УДК 004.622, 004.415.2
DOI 10.52575/2687-0932-2021-48-3-564-577

Подготовка метаданных публикаций для пакетного импорта в институциональный репозиторий на платформе DSpace

Резниченко О.С.
Белгородский государственный национальный исследовательский университет,
Россия, 308015, г. Белгород, ул. Победы, 85
E-mail: oreznichenko@bsu.edu.ru

Аннотация. Внесение метаданных о научных публикациях в институциональный репозиторий на платформе DSpace вручную занимает значительное время, даже когда данные представляют собой готовые выгрузки из реферативных баз Scopus и Web of Science, и уже имеют формат, близкий к Dublin Core. Для решения задачи преобразования и объединения данных, а также интеграции в итоговый набор метаданных оригинал-макетов публикаций с целью их пакетного импорта в университетский репозиторий, были разработаны алгоритмы использования стандартных офисных приложений и бесплатного ПО, а также созданы программные скрипты, которые позволили автоматизировать большинство рутинных операций. Использование этих алгоритмов и созданного ПО показало двадцатидевятикратное сокращение временных затрат в сравнении с ручным вводом метаданных в DSpace.

Ключевые слова: институциональный репозиторий, Web of Science, DSpace, Microsoft Excel, Python, pandas.DataFrame.

Для цитирования: Резниченко О.С. 2021. Подготовка метаданных публикаций для пакетного импорта в институциональный репозиторий, основанный на DSpace. Экономика. Информатика, 48 (3): 564–577. DOI 10.52575/2687-0932-2021-48-3-564-577.

Preparation articles metadata for batch import into DSpace repository

Oleg S. Reznichenko
Belgorod National Research University
85 Pobeda St, Belgorod, 308015, Russia
E-mail: oreznichenko@bsu.edu.ru

Abstract. Manual import of metadata records about research articles in institutional repository DSpace take a lot of time even when the input data uploads from Scopus and Web of Science databases and already has a format close to Dublin Core Metadata Element Set. To solve the problem of transforming and combining data, as well as integrating the article PDFs into the final metadata archive, some algorithms were developed. Algorithms use Microsoft Office Excel and free software. In addition, software tools by Python-scripts using "pandas" software library were created that automate most of the routine operations such as combine Scopus and Web of Science databases data export into single file, records dublicate exclude, converting authors record format and excluding records which already exist in DSpace repository. The use of these algorithms and the created software tools help to create Simple Archive Format file for batch import into DSpace repository and demonstrated a 29-fold reduction in time compared to manually metadata entering.

Keywords: Institutional Repository, Scopus, Web of Science, DSpace, Microsoft Excel, Python, pandas.DataFrame.

For citation: Reznichenko O.S. 2021. Preparation article metadata for batch import into DSpace repository. Economics. Information technologies, 48 (3): 564–577 (in Russian). DOI 10.52575/2687-0932-2021-48-3-564-577.

Введение

Институциональные репозитории открытого доступа в России продолжают активно развиваться. Главным образом существующие в стране репозитории основаны на программных платформах, использующих общий открытый стандарт OAI-PMH. Наибольшей популярностью среди российских репозиториев пользуется открытое веб-приложение DSpace [DuraSpace, 2021], однако, некоторые национальные репозитории используют в своей основе и другие программные платформы, такие как Socionet, Invenio, EPrints, vital и т. д. [Fedotova et al., 2020; SHERPA, 2021]. В Белгородском государственном национальном исследовательском университете (далее – БелГУ, университет) институциональный репозиторий также организован на программной платформе DSpace версии 5.5 и является одним из самых крупных репозиториев в России – содержит более чем 40 тысяч метаданных [Southampton, 2021]. В рамках реализации задач «Белгородской Декларации об открытом доступе к научным знаниям и культурному наследию в научно-образовательном пространстве» университетский репозиторий непрерывно пополняется сотрудниками научно-библиографического консультационного центра, которые добавляют вручную до 30 записей в течение рабочего дня. Согласно правилам, установленным в Центре, обязательным условием размещения в репозитории информации о публикации является наличие ее полного текста в виде скан-копий в формате pdf, что существенно ограничивает возможности пакетного импорта метаданных из других открытых источников. В апреле 2018 года электронный архив открытого доступа НИУ «БелГУ» (далее – университетский репозиторий) присоединился к Национальному агрегатору открытых репозиториев российских университетов (НОРА). Основатели проекта НОРА НП «НЭИКОН», анализируя общий поток статей авторов университета в наукометрических реферативных базах порталов Scopus (далее – Scopus) и Web of Science Core Collection (далее – WoS), обнаружили, что не все включенные в базы статьи присутствуют в университетском репозитории и предложили пополнить его, выгрузив тексты статей открытого доступа с целью их дальнейшего размещения в репозитории. Имея в наличии файлы с полными текстами публикаций, возникает проблема их анализа, сопоставления с метаданными, полученными на основе выгрузки из баз Scopus и WoS, а также внесения метаданных в университетский репозиторий, что при ручном выполнении всех операций будет занимать десятки часов рабочего времени. Использование существующего в репозитории механизма пакетного импорта метаданных также займет значительное время, основная часть которого уйдет на ручное формирование сводного файла для импорта. Для сокращения времени на подготовку архива публикаций в рамках данного исследования разрабатываются алгоритмы и программные инструменты, решающие задачу автоматизации процесса подготовки метаданных, описывающих публикации открытого доступа из баз Scopus и WoS, включая описание процесса интеграции в эти метаданные файлов с полными текстами исходных документов с целью последующего пакетного импорта готового архива в университетский репозиторий. Потребность в разработке алгоритмов и инструментов для предобразования экспортированных данных о публикациях в нужный формат обусловлена не только разовой необходимостью, но и возможностью их повторного применения при возникновении аналогичной задачи в будущем, в том числе, когда аналогичная задача возникает перед администраторами институциональных репозиториев в других научных организациях. Вопросы преобразования и пакетной загрузки данных в репозитории на основе DSpace рассмотрены в нескольких исследованиях. В статьях [Walsh, 2010] и [Deng, 2010] предложены инструменты для пакетной загрузки данных в коллекции DSpace, однако описанные там программные инструменты необходимо каждый раз адаптировать под новый формат входных данных, либо же адаптировать исходные данные для использования предложенного инструментария, что также займет некоторое время, даже при автоматизации большинства операцией преобразования входных данных. В отличие от описанных выше исследований, инструменты, разработанные в ходе исследования [Nash, 2016] в качестве

выходных данных, используют формат Simple Archive Format (далее – SAF), использующийся в репозиториях на основе DSpace версии 5 и выше, и предполагающие импорт архива встроенным средством пользовательского веб-интерфейса. Однако при этом входные данные, с которыми работают эти инструменты, также имеют ориентацию на собственный формат входных данных в виде экспорта из базы данных Native Health Database библиотеки Health Sciences Library & Informatics Center The University of New Mexico университета Нью-Мексико. Аналогичная ситуация имеет место при использовании инструментов, разработанных в ходе исследования [Gafurova at al., 2020], в котором осуществлялась конвертация и нормализация данных, экспортированных из таких библиотек, как EuDML, MathNet.Ru, DBLP. Использование предложенных в исследованиях [Nash, 2016] и [Gafurova at al., 2020] инструментов также требует значительных преобразований входных данных.

