SONAR Project application, Feb. 2018 (project approved) Abridged version

This document describes the project SONAR, as submitted to Program P-5 "Scientific information" in February 2018. On the 29th of June, the steering committee "Scientific Information" (P-5) approved the project.

1. Project overview

1.1 Full project title	SONAR – Swiss Open Access Repository
1.2 Short project title	SONAR
1.3 Application date	18.02.2018
1.4 Planned project duration *start/end*	01.11.2018 - 31.10.2020 (24 months)
1.5 Applicant institution	RERO – Réseau de bibliothèques de Suisse occidentale
1.6 Partner institution(s)	Haute école spécialisée de Suisse occidentale (HES-SO) Hochschule für Technik und Wirtschaft HTW Chur Università della Svizzera italiana (USI)
1.7 Project management	Miguel Moreira
1.8 Deputy PM	Patrick Ruch

2. Project classification

2.1 This is a follow-up application within the program	☐ yes ☒ no
2.2 Key area of focus addressed *Classification as per the 2017-2020 Implementation Strategy, chapter 3.3*	☒ Publications ☐ eScience ☐ Basis ☒ Services
2.3 Implementation action(s) addressed *Classification as per the 2017-2020 Implementation Strategy, chapters 4-8*	Primary implementation action addressed (Number and description): G-1 Opening up an existing service to other participants (including international participants): Service provider investment costs Any other implementation actions (numbers and descriptions): G-6 Further developing and expanding a service P-3 Data provision and collection for monitoring open access P-5 Improving the quality of open access publications P-7 Opening up and improving the quality of repositories P-13 Metadata hubs/search solutions for scientific publications and research data

3. Abstract

This project proposes to set up a Swiss Open Access Repository "SONAR", whose primary goal is to collect, promote and preserve scholarly publications by authors affiliated with Swiss public research institutions. This central repository operates as an aggregator, drawing content and metadata from existing platforms and institutional repositories (IR). In parallel, direct depositing of content by authors, or their representatives, is also possible.

Alongside the national repository, SONAR also offers autonomous IR solutions as outsourced platform to serve Swiss partnering institutions of higher education ("Institutional Repository as a Service").

The project intends to lay the groundwork for maximizing the coverage rate of open access publications by Swiss institutional repositories, with an exploratory approach. A survey conducted in 2017 by one of the applicant institutions and covering 7 major Swiss institutional repositories, shows that existing IRs make openly available about 35% of the full-text articles authored by their affiliated researchers, when compared with articles available in international or disciplinary repositories such as PMC or HAL. In contrast, the study suggests that nearly 80% of the publications could be legally available under open access conditions. The project intends to explore and develop automated procedures for tracking down and collecting from external sources, such as international subject repositories, the largest possible number of publications that, while possessing an open access status, are either not registered as such in existing institutional repositories, or the end of their embargo period has not been acknowledged, or their full-text has not been deposited. SONAR can then feed those full-text publications back to the corresponding institutions, complementing and reinforcing existing Swiss repositories. The goal is to raise the coverage rate above 70%, thus doubling the current coverage of Swiss IRs.

The dissemination on the web of the content hosted by SONAR is also an important component of the proposed solution. Given that institutional repositories, where publications are deposited, are not the only location for searching publications, another key value proposition of the project is the creation of a highly interoperable repository powered with semantically rich data, import/export and content exposure with regard to external platforms and global search engines (Google Scholar, Twitter...). It is expected that through an interaction between SONAR and other scholarly content sources, researchers should be able to make their publications highly visible, citable and openly accessible in the long term, with the least possible effort.

Another focus of the project is data normalization and analysis, with the planned creation of a database of entities such as authors, publishers, journals, research institutions, funding agencies, research projects/grants and patents (the "SONAR data hub"). Automatic content processing procedures will be investigated to support the extraction of specific metadata, as successfully explored by the applicants in Europe PMC. By uniformly collecting and normalizing such data in a central database, the project should greatly reduce discrepancies and data drifts likely to emerge from the current Swiss IR landscape.

SONAR builds on an existing service: RERO DOC, which is a multi-institutional repository operated since 2004 by RERO, the Library Network of Western Switzerland. This new development is meant to extend its range of services and its institutional coverage.

The proposed project is consistent with several principles formulated in the Swiss open access strategy commissioned by the Confederation through swissuniversities and the Swiss National Science Foundation, namely in the areas of resource coordination and pooling, as well as national monitoring. It also presents a high potential of collaboration with both existing services and ongoing projects in the framework of the program "Scientific information: Access, processing and safeguarding".

4. Project description

4.1 Background

Institutional repositories, subject repositories and academic social networks

Institutional repositories (IR) are currently a major instrument for the fulfillment of the open access goal. This is also the case in Switzerland, with many institutions of higher education (IHE) running their own IR, often associated with some form of institutional deposit mandate. Not all IHEs in Switzerland possess an IR. All cantonal and federal universities currently have some sort of IR solution, be it an independent repository or an outsourced space in a multi-institutional one (RERO DOC, Zenodo). The situation is different in the case of universities of applied sciences and universities of teacher education, with only a fraction of them possessing an IR solution. In fact, running its own IR can be unaffordable, depending on the size of the institution. Thus, a mutualized, cost-effective solution would be welcome for many IHEs.

With regard to content deposit

IRs are well established tools in general, integrated in each institution's environment, with close connections to the local IT infrastructure, the local CRIS platform (Current Research Information System) and each institution's research evaluation system. Self-archiving open access clauses included in copyright transfer agreements that authors sign when publishing, authorize depositing in some cases only in their local IR.

Although IRs provide researchers with close ties to their institution, that is less the case with regard to their research community. Indeed, international subject repositories and "academic social networks" (ASN), notable examples of which are ResearchGate1 and Academia.edu2, are highly attractive tools for research communities across several domains. These platforms offer high visibility, wide-range access statistics, as well as tools for a lively interaction among the community.

However, ASNs raise some problems: being commercial services, they offer no long-term accessibility commitment, and in fact there are several examples of such services having been acquired by other companies, usually with the aim of gaining a greater grip over the scientific communication process. For example, major publishing group Elsevier has acquired the high-profile ASN Mendeley in 2013. This issue can likewise affect SRs, as is the case of SSRN (Social Science Research Network), a leading social science and humanities repository, also acquired by Elsevier in 2016, with author-deposited content reportedly being removed without notice. Furthermore, researchers quite often deposit their articles in ASNs with no copyright concerns, raising potential conflicts with publishers. This may result in papers being forcefully removed from the platform, defeating the purpose of the deposit.

From a researcher's point of view, there should exist solutions allowing to combine both aspects: deposit in IRs in order to fulfill institutional mandates and in international platforms in order to allow interaction with their research community. Currently, this is often seen as double work, with the latter kind of deposit being in general more attractive.

