University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal) Libraries at University of Nebraska-Lincoln

Summer 6-15-2023

Batch Bitstreams and Metadata import using SAFBuilder in Dspace: A Practical Experience

Jamil Ahmed Dr.
Bennett University, jamil.ahmed@bennett.edu.in

Follow this and additional works at: https://digitalcommons.unl.edu/libphilprac

Part of the Library and Information Science Commons

Ahmed, Jamil Dr., "Batch Bitstreams and Metadata import using SAFBuilder in Dspace: A Practical Experience" (2023). *Library Philosophy and Practice (e-journal)*. 7806.
https://digitalcommons.unl.edu/libphilprac/7806

Batch Bitstreams and Metadata import using SAFBuilder in Dspace: A Practical Experience

Dr. Jamil Ahmed
Learning Resource Centre
Bennett University, Greater Noida, UP, India
jamil.ahmed@bennett.edu.in

Dr. Sanjay Kataria
Learning Resource Centre
Bennett University, Greater Noida, UP, India
Sanjay.kataria@bennett.edu.in

***Abstract*—** Digital repositories play a crucial role in organizing and preserving vast collections of digital content. Efficiently ingesting large amounts of data into these repositories is a common challenge faced by institutions. This paper explores the use of bulk upload techniques in DSpace, an open-source digital repository software, to streamline the ingestion process and enhance repository management. We discuss the benefits of bulk upload in terms of time savings, metadata consistency, and scalability. Additionally, we delve into the technical aspects of implementing bulk upload in DSpace, covering the Simple Archive Format (SAF), metadata mapping, and handling of digital files. Furthermore, we highlight real-world examples and best practices for utilizing bulk upload in DSpace. By adopting this approach, institutions can significantly improve their efficiency in managing and preserving digital content, ensuring a seamless user experience, and facilitating knowledge dissemination**.** Here, an experimental method of research/ case study technique is utilized to evaluate the efficiency and effectiveness of the design model for implementation of the bulk uploading of documents in Dspace at Bennett University is practiced. The feedback is gathered in order to identify the flaws and make the necessary improvements. Simple Archive Format (SAF) is a utility that converts Bitstream/Content files plus a metadata.csv file into a Simple Archive Format package, making bulk uploads to the DSpace repository simple. All question papers were digitized using a high-quality scanner, an Excel file with Dublin core information was created, and Excel was converted to CSV format in order to import all old question papers in bulk into the Bennett University Digital Repository Services. The study indicates that it is essential to pay close attention to the precise format of metadata leveraging the Dublin core and the file's location. It is an experiment conducted by the Bennett University Library and the research was confined to Bennett University digital repository.

Keywords- Dspace, Digital Library, Simple Archive Format (SAF), Dublin Core, Bitstream

I. INTRODUCTION

Dspace is an open-source software used to build digital archives for any kind of digital information. It is a repository that stores a specific need as a digital archives system, focusing on the long-term storage, access, and preservation of digital content.[4] Dspace is made up of numerous programs, a metadata repository, and Java web applications. Interfaces for administration, deposit, ingest, search, and access are provided by web apps. The asset store is kept up to date on a file or comparable storage system. The PostgreSQL database can be used with the metadata, which is kept in a relational database and includes access and configuration details. The main way to access Dspace holdings is through a web interface. Dspace's more recent iterations use Apache Solr to enable faceted search and browsing capability. Manual upload in Dspace is time taking process if there is a bundle of digital content.

DSpace is widely used by academic and research institutions, government agencies, and other organizations to manage their digital content, including articles, theses, dissertations, reports, datasets, images, videos, and more. DSpace is designed to be customizable and extensible and supports a wide range of metadata standards and file formats[2].

II. DSPACE SAF BUILDER

Dspace SAF Builder is a tool that creates a Simple Archive Format package from content files and a metadata spreadsheet, making it simple to batch import the package into the institutional repository Dspace. It can be used in the scenario when someone has a spreadsheet with content files and metadata that will be ingested into a repository.

