СЛИЯНИЕ АВТОРИТЕТНЫХ/НОРМАТИВНЫХ ДАННЫХ ДЛЯ РАСПРЕДЕЛЕННОГО ЭЛЕКТРОННОГО КАТАЛОГА БИБЛИОТЕК ЛЕНИНГРАДСКОЙ ОБЛАСТИ

А.А. Князева, О.С. Колобов, И.Ю. Турчановский

Институт вычислительных технологий СО РАН
e-mail: aknjazeva@ict.nsc.ru
Институт сильноточной электроники СО РАН
e-mail: okolobov@hcei.tsc.ru
Институт вычислительных технологий СО РАН
e-mail: tur@hcei.tsc.ru

В работе рассмотрена задача слияния авторитетных/нормативных баз данных для распределенного электронного каталога библиотек Ленинградской области. Приводятся основные принципы слияния авторитетных записей и описывается программный комплекс.

Введение

Необходимость идентификации объектов реального мира в библиографических данных рассматривалась бельгийским социологом Полем Отле еще в конце XIX века [1]. Такую идентификацию в библиотечном каталоге осуществляют с помощью установления связи со специальной авторитетной/нормативной записью[1], однозначно указывающей на данный объект. В качестве такой записи может выступать любой структурированный документ, содержащий информацию об объекте и удовлетворяющий требованиям, разработанным международными организациями [2]. Использовние авторитетного файла, состоящего из множества авторитетных записей, имеет важное значение для автоматизации работы библиотеки.

В России используются различные автоматизированные библиотечно-информационные системы (АБИС). Среди них есть зарубежные системы (VTLS, ALEF, Liber и др.) и российские («Руслан», «ИРБИС», «БУКИ», «Нева» и др.). Большинство АБИС позволяет осуществлять авторитетный контроль электронного библиотечного каталога. При этом под авторитетным контролем понимается процесс поддержания единообразия форм авторитетных заголовков, определяющих одно и тоже лицо, организацию, предмет и так далее в библиографическом файле, контроль за адекватностью присвоения предметных рубрик и индексов библиотечно-библиографических классификаций документам, а также контроль за последовательным соблюдением принципов, методик, инструкций и правил по представлению поисковых признаков [3].

Наличие авторитетного файла позволяет оптимизировать работу каталогизатора, облегчить поиск по имени индивидуального и коллективного автора, по названию серии, по предметной рубрике и по индексу библиотечно-библиографической классификации в электронном каталоге. Каталогизатор, создавая запись для электронного или карточного каталога, должен выбрать из списка существующих авторитетных/нормативных заголовков наиболее подходящий, что обеспечивает точную идентификацию автора, присвоение

1 Далее в рамках данной работы используется термин «авторитетная запись» (или сокращенно АЗ)

адекватной предметной рубрики и индекса классификации. Обработка документов с использованием такого средства унификации, как авторитетный файл, а также его поддержка в электронном каталоге обеспечивает точность и полноту поиска за счет установленных связей от синонимичных форм написания фамилий авторов, наименования серий, предметных рубрик, индексов библиотечно-библиографических классификаций к принятым и утвержденным в электронном каталоге формам. [3].

Таким образом, в настоящее время установление связей с авторитетными записями производится каталогизаторами вручную и однократно (только на этапе создания библиографической записи). Как следствие, при объединении ресурсов нескольких библиотек возникают задачи выявления дубликатов записей и восстановления утерянных или отсутствующих связей между авторитетными и библиографическими записями.

1. Обзор работ в области слияния баз данных

Слияние данных из разных источников является актуальной задачей, возникающей в различных областях знания. Она тесно связана с задачами идентификации сущностей (entity identification) [4], установления связей (record linkage) [5], выявления дубликатов (duplicate detection) [6-8]. Перечисленные задачи актуальны для широкого диапазона ресурсов, распределенных и локальных. В самых разнообразных данных часто встречаются упоминания одних и тех же объектов реального мира, которые необходимо связывать между собой для обеспечения более качественной работы с информацией. В частности, в нашей работе используются методы нечеткого сопоставления строк и общие принципы связывания документов.

