| Acronyme | DISSEMIN |
|---|---|
| **Titre de la proposition** | Diffusion d'articles scientifiques et implémentation de politiques institutionnelles |
| **Nom et prénom du responsable (coordinateur)** | Amarilli, Antoine |
| **Organisme du coordinateur** | Identification de l'établissement : Télécom ParisTech<br>Adresse : 46 rue Barrault, 75013 Paris |
| **Choix d'une des catégories suivantes:** | Web service development |
| **Aide demandée** | 99 360 euros |
| **Coût complet :** | 169 360 euros |
| **Durée** | 24 mois |
| **Mots clés de la proposition** | Open access, text and data mining, open repositories |

## 1. Description of the proposal

Enabling universal access to the scientific literature was the original reason why the World Wide Web was invented. Still, thirty years later, the problem is far from solved. University libraries are wrestling with academic publishers who ask for unaffordable subscription prices, and internet service providers block pirate websites which bypass paywalls.

The *open access movement* is an attempt to tackle the root of this issue by publishing articles openly on the web. To encourage this practice, universities and research funders have adopted policies that encourage their researchers to make their articles publicly available. Some editorial boards have moved away from traditional publishers and now run their journals electronically. Open access repositories have become the go-to platforms to read the latest research in many fields, even before it has been published in a peer-reviewed venue. Still, despite all these efforts, most of the scientific literature is still locked behind paywalls.

In this project, we propose to address this issue by developing the Dissemin platform. Dissemin (https://dissem.in/) is a web service to help researchers ensure that their publications are free to read. It

does so by gathering information from various metadata sources and analyzes the online availability of each paper (over 100 million, see diagram) to determine which articles are freely available online and which are not.

Articles which are free to read can be either:
- openly available from the publisher itself ;
- or self-archived in an open access repository (known as green open access).

Other articles are broken down into three categories :
- those which could be legally uploaded by the authors to an open access repository, e.g., when the publisher's policy allows it;
- those for which the publisher's policy is unclear ;
- those which cannot legally be uploaded to an open repository.

In addition to this analysis, Dissemin also makes it easy for researchers to upload their publications to a selection of open repositories (HAL, Zenodo and soon OSF preprints). Depositing articles with Dissemin is simpler than doing it directly on the repositories, because Dissemin pre-fills all article metadata (authors, title, etc.) because it has already harvested it.

The Dissemin platform is already available online, but researchers currently have little incentive to use it: most researchers are not familiar with open access and do not realize the importance of making their research widely available online. In this proposal, our goal is to change this, using the policies adopted by research institutions and funding agencies: indeed, many such actors have adopted *open-access mandates* requiring their employees and project members to deposit their research openly, and compliance with these mandates is a prerequisite for promotions and grants.

Our goal is therefore to align Dissemin with institutional policies: we want to make Dissemin a tool of choice for universities and funders to implement their open access policies and monitor their uptake. This can be decomposed in several milestones that we detail below.

### *Objectives and milestones*

#### Integrating institutional repositories in Dissemin.

Institutions generally ask their researchers to use specific open access repositories to make their publications freely available, e.g., repositories hosted by the institutions, or national repositories such as HAL in France. It is therefore crucial that Dissemin interfaces with as many of these repositories to serve the needs of researchers. To do this, we need to develop connectors to various repository vendors, and to negotiate agreements with repositories to enable direct deposit interfaces, as well as maintaining and improving existing connectors.

#### Linking publications to their funders and institutions.

There is currently no simple mechanism for funding agencies and research institutions to monitor the publications that are produced by the researchers that they employ or finance, and it is even harder to

determine whether these publications have been deposited as open-access. For this reason, these actors cannot check the outcome of the open access policies that they adopt. We plan to simplify this analysis with Dissemin by recording the link between articles institutions, which will also encourage institutions to promote Dissemin to their users and to develop connectors for their own repositories. Establishing this link will be done using existing curated data sources (Fundref) and disambiguation of unstructured ones, using an entity linking algorithm. This entity linking algorithm constitutes by itself an important deliverable as it will be usable for other text mining applications. In particular, we intend to experiment with an email campaign in ESPCI to deposit the missing publications in HAL, following our existing collaborations with the ESPCI library (see "Consortium" below).

