Upgrading CERN's Library Notifications Service

Jai Dilakshan Mohan, University College London
Supervisor: Salomé Rohr, CERN

Abstract

CERN's library serves its scientists in their research endeavours, hence it is crucial that they have access to the latest publications. This project involved improving CERN's library notification service to allows librarians to be notified of the latest releases from relevant publishers. Many publishers have existing notification channels consisting of email alerts and RSS feeds. However, these did not offer the desired customisation, so the ideal solution was to scrape publishers' websites and fetch the title lists in the format of KBART files. The web scraper was built as a Python command line interface for usability. The program fetches the latest releases successfully for the following publishers: Cambridge University Press, Springer, and Taylor and Francis. During development, Oxford University Press and World Scientific began blocking scraping attempts, so a working solution could not be made for these publishers. For the future, it would be ideal to liaise with publishers to gain permission for scraping, so that a more comprehensive notification service can be developed.

Geneva, Switzerland
September 15, 2023

Contents

1 Introduction

CERN's Scientific Information Service (SIS) has a focus on "efficiently managing, preserving and disseminating scientific information to make it openly accessible" to the CERN community [1]. With the newly renovated library opening on 28 September 2023 [2], it is especially important for the library to have access to the latest publications. This project involved upgrading CERN's existing library notifications service, so that the library will have access to the latest book releases from publishers. The library develops its collection through patron-driven acquisition, in which the CERN community requests books for the library to purchase. Finding the latest releases early will allow the scientific community at CERN to have a list of items to consider. In the scientific world, it is crucial that knowledge is easily accessible, so this project will aid scientists at CERN in their research endeavours.

2 Existing Notification Channels

The first course of action in this project was to find and analyse the notification channels that already exist for various publishers. As seen in Table 1, all publishers offered email alerts. However, there are a few caveats to this method of notification. Firstly, not many publishers offer the freedom to select what notification content is required. This will result in irrelevant content appearing in emails. The format of email content is also not easy manipulable, hence requiring more human input. Furthermore, the frequency of emails is decided by the publisher, rather than by CERN's SIS.

Some publishers offer their own APIs. Although it is of more utility than email alerts, learning how to use these APIs and maintaining them would require considerable time and effort. Additionally, RSS feeds are offered by publishers, however, for the majority of cases, this is only available for journal articles - not books.

KBART files, ("Knowledge Bases and Related Tools,"), are a popular format in which publishers release title lists [3]. This format is both human- and machine-readable, hence making it s uitable to search and identify books. Finally, there exist MARC (machine-readable cataloging) records. These are more difficult to extract data from than KBART, as they are not human-readable [4]. However, using PyMARC [5], it is possible to convert the data into a format similar to KBART files.

Publisher	Email Alerts	Title Lists	API	RSS
Cambridge University Press	Yes	KBART	Yes	No
Elsevier	Yes	KBART	Yes	Journals Only
Oxford University Press	Yes	KBART	No	Yes
Princeton University Press	Yes	Catalog	No	No
Springer	Yes	KBART	Yes	Journals Only
Wiley	Yes	KBART	Yes	Journals Only
World Scientific	Yes	MARC	No	No

Table 1: List of publishers and their respective notification channels.

3 Developing a Web Scraper

It was decided that the best way to create a notification system was to scrape the publishers' websites and download their latest releases. Web scraping is the act of extracting data from websites, and has its advantages and disadvantages. One one hand, it affords great customisation and is cost effective, however, it can also be slow and require regular maintenance.

3.1 Specification

The program would be built on Python and can be interacted with using a command line interface (CLI). Python is used for its simple syntax and maintainability, and a CLI for its ease of use - even for those without programming experience. The program retrieves subjects according to CERN's library acquisition policy. These main subjects that are focused on include: mathematics, physics, computer science, chemistry, statistics, and engineering.

3.2 Development

The primary method of scraping is using Python's BeautifulSoup library [6]. Web pages are written in HTML, and BeautifulSoup is able to extract data from these HTML files. For instance, to download the latest physics book releases from Cambridge University Press, the program searches for the phrase "physics", and downloads the KBART file from the corresponding URL, as shown in Figure 2. However, in some cases, web pages are also written in JavaScript, and BeautifulSoup cannot simply parse their HTML file - downloading a file may involve clicking a button. In this case, Selenium is used. Selenium acts as a headless browser (i.e. a browser without a graphical interface) and simulates a button click, which then leads to the download of a file [7].

```
Enter the publisher (cambridge, springer, taylor): cambridge
Enter the subject (computer science, engineering, mathematics, physics, science
and engineering, statistics): physics
Enter the directory: /Users/jaimohan/Downloads
Successfully downloaded physics.tsv
```

Figure 1: Command-line interface to fetch the latest releases

For Oxford University Press, the process involved additional steps, as the website does not offer the option to download specific subjects. Therefore, the main file had to be filtered and separated into subject-specific files. As World Scientific publishes their title lists in the form of MARC records, it was the publisher that required the most programming. Using PyMARC, it was possible to read batches of records, retrieve important information such as title and authors, and write this information into a file with a format identical that in Figure 2.

3.3 Challenges

The project underwent numerous challenges as detailed below:

- Since the beginning of the project, Wiley and Elsevier had prohibited web scraping.
- Princeton University Press's catalogue format was unsuitable for converting to a KBART format.
- At the tail end of the project, Oxford University Press and World Scientific had blocked the attempted requests for scraping. Therefore, their development had to be abandoned mid-project.
- It was not possible retrieve the KBART file from Taylor and Francis with BeautifulSoup. Selenium had to be used although it is a much slower solution.

