Research Ideas and Outcomes 2: e8462
doi: 10.3897/rio.2.e8462

OPEN ACCESS

Small Grant Proposals

Paperity Central: An Open Catalog of All Scholarly Literature

Marcin Wojnarski‡,§, Debra Hanken Kurtz|

‡ Paperity, Warsaw, Poland
§ University of Warsaw, Faculty of Mathematics, Informatics and Mechanics, Warsaw, Poland
| DuraSpace, Beaverton, OR, United States of America

Corresponding author: Marcin Wojnarski (mwojnarski@paperity.org)

Reviewable v1

Received: 12 Mar 2016 | Published: 14 Mar 2016

Citation: Wojnarski M, Hanken Kurtz D (2016) Paperity Central: An Open Catalog of All Scholarly Literature. Research Ideas and Outcomes 2: e8462. doi: 10.3897/rio.2.e8462

Abstract

The goal of this project is to build Paperity Central, a global universal catalog of Open Access (OA) literature - "gold" and "green" combined - that will ultimately include 100% of past and new open literature, and subsequently - with the advent of universal OA - will become a catalog of all scholarly literature published anywhere in the world. Paperity Central will combine automatic indexing of journals and repositories, with manual data curation via a "wiki"-type functionality; will expose open APIs for programmatic access to data and development of add-on services and applications; and will greatly facilitate the discovery and navigation in OA literature, as well as dissemination of new research.

The catalog will be built by extending an existing prototype, Paperity - an aggregator of gold & hybrid OA journals - with three key features:

1) The aggregation of green OA literature from repositories. "Green" metadata will be seamlessly merged with "gold", in a systematic and consistent way: with deduplication of repeated entries, assignment of globally unique permanent identifiers, reconnecting every item with its primary source of origin (e.g., a journal) and establishing semantic links with related objects: author profiles, institutions, funders, grants, datasets, protocols, reviews, cited/citing works...

2) The "wiki" functionality that will enable users to improve, curate and extend the catalog manually in a collaborative, community-controlled way, like in Wikipedia: with full history and transparency of edits, easy rollbacks, moderation of edits by peers. Manual curation will be particularly important for "green" metadata, which frequently contain missing or incorrect information; and for cataloguing those publications that are inaccessible for automatic harvesting, like the articles posted on author homepages only.

3) Open APIs, featuring in particular a feedback loop from Paperity to repositories (including DSpace, a partner in this proposal), enabling source repositories to pull all metadata corrections and extensions collected by Paperity.

Keywords

open access, open science, open data, wiki, API

Introduction

Literature is the fabric and the substance of Science. It communicates, preserves and gives form to all research findings, past and new, that comprise the edifice of Science. It establishes the record of past discoveries and marks the starting point for new research. It constitutes the central communication medium that connects all scholars around the world - across countries, institutions and disciplines - and makes them into a community: the scholarly community.

In the same way, *open* literature is the fabric and the substance of *Open* Science. If we are serious about opening up the system of scientific research, we must plant it on the foundation of open literature and make sure that this literature is properly organized and maintained: accessible for all in one central location, easily discoverable, available within its full context, annotated and semantically linked with related objects. Most importantly, it must be catalogued in a way that permits unique identification of every single item throughout the entire body of knowledge and deep analysis of relationships between items.

Currently, the scholarly literature - including its Open Access (OA) subset - is unordered, dispersed over thousands of different websites and disconnected from its context. Nobody really knows what elements exactly contribute to the scholarly record, where and how to find them or how to analyse deep links between them. For example, assume we want to find all articles on Zika published in 2015. We can find some of them today using services like Google Scholar or PubMed Central, but how do we know that no other exist? Or that we have not missed any important piece of literature? With the existing tools, which have incomplete and undefined coverage, we do not know and will never know for sure.