### Integrating funder policies and fostering their implementation.

Beyond linking publications to institutions, we want to represent the open access mandates adopted by these institutions, by integrating to Dissemin the existing databases for such policies (ROARMAP and SHERPA Juliet). Storing such policies in a structured format will allow us to assess their compatibility with publisher policies and with more general laws. For instance, the République Numérique bill introduces gives researchers in French public research institutions the right to deposit their articles, overriding the usual embargo periods of many publishers.

#### *State of the art*

To our knowledge, no comparable platform to Dissemin currently exists. Tracking the publications authored within a given institution is generally done with Current Research Information Systems (CRIS), such as Pure (Elsevier) or Elements (Symplectic), none of which are open source and all of which can only be used internally in a single institution. These systems are therefore not suited to implement funder policies or represent overlapping affiliation layers which are common in research.

#### *Existing achievements to justify the credibility of the proposal*

Dissemin has already been used to run a campaign to deposit the publications cited in Wikipedia. After harvesting which research articles were cited in Wikipedia, the authors of these works were contacted and encouraged to use Dissemin to upload their publication to an open archive. Thanks to this campaign, thousands of Wikipedia references were improved by adding a link to an open access copy. The improvement of these references continues since this initiative, with a crowd-sourced effort run every year at the occasion of the international Open Access Week. There is therefore evidence that Dissemin is well suited for campaigns reaching out beyond single institutions, which would be impossible with existing CRIS systems.



We have already demonstrated our ability to integrate repositories into Dissemin via three approaches. First, we have already developed some connectors allowing Dissemin users to deposit their articles in a selection of open-access repositories directly from the Dissemin website. These repositories provide long term archival, are well indexed in search engines and are run by not-for-profit organizations.

On second form of repository integration is to provide data from Dissemin in institutional repositories. For instance, the University of Bordeaux uses Dissemin directly inside its institutional repository, using the Dissemin API which provides an open and machine-readable interface to Dissemin's data for third-party systems. Researchers of the university are presented with a list of publications that they could upload, directly from the institutional repository. This also simplifies the deposit process again thanks to the metadata exposed by Dissemin.

Third, integration with repositories can also take the form of self-hosted instances of Dissemin and/or contributions to the Dissemin codebase. In this spirit, we are working closely with a group of three German universities (TU Darmstadt, University of Stuttgart and TU Braunschweig) as part of their OpenIng project, funded by the German Federal Ministry of Education and Research (grant amount: 299k€). This project focuses on engineering sciences and intends to connect Dissemin to additional metadata sources, developing custom repository connectors and integrating with the authentication mechanisms of the partner institutions.

These three approaches illustrate how the project can realistically address its goals. They also underline the importance of the open source license chosen for the project: Affero GPLv3. This license ensures that third-parties can freely reuse and improve the platform, and that they must publish their own improvements to the platform under a similar license. This means that any improvement made to Dissemin in the scope of their project will be reusable on our own instance of the software. The successful funding of the OpenIng project, and the fruitful collaboration between their development team and ours, shows that this model offers a viable route to long term sustainability.

Overall, we believe that the Dissemin project can credibly address the goals of the call "ANR Flash sciences ouvertes", by fostering the development of open science practices across many domain areas, and contributing to the development of open research data and metadata, in particular the open-access status, bibliographic metadata and institution/funder information for over 100 million research articles. We request financial help from ANR to fund the manpower required to accelerate the development of this platform and focus on the new features of integrating institutional repositories and managing institution and funder information and open-access mandates, which we believe are instrumental to the development of open data and open science for research outputs.

**Tasks**

### A. Data model design

To link publications and researchers to institutions and their open access policies, we have to design new data models to store this information in the platform. Our goals are threefold:
1. Leverage existing databases of institutions and policies, to learn from past experiences and make data ingestion easier.
2. Make it possible to infer the institutional policies applying to a given publication. Researchers can be affiliated, employed or funded by institutions in various ways, and policies have varying scopes and effects. The model should be able to represent these situations in a faithful way.
3. Minimize the curation load on the platform administrators and make it possible for users to easily fix and improve the data themselves. For instance, we want to make it easy for

researchers to add their own research laboratory, or for librarians to improve the linkage of researchers to the institution they work for.