With regard to content search

When searching for scholarly literature, IRs are seldom a starting point for researchers, as they contain only the research output of a single institution. Global search engines (Google and the like) or, better yet, scholarly-dedicated search engines and bibliographic databases are the preferred tools for that. How can IRs remain relevant with regard to research output dissemination when their local search engine is not a high-profile resource? Most IRs feature OAI-PMH export capabilities, which allows its content to be harvested and indexed by other repositories and thus be included in the results provided by the latter's search engine. But this nearly 20-year-old protocol is rather static and focused on metadata, not full-text. Meanwhile, modern web standards, tools and protocols are gradually being adopted, offering a much higher potential for data interaction, allowing for a distributed scholarly communication model and high web exposure. That includes Linked Data, OAI-ORE, ResourceSync and Signposting. Indeed, full-text web exposure is the main answer for the content dissemination challenge. Related to that, in April 2016, COAR, the Confederation of Open Access Repositories,

[1] https://www.researchgate.net
[2] https://www.academia.edu

launched a “Next Generation Repositories Working Group” to help identify functionalities and technologies for repositories and develop a roadmap for their adoption3. The final report4 has been published in November 2017, bringing useful advice for the development of SONAR.

ORCID as a major hub within the scientific communication process

ORCID is a not-for-profit organization that provides an identifier for individuals to use with their name as they engage in research, scholarship, and innovation activities5. Along with the ORCID ID (the identifier itself), ORCID can also register the record of the corresponding individual, including its education, list of employers and list of publications. The ORCID ID of a researcher is valid throughout his/her career, facilitating mobility and recording all his/her publications, regardless of affiliation. This list, potentially complete and up-to-date, can be used to include in personal web pages, resumes for project submissions, job applications and other similar occasions. ORCID therefore provides an invaluable help for authors in managing their research output, and is being increasingly acknowledged by several stakeholders in the scientific communication community.

However, ORCID can only store the metadata of publications, not their full-text content. Ideally, the metadata of each publication in the researcher's ORCID record should include a persistent identifier (e.g. DOI6 or ARK7) pointing to a permanently accessible copy. ORCID is an open initiative, and provides useful APIs for interacting with its database (although some functionalities require a paid membership), therefore nicely integrating in various automated workflows. The ORCID identifier could play a crucial role in the process of managing data related to the scientific communication activity. In the Swiss context in particular, the connection of this identifier with Swiss edu-ID, a widely deployed infrastructure for authentication and authorization providing an academic identity to every researcher, can be a key interoperability factor.

In January 2017, ORCID has established an Organization Identifier (OrgID) Working Group8 with the intent of defining an organization identifier registry to facilitate the disambiguation of researcher affiliations. The outcome of this initiative might come in handy for the implementation of certain procedures planned in this project that involve affiliation data (cf. paragraph Content tracking [WP5] in sect. 4.3).

Open access deployment

There is a clear trend towards open access, both in Switzerland and more widely in Europe and the USA, for instance, with deposit mandates being issued by research funding agencies as a precondition for funding approval (Swiss National Science Foundation (SNSF), European Union, National Institutes of Health (USA)).

Switzerland is currently developing a national open access strategy9,10, which will require suitable support tools, able to identify and showcase Swiss OA publications, and to assess OA activity. For its part, the SNSF announced in December 2017 that all publications produced in SNSF-funded projects are to be freely available in digital format as of 2020, which fits in with the national strategy11.

Swiss IRs are far from covering the full extent of Swiss OA publications, namely with respect to full-text accessibility. Tools are needed to track down publications that, although benefitting from an open access status, are not registered as such in existing IRs and funding agencies' records.

---

3 https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/

4 https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf

5 http://orcid.org/about/what-is-orcid/mission

6 https://www.doi.org/

7 https://en.wikipedia.org/wiki/Archival_Resource_Key

8 https://orcid.org/content/organization-identifier-working-group

9 https://www.swissuniversities.ch/fr/themes/politique-des-hautes-ecoles/open-access/

10 https://preview.tinyurl.com/OA-strategy-20170131

11 http://www.snf.ch/en/researchinFocus/newsroom/Pages/news-171213-100-percent-open-access-to-snsf-research.aspx

Possibility of building on an existing service – RERO DOC

RERO DOC12 is the digital library of RERO13 (Library Network of Western Switzerland). It was launched in 2004 and currently operates as institutional repository for a few Swiss universities: Université de Fribourg, Université de Neuchâtel, Università della Svizzera italiana, as well as part of the University of Applied Sciences of Western Switzerland. As a multi-institutional repository, offering IR services to several institutions, including outside the RERO network, RERO DOC already has several features that are required for SONAR, which can be developed and deployed as an extension of the existing service. In addition, RERO has a long experience of repository management, and valuable technical expertise in Invenio (see section "Technical infrastructure"), which are useful assets for the current project. RERO being the planned future operating institution of SONAR, these assets can be directly leveraged on behalf of the project.

In summary

- In the current context, taking into account the challenges cited above, a service that can complement and reinforce existing IRs would be a welcome development, one that is capable of improving their coverage and visibility.
- Solutions must be provided for the researcher to make his/her publications highly visible, citable and openly accessible in the long term, with the least possible effort. The possibility of referencing and/or transferring copies of one's publications between repositories with a single deposit operation must be fully exploited.
- National-scale services supporting ongoing efforts towards open access in Switzerland have been identified as necessary, for instance with regard to mutualized resources and monitoring.
- Several recent technical and collaborative developments strongly favor a renewed repository service. Besides the growing adoption of persistent identifiers, namely DOI or ARK for publications and ORCID for authors, as well as web technologies, such as Linked Data, some of which are already being implemented in current IRs, the implementation of distributed scholarship features and wide web interoperability offer an interesting improvement potential (cf. COAR's "Next Generation Repositories" report mentioned in sect 4.1). However, care should be taken to avoid adopting every possible technology or initiative, as the domain changes very quickly and trends come and go. A long-term vision is fundamental.

4.2 Project goals

Goals

** In the following text, references are provided to the work packages that handle the mentioned elements, [in square brackets] (cf. sect. 4.6).*

Goal 1: Science showcase

The service resulting from this project is intended to *collect*, *promote* and *preserve* scholarly publications by authors affiliated with Swiss public research institutions. These 3 roles can be formulated as follows:

- **Collection**: Besides collecting publications already registered in existing Swiss institutional repositories, a "Content tracking" project component intends to lay the groundwork for maximizing the coverage rate of Swiss publications deposited in SONAR. [WP4, WP5]
- **Promotion**: For the sake of content dissemination and visibility, the project aims to dedicate a considerable effort on interoperability, data import/export and content exposure with regard to local institutional tools, external platforms and especially the web at large. [WP2, WP6]
- **Preservation**: SONAR should provide a permanent archive for Swiss scholarly publications. Therefore, the project includes the analysis and implementation of a long-term preservation solution. [WP9]

The project should be able to support web-based distributed publication models, which allow research communication to be based on linking objects in the web, such as text, primary research data, results of experiments, multimedia content or web sites.

