# Open citation content data

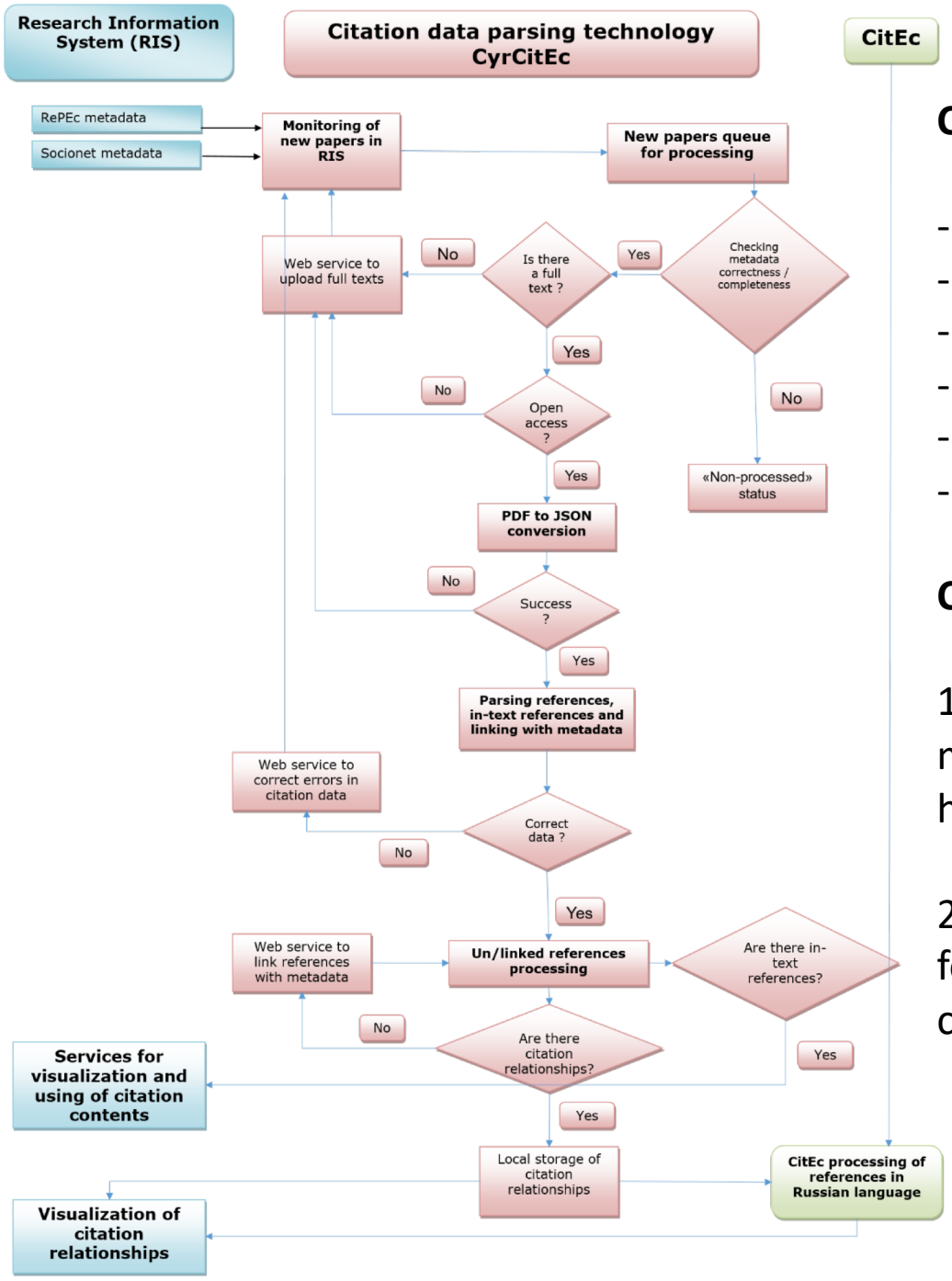
## Cirtec project
## (former CyrCitEc/CitEcCyr)

Sergey Parinov, CEMI RAS and RANEPA

Cirtec project is funded by Russian Presidential Academy of National Economy and Public Administration (RANEPA)

# Cirtec main principles

- Open infrastructure. Two initial nodes: CitEc (http://citec.repec.org/) and Cirtec systems with a specialization on processing papers in specific languages. Other nodes, e.g. specialized on processing citation data in languages, like Chinese, Japanese, Arabic, etc., could be added by the same way. There is also an intention to integrate data about references into the OpenCitations Corpus (http://opencitations.net/).
- Transparency. Cirtec allows publishers, authors and readers of papers to see how the citation data of their papers were extracted by the system. They can trace why some papers' references / in-text citations are not processed or not counted.
- Enrichment. Integration with research information system (RIS). Providing tools for authors of papers to enter additional data to correct errors of processing citations found in their papers and to enrich their citation relationships.
- Public control. Readers of papers can publicly or private react to authors misbehavior in order to increase their number of citations by using the enrichment facilities.