We plan to rely on the open knowledge base Wikidata, where a database of academic institutions is already being curated, drawing up on many established databases[1]. Our goal is to mirror this dataset on the Dissemin platform while letting our users improve this data directly on Wikidata.

### B. Institution linkage

Dissemin currently holds the metadata of more than 100 million publications from authors around the world. Our goal is to link as many of these publications to institutions, via various open data sources:

1. Curated databases such as Crossref's funding registry (Fundref), linking 3.8 million publications to funders, or OpenAIRE's funding relations. This data comes with institution identifiers that are linked to Wikidata.
2. Free-text affiliation strings stored in Crossref, ISTEX and other sources. These strings can be aligned to identifiers with entity linking tools, but possibly introducing noise in the data.
3. Employment information from ORCID, which come with time boundaries and institution identifiers but whose accuracy is also not perfect.

The goal is to perform this linkage without compromising the accuracy: coverage can then be optimized for a selection of institutions, when the interest develops.

### C. Outreach and campaigning tools

Once we have sufficient publication coverage for a given institution, we want to make it easy for this institution to use Dissemin to improve the availability of these publications. To this end, we will run a pilot project in ESPCI to establish the appropriate workflows. This means:

1. Providing simple ways to export Dissemin's data, for instance by downloading a spreadsheet of all publications in a given institution, with their status and metadata;
2. Writing tutorials to help librarians and research administrators work with this data, relate it to their own research information system and conduct email campaigns with it;
3. Develop other tools as suggested by the feedback received in the pilot project.

We also plan to work with Wikimedia Italy to explore the lessons learnt by their own campaign run with Dissemin, which focused on freeing the access to articles cited in Wikipedia.

### 2. Consortium

The consortium of this project consists of the people who are already working on the development of the Dissemin platform. The project will be administratively hosted at Télécom ParisTech and managed by Antoine Amarilli, but the bulk of the work will be performed by this existing network of collaborators and by the research engineer that we intend to hire. We stress here that, because Dissemin is a large open project involving many actors, for simplicity the consortium members were not administratively registered on the ANR website or accounted as project costs: nevertheless, all the people listed here have confirmed their implication in the project proposal.

**Project members:**

---

[1] Specifically, for institutions: RNSR, GRID, ISNI, VIAF, SUDOC, AURÉHAL, RINGGOLD and others. For policies: ROARMAP and SHERPA/Juliet.

- **Antoine Amarilli**, project director.
  - 🪙 Associate professor of computer science, Télécom ParisTech. Institut Polytechnique de Paris.
  - 🪙 With his expertise on complexity theory and database systems, Antoine specializes in uncertain data management and is therefore best placed to mitigate the risk induced by noisy metadata sources.

- **Pierre Senellart**, project director.
  - 🪙 Professor, Computer science department, École normale supérieure, PSL University. Head of the Inria Valda team.
  - 🪙 As a leading researcher in complex data management and web-scale systems, Pierre has key insights on the technological and scientific challenges required to build a web-scale information retrieval system.

- **Marie Farge**, spokesperson.
  - 🪙 Emeritus research director, LMD-IPSL, CNRS & École normale supérieure, PSL Université Paris.
  - 🪙 Marie has a deep understanding of the scholarly communications ecosystem, and is a renowned open science expert with extensive contacts in a variety of fields. She has served on many policy boards, including at the European Research Council.

- **Lucas Verney**, research engineer.
  - 🪙 PhD student, École normale supérieure, PSL Université Paris.
  - 🪙 With his extensive experience on developing open source projects and his expertise in bibliographic management systems, Lucas is best placed to enhance Dissemin's usability, reliability and feature portfolio.

- **Catherine Kounelis**, deposit process supervisor
  - 🪙 Head Librarian, Library & Center for Historical Resources, ESPCI Paris, PSL Université Paris
  - 🪙 Catherine knows what it takes to fill an institutional repository and will play a pivotal role in making Dissemin fulfill the needs of institutions by running a trial campaign at ESPCI.

- **Charles Paperman**, database specialist
  - 🪙 Maître de Conférences, University of Lille
  - 🪙 With his expertise on database systems, Charles is improving Dissemin's storage back-end to make sure it can ingest article metadata at a high speed and be easily extensible.