Описание методов и средств реализации

Выбор инструментальных средств для реализации преобразования данных обусловлен не только их доступностью, но и особенностями формата входных данных, а также выбранным способом реализации задачи. Исходными данными для импорта выступают сведения, представленные в базах Scopus и WoS в виде файлов экспорта, которые можно получить встроенными средствами соответствующих web-приложений, используя интернет-браузер. Специфика формата экспортируемых из баз Scopus и WoS данных предполагает для их анализа и обработки использовать средства процессора электронных таблиц Microsoft Office Excel (далее – Excel) и его настройки Microsoft Power Query [Microsoft, 2021]. Так как форматы данных в экспортных файлах Scopus и WoS имеют отличия, то предлагается разработать программное средство для объединения выгруженных метаданных в единый табличный файл. Разово эту задачу можно решить средствами того же Excel, затратив при этом от сорока минут, предусмотрев при этом неизбежное возникновение дублирующих записей, которые невозможно отследить средствами Excel из-за незначительных различий в написании заголовков публикаций. Однако целесообразней разработать и использовать несколько программных функций, которые бы не только решали задачи объединения данных из двух источников, но и автоматизировали другие задачи, связанные с преобразованием данных, а также формированием итогового архива для пакетного импорта. В данном исследовании для реализации поставленных задач используются средства языка программирования Python 3.9, включая функции библиотеки для анализа и обработки больших данных Pandas. Используя структуру данных Dataframe модуля Pandas [Wood, 2021], можно осуществлять необходимую обработку и преобразования данных, представленных в табличном виде. Для удобства работы со средой Python, осуществления отладки подпрограмм, а также подключения дополнительных библиотек используется интегрированная среда разработки PyCharm Community Edition [JetBrains, 2021]. Для получения списка имен файлов-макетов публикаций, а также пакетного переименования этих файлов по заданному шаблону, применяется бесплатное приложение Advanced Renamer [Jensen, 2021]. Генерация файла-архива для импорта в DSpace производится с помощью бесплатной утилиты SAFBuilder [Dietz, 2015], для работы которой требуется предустановленная виртуальная машина Java Runtime Environment [Oracle, 2021].

Описание структуры исходных данных и структуры выходного файла

Структура метаданных одной публикации в университетском репозитории соответствует набору элементов Дублинского Ядра (далее – DC) [Middleton, 2021]. Просмотреть и проанализировать конкретный набор элементов метаданных можно отобразив полное описание конкретной публикации (ресурса) в репозитории (рис. 1).

В табл. 1 представлен полный список элементов DC, входящих в состав метаданных описания ресурса в университетском репозитории, с их описанием. Этот набор метаданных будет определяющим при отборе полей в экспортных файлах из баз Scopus и WoS.

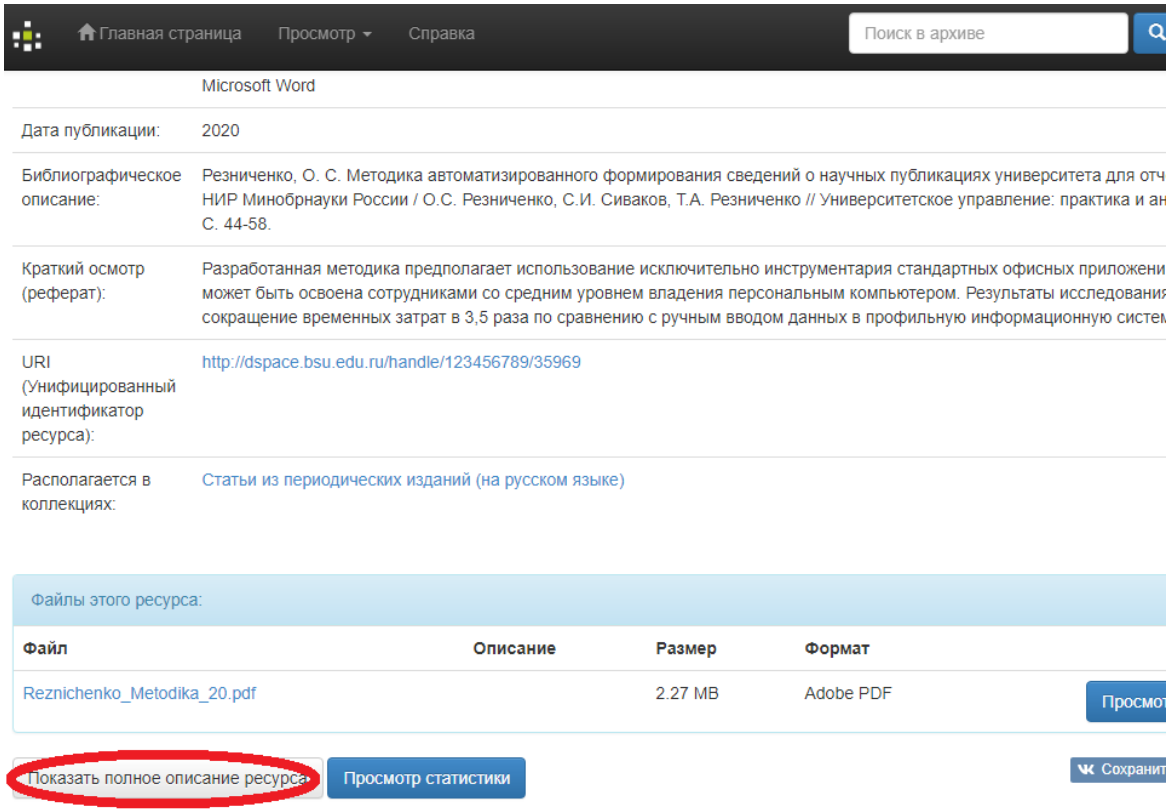

Рис. 1. Просмотр полного описания публикации в университетском репозитории
Fig. 1. Description of full article record in the University Open Access Repository

Таблица 1
Table 1

Полная запись метаданных описания публикации
Full article metadata record

Код поля в формате DC	Описание
dc.contributor.author	Первый автор
dc.contributor.author	Второй автор
…	Остальные авторы
dc.date.issued	Год опубликования
dc.identifier.citation	Информация для цитирования
dc.identifier.uri	Идентификатор ресурса в репозитории
dc.description.abstract	Аннотация
dc.description.provenance	Submitted by Администратор Ресурса (dspace@bsu.edu.ru) on 2020-05-24T14:11:53Z No. of bitstreams: 1 Moskovkin_Instrumenty.pdf: 752845 bytes, checksum: 83a9d4082047dca040d00487068e96d3 (MD5)
dc.subject	Первое ключевое слово
…	Остальные ключевые слова
dc.title	Заглавие публикации
dc.type	Тип публикации
dc.identifier.citationpublication	Название журнала\издания
dc.identifier.citationvolume	Номер тома журнала\издания
dc.identifier.citationnumber	Номер выпуска журнала\издания
dc.identifier.citationfirstpage	Номер начальной страницы публикации
dc.identifier.citationendpage	Номер последней страницы публикации
dc.language.iso	Язык публикации

Scopus [Elsevier, 2021] – это библиографическая и реферативная база данных и инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях. База данных доступна научным организациям через веб-интерфейс на условиях Национальной подписки и только с определенных подпиской IP-адресов. Поисковый аппарат Scopus интегрирован с поисковой системой Scirus для поиска веб-страниц и позволяет экспортировать до 20 000 записей, в том числе в формате «CSV Excel» (рис. 2).

