Слияние авторитетных/нормативных данных для распределённого электронного каталога библиотек Ленинградской области

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

Институт вычислительных технологий СО РАН

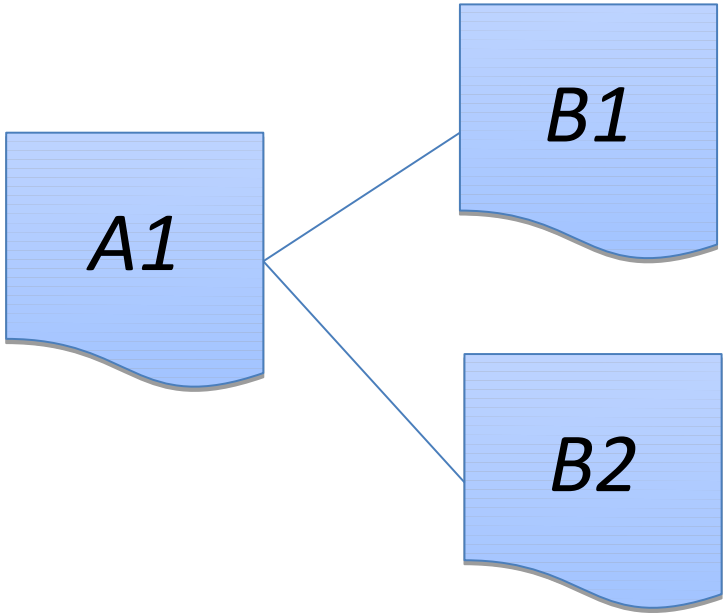
Институт сильноточной электроники СО РАН

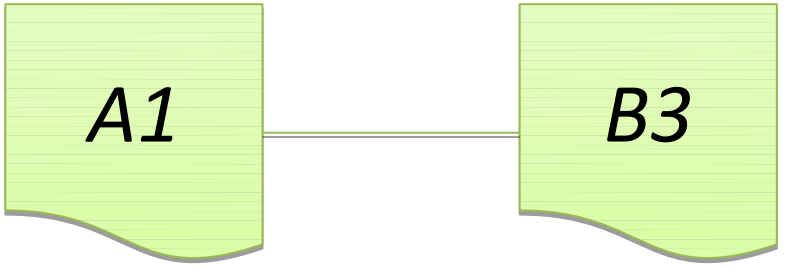
XV Российская конференция с международным участием
«Распределённые информационно-вычислительные ресурсы» (DICR-2014)

Постановка задачи

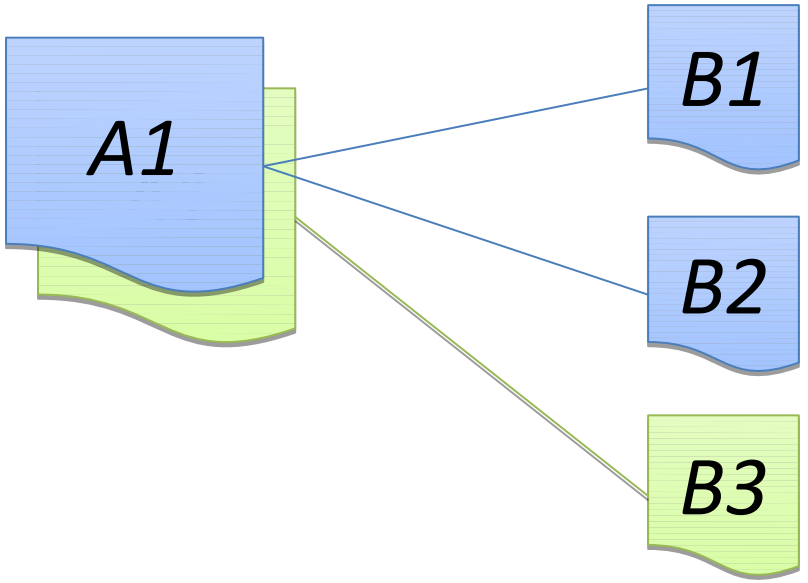
Разработка инструментария для слияния авторитетных данных с выявлением дубликатов и сохранением установленных связей без участия каталогизатора.

Простое объединение ресурсов





Слияние авторитетных данных



Близкие задачи

- Идентификация сущностей (Entity identification)
- Связывание записей (Record linkage)
- Выявление дубликатов (Duplicate detection)
- Разрешение имен авторов (Author name disambiguation)
- ...

Проект VIAF

The Virtual International Authority File (англ.) – Виртуальный авторитетный файл

Международный проект, начинался с совместной работы:

- Library of Congress (LC),
- Deutsche Nationalbibliothek (DNB),
- Bibliotheque nationale de France (BNF)
- OCLC.