[12] https://doc.rero.ch
[13] https://www.rero.ch/

Goal 2: Institutional Repository as a Service

SONAR should provide custom institutional repository solutions, as outsourced service for interested Swiss institutions of higher education [WP2]. This goal is pursued with two different approaches:

- With **dedicated repository instances**, customized to each IHE, with the look of the front-end matching the institution's branding requirements, and implementing their own self-archiving policies and management procedures.
- With a shared, **multi-institutional service** with a single, common search and navigation interface, as well as common self-archiving policies and metadata schemes. In this case, the contents of each institution are distinguished by means of collections, navigation paths in the search interface (facets) as well as some visual customization at the record level. This is the approach currently followed by RERO DOC.

Goal 3: Gather information about research output in Switzerland

The project aims at gathering relevant data about Swiss research activity, allowing the later development of uniformed key performance indicators. [WP7, WP8]

Goal 4: Foster national cooperation between the institutional repository arena and other existing services and ongoing projects in the framework of the program "Scientific information: Access, processing and safeguarding"

There is a great potential of cooperation between SONAR and several national projects and services in the scientific information domain, especially in the P-2 and P-5 area. Examples include SLSP, DLCM, ORD, Swissbib, National Licences, Swiss edu-ID, SYMPHONY, HOPE or CCDigitalLaw. Such cooperation should be explored in the course of the project and, whenever possible, actual interactions should be deployed. Technology reuse, data sharing, interoperability, 3rd-party service sharing (ex: common digital preservation solutions) are such examples.

Goal 5: Comply with national open access strategies

The proposed project is consistent with several principles formulated in the Swiss open access strategy commissioned by the Confederation through swissuniversities and the Swiss National Science Foundation, namely in the areas of resource coordination and pooling, as well as national monitoring. A detailed Action Plan, which is currently in preparation, should provide a better indication of the potential evolution of SONAR in that context. The project is also consistent with the goals defined by the KUB-CBU[14] in a recent statement supporting open access. The future governance structure of SONAR should be able to coordinate its activity and development with national policies concerning open access and scientific communication in general.

Target groups and benefits

Researchers

With SONAR, and in interaction with other scholarly content sources, researchers should be able to make their publications highly visible, citable and openly accessible in the long term, with the least possible effort. At the same time, their publications should integrate the respective local IRs in order to fulfill self-archiving mandates and interact with local tools within their institution.

Research institutions (namely institutions of higher education)

The project intends to lay the groundwork for maximizing the coverage rate of open access publications by Swiss institutional repositories, beyond those that are currently registered.

A recent study performed at the HEG Geneva[15] on a sample of Swiss IRs (N=7) reports that the coverage of full-text articles authored by Swiss affiliated researchers does not exceed 35%. In contrast, the authors of the study suggest that the coverage could approach 80% if articles were directly harvested from international IRs, such as PubMed Central, arXiv or SocArXiv. The disclosed publications could then be fed back to those repositories, improving the monitoring of the research activity of the corresponding institutions.

[14] http://www.kub-cbu.ch/dokumente-documents/kub-statement-on-open-access/

[15] PUTALLAZ, Matthieu and SCHWOB, Elodie, 2018. Enrichissement des dépôts institutionnels suisses : vers une couverture complète de la publication académique ouverte : stratégie d'automatisation du moissonnage de plein-textes. Travail de recherche. Genève: Haute école de gestion de Genève. (To be published)

Swiss institutions of higher education, namely those that do not have an institutional repository, can outsource that service to SONAR, which offers a mutualized, cost-effective solution, that can be customized to each institution's needs.

Libraries usually play an important role in the scientific information process within IHEs. One of the areas in which they are usually involved is the depositing of publications to the institutional repositories, on behalf of the authors affiliated with their institution. SONAR can help alleviate that process, with automated or assisted deposit procedures.

Funding agencies

Open access is becoming mandatory for publicly funded research. However, funding agencies can hardly monitor the progress towards the objectives of the institutions they support. The data collected in the framework of SONAR, relating publications to research grants, can be of great value for those agencies, allowing them to better track open access publications produced by their own funding, as well as adding links to the full-text from their projects' database.

A dialogue has been maintained with the Swiss National Science Foundation (SNSF) about the current project. As a result, the SNSF has phrased its point of view regarding SONAR as follows:

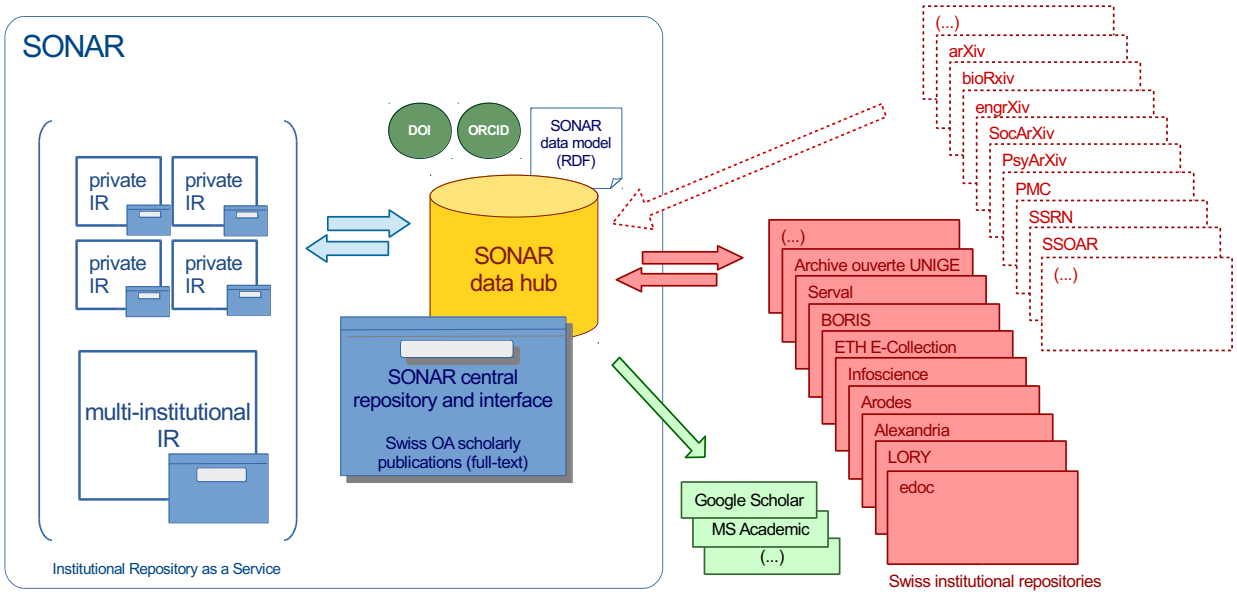
> ***Usefulness of SONAR to the Swiss National Science Foundation (SNSF)***
>
> *The SNSF requires grantees to report their publications as output of their grants. Grantees can use several databases to import their publications (Web of Science, Scopus, CrossRef, ORCID) or use BibTex or Endnote to import in the grant administration system mySNF. The imported metadata is often not complete, because required information by the SNSF like OA Status or peer-review status is not available in the external sources or is not transported in the interface or format. Often grantees also have to tweak the mapping of publication types or fields to the forms of the SNSF. Feedback from researchers shows, that they would welcome a way to import the publications from their local Institutional Repositories (IRs). Since many repositories check the submissions of their researchers for accuracy of the (meta)data, the SNSF would also profit with quality-controlled metadata. The idea to build such an import interface from Swiss IRs to mySNF exists for many years, but never was realised. SONAR is a good opportunity revive that idea providing the following benefits:*
>
> - ***One API for all Swiss Repositories****: For the SNSF it's helpful to have one API to the content of all Swiss IRs, instead of having to establish connections to each IR individually.*
> - ***Overcome the Hurdle of OAI-PMH****: In the repository world OAI-PMH is still the common exchange protocol. However, OAI-PMH only allows harvesting a whole repository or usually large sets that have been predefined by the repository owner. Specific ad hoc searches (like for an author or for a DOI are for a grant) are not directly possible with OAI-PMH. The intention of SONAR doing the harvesting and offering a more modern API (that does not require harvesting first) is highly appreciated.*
> - ***Richer Metadata****: Usually Repositories have internal information, that is not exposed in the common OAI-DC format, which currently serves as a minimal standard for data exchange. Some repositories have started to expose some other formats as well (like MODS) or some have enriched the OAI-DC with elements from OpenAIRE. SONAR offers for all stakeholder a good opportunity to examine the existing practices and find consensus about formats, fields and exchange protocols. Of great interest for the SNSF is to find a way how repositories can map publications or other output from grants of the SNSF, so SONAR can be searched by a grant number.*
> - ***Commitment and space for collaboration****: The drive of earlier attempts to realise the data exchange between the SNSF and the repository community had been lost in the daily business of everyone. A specific project with dedicated resources, including a good project management, will likely generate commitment on all side to eventually build such an exchange.*
> - ***Starting point for more international initiatives****: There are many international initiatives that would also benefit from SONAR. OPENAIRE for example has many overlapping aims and to a certain extent raises the question of duplicate efforts (especially on technical side). Experience shows however, that not all IRs have implemented the guidelines from OPENAIRE. By developing a national solution, SONAR as a more „relevant and closer" project with dedicated resources has the potential to kick off developments that will lead to to more interoperability not only useful for SONAR, but also for other initiatives like OPENAIRE, BASE, ORCID, ORG-ID, BASE, OpenAPC, DataCite, Pubmed etc.*