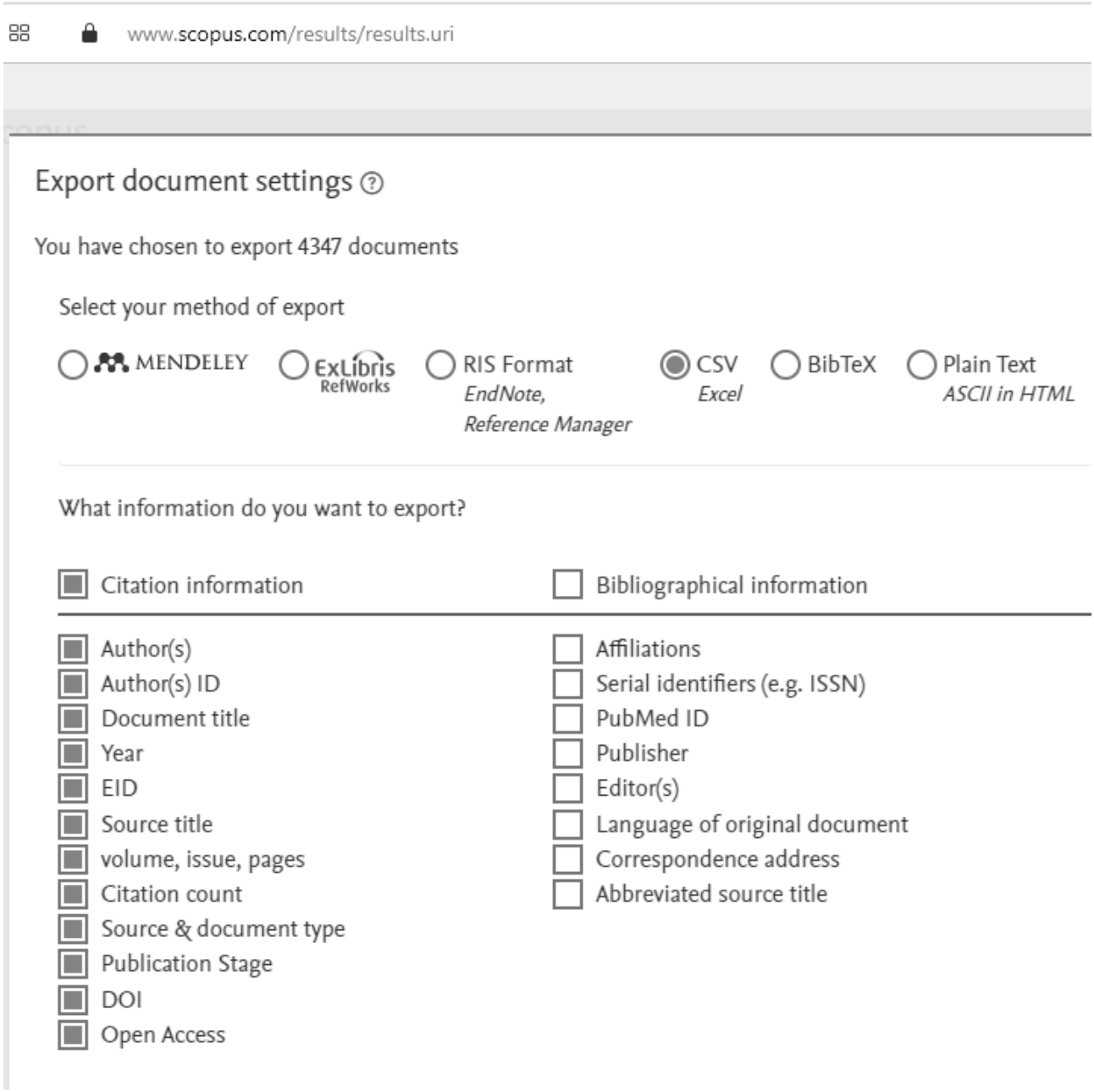

Рис. 2. Окно экспорта сведений о публикациях из базы Scopus в формат «csv»
Fig. 2. Web-interface for exporting information about articles from the Scopus database into the "csv" format

Web of Science Core Collection (WoS) [Clarivate, 2021] – поисковая интернет-платформа, объединяющая реферативные базы данных публикаций в научных журналах и патентов, в том числе базы, учитывающие взаимное цитирование публикаций. В этой платформе предусмотрены возможности поиска и анализа библиографической информации и

управления ею, а также возможность экспорта записей в формате Excel (рис. 3). База данных Web of Science Core Collection также доступна через веб-интерфейс на условиях Национальной подписки.

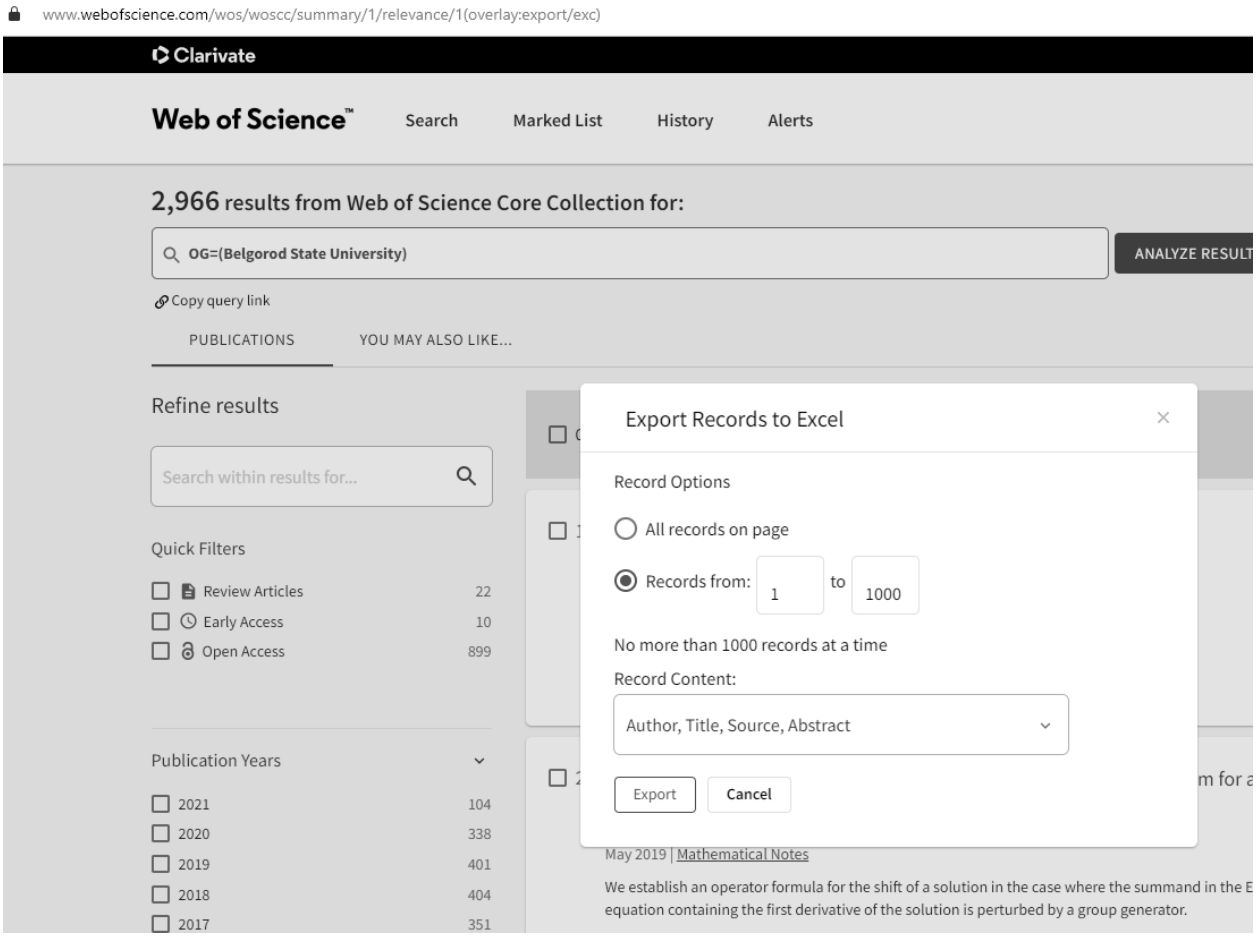

Рис. 3. Окно экспорта сведений о публикациях из базы Web of Science в формат «xls»
Fig. 3. Web-interface for exporting information about articles from the Scopus database into the Excel format