К началу 2012 г. включал 20 организаций из 16 стран

http://viaf.org/

Анализ данных (пример)

№	Фамилия (200$a)	Инициалы (200$b)	Расшифровка (200$g)	Дополнение (200$c)	Даты (200$f)	Кол-во
1	Кафка	Ф.	-	-	-	2/14
2	Кафка	-	-	Франц	-	1/2
3	Кафка	Ф.	-	Франц	-	6/21
4	Кафка	Ф.	писатель	-	-	1/2
5	Кафка	Ф.	-	Франц	1883-1924	1/2
6	Кафка	Ф.	австр. писатель	Франц	1883-1924	1/6
7	КАФКА ФРАНЦ (ПИСАТЕЛЬ)	-	О НЕМ	-	-	4/0
8	КАФКА Ф.	(«ДНЕВНИКИ»)	-	-	-	4/0
9	КАФКА Ф.	(«ПРЕВРАЩЕНИЕ»)	-	-	-	4/1
					Всего записей:	**24/48**

Эксперимент

Тестовые наборы авторитетных записей

Описание результатов работы для каждого набора

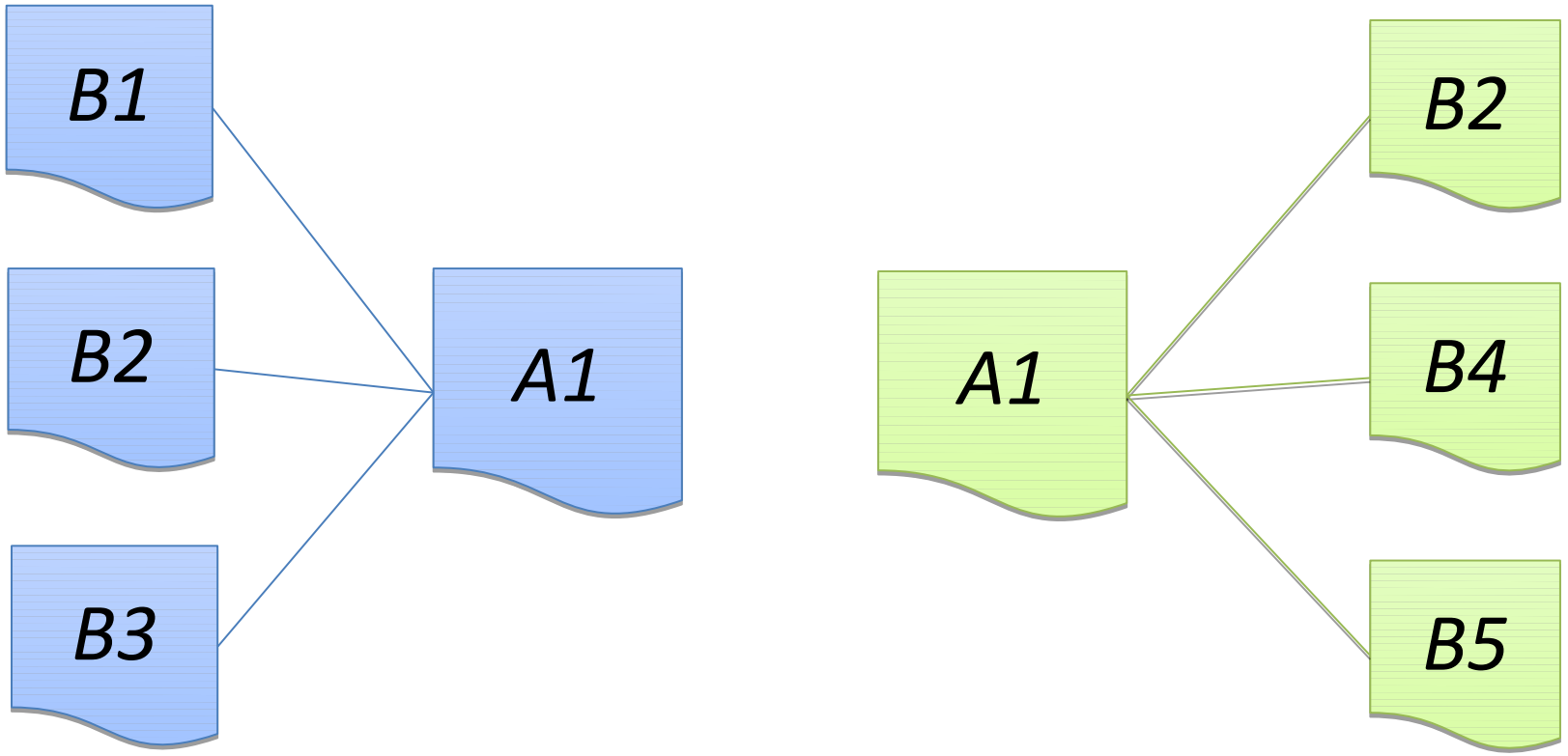
Возможные ошибки:

- I рода: ложное отрицание связи (больше записей);
- II рода: ложные связи (меньше записей).