В частности, задача выявления и слияния нескольких авторитетных записей для одного автора решалась в рамках проекта Виртуальный международный авторитетный файл (The **V**irtual **I**nternational **A**uthority **F**ile, VIAF) [9] Международной федерации библиотечных ассоциаций и учреждений (The **I**nternational **F**ederation of **L**ibrary **A**ssociations and Institutions, IFLA). Проект начинался с совместной работы Library of Congress (LC), Deutsche Nationalbibliothek (DNB), Bibliothèque nationale de France (BNF) и (OCLC). К началу 2012 года в проекте уже участвовали 20 организаций из 16 стран.

Целью проекта является обеспечение возможности автоматического сопоставления и связывания авторитетных записей из различных национальных источников. Для сопоставления записей в рамках VAIF был разработан набор правил, основанных на анализе различной информации об авторах с точки зрения её значимости для соответствия на уровне записей. Например, подчёркивается важность для сопоставления записей информации о соавторах, предметных рубриках, а также названий произведения автора, описания которых уже есть в базе данных.

В данной работе используются некоторые идеи, сформулированные в процессе работы над проектом VIAF. В частности, задействованы расширенные авторитетные записи, включающие кроме информации об авторе и информацию о его публикациях, содержащуюся в связанных библиографических записях [10].

2. Постановка задачи

2.1. Источники данных

В данной работе использовались данные распределенного электронного каталога библиотек Ленинградской области [11]. Создание распределённого каталога описывается в другой нашей статье [12]. Каталог объединяет ресурсы 20 библиотек и содержит более

300 тысяч библиографических записей в формате RUSMARC [13]. Записи распределены по библиотекам неравномерно. Из всего массива около 37% приходится на долю Ленинградской областной научной универсальной библиотеки. Остальные библиотеки частично заимствуют библиографические записи у неё, частично создают их самостоятельно.

2.2. Задача слияния авторитетных файлов

Рассмотрим ситуации, возникающие при объединении авторитетных файлов двух библиотек. При простом объединении ресурсов мы получим следующую картину (рисунок 1):

1. В объединении присутствуют записи-дубликаты (как авторитетные, так и библиографические)
2. В то же время отсутствуют некоторые необходимые связи между записями.

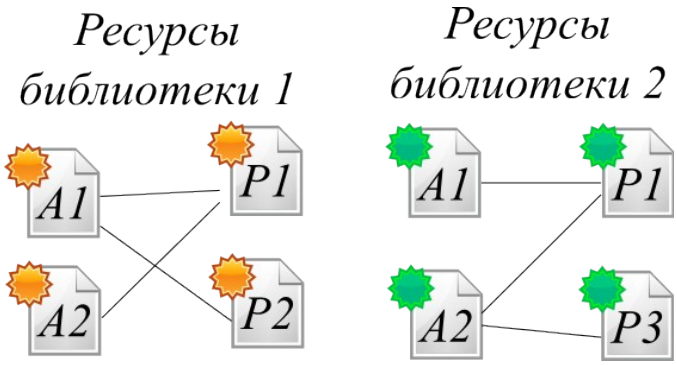

Рисунок 1.Пример простого объединения ресурсов двух библиотек

На рисунке изображены авторитетные записи для двух разных авторов (*A1* и *A2*), а также библиографические записи для трех публикаций (*P1*, *P2* и *P3*). В результате объединения получаем по записи-дубликаты для обоих авторов. Также видим, что отсутствуют связи между авторитетными записями одной библиотеки и библиографическими записями, созданными в другой библиотеке.

Таким образом, необходимо проводить процедуру не простого объединения, а слияния баз данных. При слиянии следует исключить повторяющуюся информацию и при этом сохранить всю различающуюся информацию (прежде всего, об установленных связях).

Слияние баз данных предлагается проводить в два этапа:

1. Слияние авторитетных файлов

2. Слияние библиографических баз данных

В данной работе рассматривается первый этап слияния, который касается авторитетных записей. Задача выявления дубликатов среди библиографических записей [14] существенно упрощается после проведения такого слияния.

В качестве иллюстрации текущей ситуации с авторитетными файлами в распределённом электронном каталоге библиотек Ленинградской области приведем пример. В таблице 1 представлен список авторитетных записей, которые были найдены по поисковому запросу «Кафка». Все они относятся к одному лицу — австрийскому писателю Францу Кафке. В заголовке таблицы приводится назначение основных полей, описывающих автора. В скобках указаны соответствующие поля формата RUSMARC/Autorities.