The dense network of scientific knowledge comprises millions of interconnected elements (Fig. 1): articles, theses, reports, monographs. They come with millions of collateral objects representing context of each publication: experimental datasets, source code, author &

institution profiles, reviews, readers' opinions, publication and deposition venues (journals, repositories), archival copies, funding information etc. Today, neither we have a full view of all core nodes of the network, nor do we possess any structured knowledge of collateral information that accompanies every publication item. This situation brings vast damage to Science and scientists: impedes research, hampers dissemination, prevents creation of higher-level services and obstructs fair evaluation of research.

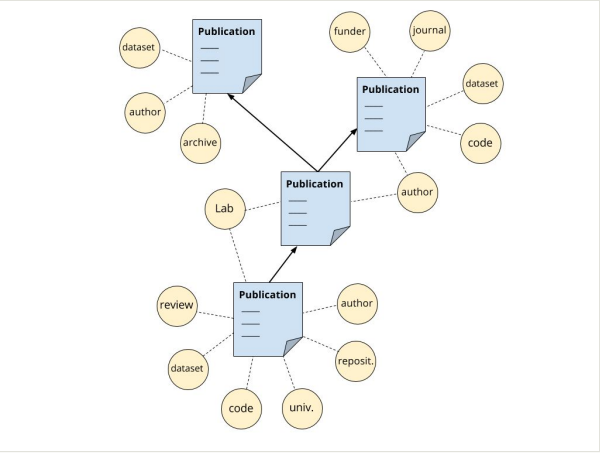

Figure 1.

The network of scientific knowledge is composed of publications in its core (blue; typically connected by citation relationship) and collateral objects of various types (yellow) that represent context of each publication. In Paperity Central, core nodes will be stored as native objects with global permanent identifiers, fully searchable, browsable and wiki-editable. Collaterals will be represented either as native objects or as annotated links to external resources elsewhere on the web.

With Paperity Central, we want to change that.

The Solution

Paperity Central will solve the above problems by creating a central catalog of all scholarly literature, one that is *complete*, *consistent* and *comprehensive*:

1. **Complete.** It will include ultimately all scholarly publications, past and new, published anywhere in the world, in any language. It will be designed in such a way that achieving 100% coverage is not only technically, but also practically possible. It will focus on OA literature and will strive to provide full texts wherever available, but will allow the addition of bibliographic records for closed publications, as well, in a hope that their full text will be found and added later on (such entries will have lower visibility to prevent distractions when navigating in full-text resources).

2. **Consistent.** The catalog will perform record deduplication and linkage, and assign globally unique identifiers and permanent URLs, which can be used subsequently in bibliographic citations, referred to by other academic services or pulled back by institutional repositories. The catalog will keep structured information about publications, with every item uniquely cross-linked to its *primary source of origin* (a journal; a proceedings object; an institution object for dissertations etc.). With these publication-origin links in place, the catalog will provide clear arrangement of all its content, easy browse & search, and precise deduplication and disambiguation of new records. Whenever these links are missing (e.g., for "green" records from repositories), the system will recreate them: automatically when enough data is available or manually with users' help.

3. **Comprehensive.** Every item will come with rich metadata and contextual information about all related objects (Fig. 1). Full text of publications will be readily available: displayed directly in Paperity or provided as links to external files if displaying is impossible for legal or technical reasons. Collaterals will be stored as semantically annotated links to external objects elsewhere on the web, thus integrating all different types of Open Science resources: implementations (e.g., from github or other services), datasets (figshare), workflows (myExperiment), author profiles (ORCID), reviews (Publons) etc.

How to achieve these goals? Given an immense volume of scholarly literature and the complexity of relationships between individual items, the catalog must combine automatic aggregation of literature - harvesting bulk data from the web on a massive scale, thus solving the problem of quantity - with manual edits contributed in a controlled way by users who would fix up any erroneous or missing data after automatic harvest, thus solving the problem of quality.

This is exactly how Paperity Central will work. It will combine (Fig. 2):

1. An **aggregator** of OA literature that automatically harvests new publications from places of their original publication (journals, conference proceedings etc.; gold OA) or from deposition sites (repositories; green OA).

