Paper ID #21224

Leveraging Python to Improve Quality of Metadata of Engineering Faculty Publication Records

Ms. Qianjin Zhang, University of Iowa

Qianjin (Marina) Zhang is Engineering & Informatics Librarian at the Lichtenberger Engineering Library. As a subject librarian, her work focuses on instruction, reference, consultation services and collection management for the engineering faculty and students. She's also interested in research data management and support Research Data Services. She holds a MA in Information Resources & Library Science from The University of Arizona (Tucson, AZ), and a BS in Biotechnology from Jiangsu University of Science and Technology (Zhenjiang, China).

Leveraging Python to Improve Quality of Metadata of Engineering Faculty Publication Records

Abstract

The Engineering Library at the University of Iowa conducted a project which consisted of reviewing metadata of engineering faculty publications in the Academic and Professional Records (APR), which is a locally branded faculty profile system. The challenge of the project was that there are thousands of records with erroneous or missing metadata, making it difficult to manually check Digital Object Identifier (DOI) and ISSN. Our strategy was to analyze the complete dataset, break it down into subsets with some common patterns and then focus on those subsets. The processes were conducted using Python. As a result, we prioritized records that have almost complete metadata but missing DOI and/or ISSN, retrieved DOI from PubMed and CrossRef online queries separately and added ISSN by matching journal titles or conference names with authorities. The implementation of Python can not only make the review process effective and efficient but also expand library services to the APR project.

Background

Faculty profile systems that capture and showcase faculty scholarly activities and accomplishments are emerging in many institutions. The platforms of faculty profile systems include commercial platforms such as Activity Insight's Digital Measures, Elsevier's Pure and Symplectic Elements and open-source platforms such as Profiles and VIVO [1]. During the trend, some university libraries have become actively involved in the implementation of faculty profile systems and expanded their roles in university leadership and stakeholders. For example, librarians from Duke University, Emory University and Georgia Institute of Technology recently reported use cases of implementation of Symplectic Elements at their home institutions and highlighted libraries' significant roles in the system adoption [1].

Like many other institutions, the University of Iowa has started migrating faculty information to Activity Insight's Digital Measures, locally branded as Academic and Professional Records (APR). The APR project is a collaborative initiative of the Office of the Provost, Information Technology Services and the University colleges to capture faculty information on teaching, research, grants, service, as well as records on professional accomplishments and interests. Since a record of publication would make a strong case for faculty excellence in scholarship especially for promotion and tenure, accuracy of publication records is significantly important. Since early 2017, the College of Engineering has steadily migrated their faculty data to the APR. The College of Engineering has 95 faculty members in five departments, including Biomedical Engineering, Chemical and Biochemical Engineering, Civil and Environmental Engineering, Electrical and Computer Engineering, and Industrial and Mechanical Engineering. Upon request by the College of Engineering and the APR project leader, we were to review engineering faculty publication records shown in Figure 1 [2] to improve the quality of metadata, especially focusing on Digital Object Identifier (DOI), ISSN, PubMed ID (PMID) and PubMed Central ID (PMCID).

Based on a rough evaluation of metadata quality, we realized that thousands of yet-to-be-identified records with erroneous and missing metadata would make a routine manual review time-consuming and costly. However, some libraries have implemented Python scripts in managing metadata for library resources. For example, the University of Minnesota Libraries used Python scripts to evaluate MARC record completeness for electronic books [3] and librarians at the University of Virginia utilized Python to perform quality control on MODS records for digital collections [4]. Both examples indicate that Python would increase efficiency in quality control of metadata. In consideration of the challenges we are facing, scripting with Python would be an appropriate approach over the manual approach.

