# Опыт идентификации персон для CRIS-систем

© А.А. Князева
© И.Ю. Турчановский
Институт вычислительных технологий СО РАН, Томск
aknjazeva@ict.nsc.ru
tur@hcei.tsc.ru

© О.С. Колобов
Институт сильноточной электроники СО РАН, Томск
okolobov@hcei.tsc.ru

О.Л. Жижимов
Институт вычислительных технологий СО РАН, Новосибирск
zhizhim@mail.ru

**Аннотация**

В данной работе приводится описание системы идентификации персон, которая создавалась в процессе разработки Единого репозитория результатов научно-технической деятельности (РНТД) в ИВТ СО РАН. Кратко описываются принципы и методы, используемые при создании системы, а также ее структура. Описан алгоритм создания авторитетной базы данных с описаниями персон в автоматическом режиме, без участия пользователя. Для выявления нечетких дубликатов в упоминаниях персон использовались индексирование по биграммам и расстояние редактирования.

## 1 Введение

Разработка информационных систем (ИС), предназначенных для сбора и хранения информации о результатах научной деятельности, в настоящее время крайне актуальна [1]. К таким системам относятся научные сети (например, Scopus [2], Web of Science [3], ResearchGate [4], SciVerse [5], Cross-Ref [6]), и целый класс информационных систем CRIS (Current Research Information Systems) [7]. Предметом рассмотрения в данной работе будут CRIS-системы.

С 2000 года существует организация *EuroCRIS*, объединяющая разработчиков и исследователей ИС текущих исследований в странах Европейского Союза. EuroCRIS занимается созданием и поддержкой стандартов и методологий создания CRIS-систем.

В настоящее время распространение CRIS-систем не ограничивается географическими рамками. В EuroCRIS более 300 делегатов из 43 стран. Членами данной организации являются пять российских организаций: Центральный экономико-математический институт РАН, Институт вычислительных технологий СО РАН, Институт нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН, Уральский федеральный университет и Научная библиотека «КиберЛенинка». Разработкой CRIS-систем занимаются также Астраханский государственный университет [8], Институт математики и механики им. Н.Н. Красовского УрО РАН, Институт вычислительных технологий СО РАН [9] и другие. Существуют также системы, взаимодействующие с локальными CRIS-системами и расширяющие их функциональность. В качестве примера можно привести систему, разрабатываемую в среде крупного отечественного онлайнового научно-образовательного пространства, поддерживаемого системой Соционет [10, 11].

Исходя из широкого диапазона пользователей возникает необходимость учета самой разнообразной научно-исследовательской информации, а также и большой набор требований к CRIS-системам.

В данной статье описывается система идентификации персон, которая разрабатывалась в рамках создания системы агрегирования данных по научным проектам в Институте вычислительных технологий СО РАН[1]. Задача идентификации персон в данных CRIS-систем, объединенных в единый репозиторий, близка к задачам идентификации сущностей (entity identification) [12], установления связей (record linkage) [13], выявления дубликатов (duplicate detection) [14–16].

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

[1] Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (Государственный контракт № 14.521.11.0004 от 14.08.2013 «Разработка системы агрегирования данных по научным проектам из различных источников для обеспечения мониторинга реализации мероприятий и программ», шифр 2013-2.1-14-521-0017-002).

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.

## 2 Обзор работ в области идентификации объектов

Для идентификации объектов в CRIS-системах в настоящее время используется подход, основанный на принципах LOD[2]. Он позволяет использовать собственные идентификаторы, которые при этом становятся внешними. Созданные идентификаторы могут однозначно разрешаться и сторонними системами благодаря использованию механизма URI. Такой подход разрабатывается в рамках проекта CERIF-Linked-Data[3]. В дальнейшем связать данные CRIS-систем, преобразованные в соответствии с принципами LOD, можно связать с другими источниками LOD (например, библиографическими) с помощью инструмента автоматического установления RDF-ссылок Silk[4].

Использование семантических связей при создании и отображении CERIF-документов рассматривается также в работах С.И. Паринова [17].

Вопросы интеграции информационных ресурсов, формирования наборов метаданных и онтологий для научных информационных ресурсов рассматриваются в работах А.Н. Бездушного, М.В. Кулагина, В.А. Серебрякова и др. [18, 19].

Задача создания собственной CRIS-системы, в которой производится идентификация персон, рассматривается также в работах А.С. Умарова и др. [1].

