Автоматизация наполнения контента репозитория JINR DOCUMENT SERVER

Т.Н. Заикина[1,a], О.В. Егорова[2], С.В. Куняев[1], Ж.Мусульманбеков[1], Р.Н. Семенов[1], П.В. Устенко[1], И.А. Филозова[1,2], Г.В. Шестакова[1]

[1]Объединенный институт ядерных исследований,
141980, г. Дубна Московской области, ул. Жолио-Кюри 6

[2]Университет «Дубна», 141980 г. Дубна Московской области, ул. Университетская, 19

E-mail: [a] ztanya@jinr.ru

Сервер Документов ОИЯИ (JDS - jds.jinr.ru) был создан и развивается как институциональный репозиторий Открытого Доступа статей, препринтов и других материалов, отражающих и содействующих научно-исследовательской деятельности в ОИЯИ. JDS учитывает все возможные аспекты управления электронной библиотекой и развернут на базе программного обеспечения Invenio. На сервере JDS также представлено много медиаматериалов, таких как видеолекции для молодых ученых, постеры, аудиолекции, новости об ОИЯИ. В будущем JDS рассматривается как часть корпоративной информационной системы ОИЯИ.

Актуализация контента такого хранилища является трудоемкой работой. Данная статья посвящена автоматизации сбора, предварительной обработки и ввода данных в информационную систему JDS.

Ключевые слова: репозиторий Открытого Доступа, автоматизация, библиографические данные, авторитетные записи, web-scraping.

Сервер документов ОИЯИ — JINR Document Server

В ОИЯИ создан и развивается Сервер документов ОИЯИ — JINR Document Server (JDS – jds.jinr.ru). Целями JDS являются хранение информационных ресурсов ОИЯИ и предоставление эффективного доступа к ним, повышение уровня информационной поддержки сотрудников ОИЯИ, предоставление доступа к другим научным архивам, оценка эффективности научной деятельности ОИЯИ [Filozova, Musulmanbekov,…, 2013; Zaikina, Musulmanbekov,…, 2012].

JDS содержит статьи, препринты, видеоматериалы с лекциями для молодых ученых, постеры, аудиолекции, новости ОИЯИ и другие материалы, отражающие и содействующие научно-исследовательской деятельности в ОИЯИ [Zaikina, Musulmanbekov,…, 2011].

Основными источниками наполнения репозитория являются: arXiv.org, SPIRES, ADS, MathSciNet, CERN Document Server, система персональной информации о сотрудниках ОИЯИ, материалы издательского отдела ОИЯИ. Наполнения и актуализация информации осуществляется посредством сбора данных в полуавтоматическом режиме и загрузки документов пользователями и контент-менеджерами.

Автоматизация наполнения контентом репозитория

Поддержка и наполнение контентом репозитория требует серьезных временных и людских ресурсов. Большие временные затраты, большая доля ручного ввода и ошибки операторского ввода приводят к низкой эффективности сбора и загрузки данных в систему. В связи с этим возникает задача максимальной автоматизации заполнения информационной системы.

Все метаданные в репозитории представлены в формате MARCXML [MARC Standarts], следовательно, задача автоматизации наполнения контентом сводится к задаче формирования MARCXML файла.

Загрузка информации в систему JDS возможна в двух режимах: 1) внесение через web-интерфейс (используется для 1 записи); 2) пакетная загрузка (для набора записей). В первом случае пользователь самостоятельно вводит (самоархивирование) данные в репозиторий, и система формирует запись в формате MARCXML. В пакетном режиме контент-менеджер загружает уже сформированный файл в формате MARCXML, содержащий весь необходимый набор метаданных.

Поскольку в будущем планируется интеграция JDS в корпоративную информационную систему ОИЯИ, как блок CRIS & OAR (Current Research Information System and Open Access Repository), необходимо обеспечить эффективное повторное использование данных о сотрудниках, проектах, грантах и т.д. [Filozova, Bashashin,…, 2016]. Для этого разрабатывается процедура управления коллекциями авторитетных записей.

