Towards Easy Deposit: Lowering the Barriers of Green Open Access with Data Integration and Automation

Hui Zhang, Oregon State University, hui.zhang@oregonstate.edu

Abstract

This proposal describes the design and development of an interoperable application that supports green open access with long-term sustainability and improved user experience of article deposit. **Introduction:** The lack of library resources and unfriendly repository user interface are two significant barriers that hinder green open access. Tasked to implement the open access mandate, librarians at an American research university developed a comprehensive system called Easy Deposit 2 to automate the support workflow of green open access. **Implementation:** Easy Deposit 2 is a web application that is able to harvest newly publications, outreach for manuscript on behalf of the library, and facilitate self-archiving to IR. It is developed and maintained by the library and integrated with the IR. **Results and Discussion:** The article deposit rate is about 25% with Easy Deposit 2, which increases significantly comparing to the previous period. It also serves as a local database for faculty publications with open access status. The lesson learned is that library cannot rely on a single commercial provider for publication data due to mismatched priorities. **Conclusion:** Recent IT developments provides new opportunities of innovation like Easy Deposit 2 in supporting open access. Academic librarians are vital in promoting "openness" in scholarly communication such as transparency and diversity in the sharing of publication data.

Introduction

We have seen a surge of interest to open access (OA) in recent year because of the combination of several major developments. First, more finding agencies, such as the US National Institutes of Health and the US National Science Foundation, have made OA publishing mandatory for grantees. Moreover, academic libraries are considering OA as a potential solution to keep access for scientific and research literatures.

A widely adopted definition of OA is "makes the research literature free to read online" (Willinsky, 2003), but what is green OA? The scholarly publishing nowadays is complex and researchers have been attempted to classify OA literatures into subtypes based on factors such as the source and license of reuse. In a recent paper (Piwowar et al., 2018), the authors divide OA literature into four categories:

- Gold OA: Published on an open-access journal that is both free to read and re-use.
- Green OA: Published in a subscription based journal, but a free copy is available in a repository. It is also referred as author self-archiving.
- Hybrid: Published in a subscription based journal but is free under some kind of license or after an embargo.
- Bronze: Free to read on publisher website but have no license for re-use.

Green OA, or self-archiving, is usually the last option for authors if they cannot afford the Article Processing Charge to publish in a gold or hybrid OA journals.

However, despite the benefits, recent studies find that only 10-20% of the articles have been self-archived even when publishers have granted the authors right to deposit to a repository (Harnad et al., 2008; Piwowar et.al, 2018). Several barriers can be blamed for the small percentage of green OA, and the most significant one of them is that only a few authors deposit their manuscripts into open access repositories due to inadequate marketing or the effort required for deposit (Björk, 2013). The project described in the article is initiated to increase green OA by lowering the barriers of deposit into institutional repository by authors.

Support for Open Access at OSU Libraries

The faculty members at Oregon State University (OSU), a public research university in the US, passed an open access policy in 2013 that mandates every faculty member to grant OSU a permission to make available his or her scholarly articles and to reproduce and distribute those articles for open access. OSU's Library is in charging of implementing the OA policy by promoting the policy to faculty members and depositing accepted manuscripts (post-peer review, pre-typeset) of their articles to ScholarsArchive@OSU, OSU's institutional repository (ScholarsArchive@OSU) also managed by OSU Libraries (OSU Faculty Senate, 2013).

An early workflow to implement the university's OA mandate requires human intervention at every step. First, OSU librarian creates a search alert in Web of Science that finds all the articles authored or co-authored by OSU faculty, and the results will send to a dedicated OSU email address. A library staff is assigned as the owner of that email account and coordinates the OA workflow. For each article listed in the Web of Science search alert, that library staff will take the following actions:

1. Searches SHERPA/RoMEO to get the corresponding publisher's policies on copyright.
2. Finds out email address of the contact author from the search alert.
3. Contacts the author and asks for archiving the article manuscript into ScholarsArchive@OSU.
4. If the author replies with manuscript in attachment, the library staff will deposit it into IR on behalf of the author.

This manual workflow produces a high deposit rate, defined as the number of deposited articles divided by the total articles authored by OSU faculty, around 43% between 2013 and 2015 (Zhang et.al, 2015). However, OSU Libraries had to abandon this workflow because it can no longer commit a full time staff on OA alone due to limited resources. From faculty feedbacks, we also learned that a main barrier for self-archiving is the time and effort required to deposit articles in IR. For example, faculty complain about the un-friendly repository UI and the number of required metadata fields.