4.3 Proposed solution

** In the following text, references are provided to the work packages that handle the mentioned elements, [in square brackets] (cf. sect. 4.6).*

Main architecture

The schema below illustrates the various components, data flows and entities of the proposed solution.



SONAR firstly consists of a **central national repository**, hosting two kinds of content:

- The metadata and the full-text files of open access scholarly publications, namely peer-reviewed material and preprints (possibly under a time-limited embargo). These are publicly accessible via human and machine (API) interfaces. [WP2]
- SONAR also collects metadata about all Swiss publications, including those that are paywalled, and stores them in a structured database, according to a defined data model. With these collected metadata, which are accessible only to certain entities and under certain conditions, it is possible to create a hub of normalized data about scholarly communication in Switzerland, which can later be used to assess Switzerland's research activity. In particular, by collecting metadata about all publications, it is possible to calculate which percentage of an institution's output is open access, which is a very important indicator for monitoring the uptake of OA. [WP2, WP8]

Secondly, the system includes individual **dedicated repositories** intended for IHEs that require their own IR, which use that service on a rental basis. These are named "**private IRs**" in the schema, in the sense that they are used by a single institution independently, and customized to its specific needs ("IR as a Service"). [WP2]

Thirdly, there is a shared **multi-institutional repository**, which is also a kind of "IR as a Service", but hosting the content of several institutions at once with a single, common search and navigation interface. The contents of those institutions are distinguished by means of "collections", navigation paths in the search interface (facets) as well as some visual customization at the record level. The approach for this multi-institutional repository roughly corresponds to what RERO DOC currently proposes, and it offers a more affordable solution than a "private IR". [WP2]

Populating SONAR

Types of content

There is a clear difference between the type of content hosted by the **central repository**, on the one hand, and by the **private and multi-institutional repositories**, on the other hand. While the former is limited to scientific publications representing the output of research activity (peer-reviewed material

and preprints – the actual content types should be defined by a community discussion), the latter admit additional types of content, such as master's thesis, working papers, research reports, posters, etc. They can even host non-scholarly content, such as digital or digitized heritage material, in that case operating as digital libraries, as currently does RERO DOC. Despite not being the main focus of the project, this feature makes SONAR an interesting solution also for heritage libraries.

Another distinction lies in the fact that the central repository operates mainly (though not exclusively) as an aggregator, drawing content and metadata from existing platforms and institutional repositories, including the private and multi-institutional IRs of SONAR [WP4], whereas the latter are populated mainly through direct content deposits by authors, library staff, etc. Human-entered metadata collected through direct deposits can provide very useful information, such as the grant number, the funding agency, or the embargo period of a publication. [WP2, WP8]

Identifiers

Persistent identifiers, such as DOIs, ARKs and URIs, are assigned to every document ingested by SONAR, in order to enable permanent accessibility and citability. Many publications tracked by SONAR may already have one or more DOIs assigned, for instance by the publisher and/or by an IR. However, for the sake of long-term persistence, a local identifier must be minted by SONAR in any case, for referencing the locally archived version. The ARK technology can be an interesting solution used for this purpose, as an alternative to DOI, in order to avoid a multiplication of DOIs for the same version of a document16,17. Local identifier assignment is also important in the case of immediate preprint deposits, in which case the minted identifier can be used by the authors for early citation of their yet-to-be-published work. [WP2]

Content sources

At an initial stage, the main source of content for the central repository consists of the publications registered in institutional repositories run by Swiss institutions of higher education, together with those publications that are directly deposited in SONAR's rented IR spaces. At a later stage, on the basis of exploratory work to be developed in the course of the project [WP5, WP7], it is expected that a considerable number of additional publications can be collected from external sources.

Monitoring data

In the context of content aggregation from Swiss IRs, it would be appropriate to agree on common metadata standards, including specifications on Green, Gold and hybrid open access, which could favor a greater control on publication costs. A requirements analysis phase amongst Swiss IR managers about the services to be developed is proposed [WP4]. Meanwhile, at the initiative of AKOA[18], an informal working group on the Monitoring of open access in Switzerland has been created in November 2017, composed of delegates from institutions concerned by open access issues (namely IHEs and funding agencies). One of the goals of this group is the definition of a common metadata schema for OA monitoring purposes, which should include information distinguishing between Green OA, Gold and Hybrid OA, as well as publication costs, APC (Author Processing Charges) and other data. RERO, the main applicant institution, participates in this group. This kind of collaboration at the national level allows SONAR to take into account the requirements of the community that it is supposed to serve.

