Открытые данные русскоязычных научных публикаций:
возможности для индексирования и наукометрического анализа

Проект CitEcCyr, РАНХиГС

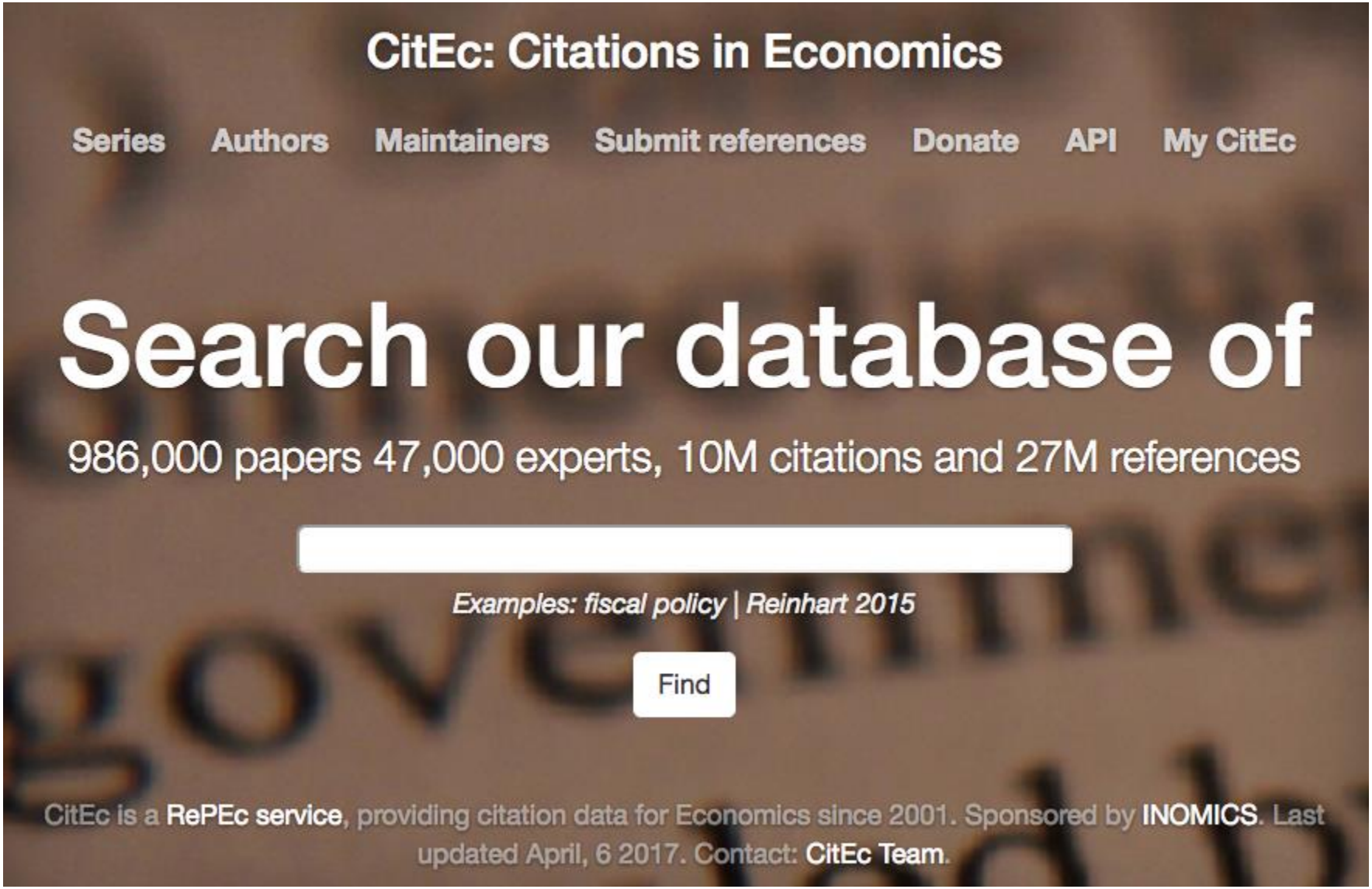
Вирус прозрачности и эпидемия открытости