In order to provide a sustainable and long-term support for green OA, several librarians and staffs at OSU Libraries initiated a project that will not only automate many steps of the OA support workflow but also make the article self-archiving as easy as clicking a button. The result

of the project is Easy Deposit 2 (https://ed2.library.oregonstate.edu/), a web application that harvests journal articles, automates OA promotion and outreach, and supports easy deposit into institutional repository. This rest of the article is outlined as:

- Implementing Green OA with Easy Deposit 2: Describe the technical details of automation at each main step of OA support workflow and how to integrate all the steps under Easy Deposit 2.
- Results and Discussion: Compare the numbers of article deposit in different periods (early workflow, with Easy Deposit 2 support, and faculty self-archiving) and review lessons learned from supporting open access with this new approach.
- Conclusion: Introduce future plan and explain why green OA is still valuable with growing of other OA publishing models.

We hope this article will provide its readers an insight of how to take advantages of latest development in digital library for a sustainable support of open access.

Implementing Green OA with Easy Deposit 2

System Design of Easy Deposit 2

We initiate the Easy Deposit 2 (ED2) project as a reboot of library's support for green OA. The project is named ED2 because EasyDeposit has already been used as the name of an open source toolkit (http://easydeposit.swordapp.org/). We design the functions of ED2 from the perspective of how to automate each step of the OA workflow.

- Harvest articles authored by university faculty: using Web of Science API and search by institution name.
- Verify publisher's policies on copyright: we believe that we can skip this step because OSU's OA policy grants the University a nonexclusive license for scholarly articles published by faculty members. We consult the General Counsel of the University and finally get the green-light with the condition that the deposit has to be initiated by faculty members.
- Get article metadata and contact author email: parse the information from Web of Science API and save locally.
- Recruit article manuscript: send out article recruiting email to contact author with a deposit link to ED2.
- Deposit article: the deposit link will direct faculty members to the deposit page at ED2, which pre-populates all the metadata fields. Faculty member will upload article manuscript and submit the article into ScholarsArchive@OSU.

The ED2 system is comprised of three modules to provide the core functions above:

- Fetch module: support article harvesting.
- Parsing and email module: support parsing records of metadata and contact email, saving records locally, and sending out emails for manuscript recruitment.
- Deposit module: support pre-populates metadata records for deposit page, accepting uploaded manuscript, and ingesting the article object into IR.

The diagram in Figure 1 illustrates the functions and modules of ED2 and how ED2 interacts with external entities such as Web of Science, authors, and institutional repository.
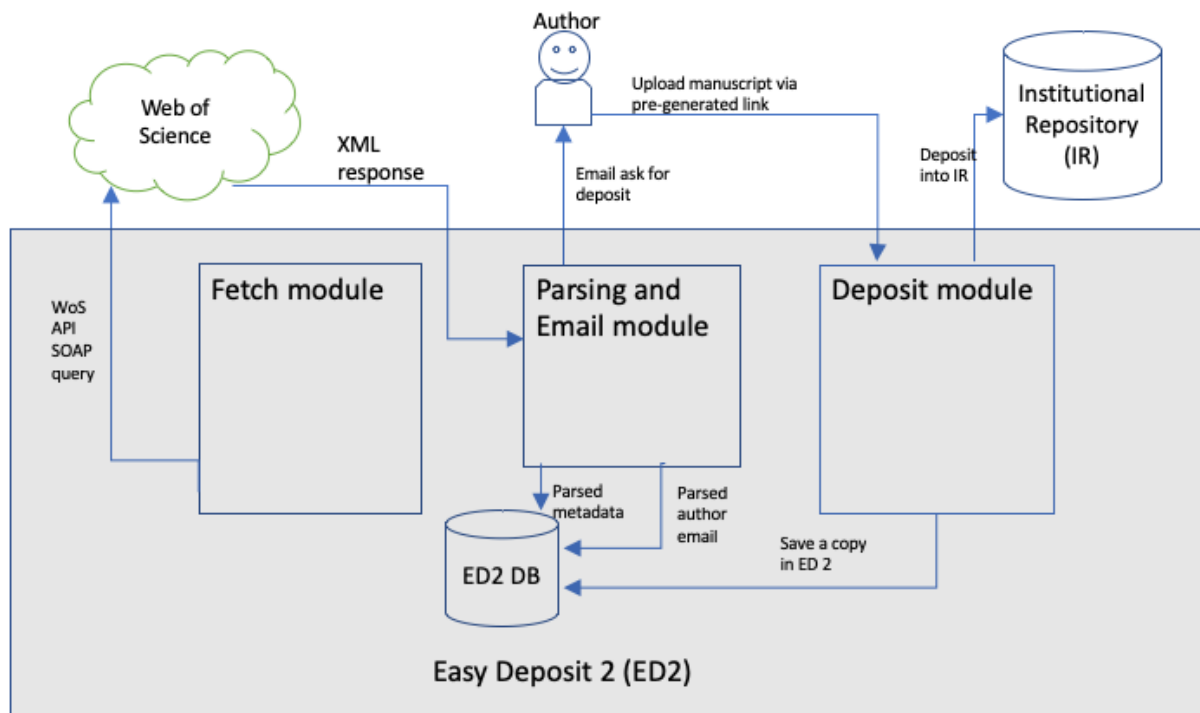

