

Открытый доступ к содержанию цитирований

Паринов Сергей,
ЦЭМИ РАН и РАНХиГС

О проекте

- Финансируется РАНХиГС с середины 2016 года
- Руководитель – Оксана Медведева (РАНХиГС)
- Основные цели проекта:
 - 1) создание открытого ресурса с данными о цитированиях в научных публикациях на русском языке
 - 2) разработка методов качественного анализа содержания цитирований и создание на этой основе новых наукометрических показателей

Основные текущие результаты

- Создана технология потокового извлечения данных о цитированиях из текстов научных публикаций в виде PDF файлов
- Извлекаемые данные о цитированиях содержат
 - источники из списка литературы
 - сноски на источники в тексте публикации
 - связи цитирования между публикациями
- Коллекции публикаций обрабатываются в ежедневном режиме
- Программное обеспечение технологии, а также результаты ее применения свободно доступны всем желающим

Обрабатываемые данные о научных публикациях

- Источником данных о публикациях является научная информационная система Соционет
 - Обеспечивает метаданные и ссылки на полные тексты в стандартизованном виде
 - Осуществляет входной контроль качества
 - Включает полное содержание RePEc
 - Позволяет добавление отдельных публикаций, коллекций публикаций (отдельные журналы) и групп коллекций (ресурсы издателя)
- Уже обрабатываются публикации НЭИКОН, РАНХиГС, Институт Гайдара, СЕФИР, Европейский Университет, НИУ ВШЭ, и др.
- Планируем сотрудничество с проектом НЭИКОН «Национальный агрегатор открытых репозиторий российских университетов»

Результаты обработки публикаций

- Текущие статистические данные об обработке публикаций тут - <http://citru.repec.org/stats.html>

series	[1] records	[2] with futli	[3] with WARCs	[4] with PDF WARCs	[5] with JSON	[6] with HTML
repec:bkr:wpaper	26	26	17		13	13
repec:cas:wpaper	26	26	14			
repec:cfr:cefirw	228	228	127		122	121
repec:eer:wpalle	240	240	135			
repec:eus:ce3swp	26	26	18		18	18
repec:eus:wpaper	53	53	29		29	28

Статистика обработки публикаций на 20.04.2018

Коллекций	203
Всего публикаций	90413
С полными текстами	82983
Имеют источники	47463
Источников	837869
Цитирований	780348
Ссылок на источники	1111150
Связи цитирования	44538
Неупоминаемые источники	123464

Только примерно половина публикаций имеют данные о цитированиях в стандартном виде.

Из них у 36% есть источники без ссылок цитирования.

Около 15% распознанных источников в списках литературы не имеют ссылок цитирования.

В частности, ссылки не распознаются из-за ошибок в оформлении.

Открытый источник данных о цитированиях

Ежедневно пополняемые и обновляемые наборы данных на <http://peren.openlib.org/>

- представляет собой набор вложенных папок, имена которых образованы из кодов публикаций
- папка RePEc – коллекции/публикации из RePEc
- папка spz – коллекции/публикации из Соционет

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 <u>RePEc/</u>	2018-04-23 00:17	-	
 <u>spz/</u>	2017-10-18 19:10	-	

Набор данных об отдельной публикации

Извлеченные данные о цитированиях - в файле **summary.xml**
Остальные файлы содержат отчеты о процессе обработки и промежуточные результаты

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 0.pdf-stream.json	2017-10-24 02:46	305K	
 0.pdf-stream.json.err	2017-10-24 02:46	0	
 0.pdf-stream.json.log	2017-10-24 02:46	0	
 0.pdf-stream.reci.xml	2018-04-22 00:37	22K	
 0.pdf-stream.reci.xml.err	2018-04-22 00:37	0	
 summary.xml	2018-04-22 04:25	22K	

Пример извлеченных данных об источниках из списка литературы

```
▼<reference num="1" start="39317" end="39532" author="Лысенко Бетанов Звягин "  
title="Теоретические основы баллистиконавигационного обеспечения космических полетов"  
year="2014">  
▼<from_pdf>  
    Лысенко Л.Н., Бетанов В.В., Звягин Ф.В. Теоретические основы  
    баллистиконавигационного обеспечения космических полетов. Монография / под ред. Л.Н.  
    Лысенко. М.: Изд-во МГТУ им. Н.Э. Баумана, 2014. 518 с.  
    </from_pdf>  
</reference>
```

Пример извлеченных данных о содержании цитирования источника

▼<intextref>

▼<Prefix>

При этом принято различать запросную технологию измерения дальностей и беззапросную технологию (БЗ–технологию) кодовых и фазовых измерений псевдодальностей. Последняя, в силу ряда причин, является предпочтительной, и потому позиционируется как основная ГЛОНАСС–технология

</Prefix>

▼<Suffix>

. Тем не менее, и запросная технология не потеряла своей значимости и продолжает широко использоваться при решении таких вспомогательных задач, как формирование ЭО резервных КА спутниковой системы, либо КА системы при нестабильной работе бортового синхронизирующего устройства.