2. A **wiki-catalog** where users can manually add records or edit existing ones, by editing metadata (title, authors, keywords etc.), uploading or backlinking to full text, linking to related external objects etc.

These two functionalities will be seamlessly combined into one system: the data of an automatically harvested item could be refined and extended by wiki-editors, while the data inserted manually by users could be extended through automatic harvesting. Both types of objects - created automatically or manually - will be accessible through the same browse & search mechanism, with the ability to narrow or prioritize search results based on origin.

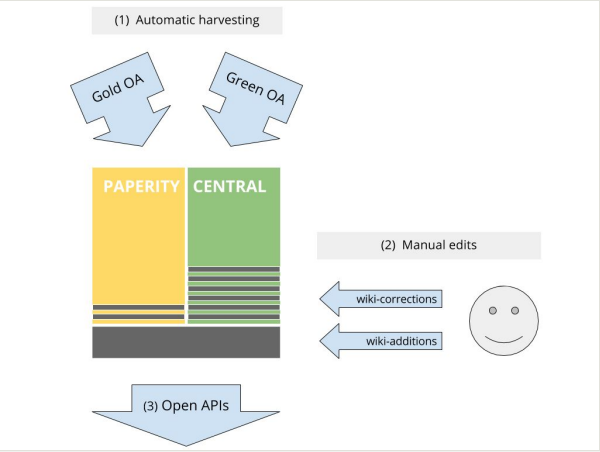

Figure 2.

Paperity Central will combine (1) automatic harvesting of "gold" and "green" OA literature with (2) manual wiki-edits by users: corrections/extensions to existing entries (grey stripes) or additions of new entries (grey box). Manual edits are more important for "green" than "gold" data, because of lower quality of "green" metadata as pulled from repositories, compared to "gold" data pulled from original sources (publisher websites). Paperity Central will also expose all the collected data to external applications through open APIs (3).

We predict that ultimately **90%** of records will be added automatically by the harvesting process, while **10%** will be created manually - those which are too dispersed to be found by an automated process. Manual edits provide the only viable solution for the *long tail* problem: the most dispersed 10% of publications would require 90% of implementation effort if we wanted to harvest them automatically, thus, to keep the project technically viable, we will rely on users' help to index them.

The "wiki" functionality will - in its principles - work in a similar way as on Wikipedia. It will be open to all users, allow collaboration (multiple edits by different users) and provide community control: full transparency guaranteed by public history of changes, easy rollbacks to quickly fix up mistakes or vandalisms, moderation of edits by peers (especially by journal editors or IR admins overlooking the items assigned to their collections).

Despite sharing the same editing principles as Wikipedia, Paperity Central will also possess many distinct features: every entry will consist of structured data (fields of various types and semantics), unlike Wikipedia pages which are basically text documents; the catalog itself will possess internal structure, with every item being assigned to higher-level objects: journals, repositories, collections - unlike Wikipedia, where the corpus is a flat list of articles. Due to these, and other, critical architectural differences, the "wiki" functionality in Paperity Central must be implemented from scratch, and reusing existing code, of, say, MediaWiki (powers Wikipedia), is not a viable solution.

Interoperability

All the data collected by Paperity Central will be made available to external applications through open APIs (Fig. 2). The OAI-PMH interface will be exposed for standardized access to article data. It will be complemented by a custom XML- or JSON-based interface for rich access to all types of information, including those which do not fit into the OAI-PMH architecture (e.g., history of edits).

Importantly, the APIs will allow taking full snapshots of the catalog. This will secure long-term community-based preservation of the data, independently of our own efforts in this regard, and will make the catalog a common good of the entire scholarly community: one that is built by and belongs to the community.

A concrete application that will make use of the APIs and will be developed by ourselves during the project, will be the Paperity-to-repositories **feedback loop**. It will allow institutional repositories (IR) - those serving as data sources for Paperity Central - to pull metadata corrections and extensions collected via "wiki", or imputed by Paperity itself (e.g., through cross-correlating "green" and "gold" metadata). In particular, IRs will be able to pull deduplicated global identifiers of publications, which for the first time will enable interoperability and content matching between different repositories. The IR administrators we have been talking to are very keen to use such a functionality when only it becomes available.

