as Archival Material and need to be preserved	Preservation needs for digital material to be considered	Important for implementation plan
Departmental Records Officers (DRO) should define and ensure the implementation of specific records management rules for the department	In the context of OC3, DROs are mainly considered for archives	Validation of actual DRO role in departments
Note: At the time of the creation of OC3, digital archiving did not exist. A revision of OC3 seems to be overdue to capture today's challenges of digital records management. A new OC3 might therefore impose additional constraints or requirements on the information management landscape.	OC3 is not adequately considering today's information realities.	New requirements might arise later

Operational Circular 5

Operational Circular 5

Specific user rights will be granted in connection with specific professional duties and only as long this is necessary	Standard CERN Account rules and processes will need to be foreseen for any new information management tool	Important for the future solution architecture
User credentials must be kept confidential and unauthorised use is prohibited		

Operational Circular 6

Operational Circular 6

Definition of *CERN Author* and *CERN Scientific Document* and other terminology can give guidance for this project	Key definitions should be revised and updated as part of this project. OC6 definitions are outdated and do not reflect today's diversity of material	Revised definitions from this project can serve as input to a revision of OC6
Approval procedures for CERN Reports and CERN Divisional Reports are still used in principle but do not fully reflect current realities anymore		
All CERN Scientific Documents must be made publicly available (responsibility of SIS)	Future processes need to enable SIS to follow this mandate	Need to be reflected in roles & responsibilities
SIS must ensure that an electronic procedure is provided to authors to submit their documents	An electronic submission process has to be part of any future solution	Important for scope & solution architecture
Definition of report numbers and a respective process	A new landscape should	Important for

to obtain them	enable the ability to systematically assign report numbers in different categories	scope & solution architecture

Operational Circular 11

Operational Circular 11

Each Service Owner is responsible for the compliance with OC11 and has to establish Records of Processing Operations	As part of the implementation plan, Service Owners of the different landscape components have to be defined and RoPOs established	To be considered in implementation plan
Personal data have to be processed in a fair, proportionate and transparent manner following a clear purpose	As part of the new solution design, an adequate level of personal data processing has to be determined	Important for solution architecture
Data retention periods have to be defined and processes established to comply with such defined periods.	As part of the new process design, data retention periods should be defined	Important for solution architecture
Data Subjects have rights such as right to information, deletion, correction, portability.	New applications as part of the future landscape to support all relevant data subject rights	Important for solution architecture

CERN Open Access Policy

CERN Open Access Policy

All CERN articles have to be published open access	OA monitoring need to be systematically possible through CERN's CRIS	Important for scope & solution architecture
SCOAP3 is the preferred outlet for physics results, external funding should be used where applicable	Type of the OA funding source should be systematically captured and monitored	Important for scope & solution architecture
If articles cannot be covered through central agreements, authors can request central funding prior to journal submission	A new preprint and journal submission process should allow for an automated request for central OA funding	Important for scope & solution architecture

CERN LHC Open Data Policy

CERN LHC Open Data Policy

All Level 1 data (related to publications) will be made available e.g. through HEPData	A new landscape has to ensure interoperability with external solutions such as HEPData and links between artefacts (data +	Important for solution architecture

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

	publication) have to be transparent	
Regular release of Level 2 and Level 3 data through CERN Open Data Portal (COD)	A new holistic solution has to include an adequate solution for open data releases (either maintaining COD or including its functionality somewhere)	Important for scope & solution architecture
All data releases will be made available applying FAIR standards	FAIR principles have to be embedded into any new solution as part of the future information landscape	Important for solution architecture

General Conditions for CERN Experiments

General Conditions for CERN Experiments

All results from experimental collaborations, including scientific data shall be published OA	Information landscape services should support experimental collaborations with publishing their papers and data (OA)	Important for scope & solution architecture
Each Collaboration shall notify CERN in writing of forthcoming publications	New processes should enable collaboration to easily comply with this notification requirement	Important for solution architecture

DOI: http://dx.doi.org/10.17181/CERN-OPEN-2023-006

Annex V: Standards for metadata schema and management

General principles

- **Validated fields, rather than manual input**
- **PID to be used instead of free-form text wherever possible, respective content to be derived from PID registry (e.g. author name from ORCiD)**
- **Provenance information of PID, keyword, abstract, or imported record**
- **Log file to reconcile the history metadata changes**

Fields

The following fields are deemed as minimum requirements for successful information management in the CERN context. The exact hierarchy and data format will need to be described, for instance as a JSON schema.

1. Title
   - Compliments to the title, such as translated title, preprint title, alternative title
   - Additional titles should be searchable but can be hidden
2. Statement of responsibility (PID)
   - individual contributor roles as described by CASRAI (Consortia Advancing Standards in Research Administration Information - https://casrai.org/credit/), namely: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualisation, Writing – original draft, Writing – review & editing
   - In case of non-individual contributor: collaboration (standardisation of naming is needed – coordination with User Office for Greybook) or other entity
   - Role for OA funding purposes: corresponding author
3. Contributor's name (PID)
   - If applicable: previous/alternative names
4. Document versioning – previous version(s) and dates of submission
5. Affiliation (PID)
   - The level of granularity needs to be defined (implications for impact assessment)
   - ISO country code required (implications for impact assessment)
6. Abstract
7. Language
8. Content type
   - Refer to here for publication type ontology.
9. Publication status
   - Example: submitted/accepted/published/withdrawn
   - Implications for impact assessment. Comment field needed.

10. Publication reference
    - Standardised journal abbreviation, page, volume, date, article ID, DOI
    - Book chapters
    - Conference proceeding contribution
11. Event (PID)
    - Distinction between event (such as conference, seminar, etc.) and publication
12. Keywords
    - Indication of used schema and value.
    - Solution should accommodate different schemas
13. Infrastructure used (PID)
    - Instrument, accelerator, testbeam, facility, etc.
14. Funding information
    - Grant number (PID)
    - Funder (PID)
    - Relevant EU project
15. Institutional Identifier of artefact (PID)
    - Report number or similar
16. Attachment
    - Attached full text/dataset/software/media type location (PID)
    - Format of attached file and its size
    - Timestamp to detect if the full text / attached files must be re-processed
17. Tag indicating if record has been withdrawn/hidden
18. For articles: Open access information
    - OA model (gold….) and funder (PID)
    - APC for cost-benefit assessment purposes
19. Licence type and version
20. Link to related documents
    - Versions, languages, same record in another repository, semantic relationships….) or to authority lists – using ISNI (isni.org) to allow further linking (Wikipedia etc.)
21. References “from” and “to”
    - Implications for impact assessment