Корпоративные репозитории научных публикаций и проблемы обмена данными

Ковязина Елена Васильевна, кандидат технических наук, научный сотрудник, Институт вычислительного моделирования Сибирского отделения Российской академии наук, г. Красноярск, elena@icm.krasn.ru

Ранжирование научных и образовательных организаций, потребность учета публикационной и иной научной активности сотрудников породили необходимость взаимодействия библиотечных и наукометрических систем. Передача данных о публикациях в сводные отчеты, заимствование данных из внешних источников в репозитории научных публикаций организации, работа с едиными базами данных персоналий выявили множество проблем, связанных с несовместимостью форматов и отсутствием единых правил обмена данными.

Репозитории научных публикаций организации (IR - institutional repositories) получили повсеместное распространение в России и активное развитие в связи с необходимостью учета публикационной активности сотрудников, инициированной процессом ранжирования научных и образовательных организаций. Прилагаются максимальные усилия для обеспечения полноты и адекватности представленных в них данных о публикациях, для чего используются как встроенные сервисы программных систем, так и личные контакты. В результате данные организаций приобретают законченный и корректный вид. Для обмена данными между репозиториями различных организаций используется российский коммуникативный формат RUSMARC, или другие форматы семейства MARC, взаимная совместимость которых обеспечивается общими структурами данных и правилами заполнения полей. Для описания документов в Интернет используется также формат Dublin Core (DC) и его расширенная модификация MODS, поддержка которых, как правило, также обеспечивается разработчиками программного обеспечения. Однако потребность в данных репозиториев не ограничивается только библиографическими данными и полными текстами публикаций. Для обеспечения необходимой информативности записи данных обогащаются дополнительными полями, такими, как поле идентификаторов публикации в индексах научного цитирования (WoS, Scopus, РИНЦ и т.д.), поле текущего импакт-фактора источника публикации, поле списка библиографии, показатель текущего цитирования. Богатый набор полей MARC-форматов позволяет найти для каждого из дополнительных атрибутов подходящее по смыслу поле. Наличие коммуникативных форматов позволяет организовать обмен данными с помощью простейших конверторов данных с взаимно однозначным соответствием полей.

Однако заинтересованность научных и образовательных организаций в возможно более широком распространении результатов их исследовательской деятельности привела к тому, что массивы метаданных полнотекстовых документов репозитория организованы, как правило, не единственным образом, а, как минимум, в виде библиографической базы данных в Системе Автоматизации

библиотек (САБ) организации и в стандартах Архивов открытого доступа (ОА). Вариант одного из таких архивов представлен в [1]. Открытые архивы используют иные структуры данных, чем САБ. Для их описания используется конструкции в синтаксисе RDF/XML. Использование DC и MODS упрощает задачу, но все же, требуются большие усилия для обмена данными. Практические аспекты преобразования данных из MARC-форматов в DC, RDF и обратно отражены во множестве зарубежных публикаций, например, в [2-5].

Для организации дифференцированного доступа к данным репозитория необходима база персоналий, а данные о персоналиях также представлены двояким образом: в САБ – в виде базы данных читателей, в ОА – как база данных LDAP. Для минимизации объема работы также, как и в предыдущем случае, был бы полезен обмен данными. Потребности обмена данными между базами персоналий различных структур не так широко отражены в публикациях, однако это не означает, что эта задача не является актуальной, особенно в аспекте рассматриваемых далее CRIS-систем.

Данные репозиториев используются часто как составные части массивов данных различных информационных систем. В библиотеках академических институтов, в частности, это могут быть региональные и корпоративные информационные системы текущих исследований (CRIS - Current Research Information System), такие, например, как представленные в публикациях [6-8]. К системам этого вида можно отнести как локальные базы наукометрических данных организации, например, АРМ Ученого секретаря, различные коммерческие CRIS-системы, во множестве предлагаемые их производителями, например, [9-10], так и международные индексы научного цитирования. CRIS-системы формально создаются в соответствии со стандартом CERIF, определяющим структуры данных и связи между этими структурами [11-12]. Однако, ряд исследователей CRIS-систем отмечает плохую совместимость данных в различных таких системах, приводящую к серьезным сложностям обмена данными между ними [13].