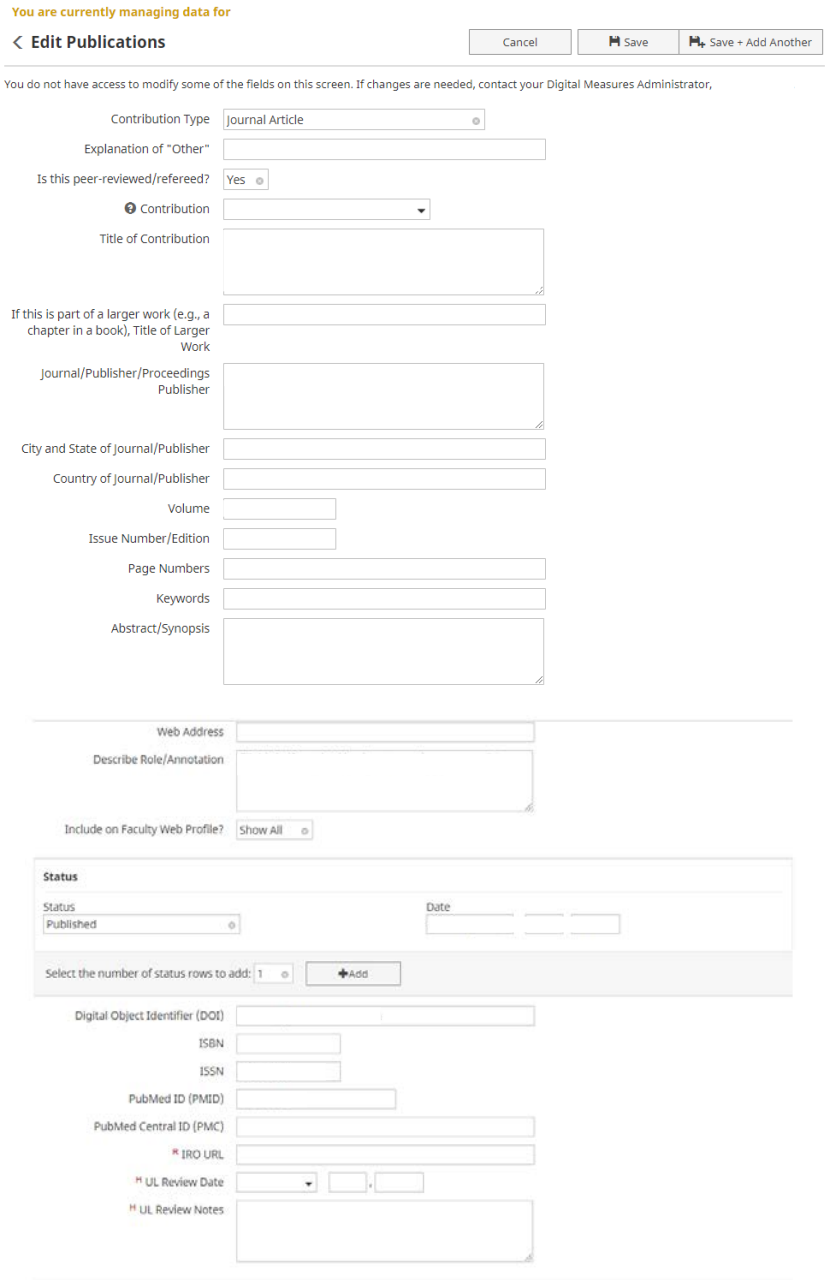

Figure 1: Faculty Publication Record User Interface in the APR

Workflows

Given that faculty publication records are imported from different sources such as PubMed, Scopus and manual input, our strategy was to evaluate record completeness, break down the large set of records into several subsets of records with some common patterns and then manipulate subsets, especially with a focus on identifying DOI, ISSN, PMID and PMCID.

First, we extracted the engineering faculty publication records in a csv file from the APR and briefly evaluated record completeness through checking for the presence or absence of DOI. We also sorted records by checking for the presence or absence of PMID or PMCID because PubMed records have PMID or PMCID, while Scopus records and manual input records have neither of them. Since a csv file is a two-dimensional spreadsheet-like object in which columns are considered as metadata elements and rows are as individual records, we analyzed data in the csv file using pandas, an open-source Python package for data structures and data analysis. All scripts were written and executed on Jupyter Notebook, an open-source web application for data cleaning, transformation, numerical simulation, statistical modeling, data visualization and machine learning [5].