Таблица 1.Результаты поискового запроса в объединённой базе авторитетных записей

№	Фамилия (200$a)	Инициалы (200$b)	Дополнение (200$c)	Расшифровка инициалов (200$g)	Связанные даты (200$f)	Кол-во записей
1	Кафка	Ф.	-	-	-	2
2	Кафка	-	-	Франц	-	1
3	Кафка	Ф.	-	Франц	-	6
4	Кафка	Ф.	писатель	-	-	1
5	Кафка	Ф.	-	Франц	1883-1924	1
6	Кафка	Ф.	австр. писатель	Франц	1883-1924	1
7	КАФКА ФРАНЦ (ПИСАТЕЛЬ)	-	О НЕМ	-	-	4
8	КАФКА Ф.	(«ДНЕВНИКИ»)	-	-	-	4
9	КАФКА Ф.	(«ПРЕВРАЩЕНИЕ»)	-	-	-	4
Всего записей:						**24**

Таким образом, мы получили 24 записи, относящиеся к одному автору. Записи имеют различную полноту. Очевидно, что вместо всех этих записей должна быть всего одна, к которой должны быть привязаны все библиографические записи.

Часть записей была создана в «основной» библиотеке и затем заимствована другими библиотеками. Такие записи достаточно легко объединяются на основе простых правил, задействующих сравнение идентификаторов записей (как правило, записи заимствуются вместе с идентификаторами). Однако даже после проведения такой работы в приведенном примере получим 12 авторитетных записей для Франца Кафки. Некоторые из этих записей, видимо были созданы в результате систематических ошибок используемой АБИС. Это предположение согласуется с тем, что с этими записями не было установлено связей в библиотеке их создавшей. При этом в других библиотеках (после заимствования) они уже используются как полноправные авторитетные записи и с ними могут быть установлены связи. Примером такой ошибочно созданной, но используемой в дальнейшем записи является запись из строки 9 таблицы 1. Очевидно, она содержит ошибочно внесенную информацию: инициал записан вместе с фамилией, а в поле, предназначенном для инициалов, указано название рассказа «Превращение». Тем не менее с этой авторитетной записью связана одна библиографическая запись. Следовательно, мы не можем её просто удалить — необходимо объединить её с более «качественной» записью и сохранить установленную связь.

Задача составления правил, по которым будет определяться «качество» записи — её безошибочность и полнота — нетривиальна. В первую очередь было принято следующее допущение: те авторитетные записи, с которыми не установлено ни одной связи, исключаются из рассмотрения. Это не означает удаления записи из электронного каталога отдельной библиотеки, это означает лишь, что она не появится в едином авторитетном файле до тех пор, пока не будет связана с библиографической записью (хотя бы одной). После этого

запись уже будет рассматриваться системой и участвовать в процессе слияния с другими авторитетными записями.

Применим принцип исключения неиспользуемых записей к приведённому примеру. Результаты поиска в этом случае сократятся до 10 записей, а с учётом заимствований до 9.

Далее необходим механизм более подробного сопоставления записей для принятия решения о том, являются ли они дубликатами. Сопоставление предлагается осуществлять на основе набора правил, разработанных на основе анализа конкретных электронных каталогов. Данные правила учитывают возможные систематические ошибки при создании записей.

В процессе выявления и слияния дубликатов авторитетных записей автоматически осуществляется чистка данных. Некорректные с точки зрения правил заполнения записи «поглощаются» корректными. При этом поглощение означает, что вся информация об авторе берётся из корректной записи, а из некорректной извлекается только информация о связях. Таким образом мы добиваемся более высокого качества данных, чем при простом удалении всех некорректных записей.

Важным моментом является использование расширенных авторитетных записей в процессе сравнения. Под расширенной авторитетной записью понимается множество, состоящее из авторитетной записи и всех связанных с ней библиографических записей. Такое расширение позволяет обрабатывать некорректные записи и записи, в которых слишком мало информации об авторе. Учёт названий произведений авторов позволяет существенно повысить качество идентификации. Принцип использования расширенной авторитетной записи был предложен в рамках международного проекта «Виртуальный авторитетный файл» [10]. Данный принцип также использовался нами при решении задачи автоматического связывания библиографических и авторитетных записей [13].