В результате экспорта получается два файла в формате «csv» и «xls». В табл. 2 представлены только значимые для решения задачи поля экспортных табличных файлов, то есть поля, определяемые структурой метаданных университетского репозитория согласно DC.

Таблица 2
Table 2

Названия и описания полей экспортных табличных файлов из баз Scopus и WoS
Names and descriptions of fields of export files from Scopus and WoS databases

Поля Scopus	Поля WoS	Описание
1	2	3
Authors	Author Full Names	Список авторов с разделителями
Title	Article Title	Заглавие публикации
Year	Publication Year	Год опубликования статьи
Source title	Source Title	Название журнала\издания
Volume	Volume	Номер тома журнала\издания
Issue	Issue	Номер выпуска журнала\издания

Окончание таблицы 2
End of the table 2

1	2	3
Art. No.	Article Number	Номер статьи в выпуске
Page start	Start Page	Номер начальной страницы публикации
Page end	End Page	Номер последней страницы публикации
EID	UT (Unique WOS ID)	Идентификаторы публикации в наукометрических базах
Abstract	Abstract	Аннотация

Все перечисленные в таблице поля и их содержимое войдет в состав результирующего табличного файла при подготовке архива для пакетного импорта. Комплект полных текстов научных публикаций представлен набором pdf-файлов оригинал-макетов, имена которых состоят из связки «Заглавие публикации» и «Название источника» на английском языке, причем слова в именах файлов разделены символом «_» (нижнее подчеркивание).

Описание алгоритмов и особенностей реализации программных инструментов

Схема, описывающая общий укрупнённый алгоритм подготовки и импорта в институциональный репозиторий метаданных о публикациях из баз Scopus и WoS, представлена на рис. 4.

Рис. 4. Общий алгоритм подготовки и импорта метаданных о публикациях
Fig. 4. General algorithm for preparing and importing articles metadata

Ниже описываются алгоритмы подпроцессов, а также программные инструменты для реализации каждого из этапов процесса формирования архивного файла для пакетного импорта в репозиторий.

Особенность экспорта из WoS заключается в том, что за один раз имеется возможность экспортировать не более одной тысячи записей о публикациях, поэтому производить экспорт при наличии подобных ограничений рекомендуется раздельно по годам, с последующим объединением экспортных файлов в единую таблицу (шаг 3 общего алгоритма).

Для автоматического объединения экспортных файлов из баз Scopus и WoS с исключением дублирующихся записей (шаг 4 общего алгоритма) был разработан скрипт на языке Python, алгоритм которого представлен на рис. 5.

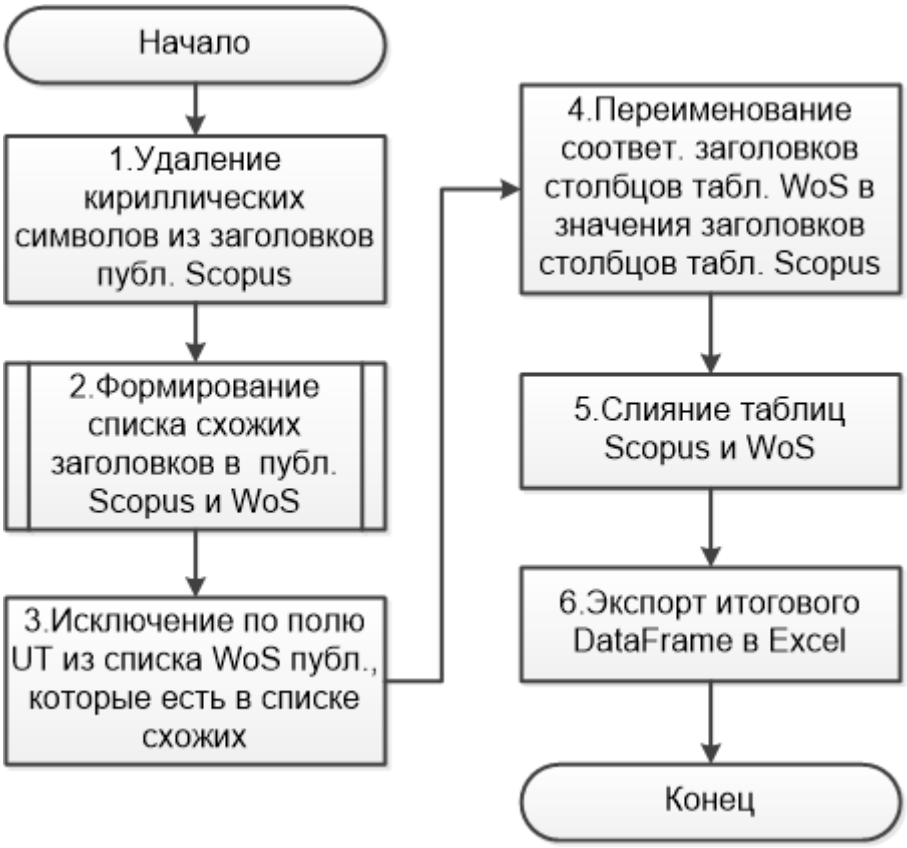

Рис. 5. Блок-схема алгоритма объединения файлов экспорта из баз Scopus и WoS в один сводный файл
Fig. 5. Flowchart of the algorithm for combining export files from bases Scopus and WoS into single table

Открытие и чтение экспортных файлов осуществляется методами «read_csv» и «read_excel» библиотеки «pandas», для работы которых необходима дополнительная библиотека «xlrd». Зачастую в таблице-экспорте из базы Scopus названия статей дублируются с названиями на национальных языках, что может привести к невозможности корректно объединить данные с экспортом WoS, исключив дубли, из-за наличия существенных отличий в написании заголовков. Поэтому, прежде чем производить сравнение названий публикаций по столбцу «Title» таблиц-экспортов из баз Scopus и WoS, необходимо удалить из этих столбцов кириллические символы. Для этого используется следующая скриптовая конструкция в виде лямбда-функции на основе следующего регулярного выражения библиотеки «Re» [Rachum, 2021]:

```
df['Title'] = df['Title'].apply(lambda x: re.sub('\s+', ' ', re.sub('[А-Яа-я]',
'', x)).strip())
```

Само сравнение заголовков публикаций производится посредством метода «ratio» библиотеки нечеткого сравнения «fuzzywuzzy» [Bicking, Leidel, 2021], при этом в качестве уровня «похожести» субъективно выбирается уровень не ниже 90 % сходства заголовков. Исходный код скрипта для объединения экспортных файлов из баз Scopus и WoS приведен в прил. 1.