4 Conclusions

In this project, CERN's existing library notification system was updated with an addition that scrapes publisher web pages to retrieve title lists. This was successful for Cambridge University Press, Springer, and Taylor and Francis. By communicating with the publishers, it may be possible to extend the solution to also work for Oxford University Press and World Scientific.

To improve on the current solution, it would be beneficial for the CLI to have an option to "download all" to retrieve title lists for all subjects for all publishers. This would drastically reduce the time required to find the latest book releases.

	A	B	C	J	K	L	M
1	publication_title	print_identifier	online_identifier	title_url	first_author	title_id	embargo_info
2	A Complex System	978-1-009258-08-1	978-1-108582-28-5	https://doi.org/10.	1017/97811085822	6CBFD8C42E0D74ED879A3D6C18836B	
3	An Introduction to	978-1-009403-29-0	978-1-009403-26-9	https://doi.org/10.	Collins	31352554E4B57429C185BF922680753	
4	An Introduction to	978-1-009401-72-2	978-1-009401-68-5	https://doi.org/10.	Cottingham	61AAC3989C199E9ADE017D899F371B	
5	Case Studies in Sta	978-1-009277-44-0	978-1-009277-43-3	https://doi.org/10.	MacKay	7F55EC5EC89D04FD25AEDDE87D3CFB	
6	Compact Matrix Qu	978-1-009345-73-6	978-1-009345-70-5	https://doi.org/10.	Freslon	9BDEDEC2A7496875F876FEF9AAF300	
7	Computational Des	978-1-108494-10-6	978-1-108643-76-4	https://doi.org/10.	Du	D52AE10C55354684F440DCD30400D59	
8	Condensed Matter	978-1-108494-60-1	978-1-108781-24-4	https://doi.org/10.	Altland	B110A3C3209A12F050664EC59F0D06F	
9	D-Branes	978-1-009401-36-4	978-1-009401-37-1	https://doi.org/10.	Johnson	07334267BC776E31D451BD43555A02I	
10	Data Modeling for	978-1-009098-50-2	978-1-009089-55-5	https://doi.org/10.	Presse	61C0E9959A38442F715287A0E2057BA	
11	Data Science for Co	978-1-108844-79-6	978-1-108953-59-7	https://doi.org/10.	Chakrabarti	304F66053C62CD439FDFA46D2D4323A	
12	Dynamics of Charg	978-1-009402-27-9	978-1-009402-28-6	https://doi.org/10.	Spohn	3045ECB81EF4C2614CD8580CB68BF3E	
13	Einstein's General	978-1-009263-70-2	978-1-009263-68-9	https://doi.org/10.	Dolan	92487D677E7D1BF45CC259CF9A86F13	
14	Electroweak Theor	978-1-009402-38-5	978-1-009402-37-8	https://doi.org/10.	Paschos	89127BFB5372ED4AA87704E0FA2D93	
15	Field Theory of Noi	978-1-108488-25-9	978-1-108769-26-6	https://doi.org/10.	Kamenev	E86E9D0A1921BA06F964DA83428F7C	
16	Finite-Temperature	978-1-009401-95-1	978-1-009401-96-8	https://doi.org/10.	Kapusta	799D4793749D3BFDEA193F97EE54339	
17	Foundations of Per	978-1-009401-83-8	978-1-009401-84-5	https://doi.org/10.	Collins	9BD2EAA4F2DD31119B10FAECDC8ED2	
18	Fundamentals of P	978-1-009171-58-8	978-1-009171-59-5	https://doi.org/10.	Paganini	FEA0FB0A4B2B158DFBF70DF9014880	
19	Gauge Symmetries	978-1-009197-22-9	978-1-009197-23-6	https://doi.org/10.	Berghofer	F309AF5640E8858C871925F70886126	
20	Gauge/String Dual	978-1-009403-49-8	978-1-009403-50-4	https://doi.org/10.	Casalderrey-Solana	37CD128195AF1DE2D806C457818181C	

Figure 2: Fetched KBART files in a .csv format for readability

Acknowledgments

I would like to express my gratitude to Salomé Rohr and Harris Tzovanakis for guiding me through this project with their expertise in their respective fields. I want to also thank Ernesta Petraityte, Lydia Pieper and Benjamin Bergia for their useful insights. Finally, I am grateful to the whole CERN community for being so welcoming, and setting up the incredible Summer Student Programme.

References

[1] CERN Scientific Information Service (2023). Mission https://scientific-info.cern/about-us/mission

[2] CERN Library (2023). The CERN Library will re-open soon! To celebrate, take part in a series of fun activities on 28 and 29 September 2023https://home.cern/news/announcement/cern/cern-library-will-re-open-soon-celebrate-take-part-series-fun-activities-28

[3] Peter McCracken, Kristina Womack (2010) KBART: Improving Access to Electronic Resources through Better Linking, The Serials Librarian, 58:1-4, 232-239, DOI: 10.1080/03615261003625927

[4] Byrne, D. J. (1998). MARC manual: Understanding and using MARC records. Bloomsbury Publishing USA.

[5] Pymarc (2023) https://pymarc.readthedocs.io/en/latest/

[6] Leonard Richardson (2023) Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[7] Baiju Muthukadan (2018) Selenium with Python. https://selenium-python.readthedocs.io/index.html