</Suffix>

<Start>2029</Start>

<End>2031</End>

<Exact>[1]</Exact>

<Reference>1</Reference>

</intextref>

Пример служебных данных

```
▼<document type="citmap" provider="RANEPA, CitEcCyr project" updated="2018-04-22T04:25:37.000000+00:00" created="2017-10-24T02:46:45.000000+00:00">
  ▼<source handle="spz:neicon:aerospace:y:2015:i:1:p:1-17">
    ▼<futli url="http://aerospace.elpub.ru/jour/article/viewFile/55/45.pdf">
      ▼<version
        url="rsync://evstu.openlib.org/warc/spz/neicon/aerospace/y%3A2015%3Ai%3A1%3A17.warc" start="2371" length="647683">
          ▼<text
            rich="rsync://evstu.openlib.org/peren/spz/neicon/aerospace/y%3A2015%3Ai%3A17/0.pdf-stream.json">
              ▼<input
                data="rsync://evstu.openlib.org/peren/spz/neicon/aerospace/y%3A2015%3Ai%3A17/0.pdf-stream.pdf">
```

Источник: <http://peren.openlib.org/spz/neicon/aerospace/y%253A2015%253Ai%253A1%253Ap%253A1-17/summary.xml>

Пример использования данных о цитированиях в Соционет

псевдодальностей.

Последняя, в силу ряда причин, является пред-
позиционируется как основная ГЛОНАСС-технология [1] 1
технология не утратила своей значимости и продолжает п-
решении таких вспомогательных задач, как формирование ЭО

[1]

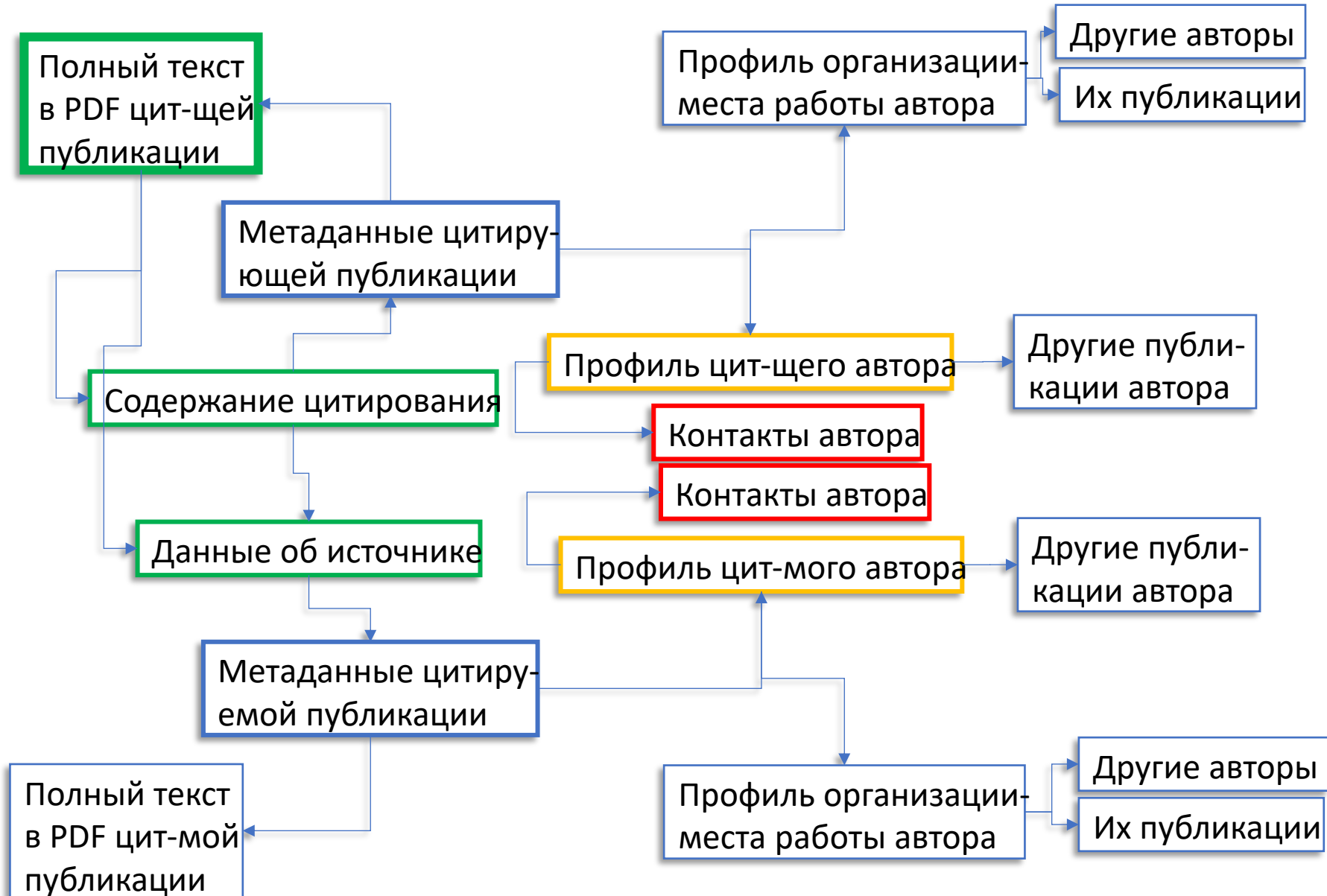
[1] -> Лысенко Бетанов Звягин (2014) Теоретические основы
баллистиконавигационного обеспечения космических полетов,
сносок в документе -1, сноска всего - 1