Licencing

Licensing is an important issue, involving two components:

- It determines how content can be handled: In some cases, the content of publications, unlike their metadata, cannot be simply copied from existing repositories into SONAR, due to licensing restrictions. Besides, some licenses prevent the full-text of publications to be exploited (text data mining – TDM). In summary, for each publication it must be determined if the full-text file can be copied to SONAR or if just a link to the remote source is admitted, and in parallel, if it full-text can be indexed for searching purposes, notably. Related to these issues, there are ongoing discussions about the current Swiss copyright law that must be watched for.

---

16 BILDER, Geoffrey, 2013. DOIs unambiguously and persistently identify published, trustworthy, citable online scholarly literature. Right? In: Crossref's Blog [online]. 20 September 2013. [Accessed 5 February 2018]. https://preview.tinyurl.com/y9todb4t

17 Also related, the Zenodo and OpenAIRE teams have co-developed an Invenio module for DOI versioning, which can be quite useful. (Invenio is the software solution adopted for SONAR.) https://blogs.openaire.eu/?p=2010

18 AKOA (Arbeitskreis Open Access), is a permanent working group of the KUB/CBU (Konferenz der Universitätsbibliotheken der Schweiz) http://www.kub-cbu.ch/projekte-projets/akoa-arbeitskreis-open-access/

- For each resource, the associated licence should be clearly indicated in a manner that is easily discoverable by both humans and machines, for example by displaying the logo of the relevant licence in each publication page and by including HTTP links that point at the URI of the license (cf. COAR's “Next Generation Repositories” report mentioned in sect 4.1).

Content tracking [WP5]

A survey conducted in 2017 by one of the applicant institutions[19] and covering 7 major Swiss institutional repositories, shows that existing IRs make openly available about 35% of the full-text articles authored by their affiliated researchers, when compared with articles available in international or disciplinary repositories such as PMC or HAL. In contrast, the study suggests that nearly 80% of the publications could be legally available under open access conditions.

In order to retrieve undisclosed Swiss publications from external sources, the project intends to harvest international open archives, such as arXiv and its derivatives (bioRxiv, engrXiv, SocArXiv, PsyArXiv), PMC, SSRN or SSOAR, as well as commercial platforms and databases such as Web of Science and Scopus. There are numerous such archives, often domain-specific, and they can generally be distinguished between preprint and postprint archives. They can contain publications from OA journals as well as from subscription-based journals. The intention, in the framework of this project, is to select a pilot subset of archives covering both preprints and postprints.

In order to identify articles authored by researchers affiliated to Swiss IHEs, two complementary approaches are considered. The first consists in starting from metadata records existing in Swiss IRs and searching the corresponding publication in the pilot set of external archives. The second approach consists in browsing those archives trying to identify authors with a Swiss affiliation. For that, a list of Swiss IHEs is needed, including the most frequently used affiliation labels. This list could be established using the openly available data at the SNSF P3 database and other similar sources. Once a record is matched, the full-text can be copied to SONAR.

In the aforementioned report (Putallaz and Schwob 2018 – cf. footnote 15), the authors test the feasibility of automating the harvesting of these missing full-text articles. They use the API of Crossref[20] to get the DOIs of the articles and then query the Open Access Button (OAB)[21], a browser extension and a web site allowing to search for legal open access versions of the full-text of an article in numerous different sources. Another tool, similar to OAB, which could be studied, is the oaDOI API[22]. These tools are useful regarding licences too, because they browse sources which are supposed to be OA. Several harvesting strategies are described, but the study was basically made possible by first extracting a sample set of articles from international platforms (PubMed Central and HAL). To make such an acquisition step successful, the authors underline the need to maintain curated affiliation names for Swiss IHE. Thus, they observe that at least 36 name variants were needed to properly track publications of a single University. Indeed, affiliation retrieval is a key aspect of the future system; the best possible automation must be pursued in this area, but some residual manual work might be needed, which does not require extensive technical expertise and could possibly be outsourced.

The appropriate frequency of such harvesting methods should be measured. The quality assessment of the automated full-text acquisition system should be made based on samples of manually curated test cases. Similarly, the effectiveness of the reconciliation strategy should be evaluated.

Another area of work in this context consists in the automatic detection of embargo periods for publications defined by the outputs of the National Licenses project or using the SHERPA/RoMEO API[23].

Visibility and interoperability [WP6]

As already mentioned, in order to improve the visibility of the Swiss scholarly publications, every document added to SONAR is assigned a DOI or ARK and a URI, for immediate and long-term citability, which is essential for researchers. In that regard, SONAR should also explore the possibility of citing not only at the document scale, but also at the paragraph scale, based on the work done in the framework of the Open Annotation Collaboration[24]. The current project does not explicitly include the implementation of Open Annotation features. However, such features are expected to soon integrate

[19] To be published.

[20] https://www.crossref.org/

[21] https://openaccessbutton.org/ :

[22] http://unpaywall.org/api

[23] http://www.sherpa.ac.uk/romeo/apimanual.php

[24] http://www.openannotation.org/

most IR software solutions, including Invenio, which already contains experimental code in that regard. Otherwise, they might be identified as an important follow-up of the project in the framework of task 3 *Identify future strategic developments: services and collaborations* of [WP1].

The assigned identifiers should be linked with external IDs, such as ORCID and Swiss edu-ID. That further increases visibility and eases the export of publication lists, for example from within authors' personal pages or resumes. This could be done through different strategies, for example, by considering a partnership with ORCID in order to push articles to its database. An API could also be exposed to enable the use of this data by the tools already employed by the researchers.

Data normalization and analysis [WP7, WP8]

The idea is to identify and store entities such as authors, publishers, journals, research institutions, funding agencies, research projects, grants and patents in a "SONAR data hub". Achieving state of the art authority files for all these entities may involve a considerable amount of work, which cannot be guaranteed in the scope of this project. However, the first steps in that direction should be made, and part of the required information already exists in different sources. In the case of manual deposits, the appropriate metadata fields can be filled by the depositor. Using existing external lists, international authority files and ontologies related to those entities, together with locally developed authority data, the system can assist the depositor with autocompletion and suggestions, thus allowing high-quality metadata to be gathered.

In addition, text mining procedures will be investigated for extracting relevant information from the content of publications. The project should evaluate available tools for extracting data from full-text documents (mainly in PDF format) related to the mentioned entities, and use them to assist in the depositing process. This is a domain where technology already exists, that should be reused as much as possible, for example the techniques developed in the framework of the Swiss National Licenses project[25], those used by PMC in order to track grant accession numbers[26], or others, stemming from other international initiatives.

The goal with this kind of data analysis is to later allow the development of value-added services, such as monitoring the implementation of open access, publication costs and providing statistics and trends about public research activity in Switzerland in general. Measuring cooperation between IHEs, researcher mobility or patent production, are examples of possible uses for the collected data. The project includes discussions with potential stakeholders with the aim of defining relevant scenarios in this domain [WP8]. The SYMPHONY project[27], in particular, has already developed highly relevant work in this domain.

Data protection issues, as well as the future governance and the possible applications of the collected data are important aspects that must be addressed in the course of the project.