- В обществе повышается интерес к контролю, создающий спрос на прозрачность и открытость:
  - Почему мы верим показателям результативности ученых, рассчитанных коммерческими организациями?
  - Можно ли сделать, чтобы формирование и обновление таких показателей было прозрачной процедурой, открытой для публичного контроля и не связанной с чьими-то коммерческими интересами?
- Да, это возможно на базе Открытой научной инфраструктуры

Открытая научная инфраструктура (ОНС)

- Bilder G, Lin J, Neylon C. (2015) Principles for Open Scholarly Infrastructure, http://dx.doi.org/10.6084/m9.figshare.1314859
- Общая идея -
  - если научные сервисы в Интернет создаются как «прозрачные» технологии с открытым бесплатным доступом к софту и создаваемым данным,
  - то, если эти сервисы кому-то нужны, научное сообщество само децентрализовано обеспечивает и поддерживает их работоспособность
- Один из близких примеров – инфраструктура RePEc и один из ее сервисов – CitEc

CitEc: http://citec.repec.org/



Проект CitEcCyr: Открытые данные о цитированиях как фрагмент ОНС

- В июне 2016 г. стартовал 3-х летний проект CitEcCyr, финансируемый РАНХиГС
- Команда - 7 человек, из них 5 – разработчики
- Цели:
  - Развитие системы CitEc для обработки цитирований в научных публикациях на русском языке
  - Создание технологии для анализа содержаний цитирований (citation content analysis)
  - Все результаты проекта (софт, технология и данные) должны быть прозрачны и открыты для общественного контроля

Zhang G., Ding Y. and Milojević S. **Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content**. arXiv:1211.6321, 2012

- В таблице 4 предложены атрибуты отношений цитирования. Например,
- Location of mentioning: 1) Abstract; 2) Introduction; 3)Literature Review; 4) Methodology; 5) Results/discussion; 6) Conclusion; 7) Others
- Frequency of mentioning: 1) Once; 2) 2 to 4 times; 3) 5 times or more
- Style of mentioning: 1) Not specifically mentioning; 2)Specifically mentioning but interpreting; 3) Direct quotation

Открытость и проверяемость результатов CCA

- Нужно показать авторам и научному сообществу результаты анализа цитирований прямо в тексте публикации
- Подходит технология аннотирования PDF документов
  - Стандарт утвержден W3C - https://www.w3.org/TR/annotation-model/
  - Софт свободно доступен - https://github.com/hypothesis
- Результаты анализа содержания цитирований показываются как аннотации при просмотре PDF публикаций

CitEcCyr: анализ содержания цитирований, конвертация PDF

- Этап 1 – конвертация PDF документов. Нужен текст, удобный для: 1) извлечения нужных данных; 2) определения текстовых координат, где должна быть выведена аннотация
- Используем потоковую конвертацию на основе node.js + PDF.js, код на github.com/citeccyr
- Результат в двух файлах: 1) в формате JSON для извлечения данных; 2) в формате plain text для подсчета координат
- Все эти файлы свободно доступны для внешнего анализа содержания цитирований

CitEcCyr: анализ содержания цитирований, извлечение данных

- Этап 2 – извлечение данных, необходимых для анализа содержания цитирований
- На данный момент мы извлекаем и анализируем:
  - Содержание списков литературы, разбивка на поля, связывание с метаданными соответствующих публикаций, определение координат
  - Ссылки на источники из списка литературы, определение координат, выделение контекста справа и слева от ссылки, источники без ссылок
- На будущее – извлечение и учет структуры разделов публикаций, классификация контекста и др.
- Все результаты находятся в открытом доступе

CitEcCyr: анализ содержания цитирований, слияние данных

- Анализ содержания цитирований может выполняться всеми желающими
- Нужен способ объединять распределенные фрагменты этих данных в единую карту цитирований для заданной публикации
- Мы разработали ХМЛ шаблон карты цитирований и механизм слияния этих данных для заданной публикации

CitEcCyr: анализ содержания цитирований, визуализация

- Этап 3 – визуализация результатов анализа содержания цитирований при просмотре PDF соответствующей публикации
  - Для каждой ссылки на источник нужно одним кликом получать подробные данные об источнике, включая общее количество ссылок на него и контекст этих ссылок
  - Для каждого источника из списка литературы нужно одним кликом получать контекст ссылок на него в данной публикации, а также во всех других