With DSpace SAF Builder, users can create a SAF package of their DSpace content, including metadata, bitstreams (i.e., the actual digital files), and any associated licenses or permissions. The SAF package can then be exported and stored in a secure location or transferred to another DSpace instance.

Using DSpace SAF Builder can be especially useful for institutions that want to migrate their DSpace content to a new version of DSpace or to a different institutional repository system. By creating a SAF package, the institution can ensure that all the necessary content and metadata are included in the transfer and that the content is preserved in a standardized, widely recognized format.

DSpace SAF Builder is a command-line tool that requires some technical expertise to use. DuraSpace provides detailed documentation and guidance on how to use the tool on its website. Additionally, there are third-party tools and services that can help institutions with the creation and transfer of SAF packages.

DIGITAL REPOSITORY SERVICES AT BENNETT UNIVERSITY

Bennett University offers digital repository services to its students, faculty, and researchers. The university's digital repository is called "BU Digital Repository" and it is hosted on DSpace, an open-source software platform for creating digital repositories.

BU Digital Repository provides a platform for storing, preserving, and disseminating the research output of the university's academic community. The repository includes a range of digital materials such as articles, preprints, book chapters, conference papers, theses and dissertations, research datasets, audio-visual materials, and more.

Search Functionality: The repository has a powerful search engine that allows users to search for materials by author, title, subject, keyword, and more.

BU Digital Repository is an important resource for the university's academic community, providing a platform for showcasing and preserving the research output of the university's faculty and students.

IV. SOFTWARE REQUIREMENTS

To perform bulk upload in DSpace, you will need the following software requirements:

1) Java JDK
2) Git
3) Maven
4) SAFBuilder

V. METHODOLOGY

All the question papers were digitized with a good quality scanner, prepared an Excel sheet with Dublin core metadata, and converted excel to CSV format to make bulk import of all old question papers in Bennett University Digital Repository Services.

Bulk uploading in DSpace is a convenient way to add a large number of items to your repository quickly. Here are some steps to follow to bulk upload items in DSpace:

1) Prepare the files: Ensure that all the files you want to upload are in the correct format and have descriptive file names. It is a good practice to organize your files in a folder with subfolders for each collection or item.