При слиянии необходимо сохранить целостность связей, установленных между исходными авторитетными и библиографическими записями. Для этого после слияния необходимо провести реиндексацию массива библиографических записей и установить связи с едиными авторитетными записями.

Описанный подход к слиянию записей приводит к значительным сложностям в случае удаления авторитетной записи из каталога отдельной библиотеки. Для внесения изменений в единый авторитетный файл в таком случает требуется удалить часть информации из новой авторитетнойз записи. При этом необходимо раскрутить цепочку изменений в обратную сторону и установить — какая именно информация из удалённой авторитетной записи была использована в записи на автора, полученной после слияния. По-видимому, наиболее простым решением в этом случае является периодическое проведение процедуры слияния баз данных заново, «с нуля».

Периодическое обновление данных распределённого каталога с проведением процедуры слияния позволяет вносить все изменения, сделанные в исходных каталогах. В то же время не приходится хранить большое количество информации о том, каким образом была сформирована единая авторитетная запись. Выполнять слияние можно off-line, не ограничивая в это время пользователей в работе с распределённым каталогом.

Для того, чтобы повысить скорость синхронизации данных предлагается использовать инкрементальное добавление новых записей в распределённый каталог. При добавлении новой авторитетной записи она сравнивается с уже существующими и при необходимости сливается с одной из них. При добавлении новой библиографической записи в системе, она

добавляется и в распределённый каталог, при этом связи, установленные в ней разрешаются в соответствии с единым авторитетным файлом.

3. Описание программного комплекса

3.1. Функциональная схема

Для решения поставленной задачи используется модифицированный программный комплекс cflib, ранее разработанный нами в рамках работы по идентификации персон для CRIS-системы ИВТ СО РАН [15].

Программный комплекс cflib состоит из нескольких функциональных блоков, взаимодействие блоков представлено на рисунке 2. На разных этапах работы блоки могут работать в различных режимах. Описание этих режимов приводится ниже.

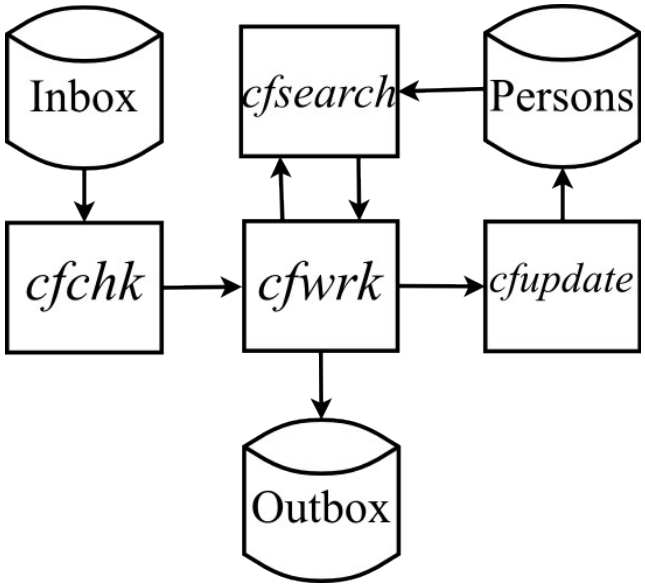

Рисунок 2.Функциональные блоки программного комплекса cflib

3.2. Работа функциональных блоков в процессе создания единого АФ

В процессе слияния АФ используются две базы данных: ***Inbox***, в которую поступают АЗ и ***Persons***, в которой формируется единый АФ. Рассмотрим работу блоков немного подробнее.

1. ***Inbox*** **–** база данных входных документов, в нее помещаются авторитетные записи (АЗ) в таком порядке:
    1. АФ главной библиотеки (записи, которые заимствовались другими библиотеками);
    2. Все остальные АФ в произвольном порядке;

Приведенный порядок позволяет обеспечить больший приоритет

2. ***Outbox*** **–** база данных, которая не используется в процессе слияния АЗ, потребуется для следующего этапа (п.п. 3.3.);
3. ***Persons*** **–** база авторитетных данных, в нее помещаются объединенные авторитетные записи, она создается в процессе работы;
4. ***cfchk*** – модуль проверки записей на корректность, проводит проверку записей на соответствие формату RUSMARC/Authority и требованиям на полноту, а также дополняет запись информацией, необходимой для индексирования (например, биграммами фамилии автора);