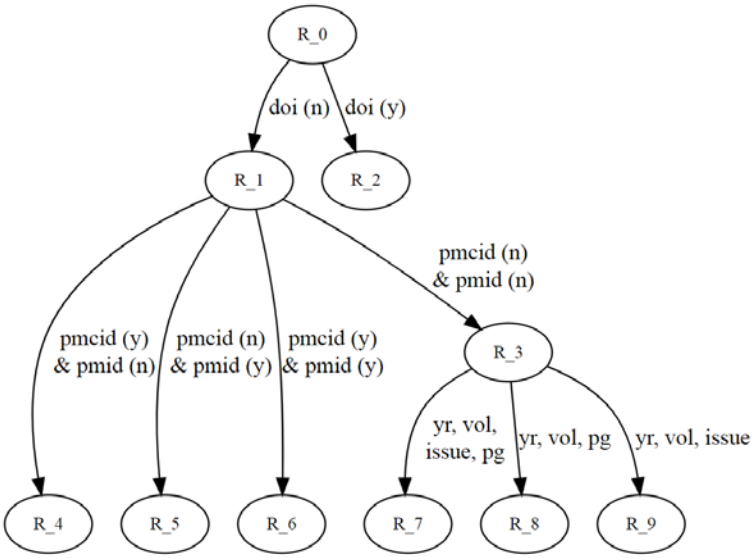

Figure 2: The Workflow of Batch-Processing Records

Figure 2 shows the workflow of batch-processing faculty publication records. The metadata elements involved in the process include record ID, title, journal title/conference name, volume, issue, page number, year of publication, DOI, ISSN, PMID and PMCID. All records in R_0 were checked for the presence or absence of DOI and sorted into two subsets R_1 and R_2 where R_1 represents records with no DOI and R_2 represents records with DOI. Next, R_1 was checked for the presence or absence of PMID and PMCID and classified into four subsets R_3, R_4, R_5 and R_6. R_3 represents records with neither PMID nor PMCID, R_4 represents records with PMCID but no PMID, R_5 represents records with PMID but no PMCID and R_6 represents records with both PMID and PMCID. For records in R_4, R_5 and R_6 that contain PMID

and/or PMCID, we retrieved DOIs using PubMed's online DOI query tool called PMCID – PMID – Manuscript ID – DOI Converter shown in Figure 3 [6].

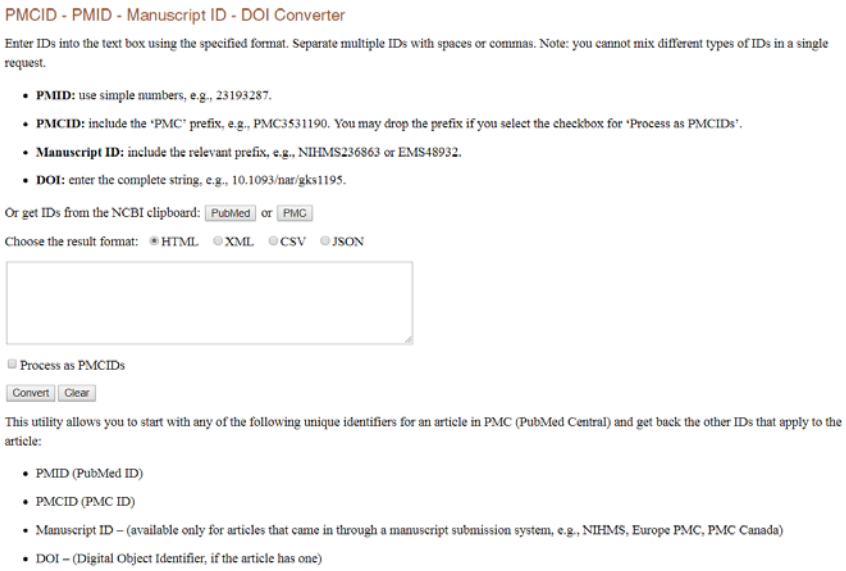

Figure 3: PMCID – PMID – Manuscript ID – DOI Converter

For records in R_3 that do not contain PMID or PMCID, we formulated references and retrieved DOIs using the CrossRef Simple Text Query shown in Figure 4 [7]. CrossRef Simple Text Query can take up to about 80 records since there are word limits.