1.csv
BCA202L_ MOBILE_ PROGRA...
BCA353L_ PROGRAM MING_US...
BCA355L_ DATA_ MINING_...
BCA356L_ APPLICATI ONS_OF_...
BCA358L_ DEEP_ LEARNIN...
BCA365_ PENETRATI ON_TEST...
BCA369L_ APPLICATI ON_OF_B...
CAREER_ SKILLS_ FOR_IT_C...
CBJE202L_ BUSINESS_ OF_FILM...
CBJE204P_ FILM_ STUDIES_...
CBJE206L_ EVENT_ MANAGE...
CBJE212L_ INDIAN_ ECONOM...
CBJE308L_ MEDIA_ LAW_AN...
CBJE316L_ CBJH316L_ PSYCHOL...
CBJH206L_ EVENT_ MANAGE...
CBJH208L_ JANSANCH AR_KE_SI...
CBJH316L_ PSYCHOLO GY.pdf
CSET205_ SOFTWARE _ENGINE...
CSET206_ DESIGN_ AND_AN...
CSET207_ COMPUTER _NETWO...
CSET208_ ETHICS_ FOR_ENG...
CSET209_ OPERATIN G_SYSTE...
CSET224_ CLOUD_ COMPUTI...
CSET226_ BLOCKCHAI N_ENGIN...
CSET227_ SYSTEM_ AND_NE...
CSET228_ DATA_ MINING_...
CSET229_ GAME_ ENGINE_...
CSET230_ DEVOPS_ PRACTIC...
CSET231_ PROGRAM MING_M...
CSET301_ ARTIFICIAL _INTELLI...
CSET308_ COMMUNI CATION_...
CSET310_ MARKETIN G_STRAT...
CSET389_ LINUX_ APACHE_...
EBTY202L_ INTRODUC TION_TO...
EBTY204L_ RDNA_ TECHNPL...
EBTY206L_ THERMODY NAMICS_...
EBTY208L_ INTELLECT UAL_PRO...
EBTY302L_ GENOMICS _AND_GE...
EBTY304L_ BIOPROCES S DESIGN...
EBTY304L_ BIOPROCES S_DESIG...
EBTY306L_ RESEARCH_ METHOD...
EBTY342L_ GENE_ THERAPY...
EBTY364L_ RATIONAL_ DRUG_DE...
EBTY382M_ CMOS_ AND_IND...
ECSE359L_ SOFTWARE _PROJEC...
ECSE382M_ APPLIED_ AI.pdf
ECSE440L_ DEEP_ LEARNIN...
ECSE446L_ ADVANCED _DATABA...
ECSE460L_ DEVOPS_ ENGINEE...
EECE202L_ ELECTRONI C_CIRCUI...
EECE282M_ DESIGN_ OF_LOGI...
EECE306L_ CONTROL_ SYSTEM...
EECE314L_ COMPUTER _NETWO...
EECE316L_ ANTENNA_ AND_WA...
EECE344L_ VLSI_ TECHNOL...
EECE345L_ INTRODUC TION_TO...
EECE350L_ WIRELESS_ COMMU...
EECE360L_ IMAGE_ PROCESS...
EECE363L_ ADVANCED _DIGITAL...
EECE382M_ SENSORS_ AND_ME...
EMAT202_ NUMERICA L_METH...
EMAT342L_ NUMERICA L_METH...
EMEC208L_ KINEMATIC S_AND_D...
EMEC212L_ FINITE_ ELEMENT...
EMEC214L_ MANUFACT URING_P...
EMEC218P_ ENGINEERI NG_SIMU...
EMEC220L_ DESIGN_ OF_MAC...
EMEC314L_ MECHATRO NICS_AN...
EMEC316L_ MECHANIC AL.pdf
EMEC342L_ COMPRESSI BLE_FLO...
EMEC458L_ REFRIGERA TION_AN...
EPHY202L_ QUANTUM_ MECHA...
EPHY204L_ THERMAL_ AND_STA...
EPHY208L_ OPTICS_ AND_LAS...
EPHY344L_ ELEMENTS_
EPHY382L_ MODERN_
EPHY441L_ COMPUTAT IONAL_M...
EPHY447L_ PARTICLE_ PHYSICS...
ESEA301L_ ENVIRONM ENT_AN...
LBAL204L_ POLITICAL_ SCIENCE_...
LBAL208L_ ECONOMIC S_II.pdf
LBBL202L_ FINANCIAL _MANAG...
LBBL204L_ INTERNATI ONAL_B...
LLLB202L_ CONSTITUT IONAL_L...
LLLB204L_ FAMILY.pdf
LLLB208L_ LLLB230L_ LAW_OF_...
LLLB216L_ INTELLECT UAL_PRO...
LLLB222L_ HUMAN_ RIGHTS_...
LLLB302L_ INTERPRET ATION_O...

Fig. 1 Files in pdf with CSV

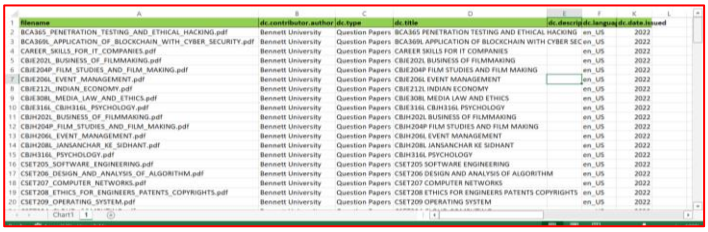
2) Create a metadata template: Create a metadata template for your items. The metadata template should include all the fields you want to use for describing your items. You can create the metadata template using the DSpace metadata registry tool [1].