Различные системы учета публикаций, например, Scopus, Web of Science, SCIENCE INDEX (на базе РИНЦ) используют различные идентификационные коды авторов[5]. Обзор различных систем идентификаторов и их сравнительный анализ приводится в работах [20, 21].

## 3 Постановка задачи

### 3.1 Источники данных

В процессе работы использовались данные Единого репозитория результатов научно-технической деятельности (РНТД), разрабатываемого в ИВТ СО РАН. РНТД объединяет данные по научным проектам нескольких организаций: ФГБНУ «Научно-исследовательский институт – Республиканский исследовательский научно-консультационный центр экспертизы», Российский фонд фундаментальных исследований (РФФИ), Российский гуманитарный научный фонд (РГНФ), ФГБНУ «Дирекция научно-технических программ», Национальный фонд подготовки кадров, ООО «Инконсалт», ФГАНУ «Центр информационных технологий и систем органов исполнительной власти».

### 3.2 Описание задачи идентификации

Имеется несколько независимых источников, содержащие данные о научно-исследовательской деятельности. Данные из источников агрегируются в выделенную базу данных (репозиторий). Агрегированные данные могут содержать дублетное описание базовых сущностей, т.е. описывать один объект реального мира в различных вариантах. Это отражается на качестве поиска, так как результаты поиска документов, относящихся к отдельной сущности, не будут достаточно полными. Необходимо решить задачу идентификации объектов реального мира в данных.

### 3.3 Варианты идентификации объектов

Идентификацию объектов реального мира можно организовать различными способами, в зависимости от наличия авторитетных данных:

1. Существует авторитетная база данных (своя или сторонняя), в ней описываются объекты, которые необходимо идентифицировать. Задача сводится к установлению связи с авторитетным документом путем указания его идентификатора.

2. Нет такой базы, необходимо создавать ее в процессе идентификации.

Первый вариант идентификации актуален для организаций, в которых существуют развитые авторитетные базы данных (библиотек, архивов) и не всегда подходит для научно-исследовательских институтов, в которых такие базы, зачастую, не сформированы на этапе разработки CRIS-систем.

Использование сторонних баз данных для идентификации сущностей может быть особенно полезным в тех случаях, когда в самих документах приводится мало информации об объекте. Тогда можно идентифицировать объект (например, персону) не столько по его описанию, сколько по его связям с другими объектами (персонами, организациями и т.п.). С этой точки зрения могут быть полезны социальные сети, которые активно развиваются в настоящее время, например, *LinkedIn*, *Facebook*, *ВКонтакте* и др. Поскольку основное внимание в них уделяется именно связям между объектами. Можно использовать профили пользователей для их идентификации при условии, что существует API. Этот подход представляется перспективным и будет развиваться в нашей дальнейшей работе.

[2] Linked Open Data.

[3] http://code.google.com/p/cerif-linked-data/

[4] Silk Link Discovery Framework – http://www4.wiwiss.fu-berlin.de/bizer/silk/

[5] ORCID, ResearcherID, SPIN-код соответственно.

В данной статье рассматривается второй вариант идентификации, при котором создаются авторитетные базы данных. Он может быть полезен в том случае, если нет готовых авторитетных баз данных и при этом нет уверенности, что пользователи уже зарегистрированы в социальных сетях или в системах учета публикаций. Для реализации данного подхода необходимо решать следующие задачи:

1. Формальный контроль входных данных.

2. Автоматическое формирование авторитетных баз данных, которые содержат документы, описывающие идентифицируемые объекты.

3. Авторитетный контроль входных данных (установление связи).

Предлагается технология связывания документов, которую можно использовать для связывания с уже существующими системами идентификаторов, при условии, что в данных системах существуют профили идентифицируемых объектов.

При этом не идет речь о создании собственной системы универсальных идентификаторов. Для технических нужд в процессе работы используются исключительно внутренние идентификаторы документов и объектов.

### 3.4 Формат данных

Данные, используемые в работе, поступают в виде документов в формате CERIF[6]. Данный формат является официальной рекомендацией для членов Европейской комиссии (European Commission). Он определяет набор обязательных и дополнительных полей, которые должны использоваться для описания научных проектов, в том числе название проекта, краткое описание, описание участников, наименование финансирующей организации и т.п. В качестве дополнительной информации могут быть указаны ссылки на другие проекты и на публикации в рамках данного проекта [22].

## 4 Идентификация персон для РНТД

### 4.1 Входные требования к документам

К входным документам предъявляются следующие требования:

• документы должны соответствовать требованиям формата CERIF;

• в упоминании персоны должны быть как минимум указаны фамилия и первый инициал на одном из двух языков (русском или английском).