Авторитетные / нормативные записи — это особый вид метаданных, представляющий собой поисковые элементы библиографических записей (имена авторов и персоналий, наименования организаций, и т. д.), представленные по определенным правилам. Авторитетные записи однозначно идентифицируют объекты и понятия, что позволяет с одной стороны, учесть при поиске информации все варианты наименований, а с другой — разграничить одинаково названные [Zaikina, Filozova,…, 2013; Zaikina, Kunyaev,…, 2014].

В рамках этой деятельности на тестовом сервере JDS-TEST3, развернутом на облачном сервере [Baranov, Balashov,…, 2016], ведутся работы по созданию коллекций авторитетных записей. Созданы и протестированы коллекции Персоналии, Организации, Журналы, Рубрики, Темы и проекты, Эксперименты, Гранты.

К настоящему моменту разработаны приложения для автоматической загрузки библиографических описаний препринтов, авторитетных записей об авторах-сотрудниках ОИЯИ и грантах с участием ОИЯИ.

Автоматическая загрузка препринтов

Для загрузки Препринтов ОИЯИ с сайта Издательского отдела было разработано приложение (рис. 1), которое формирует MARCXML файл, полностью готовый к загрузке через web-интерфейс в пакетном режиме (batch upload). Применение данного приложения значительно упрощает работу контент-менеджера и позволяет нажатием одной кнопки загрузить сразу достаточно большое количество записей.

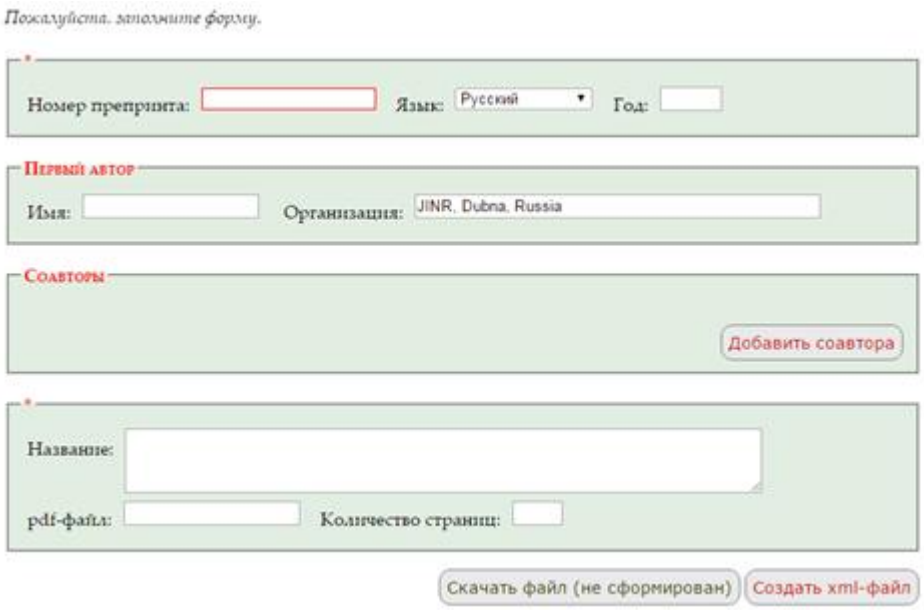

Рис. 1. Веб-интерфейс приложения для загрузки Препринтов ОИЯИ

Автоматическая загрузка авторитетных записей

Создан комплекс модулей для идентификации авторов-сотрудников ОИЯИ, представления их в метаданных, и установления связи между публикациями и авторами (PYTHON, HTML5, PHP). На текущий момент данное ПО используется как back-end решение для извлечения из метаданных публикаций имен авторов-сотрудников ОИЯИ, присвоения им уникального идентификатора и загрузки метаданных на тестовый сервер JDS-TEST3 (http://jds-test3.jinr.ru/).

Для автоматизации процедуры сбора информации о грантах и формирования полностью готового к загрузке файла было разработано приложение “GrantScrap” (рис.2) [Ryan Mitchell, 2015; Egorova, 2016].

Приложение “GrantScrap” выполняет следующие основные действия:

1. Scraping — извлечение данных о грантах с веб-страницы портала ОИЯИ;
2. обработка файлов;
3. формирование итогового MARCXML файла.