Content dissemination [WP6]

One important step to increase the visibility of Swiss scholarly publications is to have them indexed by the main scholarly search engines. SONAR can take advantage of the fact that RERO DOC has already gone through the indexing process by Google Scholar, which is usually a very long process. This experience should be extended to Microsoft Academic[28].

Also, in order to ensure high interoperability, legacy protocols like OAI-PMH should be implemented, as well as recent ones like ResourceSync[29] and OAI-ORE[30]. It means in particular to make content more "web-centric", easily accessible and understood by machines, especially the search engine crawlers (cf. COAR's "Next Generation Repositories" report mentioned in sect 4.1).

Many technical requirements mentioned in this section are already implemented by the proposed software platform – Invenio (cf. section "Technical infrastructure" below).

---

[25] http://edoc.unibas.ch/57350/. Report "Metadata Management & User Authentication", cf. sect. 5.1. *Defining related publications*

[26] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383902/

[27] http://www.htwchur.ch/digital-science/forschung-und-dienstleistung/institut-sii/projektuebersicht/symphony-english.html

[28] http://academic.research.microsoft.com/

[29] http://www.openarchives.org/rs/resourcesync

[30] https://www.openarchives.org/ore/

Long term preservation [WP9]

SONAR should provide some sort of long-term preservation solution, adequate with the kind of content to be preserved. Cost-effectiveness is an important issue to be taken into account. The project should compare existing approaches, including LOCKSS and similar, as well as OAIS-oriented solutions. Outsourcing is also an option.

Technical infrastructure

Invenio[31] is a free, open-source software suite for running a document repository on the web, among other uses. SONAR can make use of the latest release of Invenio, combined with cloud technologies such as virtualization and dockerization, in order to benefit from the high level of flexibility required for running several parallel IR instances. These advanced and highly flexible technologies, which are already deployed at RERO, are major factors allowing the kinds of services that are proposed for SONAR.

Notable features for SONAR

- Custom views for institutions and collections, in the user interface
- Recording of embargo information about publications, in order to release them automatically at the end of the embargo term
- Deduplication of publications from multiple institutions
- Focus on widespread use of persistent identifiers: DOI/ARK and URI assignment; harvesting of several kinds of relevant identifiers from external sources (DOI, URN, URI, ORCID, etc.)
- Focus on open APIs for data reuse
- Import and export of publication lists
- Connection with Swiss edu-ID for user authentication, with the possibility to link to the author's ORCID ID

Reference standards, technologies and data sources

- Linked Open Data / RDF data model
- VIVO[32] – an RDF ontology for representing scholarship
- Memento[33] – to support the temporal dimension of the web
- OAI-ORE (Open Archives Initiative Object Reuse and Exchange) – to promote scholarly communication as aggregations of web resources (distributed scholarship)
- ResourceSync and OAI-PMH – for repository content synchronization and metadata exchange
- Bibliometrics and Altmetrics
- Link to SHERPA-RoMEO – a database of publisher copyright policies

Future developments

Once the main features are implemented, further developments can be considered, either in the course of this project or in a follow-up project. These can be analyzed and prioritized in the context of Work Package 1: "Business model and governance", under the task: "Identify future strategic developments: services and collaborations".

- Some components of the current project, such as content tracking and text analysis, are exploratory, aiming at creating pilot subsystems. The findings at the end of the corresponding work packages can later be consolidated into full-fledged solutions.
- It could be interesting to propose open peer-review (OPR) features within the user interface[34]. The implementation of an OPR platform involves more than just providing the tools, but it is a development worthy of further attention. Some work on local peer-review platforms (not necessarily OPR) is already being done in Switzerland, as is the case with the HOPE project at the University of Zurich (see sect. 4.4 Environment analysis), and the BOP-Serials[35] initiative at the

[31] http://invenio-software.org/

[32] https://wiki.duraspace.org/display/VIVODOC19x/Ontology+Reference

[33] http://mementoweb.org/guide/quick-intro/

[34] http://journal.code4lib.org/articles/12171, https://blogs.openaire.eu/?p=1371, https://blogs.openaire.eu/?p=1410, http://www.openscholar.org.uk/developing-the-first-open-peer-review-module-for-institutional-repositories

[35] https://bop.unibe.ch/index.php/index/index

University of Bern, which are both based on Open Journal Systems (OJS)[36]. Collaborations might be developed in this area.

- A study might take place concerning the file formats used to publish scientific results. For legacy reasons, PDF is by far the most used one, but it might no longer be the case in a near future. Indeed, this is more a lineage coming from the print age. We are witnessing a shift towards all-digital environments, and this is also true for the scientific publication. This digital evolution does not only concern the end format of the publication, but also the process of writing and editing, often collaboratively. So, SONAR should be prepared to appropriately handle scientific publications in HTML or EPUB. And proposing researchers a web interface to write and edit articles, as Authorea[37] already does, is something that should also be considered.
- Functionalities related to social interaction could also be interesting, such as the ability to comment publications and to categorize them through folksonomies[38]. With regard to commenting, in particular, that could be implemented with ResourceSync, based on WebSub[39]. The system could also provide notifications for article changes, new publications by an author or on specific subjects. (Cf. COAR's "Next Generation Repositories" report mentioned in sect 4.1.)

4.4 Environment analysis

There are some examples of international initiatives with an approach similar to SONAR's, with some variations. In Switzerland, there are several projects that are complementary to SONAR, with which collaborations and synergies should be undertaken.

In Switzerland

- **SLSP[40]**: the Swiss Library Service Platform project is setting up an organization that will provide library services to Swiss academic institutions. Among its planned future set of services, SLSP intends to offer institutional repository solutions to its clients. This kind of service could be provided by SONAR, for example, as outsourced solution. That possibility has been discussed since 2016 between the management and strategic instances of SLSP and RERO. It has been agreed to pursue and intensify these discussions in the coming months.
- **National licenses[41]**: The National Licenses project is working on acquiring the complete archives of scientific journals used in Switzerland. In this regard, this project is changing the legal environment of entire parts of the scientific publication landscape. The current RERO DOC Digital Library, which constitutes the foundation on which SONAR should be built, already hosts since August 2017 the publications of the National licenses that benefit from secondary publication (green open access) consent[42,43,44]. SONAR should naturally inherit and extend the delivery of this kind of service to the community.
- **Swissbib[45]**: Swissbib is the national metacatalog for most of the libraries and library networks in the country. It also includes bibliographic data from some Swiss IRs, such as BORIS or ZORA, and data from the National Licenses project. On the one hand, SONAR could provide its bibliographic metadata to Swissbib, making it easier for it to harvest extensively the Swiss IRs. On the other hand, SONAR could benefit from the technologies and expertise in harvesting and de-duplicating metadata and content developed in the framework of Swissbib.