### 4.2 Краткое описание алгоритма работы

Место системы идентификации персон в РНТД представлено на рисунке 1. Система работает с данными РНТД, формирует авторитетную базу данных с описаниями персон *Persons* и сводную базу данных документов CERIF, для которых установлены связи с соответствующими персонами.

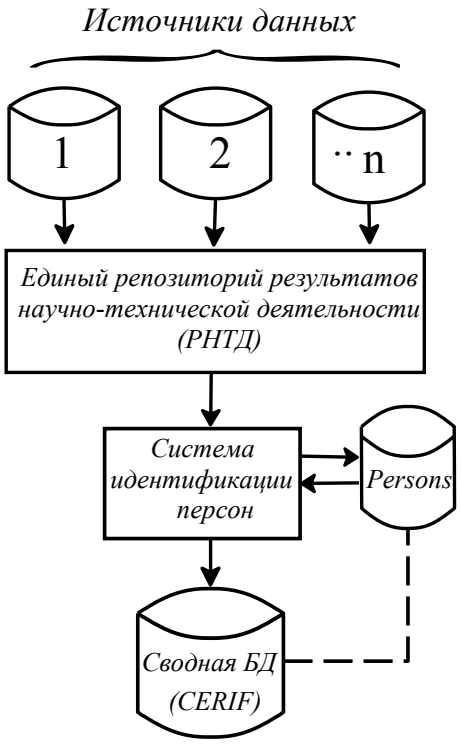


Рис. 1. Взаимодействие системы идентификации персон и РНТД

Основные принципы, которых мы придерживались при разработке модели идентификации сущностей:

• Сохранение исходных данных обеспечивает возможность возвращения входных документов к предыдущему состоянию, до того, как они были связаны. Если будет сделана ошибка при установлении связи, ее легко можно будет исправить. На практике это требование означает, что вместо того чтобы изменять документы в исходной базе данных, создается новая база данных с установленными связями.

• Использование меток содержимого, т.е. лексем, которые несут смысловую нагрузку и описывают данные. Этот принцип позволяет получать данные из системы без изменений (в их первоначальном виде) и, в то же время, при необходимости осуществлять более сложную навигацию по ним. Таким образом, мы дополняем данные и решаем проблему идентификации сущностей.

В процессе обработки входных документов создается ресурс авторитетной базы, который содержит документы с описанием персон. Авторитетные документы наполняются только той информацией, которая есть во входных данных, без привлечения сторонних источников.

В общем виде алгоритм работы программного комплекса выглядит следующим образом. Входной

[6] CERIF – формат обмена научно-исследовательскими данными, разработанный EuroCRIS

документ подвергается формальному контролю. На этом этапе необходимо проверить его корректность с точки зрения соответствия схеме данных CERIF [23]. Также возможна и проверка отдельных полей с помощью словарей допустимых значений.

Далее из входного документа извлекаем определения базовых сущностей CERIF (в данной работе рассматривается сущность типа «Персона»). Извлечение означает, что создается временный авторитетный документ, в который помещается вся информация о сущности, содержащаяся во входном документе CERIF. При этом, как правило, во входном документе упоминается более одной персоны (в проекте участвует сразу несколько исполнителей), временные документы создаются для каждой из них.

Каждый временный авторитетный документ затем сравнивается с теми, что уже находятся в авторитетной базе данных. Если такого документа (или похожего на него) в базе нет, то он загружается в нее и перестает быть «временным». Если для временного документа был найден нечеткий дубликат, то возможны два варианта:

• Во временном документе нет новой информации – авторитетная база данных остается без изменения.

• Временный документ не противоречит найденному в базе данных, но при этом содержит часть неучтенной информации – документ из базы данных дополняется этой новой информацией.

### 4.3 Поиск подобных документов в авторитетной базе данных

Сравнение временного авторитетного документа с каждым из документов, уже содержащихся в базе данных, может оказаться неоправданно трудоемким. В частности, при работе «на лету» может потребоваться сократить количество авторитетных документов, которые будут сопоставляться с временным документом. Существует множество способов ограничить круг документов для сопоставления. Приведем некоторые из них:

1. Метод стандартных блоков выделяет документы в один блок в том случае, если они содержат идентичный блочный ключ [24]. Блочные ключи формируются на основе атрибутов документов, например, первые 4 символа фамилии. Кроме того, блочный ключ может быть и составным, например, атрибут «фамилия» может сочетаться с атрибутом «год рождения». Ключи должны быть выбраны таким образом, чтобы блоки не были ни слишком большими, ни слишком мелкими.