В результирующем файле также необходимо сделать ряд последовательных преобразований данных:

1) сформировать новый столбец «Заглавие+Источник», значения в котором получены путем сцепления значений столбцов «Title» и «Source» (шаг 5 общего алгоритма);

2) получить список имен файлов с pdf-макетами публикаций, заменить в каждом имени файла символ «_» (нижнее подчеркивание) на « » (пробел) (шаг 6 общего алгоритма). Получить список можно используя функционал бесплатной программы Advanced Renamer (рис. 6);

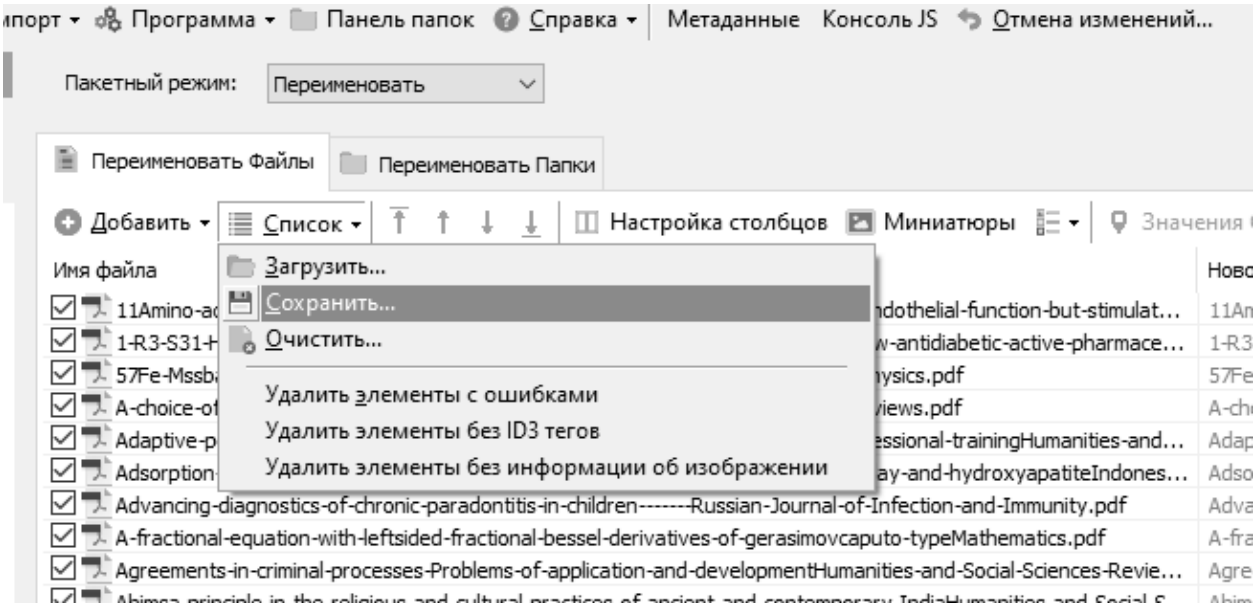

Рис. 6. Окно утилиты «Advanced Renamer» с функционалом извлечения списка файлов в папке
Fig. 6. Tool "Advanced Renamer" with the functionality of extracting a list of files in a folder

3) сопоставить получившиеся имена файлов со значениями из столбца «Заглавие+Источник», используя модифицированный алгоритм определения схожести значений, разработанный на основе ранее созданного кода. В качестве критерия уровня схожести также использован субъективный уровень в 90 % сходства (шаг 7 общего алгоритма);

4) используя функцию ВПР [Bruns, 2021], соотнести сопоставленные имена файлов с соответствующими записями в сводном экспортном файле. В качестве критерия сопоставления используются значения из поля «Заглавие+Источник». В результате имена оригинальных файлов соотносятся с соответствующими идентификаторами публикаций в базах Scopus или WoS (поля «EID» или «UT» соответственно);

5) в имена файлов посредством автозамены вместо пробела в качестве разделителя слов возвращается символ «_» (нижнее подчеркивание), а длина имен файлов усекается до 40 символов, чтобы не возникало потенциальных проблем совместимости размера имен файлов с разными файловыми системами (шаг 8 общего алгоритма). При этом, если при усечении образуются одинаковые заглавия, то к каждому такому заглавию вручную добавляется дополнительный символ, например, цифра-счетчик;

6) на основе получившегося списка с модифицированными именами фалов pdf-макетов посредством функционала все той же программы Advanced Renamer производится пакетное переименование исходных файлов.

Следующим этапом является сравнение подготовленных сводных данных со сведениями о публикациях, которые ранее уже были загружены в репозиторий сотрудниками научно-библиографического консультационного центра (шаги 9–10 общего алгоритма). В репозитории университета статьи, входящие в реферативные базы Scopus и WoS, находятся в отдельной коллекции и могут быть выгружены штатными средствами DSpace (рис. 7).

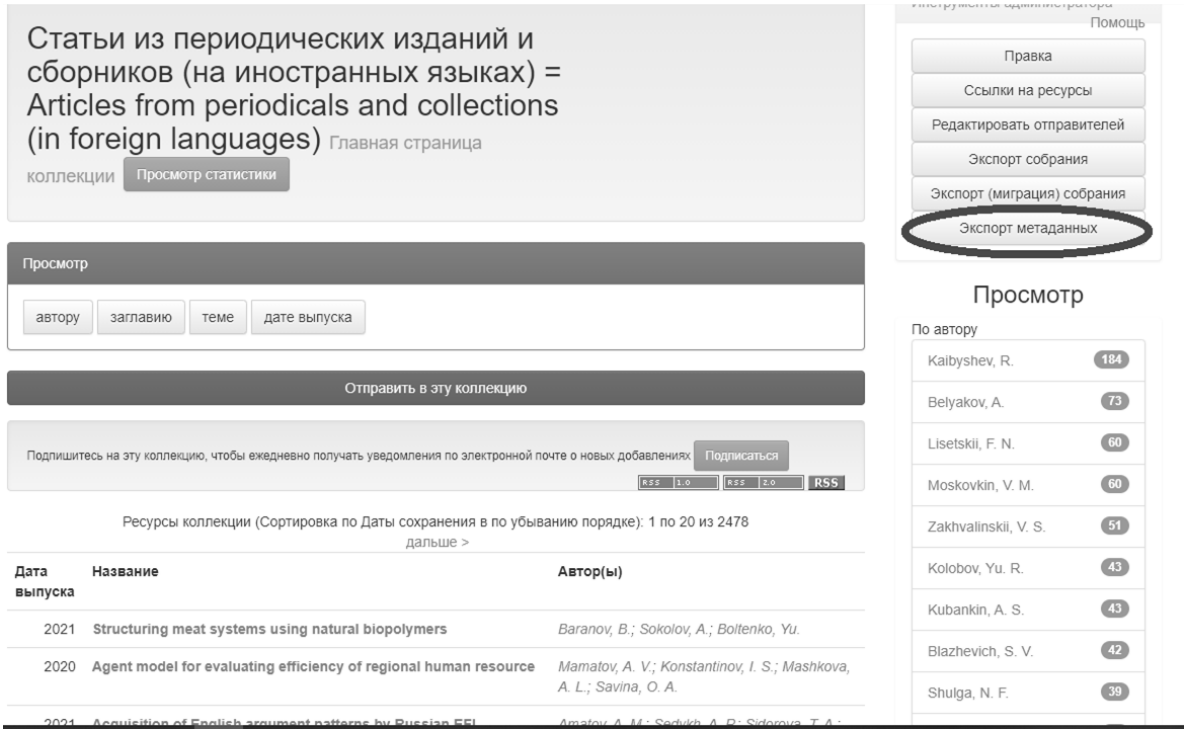

Рис. 7. Окно выгрузки метаданных коллекции из университетского репозитория
Fig. 7. Web-interface for unloading collection metadata from DSpace repository