We will develop a reference implementation of the feedback loop in DSpace, the most popular repository platform worldwide which powers over **1,500** academic repositories around the world. DSpace is an open source software project under the stewardship of DuraSpace, a not-for-profit that is a partner in this proposal and that will carry out the implementation. This implementation will serve as a reference for other IRs, who want to develop a similar feature in their own software.

Implementation

Paperity Central will be built by extending an existing prototype, Paperity (paperity.org), the first multidisciplinary aggregator of OA journals and papers (gold/hybrid OA) that indexes articles directly from journal websites and aims at including all gold OA literature. If awarded the Open Science Prize, we will extend Paperity with the remaining core features: aggregation of green OA, the "wiki" and the APIs. We will develop a basic part of each of those functionalities already in Phase 1 of the project and will continue the development in Phase 2. See Table 1 for a schedule.

Table 1.

Paperity Central development schedule.

	Phase 1	Phase 2
Green OA	generic framework for indexing OA repositories implemented;up to 100 DSpace-based IRs included	support for platforms other than DSpace; inclusion of 90% of IRs and subject repositories
Wiki	user registration;basic "wiki" functionality:editing of existing entries, history of edits, rollbacks	all types of edits possible;moderation and user permissions; wiki-text in text values
APIs	custom JSON-based API:basic access to article data	custom API fully developed; OAI-PMH; API for feedback loops; DSpace implementation of the feedback loop (pull mechanism)
Content types	journal articles only	all types of scholarly output included: articles, monographs, conference papers, dissertations, preprints, postprints etc.
Content volume	2 million fulltext items + 5+ million non-fulltext metadata as seed content for wiki-editing	Dec 2017: 10M fulltext items Dec 2018: 30M fulltext items
Features	advanced search; RSS feeds on custom searches	all other features...

Apart from the core functionality, we will develop numerous other features to maximize the utility of Paperity Central to users and the academia:

- advanced multi-criteria search (Phase 1)
- RSS feeds & email alerts on new results for a given search (Phase 1)
- categorization by discipline; custom tagging
- recommendations based on user preferences
- usage metrics per paper/journal/repository; most viewed items
- citation analysis; linking cited/citing works
- user ratings; comments; bookmarking
- community-driven detection of spam entries
- user interface in local languages; mobile application; many others...

In Phase 1, we will focus on one repository platform, DSpace, and perform pilot integration of a number of DSpace-based IRs. The following institutions have already confirmed they want to participate in the pilot:

- Duke University
- Harvard University
- University of Michigan
- Montana State University
- North Carolina State University
- Texas Digital Libraries (20 member IRs)
- University of Cambridge, UK
- Malmö University, SE
- Tomas Bata University, CZ

More will join after the project starts. Other platforms will be added in Phase 2.

Significance

Paperity Central will have a tremendous impact on scholarly communication and research. The catalog will increase discoverability of literature, facilitate communication between authors and readers, help navigate in vast amounts of literature, discover relevant publications and disseminate new results more effectively. It will accelerate progress and enable creation of plethora of add-on services that will solve numerous specific problems of the research community - very likely that even some of the applications submitted here to the Open Science Prize will benefit from our project.

Examples:

- An application that extracts scientific facts from articles could use Paperity Central API to easily browse all literature, find articles on a given specific issue (say, properties of chemicals in cancer treatment) and retrieve full text, for subsequent text mining and extraction of desired facts.
- An application that analyses social structure of academia will use Paperity Central to retrieve links between publications, authors and institutions, to build a map of scientific cooperation and subsequently spot densely connected subgraphs, key centers of collaboration, automatically recommend potential partners etc.
- Anti-plagiarism software will use Paperity Central as the primary source of data about previous works. It will help teachers and lecturers in their work, but also journal editors and reviewers who evaluate manuscripts for publication.