2. Метод ближайших соседей [25] сортирует документы на основе сортирующего ключа и затем двигает окно фиксированного размера последовательно по всем документам. Документы внутри окна составляют пары друг с другом и включаются в список пар-кандидатов. Метод может некорректно работать в том случае, если количество документов с одним значением ключа превышает размер окна, поскольку в такой ситуации будут сравниваться не все нужные документы.

3. Метод Bigram-индексирования [26] предназначен для нечеткого разбиения на блоки. Основная идея заключается в том, что значения блочных ключей конвертируются в лист биграм (подстрок, состоящих из двух символов) и затем из этих биграм формируются списки на основе заданного порога (например, выбираются все документы, в которых встречается 80% биграм).

В рамках данной работы использовался метод Bigram-индексирования на основе фамилии персоны на русском языке. Использование этого метода позволяет найти документы с опечатками в фамилии, что позволяет повысить качество идентификации.

Результаты поиска выдаются в порядке релевантности, то есть в начале списка результатов помещаются документы с точным соответствием (если они есть), а затем все менее и менее похожие (в смысле совпадения по фамилии). Таким образом, ограничение круга документов задается путем установления порога, после которого документы признаются слишком отличающимися и не передаются для более подробного анализа.

В нашей работе такой порог был установлен с помощью расстояния редактирования Левенштейна [27]. Документы с различием в фамилии больше чем в 2 символа считались непохожими исключались из дальнейшего рассмотрения. Оставшиеся документы, являющиеся потенциальными дубликатами рассматриваемого документа, формируют *множество документов для сравнения.*

### 4.4 Описание процедуры сравнения документов

Рассмотрим более подробно процедуру сравнения временного авторитетного документа с документами из множества для сравнения.

Прежде всего производится вычисление строгого соответствия для всех полей документа. В зависимости от результатов сравнения возможны следующие варианты:

1. Если все поля точно равны, делается вывод, что текущий документ является точным дубликатом найденного. Из найденного документа извлекается его идентификатор и возвращается для установления связи. Временный документ не помещается в авторитетную базу данных.

2. Если точного равенства нет, то следует нечеткое сравнение (см. таблицу 1). Допускается расхождение в одном из перечисленных полей (кроме поля с указанием пола):

(а) Если расхождение в одном поле, и не превышает границы, то делается вывод о нечетком дубликате. В найденный авторитетный документ вносится информация о вариантном наименовании, возвращается код найденного документа.

Временный авторитетный документ не помещается в базу данных;

(b) Если не было обнаружено ни точного, ни нечеткого сравнения, переходим к анализу следующего найденного документа.

Таблица 1. Способы сравнения

| Признак | Поле док-та | Способ сравнения | Порог. значение |
|---|---|---|---|
| **Фамилия** (рус. яз.) | 200$a | Расстояние редактирования | 2 |
| **Фамилия** (англ. яз.) | 400$a | | |
| **Имя** (рус. яз.) | 200$g | | |
| **Имя** (англ. яз.) | 400$g | | |
| **Пол** | 120$a | Строгое равенство | 0 |
| **Место работы** (организация) | 601$a | Относительное расстояние редактирования | 30% |

Если временный авторитетный документ прошел процедуру сравнения с каждым документом из множества для сравнения, но при этом не было найдено ни одного строгого или нестрогого дубликата, то он помещается в авторитетную базу данных.

Относительное расстояние редактирования определяется как отношение расстояния редактирования к длине первой из двух сравниваемых строк.

Пороговые значения были установлены эмпирически. В дальнейшем планируется провести более подробное исследование для выбора пороговых значений.

## 5 Оценка качества идентификации

Оценивать качество идентификации персон в рамках данной работы предлагается с помощью широко распространенных показателей: полноты и точности [28].

Показатель полноты можно рассчитать с помощью следующей формулы:

$$Recall = \frac{TruePostive}{TruePostive + FalseNegative}, \quad (1)$$

где *TruePositive* – количество верно установленных связей с созданными авторитетными документами, *FalseNegative* – количество упущенных связей.

Точность оценивается по формуле

$$Precision = \frac{TruePostive}{TruePostive + FalsePostive}, \quad (2)$$

где *FalsePositive* – количество неверно установленных связей.