- **Patricia Mirabile**, policy specialist.
  - 🪙 PhD student, Lettres Sorbonne Université (Paris IV).
  - 🪙 Patricia pushed for ENS's open access mandate and now serves on the French Open Science Council (COSO). With her expertise on open science and reproducibility, she is able to position Dissemin to match researchers' needs.

- **Antonin Delpeuch**, core developer.
  - 🪙 PhD student, University of Oxford.

🌕 Specializing in natural language processing, he brings in the expertise for the required disambiguation algorithms. He was nominated Open Access Champion by SPARC Europe for his work on Dissemin.

**Partners :**

- **OpenIng project** (TU Darmstadt, University of Stuttgart and TU Braunschweig) funded by the German Federal Ministry of Education and Research (grant amount: 299k€). This project focuses on the engineering sciences in Germany and intends to use Dissemin to foster open access in this field. This involves connecting Dissemin to additional metadata sources, developing custom repository connectors and integrating with their own authentication mechanisms.
- **Centre pour la Communication Scientifique Directe** (CCSD, UMS 3668). This partnership convened and funded by Couperin Consortium enabled us to integrate HAL deposit in Dissemin. We continue to maintain this interface in coordination with HAL's development team.
- **University of Bordeaux** has integrated Dissemin in their institutional repository via our API.
- Other partners : we already have a working relationship with other stakeholders which run services that we rely on or interface with : ORCID, Bielefeld Academic Search Engine, Crossref, the Open Access Button, Unpaywall and others.

## 3. Impact of the proposal

The proposal intends to develop the following deliverables :

A. **Fostering the archival of publications,** either directly through the Dissemin platform itself or by prompting deposits elsewhere, and more broadly by raising awareness about the issue in the research community.

This is the main goal of the platform, and the need for open archiving of scientific publications has been widely expressed across many academic communities. We can estimate success for this goal by counting how many deposits were made via the Dissemin platform, but we note that our goal is not that researchers deposit specifically via Dissemin. In particular, we do not especially intend to optimize long term user retention on Dissemin, for instance.

B. **Providing a new tool for research institutions and funders** to monitor and implement the policies they adopt, integrating with the information systems that they already use such as ORCID and their institutional repositories ;

This concentrates the new features developed as part of this project. By creating an institutional view on our data, Dissemin will become usable by research institutions to track the availability of their research output and conduct campaign to improve it. We plan to develop and document simple workflows for institutions to run their own email campaigns using our service. Our goal is to make these campaigns as accessible as possible, requiring little technical expertise. We will also ensure that they do not require institutions to share personal data (e.g., lists of employees) with us.

C. **Delivering open data and software** that can be reused for other purposes and integrated in third-party systems freely.

As a byproduct of the Dissemin project, we are curating an open datset that can be reused for other purposes. For instance, our users have created the "dissemr" R package to query our API from scripts in the R language.[2] This makes it possible to search for publications and retrieve their self-archiving policy and the repositories where they can be downloaded, for instance. Other tools produced as part of this project include a machine learning algorithm to predict the topic of a document submitted to HAL[3], and an entity linking system for Wikidata, trained on academic affiliations[4]. We anticipate that the project will continue to produce similar deliverables in the future.

### 4. Communication of the results

The results of the project will be communicated by various means :
- Presentations given at international events on scholarly communications (LIBER 2015, Open Repositories 2016, Open Science Conference 2016, submission to Open Science Fair 2019);
- Events organized around the project itself (hackathons, advocacy workshops);
- Publications about our approaches to the technical challenges involved in building the platform;
- Outreach to researchers and librarians via email campaigns and social media.

The Dissemin code is released under the GNU Affero General Public License v3.0 as recommended by DINSIC[5] and we provide an open API which can be used freely without registration. Our data sources are all open, meaning that it is possible to run the platform independently of our own instance. The source code is archived in Zenodo (DOI 10.5281/zenodo.2571984) and Software Heritage.

The platform is run by a non-profit association, the Committee for the Accessibility of Publications in Sciences and Humanities (CAPSH), which funds the technical infrastructure using donations and occasional financial support from our partners. This organization is run openly. The statutes and minutes of its meetings can be found at https://association.dissem.in/. CAPSH will continue to coordinate the development of Dissemin when the project is over: we intend to continue maintaining the platform even after the end of the project, as we have already been doing for several years.