A spreadsheet (.csv) with the following columns: [7]

TABLE I

METADATA AND DUBLIN CORE ELEMENTS

Metadata	**Dublin Core Elements**
File	PDF file name
Cover Title	dc.identifier.citation*
ISSN	dc.identifier.issn
Vol.	dc.identifier.citation*
Iss.	dc.identifier.citation*
Cover Date	dc.identifier.citation*
Year	dc.date.issued
Month	dc.date.issued
Fpage	dc.identifier.citation*
Lpage	dc.identifier.citation*
Article Title	dc.title
Author Names	dc.creator
Institution	dc.description
Abstract	dc.description.abstract
Language	dc.language.iso
Rights	dc.rights
Types	dc.types



	A	B	C	D	E	F	K
1	filename	dc.contributor.author	dc.type	dc.title	dc.descrip	dc.languag	dc.date.issued
2	BCA365_PENETRATION_TESTING_AND_ETHICAL_HACKING.pdf	Bennett University	Question Papers	BCA365 PENETRATION TESTING AND ETHICAL HACKING		en_US	2022
3	BCA369L_APPLICATION_OF_BLOCKCHAIN_WITH_CYBER_SECURITY.pdf	Bennett University	Question Papers	BCA369L APPLICATION OF BLOCKCHAIN WITH CYBER SEC		en_US	2022
4	CAREER_SKILLS_FOR_IT_COMPANIES.pdf	Bennett University	Question Papers	CAREER SKILLS FOR IT COMPANIES		en_US	2022
5	CBJE202L_BUSINESS_OF_FILMMAKING.pdf	Bennett University	Question Papers	CBJE202L BUSINESS OF FILMMAKING		en_US	2022
6	CBJE204P_FILM_STUDIES_AND_FILM_MAKING.pdf	Bennett University	Question Papers	CBJE204P FILM STUDIES AND FILM MAKING		en_US	2022
7	CBJE206L_EVENT_MANAGEMENT.pdf	Bennett University	Question Papers	CBJE206L EVENT MANAGEMENT		en_US	2022
8	CBJE212L_INDIAN_ECONOMY.pdf	Bennett University	Question Papers	CBJE212L INDIAN ECONOMY		en_US	2022
9	CBJE308L_MEDIA_LAW_AND_ETHICS.pdf	Bennett University	Question Papers	CBJE308L MEDIA LAW AND ETHICS		en_US	2022
10	CBJE316L_CBJH316L_PSYCHOLOGY.pdf	Bennett University	Question Papers	CBJE316L CBJH316L PSYCHOLOGY		en_US	2022
11	CBJH202L_BUSINESS_OF_FILMMAKING.pdf	Bennett University	Question Papers	CBJH202L BUSINESS OF FILMMAKING		en_US	2022
12	CBJH204P_FILM_STUDIES_AND_FILM_MAKING.pdf	Bennett University	Question Papers	CBJH204P FILM STUDIES AND FILM MAKING		en_US	2022
13	CBJH206L_EVENT_MANAGEMENT.pdf	Bennett University	Question Papers	CBJH206L EVENT MANAGEMENT		en_US	2022
14	CBJH208L_JANSANCHAR_KE_SIDHANT.pdf	Bennett University	Question Papers	CBJH208L JANSANCHAR KE SIDHANT		en_US	2022
15	CBJH316L_PSYCHOLOGY.pdf	Bennett University	Question Papers	CBJH316L PSYCHOLOGY		en_US	2022
16	CSET205_SOFTWARE_ENGINEERING.pdf	Bennett University	Question Papers	CSET205 SOFTWARE ENGINEERING		en_US	2022
17	CSET206_DESIGN_AND_ANALYSIS_OF_ALGORITHM.pdf	Bennett University	Question Papers	CSET206 DESIGN AND ANALYSIS OF ALGORITHM		en_US	2022
18	CSET207_COMPUTER_NETWORKS.pdf	Bennett University	Question Papers	CSET207 COMPUTER NETWORKS		en_US	2022
19	CSET208_ETHICS_FOR_ENGINEERS_PATENTS_COPYRIGHTS.pdf	Bennett University	Question Papers	CSET208 ETHICS FOR ENGINEERS PATENTS COPYRIGHTS		en_US	2022
20	CSET209_OPERATING_SYSTEM.pdf	Bennett University	Question Papers	CSET209 OPERATING SYSTEM		en_US	2022