Для расчета описанных показателей необходима тестовая выборка, на основе которой можно было бы рассчитать количество ошибок и верно установленных связей между документами.

При этом можно оценивать полноту и точность в двух вариантах. Если рассматривать все возможные комбинации документов как потенциальные связи, то в результате получим оценку метода в целом. А если среди связей рассматривать только те, которые были отобраны как потенциальные с помощью процедуры поиска подобных документов, то получим оценку работы механизма сопоставления документов. Второй вариант подходит в том случае, если на этапе сужения круга документов для сравнения не происходит потери нужных связей.

## 6 Описание программного комплекса

### 6.1 Функциональное описание программного комплекса *cflib*

В качестве системы идентификации персон (рис. 1) в данной работе выступает программный комплекс *cflib*, состоящий из следующих модулей:

- *cfchk* – проверка и коррекция входных документов, внедрение временных меток содержимого;
- *cfwrk* – сравнение временного и найденного документов;
- *cfsearch* – поиск в авторитетной базе данных;
- *cfupdate* – дополнение документа из авторитетной базы данных.

Основные модули программного комплекса *cflib* представлены на рисунке 2.

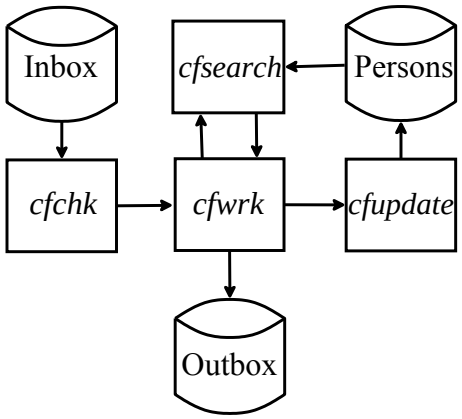


Рис. 2. Основные модули программного комплекса

Модуль *cfchk* кроме проверки входного документа на корректность также осуществляет внедрение меток содержимого, используемых для работы. К таким меткам относятся временные идентификаторы для отдельных персон, наборы биграмм и другая информация, которая понадобится при дальнейшей работе.

При помещении документа в базу данных Persons он индексируется в соответствии с подготовленным профилем индексирования. В этот профиль включено индексирование по биграммам, что позволяет модулю *cfsearch* извлекать подобные

документы для поиска среди них нечетких дубликатов.

В работе модуля *cfwrk* предусмотрены несколько этапов сравнения для выявление нечетких дубликатов. При этом можно изменить функции сравнения и использовать другие методы сравнения, не нарушая логики работы программного комплекса.

Модуль *cfupdate* предназначен для дополнения авторитетного документа отсутствующей информацией или вариантными значениями отдельных полей.

Программный комплекс *cflib* является платформо-независимым и может работать под управлением различных операционных систем или сред, включая Cygwin для MS Windows. Комплекс *cflib* написан на нескольких языках программирования: C, Perl и XSLT

## 7 Заключение

В данной работе приводится описание системы идентификации персон, которая создавалась в процессе разработки Единого репозитория результатов научно-технической деятельности (РНТД) в ИВТ СО РАН.

Особенностью данной системы является то, что в процессе ее работы создается авторитетная база данных с описаниями персон в автоматическом режиме, без участия пользователя. Первое встреченное упоминание ложится в основу авторитетного документа, а последующие могут при необходимости дополнять этот документ. Такой подход был выбран из-за того, что при создании системы в нашем распоряжении не было готовой авторитетной базы данных. Однако допускается и возможность подключения готовой базы данных, если она доступна.

В качестве формата авторитетных данных был выбран формат RUSMARC/Authorities [29], широко распространенный в библиотечном сообществе. Такой выбор позволяет в дальнейшем осуществлять простую интеграцию с библиотечными данными.

Для выявления нечетких дубликатов в упоминаниях персон использовались индексирование по биграммам и расстояние редактирования. Сравнение состоит из нескольких этапов. При необходимости можно предусмотреть и больше вариантов сравнения документов, а также изменить используемые для сравнения методы – логика работы системы от этого не пострадает.

В дальнейшей работе планируется рассмотреть различные методы сравнения документов, исследовать возможность построения обучающих выборок из документов в формате CERIF, а также использовать различные сторонние системы для идентификации персон (в частности, систему Silk).

## Литература


[1] Умаров А.С., Попова Н.В., Зелепухина В.А. Некоторые аспекты создания информационных систем для сбора и хранения научной и наукометрической информации // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 3 (23). – С. 111–118.