Figure 1. ED2 System Diagram

ED2 is developed using Ruby on Rails (https://rubyonrails.org/) with a MySQL database running as the backend storage. For the purpose of self-archiving into IR, a faculty member only needs to interact with ED2's article deposit page with two actions of uploading file(s) and clicking the "Submit" button. Figure 2 shows the ED2 deposit page for a sample article.

Figure 2. ED2 Sample Deposit Page

Furthermore, faculty members don't need an ED2 account for self-archiving. The deposit link embedded in the article recruit email includes a SHA2 "key" generated by the email of the contact author. When faculty member clicks that link, ED2 will compare the SHA2 key with the author email stored in its database for authentication. ED2 also has a dashboard for system administrators and librarians providing information such as the total number of journal articles have been harvested and how many of them have been self-archived by faculty members. The access to administration dashboard is restricted to OSU Libraries librarians and staffs.

OA Workflow with Easy Deposit 2
ED2 uses Web of Science Extended API (https://clarivate.com/webofsciencegroup/solutions/xml-and-apis/) to retrieve journal articles published by faculty members in the last four weeks from Web of Science Core Collections. The Extended API returns found articles in XML with metadata fields such as title, author, abstract, and contact email of corresponding author. ED2 will parse and save the metadata for each article in the local database together with the "author publication pair", which is a data structure storing the article identifier and author email. The primary purpose for creating the author publication pair is to retrieving the correct metadata record while article depositing. If an article has several corresponding authors, ED2 will generate an author publication pair for every author. However, starting April 2019, email of corresponding author is no longer available from Extended API. This surprise change devastates the ED2 workflow and we have to act quickly and look for a different approach. The current solution is to use the author field and search OSU's directory to find match name(s), and retrieve the email for the matching person. We will provide more details on the impact of commercial API in the Results and Discuss.

ED2 automates the workflow of self-archiving into institutional repository. For each new article that ED2 can find author email, ED2 will contact the author asking for depositing a copy the article manuscript into ScholarsArchive@OSU. A pre-generated deposit link is included in the email, and author will be directed to ED2's article deposit page after clicking the link. Because metadata fields have been pre-populated, the author only need to upload the manuscript and commit the submission to IR. The uploaded file and metadata will be packed and ingested into ScholarsArchive@OSU over HTTP. ScholarsArchive@OSU is built on the Hyrax framework, a front-end based on the Samvera software (https://github.com/samvera/hyrax). Hydra support ingesting over HTTP with content wrapped in JSON format, and ED2 takes advantage of this Hydra feature.

Results and Discussion

Results

We use and report deposit rate, defined as the number of articles deposited into ScholarsArchive@OSU divided by the total number of articles published by OSU faulty members, as the metric of green OA. The deposit rates are calculated over four phases:

- Pre-WoS: The period before OSU Libraries implement the manual workflow using the Web of Science index (WoS).
- WoS manual: The period when OSU Libraries are able to commit one fulltime staff to recruit and deposit manuscripts on behalf on faculty.
- In between: The period in between the launch of ED2 and the stop of manual OA workflow.
- ED2 OA: The period when OSU Libraries are using ED2 to support OA.