In the CRIS-2010 conference paper [1] some challenges for a CRIS-CERIF de 1
relation with the SocioNet project:

- How should we construct in form and function a system for shaping and sharing
  individual researchers stored at institutional repositories (IR) or/and CRISs s
  scientific circulation and necessary conditions for its maximal usage? [1]
- How should we organize a process of research outputs/results usage, and desig
  to provide maximally comprehensive and accurate statistics on the uptake, usag 3
  results? [1]

Now, six years later, it has become clear that there are many successful approacne 1

Parinov S. A CRIS driven by research community: benefits and perspectives. In the proceedings of the 10th International Conference on Current Research Information Systems (CRIS-2010), 2010

> *Cited fragment: "1. How should we organize a process of research outputs/results usage, and design necessary metrics and tools to provide maximally comprehensive and accurate statistics on the uptake, usage and impact of research results?"*

Research results in this paper related with the global scholarly communication 23
funded by the Russian Foundation for Basic Research, grant 15-07-01294-a. This
experimental interactive tools for assessment of non-material assets and conducting applie
funded by the Russian Science Foundation, grant 14-18-01999.

References

1. Parinov S. A CRIS driven by research community: benefits and perspectives. In the proceedings of the *10th Inte*
Current Research Information Systems (CRIS-2010), 2010.
http://eurocris.org/Uploads/Web%20pages/cris2010_papers/Papers/cris2010_Parinov.pdf
2. Kramer B, Bosman J. Survey of scholarly communication tool usage, 2016. https://101innovations.wordpress.c 3
3. Parinov S. Open Repository of Semantic Linkages. In the proceedings of the *11th International Conference on*
Systems (CRIS-2012). 2012.

Only me

> *1. Parinov S. A CRIS driven by research community: benefits and perspectives. In the proceedings of the 10th International Conference on Current Research Information Systems (CRIS-2010), 2010.*
> *http://eurocris.org/Uploads/Web%20pages/cris2010_papers/Paper*

In-text references for the document, total: 3

```
1. In the CRIS-2010 conference paper **[1]** some
challenges for a CRIS-CERIF

2. scientific circulation and necessary conditions for
its maximal usage? **[1]** •How should we organize a
process of research outputs

3. on the uptake, usage and impact of research results?
**[1]**. Now, six years later, it has become clear
```

Основные этапы работ по проекту CitEcCyr

- До конца года - переход от единичных примеров к обработке большого тестового массива публикаций (около 100 тыс. публикаций)
- Первая половина 2018 г. - запуск созданной системы в работу и обработка ежедневно поступающего массива новых публикаций, открытый доступ ко всем данным
- Вторая половина 2018 г. – доработка системы визуализации данных анализа содержания цитирований
- 2019 г. – разработка на этой базе новых наукометрических показателей и т.д.

Создаваемые возможности для наукометрического анализа

- Учет количества ссылок в тексте статьи на источники из списка литературы. Отделение источников без ссылок
- Обработка контекста вокруг ссылок для классификации содержания цитирований источников
- Ранжирование ссылок на статьи по их месту в структуре статьи (ранг выше в разделе с результатами, ранг ниже в разделе обзор литературы)
- и т.д.

Приближается конец эпохи публикаций как средства научных коммуникаций?

- Сначала: ученые выбирают из чужих публикаций интересные фрагменты, связывают их со своими идеями, чтобы создать новое научное знание, и делают из все этого публикации
- Затем: мы конвертируем эти PDF публикации в текст, пытаемся в нем разобраться какие результаты из этих работ и каким образом ученые использовали для подготовки публикации
- Зачем нужно делать публикации, когда можно организовать процесс использования результатов и учет этого использования более эффективно?

Контакты

- Оксана Медведева, административный руководитель проекта, РАНХиГС
  - oxana.medvedeva.1984@gmail.com
- Сергей Паринов, руководитель группы разработчиков, ЦЭМИ РАН и РАНХиГС
  - sparinov@gmail.com