### 5. Auto-evaluation processes

We propose to evaluate our project on the following criteria :
- Coverage and quality of links between articles and institutions. These links are ultimately used to link articles to the relevant open access policies adopted by the funders and employers and we therefore need to ensure that they are reliable.
- Accuracy of access statistics for articles. This is important to provide an accurate view of the accessibility of an institution's research output.
- Usage of our deposit interface for repositories. Experience has shown that it can rise sharply when running targeted campaigns.

### 6. Budget

---

[2] https://docs.ropensci.org/dissemr/
[3] https://github.com/wetneb/haltopics
[4] https://github.com/wetneb/opentapioca , described in arXiv:1904.09131
[5] https://www.numerique.gouv.fr/publications/politique-logiciel-libre/pratique/

For simplicity, we intend that Télécom ParisTech will handle the whole budget of the project. Our main need is to secure developer time to improve the platform. We aim to hire a research engineer to carry out all proposed improvements, which will be employed by Télécom ParisTech and collaborate with consortium members. The breakdown of the cost is as follows:

- One full time research engineer for 1 year and 3 months, 82k€
- Equipment (e.g., computer for the engineer), 1.5k€
- Web hosting costs (e.g., server rental), 5k€
- Travel costs (e.g., presentation in conferences), 3.5k€

Total cost: 92 000 EUR
Handling fee for Telecom ParisTech: 8 %
Total cost including handling fee: 99 360 EUR

Administratively, the cost of the project which is not covered by ANR is of 11747 EUR corresponding to 10% of the salary of the project coordinator over 24 months. However, we underline hat much of the work on the project will also be performed by other consortium members (see "Action plan" below) which are not administratively accounted for in the proposal's budget.

## 7. Action plan

Timeframe of the project: January 2020 — January 2021

| Task | Members | Timeframe |
|---|---|---|
| 1 - Data model design | L. Verney, C. Paperman | Jan 2020 — May 2020 |
| 2 - Software development | Research engineer, L. Verney | Apr 2020 — Dec 2021 |
| 3 - Documentation | Research engineer, L. Verney | Jun 2020 — Dec 2021 |
| 4 - Assessment | L. Verney, A. Delpeuch, P. Mirabile | Sep—Oct 2020 and 2021 |
| 5 - Communication, outreach | M. Farge, A. Amarilli, C. Kounelis | Oct 2020 — Dec 2021 |
| 6 - Project management | A. Amarilli, A. Delpeuch, P. Senellart | Jan 2020 — Dec 2021 |

1 - **Data model design**
- Improve the current database schema to enhance scalability and maintainability;
- Design the data model to represent institutions and their policies;
- Create interface mocks for institution and policy integration in the platform;
- Linking with the Wikidata community to synchronize data models.

2 - **Software development**
- Implement the new features as defined by the goals of the project;
- Test and document the code as part of the development process;

- Carry out general maintenance of the existing code base;
- Respond to user feedback to improve the workflows.

3 - **Communication, outreach**
- Run focused email campaigns in partner institutions to encourage the deposit of articles;
- Present the project and the platform in open science conferences;
- Participate in policy discussions at the national and European level;
- Communicate with users and report the feedback to development team.

4 - **Assessment**
- Run evaluation campaigns to evaluate the quality and coverage of the curated data;
- Monitor and advise development workflows;
- Conduct focused user studies to identify usability issues;
- Communicate assessment results to the team for inclusion in funding reports and external communication.

5 - **Documentation**
- Write documents explaining the general architecture of the system, integrated in the existing documentation for the platform;
- Write articles explaining the challenges encountered and the technological solutions adopted, aimed at a wider audience;
- If relevant, present the architecture in a technical conference.

6 - **Project management**
- Hire and supervise the research engineer;
- Report to funders;
- Communicate with partners;
- Organize project meetings.

| 2020 | | | | | | | | | | | | 2021 | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 |

**1:** Design

**2:** Software development

**3:** Documentation

**4:** A.            **4:** A.

*Deployment of the institutional view*

**5:** Communication, outreach

**6:** Project management