Выявление дубликатов происходит посредством все той же подпрограммы определения похожести заголовков, степень похожести при этом субъективно выбрана на уровне 80 %, чтобы охватить более существенные различия в написании заголовков публикаций. Особенностью именно этого алгоритма сравнения является необходимость предварительного приведения символов в значениях поля «Title» в выгрузке из коллекции к среднему уровню верхнего и нижнего индекса, а также приведение значений полей «Title» сводного файла и экспорта к единому (верхнему) регистру. Описанный функционал реализуется посредством следующего Python-скрипта:

```
SUB = str.maketrans("₀₁₂₃₄₅₆₇₈₉", "0123456789", )
dfw1['dc.title[ru]'] = dfw1['dc.title[ru]'].str.translate(SUB)
dfw1['dc.title[ru]'] = dfw1['dc.title[ru]'].str.upper()
dfs1['dc.title'] = dfs1['dc.title'].str.upper()
```

Так как в экспортных таблицах из баз Scopus и WoS форматы записи имен соавторов публикаций имеют некоторые отличия в части использования разделителей между фамилией и инициалами соавторов, а также следуя требованию DC, для приведения списка соавторов к формату «Фамилия, И. О.» необходимо в сводной таблице средствами Excel применить к столбцу «Authors» следующую последовательность автозамен (шаг 11 общего алгоритма):

1) заменить последовательность символов «.,» (точка и запятая) на символ «%»;

2) заменить символ «.» (точка) на последовательность символов «. » (точка и пробел);

3) заменить символ «%» на последовательность символов «.,» (точка и запятая);

4) заменить последовательность символов « ;» (пробел и точка с запятой) на «;»;

5) заменить последовательность символов «; » (точка с запятой и пробел) на последовательность «.;» (точка и точка с запятой);

6) заменить последовательность символов «..» (две точки) на «.» (одну точку);

7) заменить последовательность символов «., » (точка, запятая и пробел) на последовательность «.;» (точка и точка с запятой);

8) заменить два подряд идущих пробела на один.

Для разбиения значений строк, содержащих списки соавторов публикаций, и распределения соавторов по столбцам согласно структуре, представленной в табл. 1, применяется функция Excel «Текст по столбцам», находящаяся на вкладке «Данные» ленты [Weterings, 2021]. При этом в качестве разделителя используется символ «;» (точка с запятой).

В итоге выполнения всех операций итоговый файл в формате таблицы Excel имеет вид, представленный на рис. 8.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	dc.title	dc.date	dc.ider	dc.ider	dc.ider	dc.ider	dc.ider	dc.des	dc.lang	dc.type	dc.righ	filenan	dc.identifier.citati	dc.cont	dc.cont	dc.cont	dc
2	Tempformi	2020	Metals	10	12	1	20	The micros	en	Review	2-s2.0-850	Tempformi	Tempforming as an ad	Dolzhenko	Kaibyshev	Belyakov, A.	
3	Microstruc	2020	Metals	10	12	1	18	The micros	en	Article	2-s2.0-850	Microstruc	Microstructural change	Odnobokov	Belyakov,	Enikeev, N	Ka
4	Understan	2020	Global Ecc	24				Analysis o	en	Article	2-s2.0-850	Understan	Understanding global s	Lamchin, M	Wang, S. '	Lim, C. H.	Oc
5	Exceptiona	2020	Scientific F	10	1			Ti-rich bod	en	Article	2-s2.0-850	Exceptiona	Exceptionally high stre	Eleti, R. R	Klimova, M	Tikhonovsk	Ste
6	Functional	2020	Scientific F	10	1			Functional	en	Article	2-s2.0-850	Functional	Functional lateralizatic	Artemenkc	Sitnikova,	Soltanlou,	Dre
7	Two-dimen	2020	Crystals	10	11	1	12	Features ir	en	Article	2-s2.0-850	Twodimens	Two-dimensional surfa	Zakhvalins	Nikulichev	Pilyuk, E.	Iva
8	The cytoge	2020	Internation	21	21	1	13	Mechanisr	en	Article	2-s2.0-850	The_cytog	The cytogenomic "the	Iourov, I. Y	Vorsanova	Yurov, Y. E	Zel
9	On the fati	2020	Materials	13	19			This work	en	Article	2-s2.0-850	On_the_fat	On the fatigue perform	Malopheye	Vysotskiy,	Zhemchuz	Mir
10	Peptides: I	2020	Molecules	25	19			There is a	en	Review	2-s2.0-850	Peptides_I	Peptides: Prospects fo	Khavinson	Linkova, N	Dyatlova, A	Ku
11	Non-hemat	2020	Research I	6	3	75	86	Relevance	en	Article	2-s2.0-850	Nonhemat	Non-hematopoietic ery	Belyaeva,	Stepenko,	Lyubimov,	Ku
12	Review of a	2020	Research I	6	3	1	5	General as	en	Review	2-s2.0-850	Review_of_	Review of a new conce	Dolzhikov,	Shevchenk	Pobeda, A	Pe
13	Erythropoi	2020	Internation	21	18	1	20	Preeclamp	en	Article	2-s2.0-850	Erythropoi	Erythropoietin mimetic	Korokin, M	Gureev, V.	Gudyrev, C	Go
14	(1 R,3 S)-3	2020	Acta Cryst	76		1407	1411	The chiral	en	Article	2-s2.0-850	1_R3_S31	(1 R,3 S)-3-(1 H -Benz	Kovalenko	Konovalova	Merzlikin,	Ch
15	Dataset of	2020	Data in Bri	31				Data on th	en	Data Pape	2-s2.0-850	Dataset_of	Dataset of allele, geno	Eliseeva, M	Ponomarer	Reshetnik	Po
16	Mechanisr	2020	Crystals	10	7	1	16	The as-que	en	Article	2-s2.0-850	Mechanisr	Mechanisms of grain s	Panov, D.	Dedyulina,	Shaysultar	Ste
17	Microstruc	2020	Metals	10	7	1	12	The preser	en	Article	2-s2.0-850	Microstruc	Microstructural charac	Mironov, S	Sato, Y. S	Kokawa, H	Hir
18	Manageme	2020	Journal of I	243	3	285	292	Undergrou	en	Article	2-s2.0-850	Manageme	Management of harder	Golik, V. I	Dmitrak, Y	Komashch	Ka
19	Socio-ecor	2020	E3S Web	175				The paper	en	Conferenc	2-s2.0-850	Socioecon	Socio-economic aspe	Samarina,	Samarin, A	Skufina, T.	
20	Digitalizati	2020	E3S Web	176				The article	en	Conferenc	2-s2.0-850	Digitalizati	Digitalization of the ag	Poletaev, A	Narozhnya	Kitov, M.	
21	Using GIS	2020	E3S Web	176				The article	en	Conferenc	2-s2.0-850	Using_GIS	Using GIS technology	Buryak, Z.	Marinina, O.		
22	The influen	2020	Materials	13	12	1	23	Nanocryst	en	Article	2-s2.0-850	The_influer	The influence of co ad	Goldberg,	Obolkina,	Smirnov, S	Pr
23	Daily asse	2020	Journal of I	20	4	1673	1680	The purpos	en	Article	2-s2.0-850	Daily_asse	Daily assessment of p	Kondakov,	Voloshina,	Kopeikina,	Ka
24	On the wa	2020	Research I	6	2	1	7	The corona	en	Review	2-s2.0-850	On_the_wa	On the way from SAR	Soldatov, \	Kubekina,	Silaeva, Y.	Bru
25	On the stre	2020	Materials	13	9			The ultrafir	en	Article	2-s2.0-850	On_the_st	On the strength of a 3	Odnobokov	Yanushkev	Kaibyshev	Be
26	Dental con	2020	Polymers	12	5			A modifier	en	Article	2-s2.0-850	Dental_cor	Dental composition mo	Chistyakov	Kolpinskay	Posokhova	Ch
27	Dataset of	2020	Data in Bri	29				Data on th	en	Data Pape	2-s2.0-850	Dataset_of	Dataset of allele, geno	Belyaeva,	Ponomarer	Reshetnik	Po
28	Sustainabl	2020	E3S Web	159				The article	en	Conferenc	2-s2.0-850	Sustainabl	Sustainable developm	Sapryka, \	Shmigirilov	Vavilov, A.	Pa
29	Peculiaritie	2020	E3S Web	159				The article	en	Conferenc	2-s2.0-850	Peculiaritie	Peculiarities of urban y	Babintsev,	Gaidukova	Ushamirsk	Sh
30	Special fea	2020	E3S Web	159				The purpos	en	Conferenc	2-s2.0-850	Special_fe	Special features of cor	Bondarenk	Panaedova	Gurieva, L.	Be
31	Correction	2020	Research I	6	1	29	40	Introductio	en	Article	2-s2.0-850	Correction_	Correction of morphofu	Lokteva, T	Rozhkov, I	Gureev, V.	Gu
32	Retinoprot	2020	Biology	9	3			An importa	en	Article	2-s2.0-850	Retinoprot	Retinoprotective effect	Peresypkir	Pazhinsky	Danilenko,	Lu
33	Sugar beet	2020	Climate	8	3			The weath	en	Article	2-s2.0-850	Sugar_bee	Sugar beet harvests u	Lebedeva,	Lupo, A. R	Solovyov, A	Ch
34	Dataset of	2020	Data in Bri	28				Data on th	en	Data Pape	2-s2.0-850	Dataset_of	Dataset of allele and g	Reshetnik	Abramova,	Ponomarer	Po
35	Prevalence	2020	Internation	12	1	606	611	The acade	en	Article	2-s2.0-850	Prevalence	Prevalence and dynam	Ruzhenko	Ruzhenko	Rzhevskay	Mc
36	The health	2020	Internation	12	1	624	629	The axiolo	en	Article	2-s2.0-850	The_health	The health in the value	Vangorods	Babintsev,	Shmarion,	Ko
37	Great patri	2020	Internation	12	1	594	598	The paper	en	Article	2-s2.0-850	Great_patr	Great patriotic war 194	Lebedev, S	Shapovalo	Kisilenko,	Ko
38	Rex Eris S	2020	Internation	12	1	612	617	The proble	en	Article	2-s2.0-850	Rex_Eris_	Rex Eris Si Recte Fac	Penskaya,	Lopin, R. A	Lykov, E. I	No
39	Attitude to	2020	Internation	12	1	599	605	The stigma	en	Article	2-s2.0-850	Attitude_to	Attitude to diseases a	Rzhevskay	Ruzhenko	Ruzhenko	Re
40	Social cog	2020	Internation	12	1	618	623	Relevance	en	Article	2-s2.0-850	Social_cor	Social cognition and it	Shvets, K.	Ruzhenko	Ruzhenko	Rz