В ходе этапа scraping в файл сохраняется содержимое веб-страницы с грантами ОИЯИ, затем выделяются ссылки на гранты и создаются файлы с информацией, взятой со страниц по выделенным ссылкам. Обработка файлов происходит в цикле построчного считывания: поиск фрагментов текста по шаблонам для идентификации характеристик гранта (номер, руководитель и т.п.), формирование фрагмента MARCXML файла (chunk) для каждого MARC-поля и его запись в итоговый файл (рис. 3).

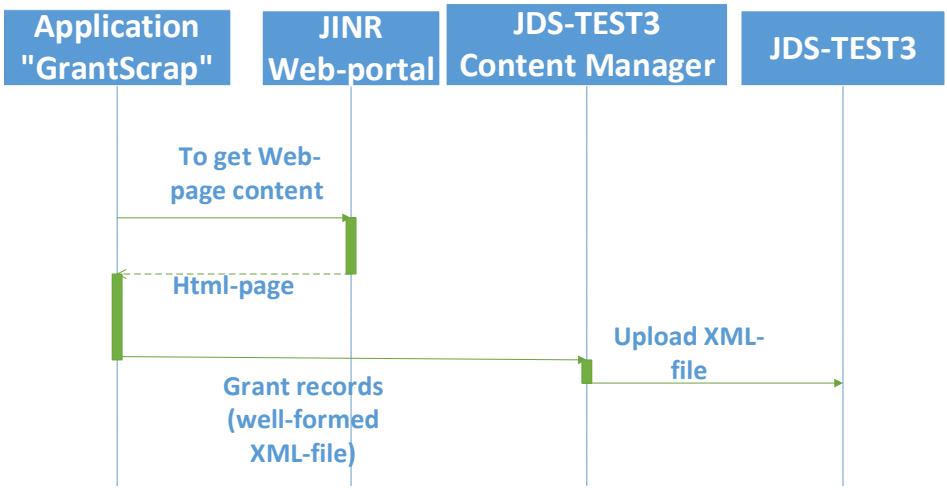

Рис.2. Диаграмма взаимодействия объектов

```
- <record>
  - <datafield ind2=" " ind1=" " tag="035">
      <subfield code="a">GRANT|(RuDuJINR)g00004</subfield>
    </datafield>
  - <datafield ind2=" " ind1=" " tag="088">
      <subfield code="a">14-02-00896</subfield>
      <subfield code="b">A</subfield>
      <subfield code="c">РФФИ</subfield>
    </datafield>
  - <datafield ind2=" " ind1=" " tag="110">
      <subfield code="a">Объединенный институт ядерный исследований</subfield>
      <subfield code="g">ОИЯИ</subfield>
      <subfield code="b">ЛАБОРАТОРИЯ ФИЗИКИ ВЫСОКИХ ЭНЕРГИЙ</subfield>
    </datafield>
  - <datafield ind2=" " ind1=" " tag="160">
      <subfield code="a">Корреляционные исследования холодной сверхплотной барионной материи</subfield>
      <subfield code="d">2014-2016</subfield>
    </datafield>
  - <datafield ind2=" " ind1=" " tag="906">
      <subfield code="a">Малахов А.И.</subfield>
    </datafield>
  - <datafield ind2=" " ind1=" " tag="980">
      <subfield code="a">GRANT</subfield>
    </datafield>
  - <datafield ind2=" " ind1=" " tag="980">
      <subfield code="a">AUTHORITY</subfield>
    </datafield>
</record>
- <record>
  - <datafield ind2=" " ind1=" " tag="035">
```

Рис. 3. MARCXML-file (фрагмент)

С помощью данного приложения были загружены метаданные о грантах РФФИ с участием ОИЯИ на тестовый сервер JDS-TEST3. Время выполнения приложения составило 9.409 сек секунд для обработки 56 записей. Расчёт затраченного времени на обработку и загрузку одной записи показал, что с помощью приложения контент-менеджеру на это требуется от 3.17 до 5.17 сек. с учетом его действий на web-интерфейсе при загрузке (навигация по меню, выбор файла). В ручном режиме на это затрачивается 2-3 минуты. Таким образом, сокращение временных ресурсов составляет примерно в 23-38 раз (120 / 5,17 ≈ 23; 120 / 3,17) [Egorova, 2016].