5. ***cfwrk*** – модуль сравнения и обработки формирует запрос на поиск похожих АЗ, который передается в модуль ***cfsearch*****,** результаты поиска поступают на вход процедуры сравнения (она тоже выполняется модулем ***cfwrk***, описана ниже);
6. ***cfsearch* –** модуль поиска в базе Persons осуществляет поиск в формирующейся базе АЗ по биграммам фамилии, результаты поиска ранжируются по релевантности;
7. ***cfupdate*** – модуль обновления данных в базе ***Persons***.

Дополнительные индексы для базы данных ***Persons*** необходимы для поиска близких авторитетных записей, а также поиска объединенных АЗ. Они создаются для следующих значений:

- Биграммы фамилии автора;
- Дополнительные идентификаторы записи (все идентификаторы, которые использовались в АЗ до их слияния).

Процедура сравнения записей, реализованная в блоке cfwrk позволяет задавать несколько степеней соответствия записей и рассматривать несколько вариантов соответствия:

1) Заимствованная запись (совпадают идентификаторы записей за исключением префикса и при этом нет противоречия в основной информации об авторе);

2) Точное соответствие (все поля за исключением идентификаторов записей и служебных полей в точности совпадают)

3) Вхождение (записи не противоречат друг другу, при этом в одной из них содержится больше информации, чем в другой)

4) Частичное совпадение (записи не противоречат друг другу в основной информации, но содержат различную дополнительную информацию).

Для последних двух вариантов требуется дополниетельное подтверждение того, что записи описывают одного автора. Для такого подтверждения используются связанные библиографические записи.

3.3. Работа функциональных блоков в процессе реиндексации

Как уже говорилось, после слияния авторитетных файлов необходимо провести реиндексацию библиографических записей и указать в них идентификаторы новых АЗ, полученных в результате слияния. Эту задачу также можно решать с помощью программного комплекса ***cflib***. Работа функциональных блоков при этом изменится следующим образом:

1. ***Inbox* –** база данных входных документов, в нее помещаются библиографические записи (БЗ) всех библиотек, порядок неважен;
2. **Outbox** – база результирующих документов, в нее помещаются БЗ с измененными идентификаторами авторитетных записей;
3. **Persons –** база авторитетных данных, которая была создана в процессе слияния АФ (п.п. 3.2);
4. **cfchk** – модуль проверки записей на корректность, проверяет записи на соответствие формату RUSMARC;
5. **cfwrk** – модуль сравнения и обработки формирует запрос на поиск АЗ, которая была указана в старых авторитетных идентификаторах, сравнения записей не требуется;
6. **cfsearch –** модуль поиска в базе ***Persons***, поиск осуществляется по идентификатору АЗ. Поиск идентификатора ведется не только в поле идентификатора записи 001, но и в специальном поле, в котором указываются все идентификаторы, использовавшиеся до слияния АФ;

7. **cfupdate** – модуль обновления данных в базе ***Persons*** в процессе реиндексации не участвует.

Реиндексация библиографических записей не предполагает сложных алгоритмов сравнения записей, её можно считать рутинной операцией. Тем не менее её важность от этого не теряется.

Заключение

В работе рассмотрена задача слияния авторитетных/нормативных баз данных для распределенного электронного каталога библиотек Ленинградской области. Описаны проблемы, возникающие в процессе объединения библиографических данных из нескольких источников.

Результаты данной работы могут использоваться при слиянии библиографических баз данных, в процессе создания сводного каталога. Такой каталог позволит предоставить полную информацию о хранимых экземплярах в удобном для позователей виде, без дублирования информации.

Как показал первоначальный анализ качества библиографических и авторитетных записей из распределённого каталога, потребуется также проведение работы по «чистке» данных. Чистка подразумевает избавление от некоторых систематических ошибок, работу с недостаточно полными записями (вероятно, в полуавтоматическом режиме) и другие методы улучшения качества данных.

