IFLA WLIC 2017 WROCŁAW

Submitted on: 26.05.2017

Data Curation with Autonomous Data Collection: A Study on Research Guides at Korea University Library

Young Ki Kim
Korea University Library, Korea University, Seoul, Republic of Korea

Ji-Ann Yang
Korea University Library, Korea University, Seoul, Republic of Korea

Jong Min Cho
Korea University Library, Korea University, Seoul, Republic of Korea

Seongcheol Kim
Korea University Library, Korea University, Seoul, Republic of Korea

Introduction

As the format and medium of research information diversifies with the development of scholarly communication channels, research information is generated and distributed more frequently via manifold media. While traditional library services focused primarily on offline information, there is an increasing demand from the research community to create more organized digital channels for accessing information. Libraries must comprehensively collect information from various sources, including online media, and selectively curate relevant research information from amidst a mass of data.

Some libraries have provided online subject guides to deal with online media and to provide subject-specific research information. These guides are gateways and tools for resource discovery, with well-organized categorization and classification of various sources of information (Bawden & Robinson, 2002). However, many of them are mere collections of hyperlinks to other web pages, and lack in-depth and detailed information or content. Others provide more detailed and organized information, but they seldom update their information because the constant collection and classification of information requires a huge amount of time and resources.

One of the major challenges to effectively providing relevant research information lies in collecting and selecting information in a sustainable manner. Another challenge is how to present collected information in a way that maximizes its accessibility.

In this paper, we propose and describe a research guide system that automatically collects information from various sources and distributes it via subject-specific online research guides, thereby enabling efficient and continuous provision of up-to-date information.

System developed

Our aim is to develop a system that automates the procedures for curating research data so that users may enjoy sustainable and easily digestible distribution of research information. To accomplish this, we have developed a research guide system that provides data collection, data management, data edit, data transport, data presentation, and data monitoring features.

We implemented a feature that automatically collects data from various information sources on a daily basis. We combined the Rich Site Summary (RSS) and web crawling techniques to develop this feature. RSS is a family of standard web formats to publish frequently updated information (Jones & Neubert, 2017) and eliminates the need for the user to manually check websites for new content. Users use RSS aggregators to collect and monitor feeds in one centralized program or location (Dey & Sarkar, 2009). We used an RSS aggregator as one building block for our implementation. This RSS aggregator retrieves information provided from various information sources in RSS format and relays it to our data presentation feature, or web service. However, using only RSS, we would not be able to get information from sources that do not support RSS. Also, because RSS takes information from a provider, it only allows access to information that providers serve via RSS, while there could be other valuable information on providers' web pages or other publishing formats. Moreover, how the content of RSS is arranged and organized varies depending on the providers' implementation. To address this, we used web crawlers to supplement RSS gathering. A web crawler is a computer program that inspects the web in a methodical manner and retrieves targeted documents (Faheem, 2012). While crawlers can extract information from web pages, which usually contain more information than the brief subsets of information available on RSS feeds, it takes more effort to build them. Each crawler needs to be set up specifically for each website because each website has different internal structures. Thus, we maximized the use of RSS and implemented web crawlers for information sources that do not support RSS or provide RSS services but have limitations in our use as stated above. Both RSS and crawled information is re-formatted into a uniform format so that it is accessible via a single interface.

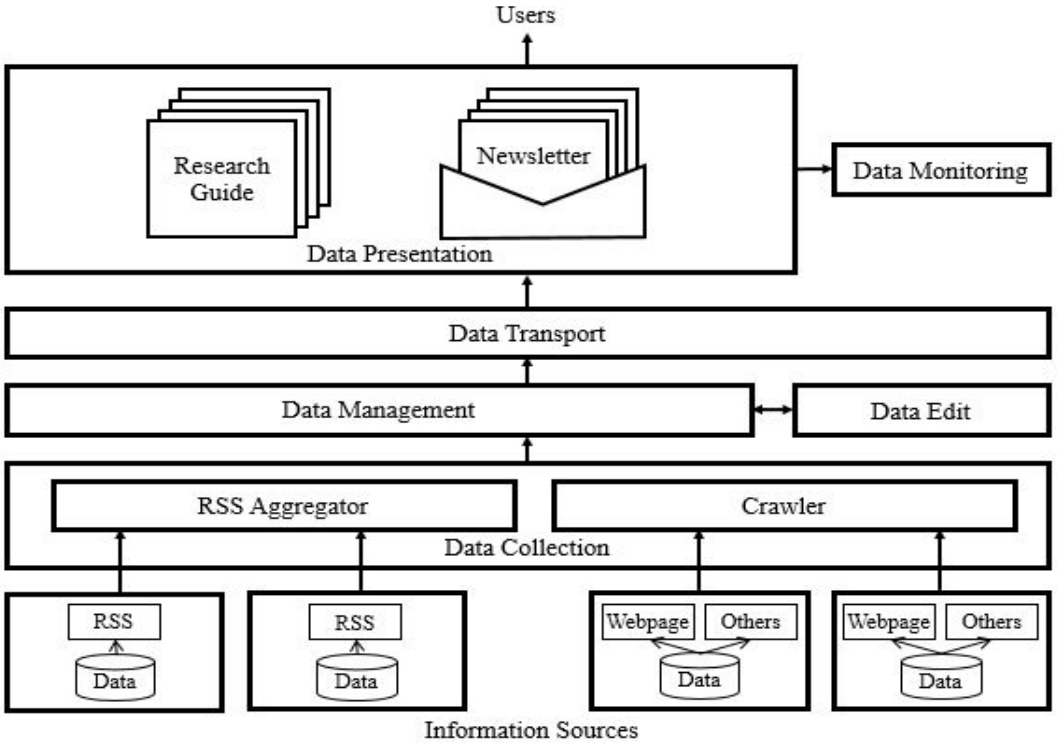
Our system's data edit feature provides a way for librarians to add, edit, remove, or confirm individual pieces of information. While the data collection feature automates the process of gathering information, it does not fully guarantee that the gathered information is adequate. Even if we can provide some rules to classify information, such as including or excluding specific keywords, it is difficult to completely categorize every piece of information. For example, a crawler that is designed to extract books of a DDC code with specific keywords can still include books for beginners, which are not appropriate for scholarly research. Or, librarians might want to add supplementary information that is not provided by the information sources that the data was extracted from. The data edit feature presents the collected information to librarians with edit interfaces and re-constructs the data with the