	Calendar Year	Total Articles	Article Deposited	Deposit Rate
Pre-WoS	2011	1615	197	**12.20%**
WoS Manual	2012	1779	799	44.91%
	2013	1682	747	44.41%
	2014	1731	742	42.87%
WoS Manual Total		5192	2288	**44.07%**
In Between	2015	2113	325	15.38%
	2016	2334	152	6.51%
	2017	2450	105	4.29%
	2018-01 to 2018-09	1885	68	3.60%
In Between Total		8782	650	**7.40%**
ED2 OA	2018-10 to 2019-03	1301	333	**25.60%**

Table 1. The Trend of Article Deposit Rate with Type of OA Support

We get the numbers in the spreadsheet above by search the underlying database with appropriate metadata fields. For the numbers of "Article Deposited", we search ScholarsArchive@OSU with resource type as "Article" then separated the results by year with values in the "Date Created" field. For the numbers of "Total Articles", we first submit an address search to WoS with the query of "AD='Oregon State University'" then use the "Year" facet to estimate the total number of articles published by the faculty.

Discussion

The Effect of a Comprehensive Green OA System

We are satisfied with the results of ED2. Six months after the launching of ED2, article deposit rate increased significantly, more than tripled, to 25.60%. Our data shows that without outreach and matching repository service, few authors will bother to deposit manuscripts into IR even there is a University mandate. However, due to limited resource, most academic libraries in the US cannot afford to have a fulltime staff dedicated on supporting self-archiving. The ED2 system is developed as an extension of OSU Libraries' repository service and integrated with the university IR. There are existing efforts to harvest publications using Web of Science API. One example is the Bibliographic Management System developed by Stanford University Library (https://github.com/sul-dlss/sul_pub) and it also inspires us for initiating the ED2 project. ED2 is an innovation as the first comprehensive system specially developed to support green OA. It is the article deposit portal of ScholarsArchive@OSU, the software agent that automates the workflow of green OA, and the database of faculty publications with their OA status. A librarian can use ED2's dashboard to answer important questions such as how many articles faculty members published last year and how many of them have been self-archived in the IR.

The Availability Issue of Publication Data

Access to metadata records of scholarly publications has become the bottleneck of running ED2. Author email is vital for ED2 because it is required for outreach, recruiting manuscript, and authentication for deposit. ED2 was able to pull out corresponding author email(s) from Web of Science Extended API until April 2019, when the company underlying Extended API decided to exclude author email(s) from the output. We develop an alternative solution to obtain author emails by looking for author names in Oregon State University's staff directory. However, for all the articles harvested by ED2, we can find a matching faculty for less than 50% of them. The lesson learned is that we cannot rely on a single and commercial source for data because the library has little influence over the vendor's decision and the priorities of both parties are not aligned in many situations. We have tested APIs of Elsevier, Crossref, and ORCID, for whether they can replace the API from Web of Science. Our results show that none of the API include author contact and the availability of metadata such as abstract and copyright in the API output is inconsistent. A long-term solution for the problem of data accessibility maybe that non-for-profit organizations, such as Crossref and ORCID, will collaborate and provide critical information such as copyright and contact through their APIs.

Conclusion

Recent technology developments in digital repository and web services provide new opportunities for innovation in supporting open access publishing. Easy Deposit 2, or ED2, is a comprehensive system developed to prompt open access and facilitate author self-archiving. ED2 is designed to automate major OA tasks such as harvesting faculty published articles, outreach for manuscripts, and article depositing into IR. ED2 significantly increases the article deposit rate to about 25%, which meets our expectation. We will improve ED2 with features such as sending out a second email to remind authors who have not achieved their manuscripts.

The first priority for future development of ED2 will be diversifying the data sources with preference for community-based or non-for-profit organizations. Proposed methods include harvesting article metadata from Crossref and looking for author contact in ORCID. Breaking the monopoly of commercial indices on publication data is critical for long-term sustainability of supporting open access and promoting of "openness" in scholarly communication. Academic librarians have vital roles in more open and transparent sharing of publication data by collaboration with scientific communities and non-for-profit publication data aggregators.

References

Björk, B.-C. (2013). Open Access—Are the Barriers to Change Receding? Publications, 1(1), 5–15. https://doi.org/10.3390/publications1010005

Oregon State University Open Access Policy | Libraries | Oregon State University. (n.d.). Retrieved February 11, 2020, from https://library.oregonstate.edu/open-access

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375