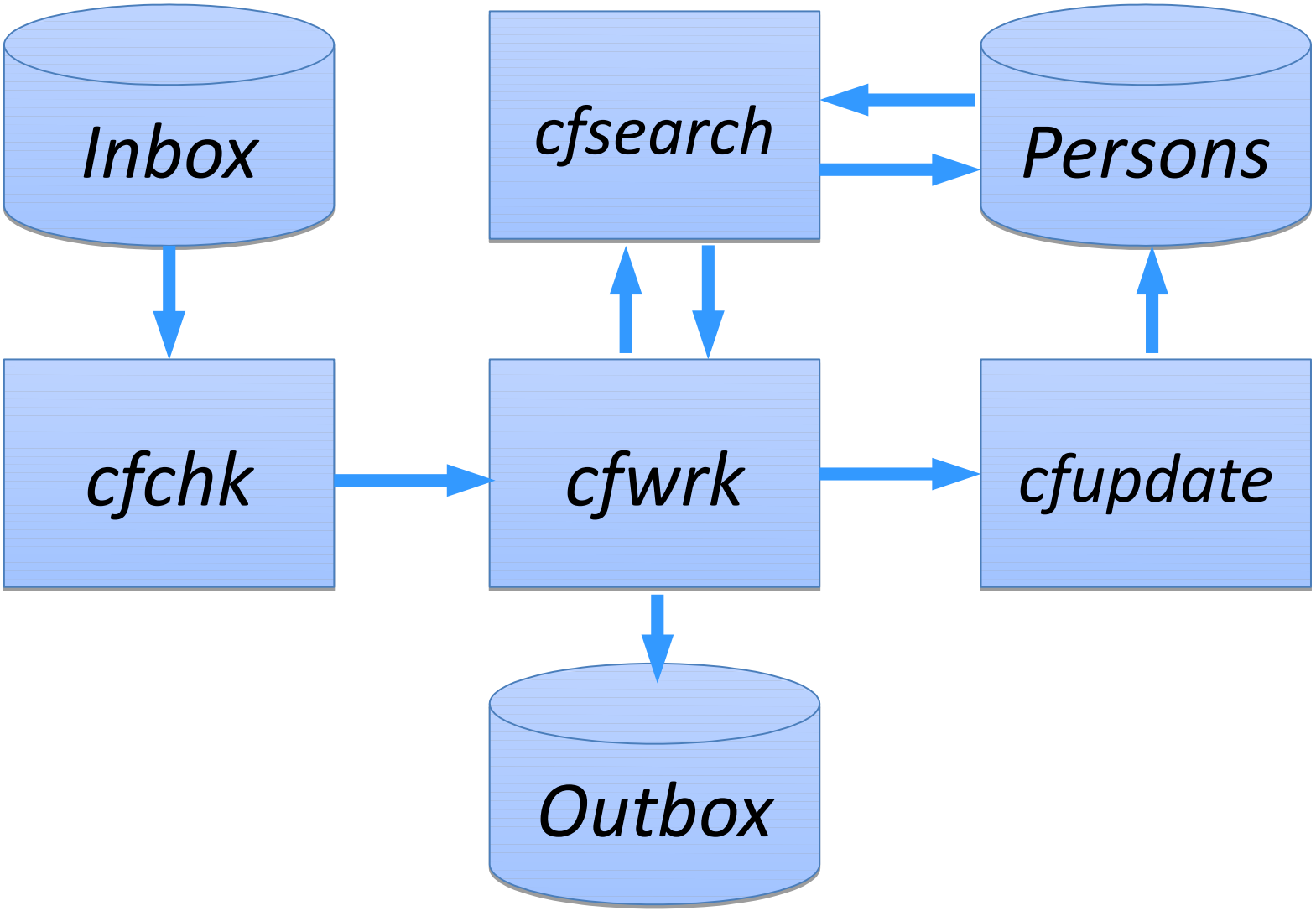
Основные принципы, используемые в работе

- Учёт систематических ошибок;
- Учёт разного качества входных записей (основные и дополняющие записи);
- Использование расширенных авторитетных записей;
- Работа в различных режимах (слияние и реиндексация).

Расширенные авторитетные записи



Функциональная схема *cflib*



Спасибо за внимание!

Слияние авторитетных/нормативных данных для распределённого электронного каталога библиотек Ленинградской области

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

Институт вычислительных технологий СО РАН

Институт сильноточной электроники СО РАН

aknyazeva@ict.nsc.ru