ЛИТЕРАТУРА

[1]. Отле П. Библиотека, библиография, документация [Текст] : Избранные труды пионера информатики / ПольОтле. -- Москва: ФАИР-ПРЕСС: Пашков Дом, 2004. -- 348, [1] с. -- (Специальный издательский проект для библиотек). -- Библиогр.: с. 312-327. -- Имен. указ.: с. 340-342. -- ISBN 5-8183-0624-0 (в пер.).

[2]. Функциональные требования к авторитетным данным : концептуальная модель : заключительный отчет, декабрь 2008 / под ред. Гленна Е. Патона ; Рабочая группа ИФЛА по разработке функциональных требований к авторитетным записям и их нумерации (FRANAR); одобрено Постоянными комитетами Секции по каталогизации и Секции по классификации и индексированию ИФЛА, март 2009; Междунар. федерация библ. ассоц. и учреждений, Рос. библ. ассоц.; [пер. с англ. О.А. Лаврёнова]. -- Санкт-Петербург: Российская национальная библиотека, 2011. -- 115 с. : илл., граф.

[3]. Муктепавел А.В. Авторитетные файлы предметных рубрик в условиях автоматизированной каталогизации: проблемы создания и ведения : дис. ..канд. пед. Наук: 05.25.03 / Муктепавел Айна Вольдемаровна. -- М., 1999. -- 179 с.

[4]. Talburt J. Entity resolution and information quality / John R. Talburt. — San Francisco : Morgan Kaufmann/Elsevier, 2011. — 256 p.

[5]. Winkler W.E. Overview of record linkage and current research directions [Electronic resource] : tech. report / W.E. Winkler ; U.S. Census Bureau, Stat. res. div. — Washington : [s. n.], 2006. — 44 p. \url{http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf}

[6]. Elmagarmid A., Ipeirotis P., Verykios V. (2007). Duplicate Record Detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1):1-16.

[7]. Bilenko M. Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases / M. Bilenko, R. Mooney. Technical Report AI-02-296, Artificial Intelligence Lab, University of Texas at Austin, 2002.

[8]. Sarawagi S. Interactive deduplication using active learning / S. Sarawagi, A. Bhamidipat // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. — P. 269—278.

[9]. VIAF: The virtual international authority file [Electronic resource] : [offic. Site] / OCLC: the world's libraries — Dublin, 2010—2012. — URL: http://viaf.org, free. Tit. from the screen (usage date: 04.06.2013).

[10]. Bennett R. VIAF (Virtual international authority file): linking the Deutsche Nationalbibliothek and Library of Congress name authority files / R. Bennett [et al.] // Int. cataloging and bibliographic control. — 2007. — Vol. 36, № 1. — P. 12-19.

[11]. Ленинградская областная универсальная научная библиотека. http://www.reglib.ru/

[12]. Колобов О.С. Зверев А.И. Князева А.А. Турчановский И.Ю. Распределенный электронный каталог библиотек Ленинградской области [Электронный ресурс] // XV Российская конференция с международным участим “Распределенные информационно-вычислительные ресурсы (DICR’2104)”: материалы конференции. -- Электронные данные. -- Новосибирск: ИВТ СО РАН, 2014. -- URL: http://conf.ict.nsc.ru/dicr2014/reportview/246614

[13]. Князева А.А., Турчановский И.Ю., Колобов О.С. Автоматический авторитетный контроль для распределенных библиографических баз данных [Электронный ресурс] // XIII Российская конференция с участием иностранных ученых "Распределенные информационные и вычислительные ресурсы"(DICR'2010): материалы конф. - Электрон. дан. - Новосибирск: ИВТ СО РАН, 2010. - 1 электрон. опт. диск (CD-ROM).- № гос. регистрации 0321100051.- http://conf.nsc.ru/dicr2010/ru/reportview/29244.

[14]. Князева А.А., Турчановский И.Ю., Колобов О.С. Выявление дубликатов в библиографических базах данных // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL'2013) : Труды 15-й Всероссийской научной конференции, Ярославль 14-17 окт. 2013 г. – Ярославль, 2013. – С. 276–282.

[15]. Князева А. А., Колобов О.С., Турчановский И.Ю. Опыт идентификации персон для CRIS-систем // Электронные библиотеки~: перспективные методы и технологии, электронные коллекции (RCDL'2014): Труды 16-й Всероссийской научной конференции, Дубна, 13-16 окт. 2014 г. — Дубна : ОИЯИ, 2014. — С. 207-213.