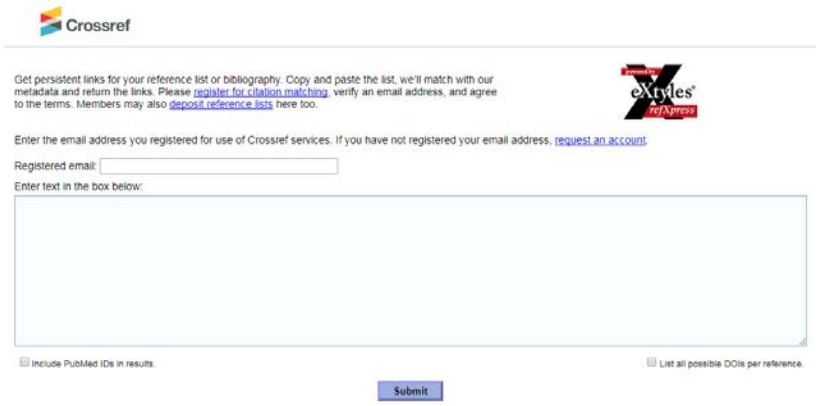

Figure 4: CrossRef Simple Text Query

Since accuracy of a reference is critical to retrieval success, completeness of metadata fields including title, journal title/conference name, year of publication, volume number, issue number and page number were checked. Records with almost complete metadata fields were extracted and formulated to references in the following formats.

- Records in R_7 that have title, journal title/conference name, year of publication, volume number, issue number and page number are in the format of:
  Last Name, First Name (Year of Publication) Title. Journal title/Conference Name Volume# (Issue#): Start Page# - End Page#

- Records in R_8 that have title, journal title/conference name, year of publication, volume number, page number but have missing issue number are in the format of:
  Last Name, First Name (Year of Publication) Title. Journal Title/Conference Name Volume#: Start Page# - End Page#

- Records in R_9 that have title, journal title/conference name, year of publication, volume number, issue number but have missing page number are in the format of:
  Last Name, First Name (Year of Publication) Title. Journal Title/Conference Name Volume# (Issue#)

A final step involved looking up ISSNs and cleaning up data. Titles, title abbreviations and ISSNs were extracted from PubMed [8] and Scopus [9] and then formatted into a csv authority file shown in Figure 5. The authority file included title abbreviations because some records might have used title abbreviations other than full titles. For records that do not have an ISSN, we filled in the empty ISSN column with journal titles or journal abbreviations. If journal titles or title abbreviations were found as an exact match against the authority file, then they were substituted with an ISSN.

JournalTitle	Abbr	ISSN
Annals of work exposures and health	Ann Work Expo Health	2398-7308
Asian journal of anesthesiology	Asian J Anesthesiol	2468-824X
Birth defects research	Birth Defects Res	2472-1727
Complementary medicine research	Complement Med Res	2504-2092
Journal of gynecology obstetrics and human reproc	J Gynecol Obstet Hum Reprod	2468-7847
Journal of stomatology, oral and maxillofacial surg	J Stomatol Oral Maxillofac Surg	2468-7855
Musculoskeletal science & practice	Musculoskelet Sci Pract	2468-7812
Polish archives of internal medicine	Pol Arch Intern Med	0032-3772
SLAS discovery	SLAS Discov	2472-5552
SLAS technology	SLAS Technol	2472-6303

Figure 5: Title, Title Abbreviation and ISSN Authority File

After data clean-up, identified DOIs and ISSNs were provided with Record IDs and Username in a csv file to be uploaded into the APR.

Results and Limitations

Figure 6 shows the exact number of records that we processed in the workflow. We found that 14,506 records in R_1 of the whole records do not have DOIs while 4,513 records in R_2 have DOIs. Among the records with no DOIs, only a small portion of records have PMID and/or PMCID. In other words, the records in R_4, R_5 and R_6 are from PubMed while the records in R_3 are from Scopus or manual input. With regard to the records in R_4, R_5 and R_6, we identified 1293 DOIs and 1197 ISSNs. For the records in R_3, we identified 1445 DOIs and 1698 ISSNs. We also identified 2485 ISSNs for the records in R_2. As a result, we successfully identified 2,738 DOIs and 5,380 ISSNs for records that have missing DOI or ISSN. However, this approach could not handle the remaining 7,831 records in R_10 partially due to incompleteness of metadata. We provided the results and discussed further about the project with the College of Engineering and the APR project leader.