librarian's input. Its implementation can vary depending on the system environment as long as it provides access to the information collected by the data collection feature.

The information then needs to be transported to user services by the data transport feature. It would be time-consuming if librarians had to type in each piece of information onto user services. The data transport feature resolves this issue by relaying the collected data to actual services in real-time when users access them, removing the need for librarians to input data manually. Its implementation is also flexible. However, we suggest providing APIs (Application Programming Interfaces) that allow access from other external services, which will allow for continued flexibility in future applications. Providing RSS support will also be useful, as it allows users to access the gathered data with the RSS tools that they are familiar with.

We provide information to users through subject-specific research guides. Each guide exists as a website that presents all the information related to its subject – users can access information from a variety of resources on each website. We designed the guides to have several subpages, which are organized with consideration to the characteristics and composition of the information they contain. We provide a summary of the newest information on the home page, including newly arrived books (from the Book & e-book page), recently updated journals and articles (from the Journal & article page), the latest news (from the News page), recently published reports (from the Report page), upcoming calls for papers (from the Call for papers page), and so on. On the Book & e-book page, we provide information about books that have recently arrived at our library and information about newly released books that have not yet been added to our library, with a link to our book request form. On the Journal & article page, we provide lists of SCI/SSCI-indexed journals of related categories. Users can browse this page for the latest articles from each journal, accessed by clicking on their journals of interest. We provide the latest academic news articles on the News page and the latest research or technical reports on the Report page. Information about upcoming conferences and calls for papers is given on the Call for papers page. Video clips of the lectures from OpenCourseWare, TED talks, and other academic contents are given on

Figure SEQ Figure * ARABIC 1. Features of System Developed



the Video page. Each page has user interfaces designed specifically for the related information – for example, the book page displays book covers, and the video page has embedded video clips that can be played instantly.

We also deliver newsletters on each subject. While dealing with daily-updated information, we assess the need for retrieving information that might have important implications in the field, from users who are not able to access our service frequently. Sending newsletters is an efficient means to inform researchers of our services and how they differ from traditional library services. We write monthly newsletters by selecting information from data collected by our system throughout the month and distribute them to researchers via email.

The data monitoring feature checks the availability of each piece of research information. As our system uses information from various sources, it is more likely to contain inaccessible information at any time; for instance, the web servers of some information sources may be temporarily unavailable or under maintenance. The data monitor feature retrieves all information on our research guides and checks whether there is any error reading each piece of information on a daily basis. It notifies librarians of the results via e-mail so that they can take early measures to improve the readiness, availability, and reliability of our services.

Results and discussion

We have implemented and run a research guide system that automates most procedures of research data curation, including data collection, data management, data edit, data transport, data presentation, and data monitoring. Currently, our research guides provide 108,000 pieces of information including journals, books, reports, news articles, calls for papers, and videos

from 22,000 information sources on 4 subjects: computer science, law, media and communication, and psychology. Each day, the guides automatically collect and update information from more than 21,000 sources.

This research can have significant implications for libraries that are in search of a feasible model of research data curation. We have designed and implemented a new system that provides automated data curation procedures, and thus enhances the volume and timeliness of information distribution, while also filtering it for relevance and efficiency. Libraries can use this or similar systems to collect and provide comprehensive research information with reasonable resources.

References

Bawden, David and Robinson, Lyn. (2002). Internet subject gateways revisited. International Journal of Information Management, Vol. 22, Issue 2, pp. 157-162.

Jones, Gina M. and Neubert, Michael. (2017). Using RSS to improve web harvest results for news web sites. Journal of Western Archives, Vol. 8, Issue 2, Article 3.

Dey, Nabin Chandra and Sarkar, Pronab. (2009). RSS feeds and its application in library services. 7th International CALIBER-2009, pp. 342-349.

Faheem, Muhammad. (2012). Intelligent crawling of web applications for web archiving. WWW, Apr 2012, Lyon, France. ACM, pp. 127-131.