Рис. 8. Содержимое итогового сводного файла для формирования архива для импорта
Fig. 8. Content of the summary file for forming an archive for import

Заключение

В статье был разработан алгоритм подготовки сведений для пакетного импорта метаданных о публикациях из реферативных баз Scopus и WoS в институциональный репозиторий на платформе DSpace, а также разработан программный инструментарий для реализации некоторых из этапов этого алгоритма. Итоговый табличный файл следует преобразовать в формат «csv» средствами того же Excel (шаг 12 общего алгоритма). Так как функционал пакетного импорта в репозиторий DSpace работает с архивами формата SAF, то необходимо воспользоваться готовой утилитой для сборки и генерации подобного архива [Dietz, 2015]. Ниже приводится скрипт для оболочки командной строки Windows для формирования требуемого SAF-архива.

```
java -jar safbuilder-1.6.jar -c e:\temp\toDS\_bsu.csv -z
```

После выполнения данного скрипта в состав SAF-архива включаются файлы с метаданными о публикациях и файлы-макеты публикаций. Процесс пакетного импорта итогового SAF-архива показан на рис. 9.

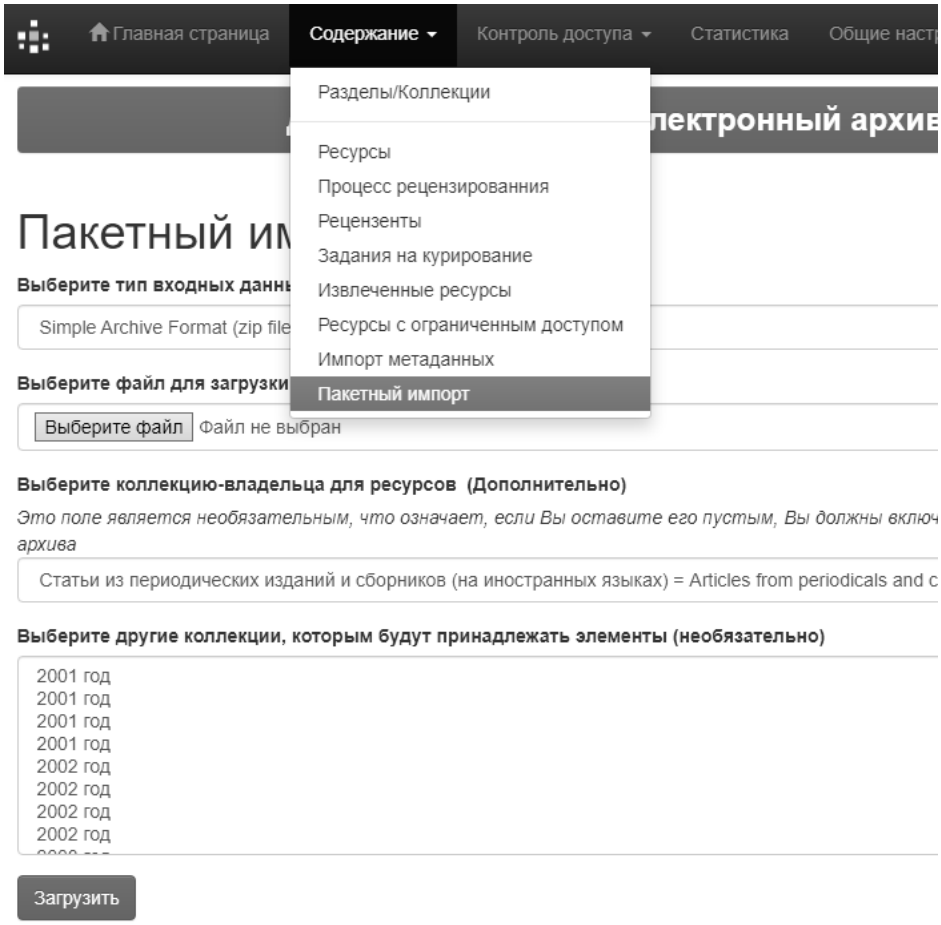

Рис. 9. Окно пакетного импорта SAF-архива с метаданными о публикациях
Fig. 9. Web-interface for batch import SAF-file with articles metadata to DSpace

Сотрудники научно-библиографического консультационного центра затрачивают на внесение метаданных одной публикации от 5 минут. Сравнительные данные о временных затратах на ручное внесение метаданных или их автоматизированную подготовку и пакетный импорт представлены в табл. 3.