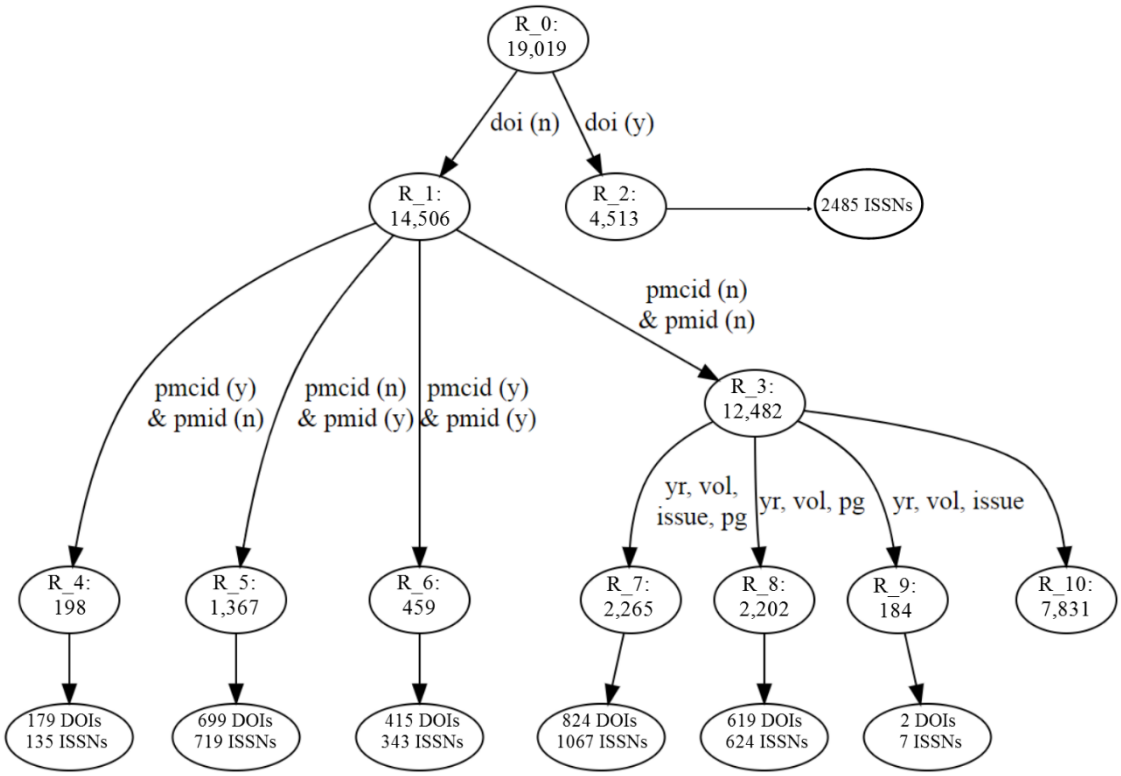

Figure 6: The Workflow of Processing the Number of Records

Conclusions

We found the implementation of Python in engineering faculty publication records review process improves the effectiveness and efficiency of the review process, saves our library staff's time and makes a contribution to the College of Engineering's APR migration as the University of Iowa Libraries is increasingly involved in this campus-wide initiative.

References

1. Givens, M., L.A. Macklin, and P. Mangiafico, *Faculty Profile Systems: New Services and Roles for Libraries*. Portal-Libraries and the Academy, 2017. **17**(2): p. 235-255. Available from https://muse.jhu.edu/article/653202

2. Andrews S. 2017. APR Publications Training [PowerPoint]. [cited 2017 Feb 1].

3. Thompson, K.T., and S. Traill. *Leveraging Python to improve ebook metadata selection, ingest, and management*. Code4Lib Journal, 2017 (38). Available from: http://journal.code4lib.org/articles/12182

4. Bartczak, J., and I. Glendon. *Python, Google sheets, and the thesaurus for graphic materials for efficient metadata project workflows*. Code4Lib Journal, 2017 (35). Available from: http://journal.code4lib.org/articles/12182

5. Jupyter Notebook [Internet]. [cited 2017 Feb 1]. Available from: http://jupyter.org/index.html

6. Finding Article Identifiers [Internet]. [cited 2017 Feb 1]. Available from: https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/

7. CrossRef Simple Text Query [Internet]. [cited 2017 Feb 1]. Available from: https://apps.crossref.org/SimpleTextQuery

8. PubMed Help Journal Lists [Internet]. [cited 2017 Feb 1]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.journal_lists/

9. Scopus Source List [Internet]. [cited 2017 Feb 1]. Available from: https://www.scopus.com/sources.uri?zone=TopNavBar&origin=sbrowse