CyrCitEc Project, RANEPa, 2018-04-05, More..

Анализ содержания цитирований

- Цели: а) создать открытый ресурс данных о содержании цитирований; б) разрабатывать методы качественного анализа этих данных
- Первичный массив данных – 780348 строчек с контекстом цитирований (на 20.04.2018)
- Возможные дополнительные данные из Соционет:
 - метаданные о цитирующей и цитируемой публикациях (иногда и полный текст цитируемой публикации)
 - данные персональных профилей авторов цитирующей и цитируемой публикаций
 - данные о других публикациях авторов и их цитированиях
 - данные о местах работы авторов и других ученых, работающих в этой же организации, включая их публикации, персональные профили и т.п.

Существующие связи между данными



Открытые данные для анализа содержания цитирований

- Особенности текущего этапа – создание ежедневно обновляемых массивов текстовых данных для:
 - анализа тональности и функций цитирований
 - создания качественных характеристик цитирующих и цитируемых авторов
 - анализа близости слов из контекстов цитирований
- Все данные доступны на <http://citru.repec.org/Word2Vec/>

Нормализация слов в контексте цитирований

- Исходный текст:
 - В связи с этим использование препаратов, блокирующих действие ФНО α (в частности, моноклональных антител к ФНО α), могло бы замедлить прогрессирование ХСН и снизить смертность.
- Нормализованный текст:
 - препарат_NOUN': 2, 'действие_NOUN': 2, 'мочь_VERB': 2, 'смертность_NOUN': 2, 'исследование_NOUN': 2, 'связь_NOUN': 1, 'использование_NOUN': 1, 'частность_NOUN': 1, 'моноклональный_ADJ': 1, 'антитело_NOUN': 1, 'прогрессирование_NOUN': 1,

Наиболее употребительные слова в контексте цитирований

Глаголы

241175 быть
169123 являться
144247 мочь
65044 иметь
41485 позволять
38782 приводить
38274 составлять
32534 показать
31208 следовать
26413 использоваться

<http://citru.repec.org/Word2Vec/verb.freq.txt>

Существительные

299072 год
239814 исследование
195462 пациент
152217 развитие
131149 уровень
130792 результат
123318 время
117531 лечение
114689 группа
111873 система

<http://citru.repec.org/Word2Vec/noun.freq.txt>

Предложения для научно-образовательного сообщества

- Проект приглашает к сотрудничеству по следующим направлениям:
 - Добавлять свои публикации для анализа содержания цитирований
 - Использовать созданную технологию и массивы данных в собственных целях
 - Участвовать в разработке методов качественного анализа содержания цитирований
- РАНХиГС планирует развивать экспертное сообщество в области качественной наукометрии
- Контакты:
 - Оксана Медведева (oxana.medvedeva.1984@gmail.com)
 - Сергей Паринов (sparinov@gmail.com)