Fig. 2 CSV file with Dublin Core Metadata

3) Use the DSpace batch import tool: DSpace provides a batch import tool that allows you to upload items in bulk. To use the batch import tool, you will need to create a CSV file that lists the metadata for each item. You can use the metadata template you created earlier to create the CSV file.

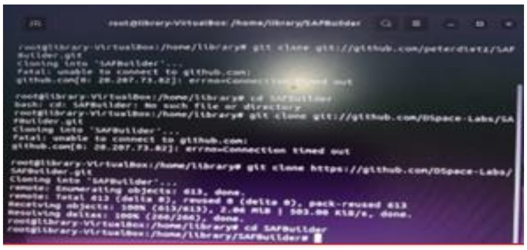
Install Java and Ant on your system if they are not already installed
Clone the SAFBuilder repository from GitHub using the following command:
git clone https://github.com/DSpace-Labs/SAFBuilder.git

Navigate to the SAFBuilder directory using the following command:
To Install and generate an Item Import package

To install SAFBuilder from GitHub and use it to upload a SAF package via the terminal or command prompt, you can follow these steps:

Open the terminal and apply the following commands

- ✓ sudo su
- ✓ sudo apt-get install default-jdk
- ✓ sudo apt-get install maven
- ✓ sudo apt-get install git

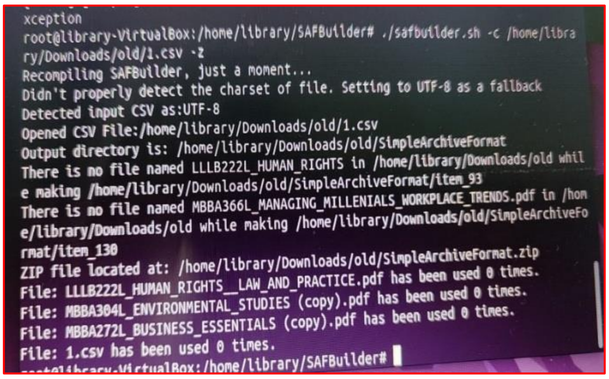


```
root@library-VirtualBox: /home/library/SAFBuilder
root@library-VirtualBox:/home/library# git clone git://github.com/peterdietz/SAFBuilder.git
Cloning into 'SAFBuilder'...
fatal: unable to connect to github.com:
github.com[0: 20.207.73.82]: errno=Connection timed out

root@library-VirtualBox:/home/library# cd SAFBuilder
bash: cd: SAFBuilder: No such file or directory
root@library-VirtualBox:/home/library# git clone git://github.com/DSpace-Labs/SAFBuilder.git
Cloning into 'SAFBuilder'...
fatal: unable to connect to github.com:
github.com[0: 20.207.73.82]: errno=Connection timed out

root@library-VirtualBox:/home/library# git clone https://github.com/DSpace-Labs/SAFBuilder.git
Cloning into 'SAFBuilder'...
remote: Enumerating objects: 613, done.
remote: Total 613 (delta 0), reused 0 (delta 0), pack-reused 613
Receiving objects: 100% (613/613), 2.06 MiB | 503.00 KiB/s, done.
Resolving deltas: 100% (266/266), done.
root@library-VirtualBox:/home/library# cd SAFBuilder
root@library-VirtualBox:/home/library/SAFBuilder# 
```

Fig. 1 Install SAFBuilder

To run:
cd SAFBuilder
./safbuilder.sh -c /path of the folder/name of the CSV file.csv -z