Заключение

В рамках работы по автоматизации наполнения контента репозитория JDS был разработан комплекс приложений:

- набор модулей для идентификации авторов-сотрудников ОИЯИ;
- приложения для сбора информации по грантам для авторитетных записей;
- приложение для формирования библиографических описаний препринтов ОИЯИ.

Данное ПО используется как back-end решение, позволяющее повысить эффективность загрузки данных в институциональный репозиторий JINR Document Server.

Список литературы

Baranov A.V., Balashov N.A., Kutovskiy N.A., Semenov R.N. JINR cloud infrastructure evolution. // Particles and Nuclei Letters, — 2016. — V.13, no 5. — PP.1046-1050.

Filozova I.A., Bashashin M.V., Korenkov V.V, Kuniaev S.V., Musulmanbekov G., Semenov R.N., Shestakova G.V., Strizh T.A., Ustenko P.V., Zaikina T.N. Concept of JINR Corporate Information System. // Particles and Nuclei Letters. — 2016. — V.13, no 5. — PP.980-985.

Filozova I.A., Musulmanbekov G, Semenov R.N., Shestakova G.V., Zaikina T.N. JDS: Digital Library of JINR Information Resources. Scientific Report 2012-2013. LIT JINR. 2013. [Electronic resource]: http://lit.jinr.ru/Reports/SC_report_12-13/p60.pdf.

MARC Standarts: MARC in XML. Library of Congress [Electronic resource]: URL:https://www.loc.gov/marc/marcxml.html.

Ryan Mitchell. Web Scraping with Python: Collecting Data from the Modern Web.– O'Reilly Media. 2015.

Zaikina T. N., Kunyaev S. V., Semenov R. N., Ustenko P. V., Filozova I. A., Shestakova G. V. Integration of JINR Scientific Information Resources on the JDS Platform. Proceedings of the 16th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections". — Dubna, JINR, 2014. — PP. 349-354, [in Russian].

Zaikina T.N., Filozova I.A. Monitoring and Statistics System for Impact Evaluation of Scientific Activities of OAI-repository JINR Document Server // Proceedings of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections". 2013. [Electronic resource]:http://rcdl.ru/doc/2013/paper/s9_2.pdf, [in Russian].

Zaikina T.N., Musulmanbekov G., Filozova I.A. Using Information Visualization at JINR Document Server // Proceedings of the 14th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections". 2012. [Electronic resource]: http://ceur-ws.org/Vol-934/paper22.pdf, [in Russian].

Zaikina T.N., Musulmanbekov G., Filozova I.A. Semenov R.N., Shestakova G. V. GUIDE FOR USERS of JINR DOCUMENT SERVER — Dubna: JINR, 2011 [in Russian].

Egorova O.V. Development of applications to collect information about grants with participation-of JINR // 66th International Student Scientific and technical conference, April 18-22, 2016 [electronic resource]: materials / Astrakhan. state. tehn. Univ. — Astrakhan: Publishing House of the Astrakhan State Technical University. — 2016. Access: 1 disk (CD-ROM), [in Russian].

The Automation of the Content Filling for JINR DOCUMENT SERVER

T.N. Zaikina[1,a], O.V. Egorova[2], I.A. Filozova[1,2], S.V. Kunyaev[1], G. Musulmanbekov[1], R.N. Semenov[1], G.V. Shestakova[1], P.V. Ustenko[1]

[1]Joint Institute for Nuclear Research, 6 Joliot-Curie, 141890, Dubna, Russia

[2] University "Dubna", 19 Universitetskaya, Dubna, Moscow reg., 141980, Russia

E-mail: [a] ztanya@jinr.ru

JINR Document Server (JDS – jds.jinr.ru), has been launched and developed as an institutional Open Access (OA) repository of articles, preprints and other materials that reflect and promote research activities at JINR. JDS possesses a digital library functionality which is provided by the Invenio software. JDS also includes collections of video lectures for young scientists, posters, audio lectures and news about JINR. In the future JDS is considered as a part of the JINR corporate information system.

The filling of the content for such a big repository is a hard work that takes a lot of time.

This article is dedicated to the automation of the content filling for JDS.

Keywords: Open Access Archives, automation, bibliographic data, authority records, web-scraping.