[36] https://pkp.sfu.ca/ojs/
[37] https://www.authorea.com/
[38] https://en.wikipedia.org/wiki/Folksonomy
[39] https://www.w3.org/TR/websub/
[40] https://blogs.ethz.ch/slsp
[41] http://www.consortium.ch/nationallizenzen/?lang=en
[42] http://www.consortium.ch/open-access/?lang=en
[43] http://edoc.unibas.ch/57350/
[44] https://doc.rero.ch/collection/NATIONAL_LICENCES
[45] https://www.swissbib.ch/

- **HOPE**[46]: HOPE stands for Hauptbibliothek Open Publishing Environment. HOPE provides a platform for researchers of the University of Zurich to publish in Open Access newly founded journals as well as to migrate existing ones. In this project, the "Journal für Psychoanalyse" was converted to Open Access. The findings from this transition can be utilized for other Swiss researchers who would like to publish their own journals.
- **DICE**[47]: This project provides information, resources and learning tools related to copyright matters in higher education. These resources could help the researchers and the libraries staff to cope with the difficult task consisting to determine the legal situation around the deposit of a specific publication, for instance.
- **DLCM**[48], **ORD**[49]: These projects are providing information, resources, learning tools and hosting for research data. SONAR should collaborate with such projects in order to link publications and data sets.
- **SYMPHONY**: Swiss System for Monitoring bibliographic data and Holistic publication behavior analysis – that is the name of a project led in 2015 at the HTW Chur whose findings may be valuable for SONAR in the framework of data analysis and metrics.

International

- **OpenAIRE+**[50]: It is a repository collecting all the publications of the research projects funded by the H2020 program. On the one hand, OpenAIRE+ promotes and significantly improves the access to Open Access content. On the other hand, it aims to systematically link Open Access publications to the data sets and the funding information related to the same research project. OpenAIRE+ seeks also to improve the interoperability of european IRs through the adoption of guidelines. This project provides both inspiration for SONAR and potential for interoperation.
- **DABAR**[51]: DABAR provides IRs as a service for all institutions in Croatia, namely universities, institutes or faculties. At present, DABAR is thesis oriented, but it's still an interesting project as inspiration for SONAR.
- **HAL**[52] and **HAL-SHS**[53]: HAL and HAL-SHS are the French national IR. Most of the publications deposited in HAL are then pushed to the appropriate disciplinary repository, such as arXiv or PMC. HAL provides portals for different institutions, either as a subdomain of "archives-ouvertes.fr", or under the institution's own domain name. HAL also offers views related to thematic collections. Some services based on publication data are available, such as extracting data related to specific laboratory or institutions, and even to a specific researcher, as long as the data has been correctly reported when deposited.

4.5 Expected national benefits

Science has long been a globalized activity, conducted across borders, institutions, cultures and languages. Specific practices and traditions do exist within the scientific community, but they depend mainly on the domain, whereas national distinctions are in general less relevant. In this context, the question may be asked: is it reasonable to propose a project focused on a national perspective of science, given the existence of institutional repositories in Switzerland, as well as international subject repositories? The applicants respond positively, for the following reasons.

- A national perspective of scientific research does exist, with national funding agencies, infrastructure, policies and organizations. This justifies the existence of tools for assembling, pro-

46 https://www.hope.uzh.ch
47 http://www.diceproject.ch/
48 https://www.dlcm.ch/
49 http://openresearchdata.ch/
50 https://www.openaire.eu/
51 https://dabar.srce.hr/en
52 http://hal.archives-ouvertes.fr/
53 https://halshs.archives-ouvertes.fr/

moting and safeguarding Swiss research publications as a whole, something that does not currently exist, and that can help assess and develop Swiss science. A Swiss perspective of scientific information clearly appears in the vision of SUC P-2: *"The P-2 program envisions a future where academic needs for information handling and processing are seamlessly supported by a Swiss information provisioning and processing infrastructure that transcends the borders of individual institutions. The program shall strengthen Switzerland's reputation as a top location for education and research and as an attractive partner in international research collaboration"* (in: "Program SUC 2013-2016 P-2 „Scientific information: Access, processing and safeguarding", White Paper for a Swiss Information Provisioning and Processing Infrastructure 2020", sect. 1.2 *Vision*[54])

- SONAR does not intend to become a central, mandatory deposit location in Switzerland, nor does it aim at replacing current IRs. It does intend to assemble and promote possibly all open access publications by authors affiliated with Swiss public research institutions, a considerable part of which are not available in current IRs, and feed them back to those IRs.
- In addition to collecting open access full-text publications, SONAR also collects (though not exposing them in its search interface) metadata about all publications, including those that are paywalled, and stores them in a structured database, according to a defined data model. This creates a hub of normalized data about scholarly communication in Switzerland, which can later be used to assess Switzerland's research activity. In that sense, SONAR has the potential to become an important tool for supporting national research policies, and in particular the recent national open access strategy.
- In principle, such developments could be done independently by each institution, as extended features of its own IR. However, a single institution would lack the critical mass and cost effectiveness that a national-scale project can achieve. Indeed, data collected and intersected on a national scale, provides greater analytic potential. Besides, this project has an exploratory component that requires a research effort, an investment that IHEs would be reluctant to support alone for their own IRs, despite possessing the required expertise.
- Alongside the national service for publication collection, promotion and safeguarding, this project also proposes a mutualized (but customizable) IR solution for interested Swiss institutions, namely those that either do not yet possess one or prefer to replace their existing IR by an outsourced solution, thus creating economies of scale and collaboration opportunities at a national level.
- The relation and potential for cooperation with national projects and services in the scientific information domain is considerable, especially in the P-2 and P-5 area, e.g. SLSP, DLCM, ORD, Swissbib, National Licences, Swiss edu-ID, SYMPHONY, HOPE or CCDigitalLaw. This is another indicator of the national relevance of the current project.

[54] https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/WhitePaper_V1.1-EN.pdf

4.6 Implementation

The project is composed of 10 work packages, to be carried out over a period of 24 months.

Sections 4.2 “Project goals” and 4.3 “Proposed solution” show contextual references to the various work packages [in square brackets], providing textual descriptions. Below are provided brief portraits of each WP, including estimated effort, start and end dates, tasks and deliverables.

	2018		2019												2020									
	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10
WP1					x	x	x	x	x	x														
WP2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
WP3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x						
WP4	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x								
WP5	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x						
WP6											x	x	x	x	x	x	x	x	x	x	x	x		
WP7							x	x	x	x	x	x	x	x	x	x	x	x						
WP8												x	x	x	x	x	x							
WP9	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
WP10	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24

WP 1: Business model and governance

Effort: 4 man/months.