Fig. 2 Creating SAFBuilder ZIP File



```
xception
root@library-VirtualBox:/home/library/SAFBuilder# ./safbuilder.sh -c /home/libra
ry/Downloads/old/1.csv -z
Recompiling SAFBuilder, just a moment...
Didn't properly detect the charset of file. Setting to UTF-8 as a fallback
Detected input CSV as:UTF-8
Opened CSV File:/home/library/Downloads/old/1.csv
Output directory is: /home/library/Downloads/old/SimpleArchiveFormat
There is no file named LLLB222L_HUMAN_RIGHTS in /home/library/Downloads/old whil
e making /home/library/Downloads/old/SimpleArchiveFormat/item_93
There is no file named MBBA366L_MANAGING_MILLENIALS_WORKPLACE_TRENDS.pdf in /hom
e/library/Downloads/old while making /home/library/Downloads/old/SimpleArchiveFo
rmat/item_130
ZIP file located at: /home/library/Downloads/old/SimpleArchiveFormat.zip
File: LLLB222L_HUMAN_RIGHTS__LAW_AND_PRACTICE.pdf has been used 0 times.
File: MBBA304L_ENVIRONMENTAL_STUDIES (copy).pdf has been used 0 times.
File: MBBA272L_BUSINESS_ESSENTIALS (copy).pdf has been used 0 times.
File: 1.csv has been used 0 times.
root@library-VirtualBox:/home/library/SAFBuilder# 
```

- C for CSV: Filename the path of the CSV spreadsheet. This must be in the same directory as the content files
- Z for Zip files

The Dspace SAF builder tool can be installed using the GitHub cloning method or we can download the zip file of

the tool. The tool gets installed successfully when the pre-required software is installed successfully [3].

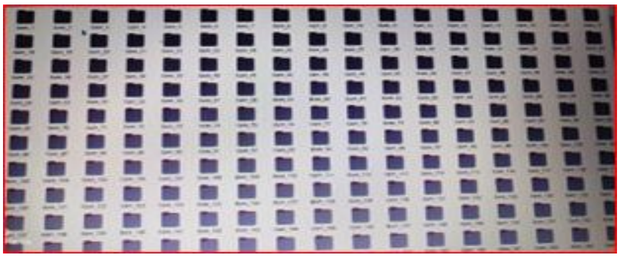


Fig. 3 SimpleArchiveFormat

Upload the files: Once you have prepared your files and metadata, you can use the batch import tool to upload the files to DSpace. The batch import tool will read the CSV file and use the metadata to create new items in the repository.

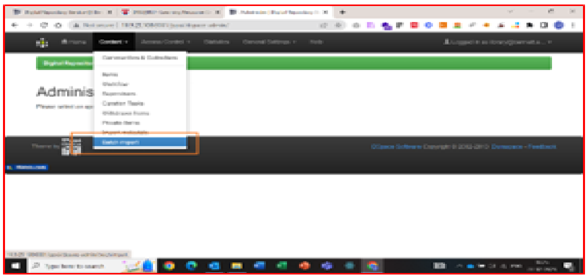


Fig. 4 Batch Import

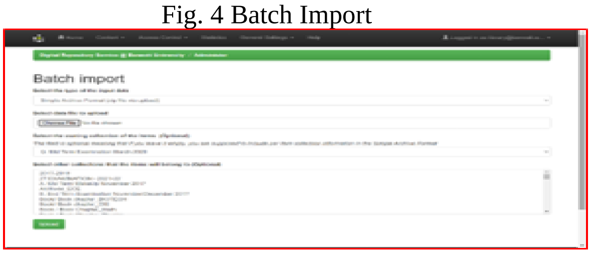
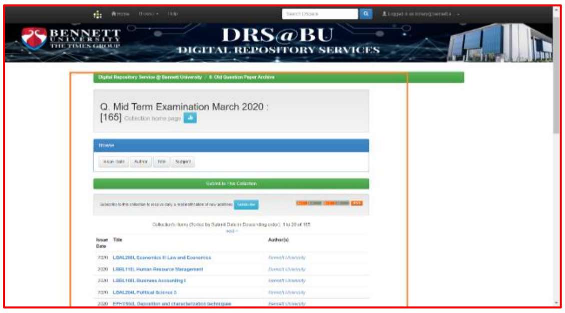


Fig. 5 Select Collection in Dspace

Result and Discussion

After the files are uploaded, you can review them and make any necessary edits or corrections. Once you are satisfied with the items, you can publish them to make them available in the repository.