Already now, we receive numerous inquiries from academics who want to build add-on services on top of Paperity and ask for an API to do this. This is a confirmation that Paperity Central is very needed. An example of a recent discussion on Twitter is presented in Fig. 3.

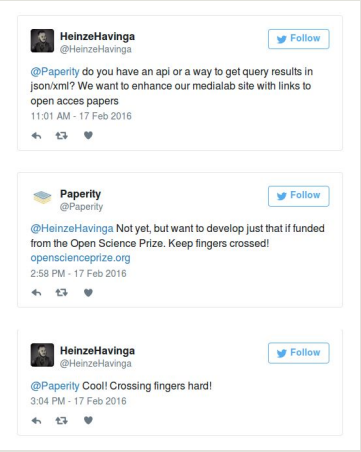

Figure 3.

One of recent discussions on Twitter where a Paperity user asks for an API to build add-on services on top of Open Access literature.

Paperity Central will benefit biomedical research most, as Open Access is most pervasive in life sciences: for instance, bio-med accounts for as much as 80% of the current Paperity content.

Importantly, Paperity Central will finally bridge the gap between "green" and "gold" routes to Open Access. The long-lasting division between these two approaches has frequently undermined the efforts to promote openness in science. By combining literature of both types in one catalog and leveraging distinct advantages of each of them to the benefit of academia, Paperity Central will put an end to this division.

Novelty

No service like Paperity Central exists today. A number of tools are somehow related or try to achieve similar goals:

- subject repositories: PubMed Central, arXiv, RePEc
- aggregators of repositories: SHARE, BASE, OpenAIRE, CORE, OneRepo
- directories: DOAJ
- search engines: Google Scholar

- social networks: Academia.edu, ResearchGate, Mendeley.

However, they are too limited to ever become a central catalog of literature. The two most distinct features of Paperity Central, missing in the above services, are the "wiki" functionality and the aggregation of OA journals.

Viability

Already now, Paperity indexes **30%** of all newly published gold/hybrid OA papers; includes **920,000** articles from **2,500** journals, a 6-fold increase from 150,000 papers upon launch in Oct 2014; receives **10,000** visits/day; partners with **EBSCO** and **Altmetric**.

These achievements show that our team is capable of developing complex software and that building Paperity Central is within our reach. So far, Paperity has not received any external funding, only small support from individual journals. When properly funded, we can build a system many times more complex than that and bring ground-breaking features to the academia.

Paperity is led by Marcin Wojnarski: a programmer and data scientist, winner in the EU Contest for Young Scientists, medalist of the International Mathematical Olympiad. Five years ago, Marcin built another innovative website, TunedIT, which aimed to improve reproducibility in data science, provided tools for data & code sharing and enabled crowdsourcing of machine learning algorithms, in the spirit of Open Science. Marcin has published a dozen of papers and contributed own packages to the open source ecosystem. His track record as a programmer, scientist, team lead and Open Science supporter guarantees that the development of Paperity Central will be completed successfully. The team consists now of 5 members and will be extended when the Prize is awarded.

DSpace is a well-established mature project, under active development since 2002. It has a very experienced management team in DuraSpace and a large community of committers, which guarantee successful completion of the tasks outlined in this proposal.

Currently, Paperity runs on a single commodity server. After extensions, it may require up to 5 servers (in 2018) to handle the growing volume of data and traffic.

Open source

We are committed to the principles of openness and if the Prize is awarded to our project we will release Paperity Central as open source under the **GNU Affero GPL**, and extensions to DSpace under the **BSD** license.

An introductory video presenting our project can be viewed in Fig. 4 and on YouTube:

Figure 4.

An introductory video presenting our Open Science Prize project.

https://www.youtube.com/watch?v=4YnLZguvtHg

Funding program

The Open Science Prize.

Supplementary material

Suppl. material 1: Paperity Central Overview

Authors: Marcin Wojnarski
Data type: animation
Brief description: An introductory video presenting our Open Science Prize project.
Filename: Animation.mp4 - Download file (13.65 MB)