[2] Scopus. http://www.scopus.com

[3] Thomson Reuters Web of Science. http://thomsonreuters.com/products_services/ science/science_products/a-z/ web_of_science/

[4] ResearchGate. http://researchgate.net

[5] SciVerse. http://www.info.sciverse.com

[6] CrossRef. http://crossref.org

[7] CRIS concept and CRIS benefits. http://www.eurocris.org/ Index.php?page=concepts_benefits&t=1

[8] Астраханский государственный университет. Результаты научной деятельности. http://science.aspu.ru

[9] Guskov A.E., Zhizhimov O.L., Kikhtenko V., Skachkov D.M., Kosyakov D. RuCRIS: A Pilot CERIF based System to Aggregate Heterogeneous Data of Russian Research Projects // Procedia Computer Science. – 2014. – Vol. 33. – P. 163–167. – ISSN 1877-0509. – http://www.sciencedirect.com/science/article/ pii/S1877050914008175/pdf?md5= d74bdd8e7724f217d214b6aaff40c1eapid= 1-s2.0-S1877050914008175-main.pdf

[10] Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек // Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2011, Воронеж, 19–22 окт. 2011 г. – Воронеж: Воронежский государственный ун-т, 2011.

[11] Когаловский М.Р., Паринов С.И. Классификация и использование семантических связей между информационными объектами в научных электронных библиотеках // Информатика и ее применения, 2012. Т. 6, вып. 3. С. 31–41.

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

[13] 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. – http://www.census.gov/ srd/papers/pdf/ rrs2006-02.pdf

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

[15] 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.

[16] 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.

[17] Parinov S. Open Repository of Semantic Linkages. In: Proceedings of 11th International Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012), Prague 2012, http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29.

[18] Бездушный А.Н., Кулагин М.В., Серебряков В.А. и др. Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии. – 2005. – Т. 10. – С. 29–48.

[19] Кулагин М.В., Лопатенко А.С. Научные информационные системы и электронные библиотеки. Потребность в интеграции // Сборник трудов Третьей Всероссийской конференции по электронным библиотекам – RCDL'2001, Петрозаводск, 11–13 сент. 2001 г. – С. 14–19.

[20] Мазов Н.А., Гуреев В.Н. Проблемы идентификации метаданных в наукометрических базах данных Web of Knowledge, Scopus и РИНЦ на примере профилей авторов // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 19-я междунар. конф. «Крым 2012» (Судак, 2–10 июня 2012 г.): Труды конф. – М.: Изд-во ГПНТБ России, 2012. – С. 1–4. – http://www.gpntb.ru/win/inter-events/crimea2012/disk/124.pdf

[21] Гуреев В.Н., Мазов Н.А. Идентификация в информационных библиографических системах: проблемы и решения // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 21-я междунар. конф. «Крым 2014» (Судак, 7–15 июня 2014 г.): Труды конф. – М.: Изд-во ГПНТБ России, 2014. – С. 1–7. – http://www.gpntb.ru/win/inter-events/crimea2014/disk/066.pdf

[22] CERIF. http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

[23] CERIF in Brief. http://cerifsupport.org/cerif-in-brief/

[24] Jaro M. A. Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society, 84(406): 414–420, 1989.

[25] Hernandez M. A., Stolfo S. J. Real-world data is dirty: data cleansing and the merge/purge problem. Journal of Data Mining and Knowledge Discovery, 1(2), 1998.

[26] Christen P., Churches T. Febrl: Freely extensible biomedical record linkage Manual, release 0.2.2 edition, November 2003.

[27] Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. Акад. наук СССР. – 1965. – Т. 163, № 4. – С. 845–848.

[28] Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, 2008. – ISBN 0-521-86571-9.

[29] Российский коммуникативный формат (RUSMARC) [Электронный ресурс] : [сайт] / Мин-во культуры Рос. Федерации, Рос. библ. ассоц., Нац. служба развития системы форматов RUSMARC. http://www.rusmarc.ru/index.html

## Experience of Person Identification for CRIS-Systems

Anna A. Knyazeva, Igor Y. Turchanovsky, Oleg S. Kolobov, Oleg L. Zhizhimov

The system of persons identification which was created in the process of development of a unified repository of scientific and technical activities (RSTA) in ICT SB RAS is described in this paper. The principles and methods used to create the system as well as its structure are briefly described. An algorithm for establishing an authoritative database with descriptions of persons automatically, without user intervention, is given. Indexing with bigrams and editing distance were used for detecting near-duplicate references to persons.