It is important to note that the exact steps for bulk uploading in DSpace may vary depending on the version of DSpace you are using and your specific configuration settings. It is recommended that you consult the DSpace documentation or seek support from your DSpace administrator if you encounter any issues during the bulk upload process.


Home
Browse
Help
Search DSpace
BENNETT UNIVERSITY THE TIMES GROUP
DRS@BU
DIGITAL REPOSITORY SERVICES
Digital Repository Service @ Bennett University / 8. Old Question Paper Archive
Browse
Issue Date
Author
Title
Subject
Submit to This Collection
Subscribe
next &gt;

Q. Mid Term Examination March 2020 :
[165] Collection home page

Collection's Items (Sorted by Submit Date in Descending order): 1 to 20 of 165

Issue Date	Title	Author(s)
2020	LBAL208L Economics III Law and Economics	Bennett University
2020	LBBL110L Human Resource Management	Bennett University
2020	LBBL108L Business Accounting I	Bennett University
2020	LBAL204L Political Science 3	Bennett University
2020	EPHY504L Deposition and characterization techniques	Bennett University

Fig. 6 Files Successfully Uploaded in Dspace

Conclusion

It is an experiment conducted by the Bennett University Library and successfully uploaded all the question papers to Bennett University digital repository. SAF packages have emerged as a standardized way to transfer content and DSpace includes functionality for importing content in SAF format. Overall, the study suggests that the use of SAF packages for importing content into Dspace can be an effective and useful approach, but that it requires careful attention to metadata and data format issues to ensure that the imported content is properly preserved and accessible. The authors conclude by noting that additional research is needed to explore the use of SAF packages for other types of content and in other repository systems.

References

[1] DSpace-Labs. (n.d.). GitHub - DSpace-Labs/SAFBuilder: Builds a Simple Archive Format package from files and a spreadsheet. GitHub. https://github.com/DSpace-Labs/SAFBuilder

[2] Johnson, R," DSpace: An Open-Source Dynamic Digital Repository". D-Lib Magazine, 12(7/8), 1-6. doi:10.1045/july2006-johnson

[3] Kavitha. K ."Mapping Metadata to Doublincore in Dspace," 2020.Summer Fellowship Report on Mapping Metadata to Doublincore in Dspace. Indian Institute of Technology Bombay. https://static.fossee.in/fossee/fellowship2020/Fellowship-Reports/Koha/KohaAndDspace-KavithaK-FSF-2020

[4] Rosa, C. A., Craveiro, O., & Domingues, P, "Open Source Software for Digital Preservation Repositories: . International Journal of Computer Science & Engineering Survey," D-Lib Magazine, vol. 8, no. 3, 2017, doi: https://doi.org/10.5121/ijcses.2017.8302

[5] Walsh, M. P, "Batch Loading Collections into DSpace: Using Perl Scripts for Automation and Quality Control," Information Technology and Libraries, vol. 29, no. 127, 2017, doi: http://doi.org/10.6017/ital.v29i3.3137

[6] James, Creel, "Batch Importing into DSpace with the SAFCreator.," Texas Conference on Digital Libraries, 2016, doi: https://www.virtualbox.org/wiki/Downloads

[7] Michel, Castagné, "Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra.," University of British Columbia, 2013,

[8] Arora, Dipti. "Building Digital Library of Technology Focus Using Dspace." In 2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), pp. 83-86. IEEE, 2018.