На фоне указанных особенностей обмена данными внутри различных CRIS-систем, проблемы обмена данными между IR- и CRIS-системами выявлены достаточно давно и отражены в целом ряде публикаций [14-17]. По-видимому, широкое обсуждение этих проблем побудило крупнейших разработчиков CRIS-систем предпринять усилия для их решения. Так, набор сервисов Web of Science и Scopus позволяет заимствовать данные этих индексов и интегрировать их в САБ. Технология заимствования записей из WoS и Scopus широко применяется многими библиотеками научных организаций, например, [17-18]. В указанных публикациях отражена последовательность заимствования:

а) поиск с помощью текущих или хранимых запросов – по автору, месту работы, выделение нужных публикаций;
б) экспорт описаний в табличном или текстовом виде;
в) табличное преобразование данных в форму пригодную для интеграции в БД репозитория.
г) динамическое извлечение данных о цитировании и интеграция их в имеющиеся записи.

Неполная совместимость данных Web of Science и Scopus приводит к необходимости детальной проработки процедуры слияния извлеченных данных для устранения дублирования. В конечном итоге процесс слияния превращается в трехэтапную процедуру:

а) слияние по DOI;

б) слияние по свертке САБ;

в) визуальная проверка данных.

Набор указанных сервисов существенно облегчает работу наполнения базы данных репозитория и позволяет отслеживать показатели публикуемости организации в зарубежных источниках. Российский индекс научного цитирования (РИНЦ) также предоставил пользователям сервиса Science Index возможность извлечения данных, однако только в XML-виде.

В ряде зарубежных публикаций отражены особенности и способы практической реализации обмена данными между IR и CRIS-системами, например, [19-20]. Есть публикации о практической реализации моделирования данных о персоналиях и работе с ними в OAI [21]. Однако в целом, IR и CRIS все еще «история двух культур» [20], параллельное существование во многом пересекающихся массивов данных, с трудом взаимодействующих между собой.

Отметим в итоге, что ведение и обслуживание репозитория перестало быть обслуживанием только библиографических данных. Записи пополняются наукометрическими данными, данными о персоналиях. Да и сам факт существования обширных массивов данных о научных публикациях, связанных с их полными текстами, диктует логику преемственности данных – от репозитория к CRIS. Однако развитие систем сдерживается отсутствием единых стандартов обмена данными, которые заменяются многообразными конверторами, созданными инициативными группами «снизу» и изощренными методами извлечения данных CRIS-систем без какого-либо желания содействия их разработчиков. Результатом является бесконечное дублирование данных их непосредственными создателями, а также копирование и конвертация больших массивов из одной крупной системы в другую, и обратно.

Литература

1. Екабсоне, М. Open Access в Латвии / М. Екабсоне // Научная периодика: проблемы и решения. – 2014. - № 4 (22). – С. 4-8. - URL: http://nppir.ru/index.php/nppir/article/view/138/215.
2. IIik, V. Notes on Operations Metadata Makeover: Transforming MARC Records Using XSLT / V.IIik, J.Storlien, J.Olivarez // Library Resources & Technical Services. – 2013. - V.58. - № 3. – P. 187-208. – Режим доступа: http://journals.ala.org/lrts/article/viewFile/5262/6394.
3. IIik, V. Notes on Operations Metadata Makeover: Transforming MARC Records Using XSLT / V.IIik, J.Storlien, J.Olivarez // Library Resources & Technical Services. – 2013. - V.58. - № 3. – P. 187-208. – Режим доступа: http://journals.ala.org/lrts/article/viewFile/5262/6394.
4. Myntti, J. Authority Control in Digital Repository: Preparing for Linked Data / J.Myntti, N. Cothran // Journal of Library Metadata. - 2013. - 13:2-3. - P.95-113. - DOI: 10.1080/19386389.2013.826061. - Режим доступа: http://www.tandfonline.com/doi/full/10.1080/19386389.2013.826061#abstract.