Start: 01.03.2019 – End: 31.08.2019 (6 months)

Tasks:

1. Identify possible business model types for prospective services
2. Quantify operational costs for prospective services
3. Identify future strategic developments: services and collaborations
4. Define present and future stakeholders: clients, partners
5. Propose a business model for prospective services
6. Identify key criteria and scenarios for a governance structure

Deliverables:

1. [D1.1] Business model proposition for prospective services [document]: 31.07.2019
2. [D1.2] Governance structure scenarios [document]: 31.08.2019

WP 2: Main IT solution: tools and procedures

Effort: 10 man/months

Start: 01.11.2018 – End: 30.10.2020 (24 months)

Tasks:

1. Prepare the hardware infrastructure
2. Install and configure Invenio3, using a setup that allows for the creation of multiple, independent but interoperable instances (SONAR main, multiple client IRs)
3. Define the procedure for creating and configuring future additional client IR instances, on demand
4. Define the underlying data model: internal schema and Linked Data (including authors, publishers, journals, research institutions, funding agencies, research projects/grants and patents)
5. Define and implement authority data management procedures
6. Define and implement open APIs
7. Define and implement persistent identifiers: DOI/ARK and URI assignment; harvesting of all kinds of relevant identifiers from external sources (DOI, URN, URI, ORCID…)
8. Define and implement import and export formats and protocols
9. Implement document deduplication

10. Migrate structure and data from RERO DOC

Deliverables:

1. [D2.1] SONAR hardware and software (Invenio3) basic installation ready for development [system]: 02.02.2019
2. [D2.2] Procedures for creating and configuring additional “IR as a Service” instances [tools and documentation]: 03.09.2019
3. **[D2.3] Stabilized SONAR system in production [system]: 28.10.2020**

WP 3: User interface

Effort: 13 man/months

Start: 01.11.2018 – End: 30.04.2020 (18 months)

Tasks:

1. Main content search and navigation
2. IRs content search and navigation
3. Content deposit
4. Conceive the user interface design
5. User’s space
6. User authentication (Swiss edu-ID, ORCID)
7. Author pages
8. User documentation
9. Multilingualism
10. Usage statistics page
11. Tests

Deliverables:

1. [D3.1] Mockups [document]: 30.04.2019
2. [D3.2] User interface implemented [system]: 29.04.2020

WP 4: Interaction with Swiss IRs

Effort: 8 man/months

Start: 01.11.2018 – End: 28.02.2020 (16 months)

Tasks:

1. Define the goals and requirements for the interaction with Swiss IRs
2. Establish the procedure for data exchange between SONAR and Swiss IRs
3. Negotiate with Swiss IRs representatives
4. Implement the data exchange process

Deliverables:

1. [D4.1] Agreements with Swiss IRs representatives [document]: 30.04.2019
2. [D4.2] Data exchange process implemented [system]: 25.02.2020

WP 5: Recovering of full-text from 3rd-party OA (Feasibility study)

Effort: 12 man/months

Start: 01.11.2018 – End: 30.04.2020 (18 months)

Tasks:

1. Define a pilot subset of international open archives
2. Identify articles authored by Swiss affiliated researchers
3. Synchronization and evaluation
4. Define APIs for the harvesting of the full-text

Deliverables:

1. [D5.1] Evaluation of acquisition accuracy [document]: 29.10.2019
2. [D5.2] JSON/XML services [system prototype]: 29.04.2020

WP 6: Content dissemination

Effort: 2 man/months

Start: 01.09.2019 – End: 31.08.2020 (12 months)

Tasks:

- Extend RERO DOC's content indexing by Google Scholar to SONAR
- Analyze and launch content indexing by Microsoft Academic
- Implement metadat export to ORCID

Deliverables:

1. [D6.1] SONAR's content ready to be indexed by academic search engines (Google Scholar and Microsoft Academic) [system]: 02.09.2020
2. [D6.2] Metadata export process to ORCID implemented [system]: 02.09.2020

WP 7: Transformation and standardization of full-text contents

Effort: 5 man/months

Start: 01.05.2019 – End: 30.04.2020 (12 months)

Tasks:

1. Analyze existing standards and transformation tools
2. Customize and integrate tools for document transformation

Deliverables:

1. [D7.1] Reviews of formating standards and resources for document representation [document]: 30.07.2019
2. [D7.2] Extraction of full-text from PDF, XML and HTML [system]: 28.01.2020
3. [D7.3] Metadata extraction from the full-text [system]: 29.04.2020

WP 8: Analytics

Effort: 3 man/months

Start: 01.10.2019 – End: 31.03.2020 (6 months)

Tasks:

1. Explore possible uses of collected data
2. Identify needs from potential 3rd parties

Deliverables:

1. [D8.1] Report on analytics [document]: 31.03.2020

WP 9: Content preservation

Effort: 8 man/months

Start: 01.11.2018 – End: 31.08.2020 (22 months)

Tasks:

1. Identify long-term preservation requirements
2. Concept for long term preservation: evaluation of solutions
3. Implement the selected solution

Deliverables:

1. [D9.1] Study of the requirements and selection of the solution [document]: 02.07.2019
2. [D9.2] Content preservation solution implemented [system]: 02.09.2020

WP 10: Coordination and promotion

Effort: 6 man/months

Start: 01.11.2018 – End: 30.10.2020 (24 months)

Tasks:

1. Project management

2. Meetings
3. Communication about the project: blog, social networks, professional journals, conferences
4. Customer prospecting

Deliverables: [none]

4.7 Milestones

No.	End date	Work package/ project phase (cf. 4.6)	Description (result/deliverable)
1	02.02.2019	WP2. Main IT solution: tools and procedures	D 2.1 SONAR hardware and software (Invenio) basic installation ready for development
2	30.04.2019	WP1. Business model and governance	D 1.2 Governance structure scenarios
3	30.04.2019	WP3. User interface	D 3.1 Mockups (User Interface)
4	30.04.2019	WP4. Interaction with Swiss IRs	D 4.1 Agreements with Swiss IRs representatives
5	02.07.2019	WP9. Content preservation	D 9.1 Study of the requirements and selection of the solution
6	30.07.2019	WP7. Transformation and standardization of full-text contents	D 7.1 Reviews of formating standards and resources for document representation
7	31.07.2019	WP1. Business model and governance	D 1.1 Business model proposition for prospective services
8	03.09.2019	WP2. Main IT solution: tools and procedures	D 2.2 Procedure for creating and configuring additional "IR as a Service" instances
9	29.10.2019	WP5. Recovering of full-text from 3rd-party OA (Feasibility study)	D 5.1 Evaluation of acquisition accuracy
10	28.01.2020	WP7. Transformation and standardization of full-text contents	D 7.2 Extraction of full-text from PDF and XML
11	25.02.2020	WP4. Interaction with Swiss IRs	D 4.1 Data exchange process implemented
12	31.03.2020	WP8. Analytics	D 8.1 Report on Analytics
13	29.04.2020	WP3. User interface	D 3.2 User interface implemented
14	29.04.2020	WP5. Recovering of full-text from 3rd-party OA (Feasibility study)	D 5.2 JSON/XML services
15	29.04.2020	WP7. Transformation and standardization of full-text contents	D 7.3 Metadata Extraction
16	02.09.2020	WP6. Content dissemination	D 6.1 SONAR's content ready to be indexed by academic search engines (Google Scholar, Microsoft Academic)
17	02.09.2020	WP6. Content dissemination	D 6.2 Metadata export process to ORCID implemented
18	02.09.2020	WP9. Content preservation	D 9.2 Content preservation solution implemented
19	28.10.2020	WP2. Main IT solution: tools and procedures	D 2.3 Stabilized SONAR system in production