Таблица 3
Table 3

Данные о временных затратах на внесение сведений о метаданных при их ручном и автоматизированном внесении
Times spent on manual and automatically entering information about articles metadata

№ п\п	Содержание этапа	Время, затраченное на ручной ввод, сек.	Время, затраченное на автоматизированный ввод, сек.
1	2	3	4
1	Добавление метаданных одной публикации	360×337	-
2	Экспорт данных из баз Scopus и WoS		300
3	Объединение выгрузок из баз Scopus и WoS в единый свод		60
4	Преобразование ФИО соавторов		120

Окончание таблицы 3
End of the table 3

1	2	3	4
5	Формирование списка с именами файлов pdf-макетов		30
6	Сопоставление списка с именами файлов-макетов с данными свода		20
7	Выявление неверно сопоставленных файлов		600
8	Усечение и переименование имен файлов-макетов		60
9	Экспорт коллекции из DSpace		30
10	Сопоставление свода с экспортом коллекции		900
11	Распределение соавторов по столбцам		30
12	Заключительные преобразования, преобразования в формат «csv»		300
13	Импорт SAF-архива в репозиторий		1800
	Всего:	121320	4250

Результаты данного исследования демонстрируют почти двадцатидевятикратное сокращение временных затрат на подготовку и импорт данных в институциональный репозиторий DSpace при использовании разработанного программного инструментария, применении стандартных офисных приложений и специализированного бесплатного программного обеспечения, имеющегося в свободном доступе. На основе разработанных скриптов [Reznichenko, 2021] автор планирует создать приложение с графическим интерфейсом, которое бы в качестве входных данных использовало три экспортных файла, и, в результате работы, формировало бы готовый сводный Excel-файл с метаданными о публикациях, пригодный для конвертации в формат «csv» и создания конечного SAF-архива для пакетного импорта в DSpace.

Referens

1. Clarivate Analytics Web of Science. Available at: https://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=C3Qtws6Zp9bRCWtj7S7&preferencesSaved= (accessed 2 June 2021)

2. Deng Sai. 2010. Optimizing Workflow through Metadata Repurposing and Batch Processing. Journal of Library Metadata, 10(4): 219-237. Available at: https://www.tandfonline.com/doi/abs/10.1080/19386389.2010.524862 (accessed 2 June 2021). DOI: 10.1080/19386389.2010.524862

3. Dietz Peter. 2015. Simple Archive Format Packager. Available at: https://wiki.lyrasis.org/display/DSPACE/Simple+Archive+Format+Packager (accessed 2 June 2021)

4. DuraSpace DSpace – A Turnkey Institutional Repository Application. Available at: https://duraspace.org/dspace/ (accessed 2 June 2021)

5. Dublin Core™ Metadata Initiative. Available at: http://dublincore.org (accessed 2 June 2021)

6. Elsevier Scopus. Available at: https://www.scopus.com/search/form.uri?display=basic= (accessed 2 June 2021)

7. Bruns Dave. 2021. EXCELJET. Quick, clean, and to the point. Excel VLOOKUP Function. Available at: https://exceljet.net/excel-functions/excel-vlookup-function (accessed 2 June 2021).

8. Fedotova O.A., Fedotov A.N., Zhizhimov O.L., Sambetbayeva M.A. 2020. DIGITAL REPOSITORY FOR RESEARCH AND EDUCATION INFORMATION SYSTEMS. Proceedings of SPSTL SB RAS, 3: 23-28. Available at: https://proceedings.gpntbsib.ru/jour/article/view/7 (accessed 2 June 2021). DOI: 10.20913/2618-7515-2019-3-23-28 (in Russian)

9. Bicking Ian, Leidel Jannis. 2021. fuzzywuzzy PyPI. Available at: https://pypi.org/project/fuzzywuzzy/ (accessed 2 June 2021)

10. Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M. 2020. Metadata Normalization Methods in the Digital Mathematical Library. CEUR Workshop Proceedings, 2543: 136–148. Available at: http://ceur-ws.org/Vol-2543/rpaper13.pdf (accessed 2 June 2021)

11. Kim Jensen. 2021. Advanced Renamer. Batch file renaming utility for Windows. Available at: https://www.advancedrenamer.com (accessed 2 June 2021)

12. JetBrain PyCharm: The Python IDE for Professional Developers. Available at: https://www.jetbrains.com/pycharm/ (accessed 2 June 2021)

13. Nash Jacob L., Wheeler Jonathan. 2016. Desktop Batch Import Workflow for Ingesting Heterogeneous Collections: A Case Study with DSpace 5. D-Lib Magazine, 22 (1–2). Available at: http://www.dlib.org/dlib/january16/nash/01nash.html (accessed 2 June 2021). DOI: 10.1045/january2016-nash

14. OpenDOAR. Browse by Country and Region. Available at: https://v2.sherpa.ac.uk/view/repository_by_country/Russian_Federation.software_name.html (accessed 2 June 2021)

15. Oracle Java SE Runtime Environment 8. Available at: https://www.oracle.com/java/technologies/java-se-glance.html (accessed 2 June 2021)

16. Wood Andrew. 2021. pandas.DataFrame. Available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html (accessed 2 June 2021)

17. Rachum Ram. 2021. re – Regular expression operations. Available at: https://docs.python.org/3/library/re.html (accessed 2 June 2021)

18. Registry of Open Access Repositories. Available at: http://roar.eprints.org/cgi/roar_search/advanced?location_country=ru&software=&type=&order=-recordcount%2F-date (accessed 2 June 2021)

19. Weterings Niels. 2021. Text to Columns – Easy Excel Tutorial. Available at: https://www.excel-easy.com/examples/text-to-columns.html (accessed 2 June 2021)

20. Walsh Maureen P. 2010. Batch Loading Collections into DSpace: Using Perl Scripts for Automation and Quality Control. Information Technology and Libraries 29, no. 3 (2010): 117–127. Available at: https://ejournals.bc.edu/index.php/ital/article/view/3137 (accessed 2 June 2021). DOI: https://doi.org/10.6017/ital.v29i3.3137

21. What is Power Query? Available at: https://powerquery.microsoft.com/en-us/ (accessed 2 June 2021)

22. Reznichenko Oleg. 2021. Appendix to article "Preparation articles metadata for batch import into DSpace repository" Available at: https://github.com/leo-phoenix/dspace_batch_import (accessed 2 June 2021)

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest related to this article was reported.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Резниченко Олег Сергеевич, старший преподаватель кафедры прикладной информатики и информационных технологий института инженерных и цифровых технологий НИУ «БелГУ», г. Белгород, Россия

INFORMATION ABOUT THE AUTHOR

Oleg S. Reznichenko, Senior Lecturer of the Department of Applied Information Science and Information Technologies, Institute of Engineering and Digital Technologies, Belgorod State University, Belgorod, Russia