5. Cole, T.W. Library MARC Records Into Linked Open Data: Challenges and Opportunities / T.W. Cole, M.-J. Han, W.F.Weathers, E. Joyner // Journal of Library Metadata. - 2013. - 13:2-3. - P.163-196. - DOI: 10.1080/19386389.2013.826074. – Режим доступа: http://www.tandfonline.com/doi/full/10.1080/19386389.2013.826074#abstract.
6. Копысов С.П. Интеграция информационных систем / С.П.Копысов, В.Н.Рычков // Вестник Уральского отделения РАН. – 2013. - № 1 (43). – С.44-50. – Режим доступа: http://www.iie-uran.ru/doc/43/44-50.pdf.
7. Жижимов О.Л., Федотов А.М., Шокин Ю.И. Платформа ZooSPACE - организация доступа к разнородным распределенным ресурсам // Электронные библиотеки: российский научный элек-тронный журнал. - 2014. - Т.17. - № 2. - ISSN 1562-5419. – Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2014/part2/ZFS.
8. Федотов, А.М. Модель информационной системы для поддержки научно-педагогической деятельности / А. М. Федотов, В. Б. Барахнин, О. Л. Жижимов, О. А. Федотова // Вестник НГУ. Серия Информационные технологии. – 2014. – Т.12. – Вып.1. С.89-101. – Режим доступа: http://www.nsu.ru/xmlui/bitstream/handle/nsu/1305/2014_V12_No1.pdf.
9. Касьянов П. CRIS-системы: для кого и для чего они существуют? [Электронный ресурс] / П.Касьянов. – Режим доступа: http://www.library.spbu.ru/blog/wp-content/uploads/2015/03/CRIS-systems.pdf .
10. Фатхуллин М. Samara State Aerospace University – Elsevier. Перспективы сотрудничества [Электронный ресурс] / М.Фатхуллин – Самара, 2014. - Режим доступа: http://lib.ssau.ru/uploaded/Publ/ease_recom.pdf .
11. Общеевропейский формат для исследовательской информации. CERIF-2004. [Электронный ресурс] . – Режим доступа: http://window.edu.ru/catalog/pdf2txt/904/37904/15711.
12. CERIF 1.5 XML. Data Exchange Format Specification. euroCRIS. - 13 Feb 2013. [Электронный ресурс]. Режим доступа: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF1.5_XML.pdf.
13. Pinto, K.S. CERIF – Is the standard helping to improve CRIS? / K.S.Pinto, C. Simoes, L. Amaral // Procedia Computer Science. – 2014. – 33. – P.80-85. Режим доступа: http://www.sciencedirect.com/science/article/pii/S1877050914008035#.
14. Кулагина, М.В. Научные информационные системы и электронные библиотеки. Потребность в интеграции / М.В.Кулагина, А.С.Лопатенко // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Сборник докладов Третьей Всероссийской конференции. RCDL`2001, Петрозаводск, 11-13 сентября 2001 г. - Карельский научный центр РАН, 2001. - С. 14-19. – Режим доступа: http://rcdl2001.krc.karelia.ru/papers/papers/kulagin_lopatenko/kulagin_lopatenko_paper.rtf.
15. Asserson, K. CRIS and Institutional Repositories / A. Asserson, K. Jeffery // Data Science Journal. - 2010. – V.9. - P.14-23. - Режим доступа: http://www.researchgate.net/profile/Keith_Jeffery/publication/45347066_CRIS_and_Institutional_Repositories/links/0c96052e8d9a85b7a7000000.pdf
16. IIva, J. Integrating CRIS and repository – an overview of the situation in Finland and in three other Nordic countries / J. IIva. – 2014. – Режим доступа: http://www.doria.fi/bitstream/handle/10024/97606/OR2014_CRIS%2Brepositories_in_Nordic_Countries-final.pdf?sequence=3.
17. Ковязина, Е.В. Российские библиотеки в "облаках": проблемы обмена данными [Электронный ресурс] / Е.В.Ковязина //Материалы XV российской конференции с международным участием "Распределенные информационные и вычислительные

ресурсы (DICR-2014)". - Новосибирск: ИВТ СО РАН, 2014 г. - CD-ROM. - ISBN 978-5-905569-07-4.
18. Баженов, С.Р. Интеграция базы данных публикаций организации с индексами научного цитирования: реализация средствами САБ ИРБИС64 [Электронный ресурс] / С.Р. Баженов, О.А.Рогозникова, М.В.Данилин // Материалы XXII Международной конференции «Крым-2015» «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса». – М.: ГПНТБ России, 2015. – CD-ROM.
19. Gartner, R. The Digital Object in Context: Using CERIF with METS / R. Gartner // Journal of Library Metadata. - 2012. - 12:1. - P.39-51. - DOI: 10.1080/19386389.2012.661689. – Режим доступа: https://kclpure.kcl.ac.uk/portal/files/6154736/cover_jlmapri2011_revised.pdf.
20. Simons, E. DC, MODS and CERIF-XML: A Tale of Two Cultures / E.Simons. – IRPPS, 2014. – Режим доступа: http://www.irpps.cnr.it/it/system/files/EdSimon.pdf.
21. Князева, А.А. Опыт идентификации персон в CRIS-системах / А.А.Князева, И.Ю.Турчановский, О.С.Колобов, О.Л.Жижимов // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г. – Дубна, 2014. – Режим доступа: http://rcdl.ru/doc/2014/paper/RCDL2014_207-213.pdf.