## Cirtec Technology:

- Takes papers from RePEc and Socionet
- Returns citation data to RePEc/Socionet
- Integrated by data with CitEc/RePEc
- Uses PDF.js to convert PDF to JSON
- Stores citation data as XML files
- Provides open access to produced data

## Cirtec Outputs (2 of 4):

1. Open source software to parse papers' metadata and full text PDFs available at https://github.com/citeccyr

2. Open service to process papers' PDFs for extracting citation data including citation contexts

# 3. Open dataset at http://cirtec.ranepa.ru/data/


```
▼<document type="citmap" provider="RANEPA, CitEcCyr project" updated="2018-05-21T07:15:00.000000+00:00"
 created="2018-03-23T00:16:15.000000+00:00">
  ▼<source handle="repec:hig:fsight:v:11:y:2017:i:4:p:84-95">
    ▼<futli url="https://foresight-journal.hse.ru/data/2018/01/14/1160539232/8-Kuzyk-84-95.pdf">
      ▼<version url="rsync://citru.repec.org/warc/RePEc/hig/fsight/v%3A11%3Ay%3A2017%3Ai%3A4%3Ap%3A84-
       95.warc" start="2501" length="258033">
```

```
▼<intextref>
  ▼<Prefix>
     Still, in this regard Russia remains far behind not only the countries that traditionally
     adhere to the university-based R&D model, but even behind some of the former socialist
     countries and Soviet republics
   </Prefix>
  ▼<Suffix>
     ; Gokhberg, Kuznetsova, 2011]. Russian universities' R&D cooperation with businesses does
     not look very impressive either (Figure 2). However, the current Russian situation with
     universities' research and innovation activities is not at all unique.
   </Suffix>
   <Start>8757</Start>
   <End>8781</End>
   <Exact>[Gokhberg et al., 2009</Exact>
   <Reference>20</Reference>
 </intextref>
```

```
▼<reference num="20" start="54464" end="54654" author="Gokhberg Kuznetsova Zaichenko "
 title="Towards a New Role of Universities in Russia Prospects and Limitations" year="2009"
 handle="repec:oup:scippl:v:36:y:2009:i:2:p:121-126">
  ▼<from_pdf>
     Gokhberg L., Kuznetsova T., Zaichenko S. (2009) Towards a New Role of Universities in
     Russia: Prospects and Limitations. Science and Public Policy, vol. 36, no 2, pp. 121–126.
   </from_pdf>
 </reference>
```

# 4. Statistics and a monitoring tool on the citation data extraction process

- To monitor everyday changes, missed/damaged papers, processed/unprocessed citation data, etc.
- A fragment of the main page -

| series | [1] records | [2] with futli | [3] with WARCs | [4] with PDF WARCs | [5] with JSON |
|---|---|---|---|---|---|
| RePEc:bkr:wpaper | 26 | 26 | 21 | 17 | 17 |
| RePEc:cas:wpaper | 26 | 26 | 20 | | |
| RePEc:cfr:cefirw | 228 | 228 | 170 | 164 | 160 |
| RePEc:eer:wpalle | 240 | 240 | 194 | | |
| RePEc:eus:ce3swp | 26 | 26 | 23 | 23 | 23 |
| RePEc:eus:wpaper | 53 | 53 | 42 | 42 | 40 |
| RePEc:gai:gbchap | 37 | 37 | 37 | 36 | 36 |

http://cirtec.ranepa.ru/stats.html

# Statistics on dataset of citation data

| Statistics on 2018.09.01 | Totals |
|---|---|
| processed collections of papers | 317 |
| metadata records available | 144,250 |
| records with links to paper's full text | 132,035 |
| PDF files in Web ARChive | 108,823 |
| JSON files with found reference sections | 74,268 |
| total references | 1,272,126 |
| total citation contexts | 1,203,358 |
| total mentioned references | 1,091,996 |
| total citation relationships (including DOI) | 166,976 |
| total non-mentioned references | 180,130 |

50%

15%

We accumulate and store all Cittec statistics from 2018-07-05

Source: http://cirtec.ranepa.ru/stats.html

# Current Cirtec activities: citation contexts analysis

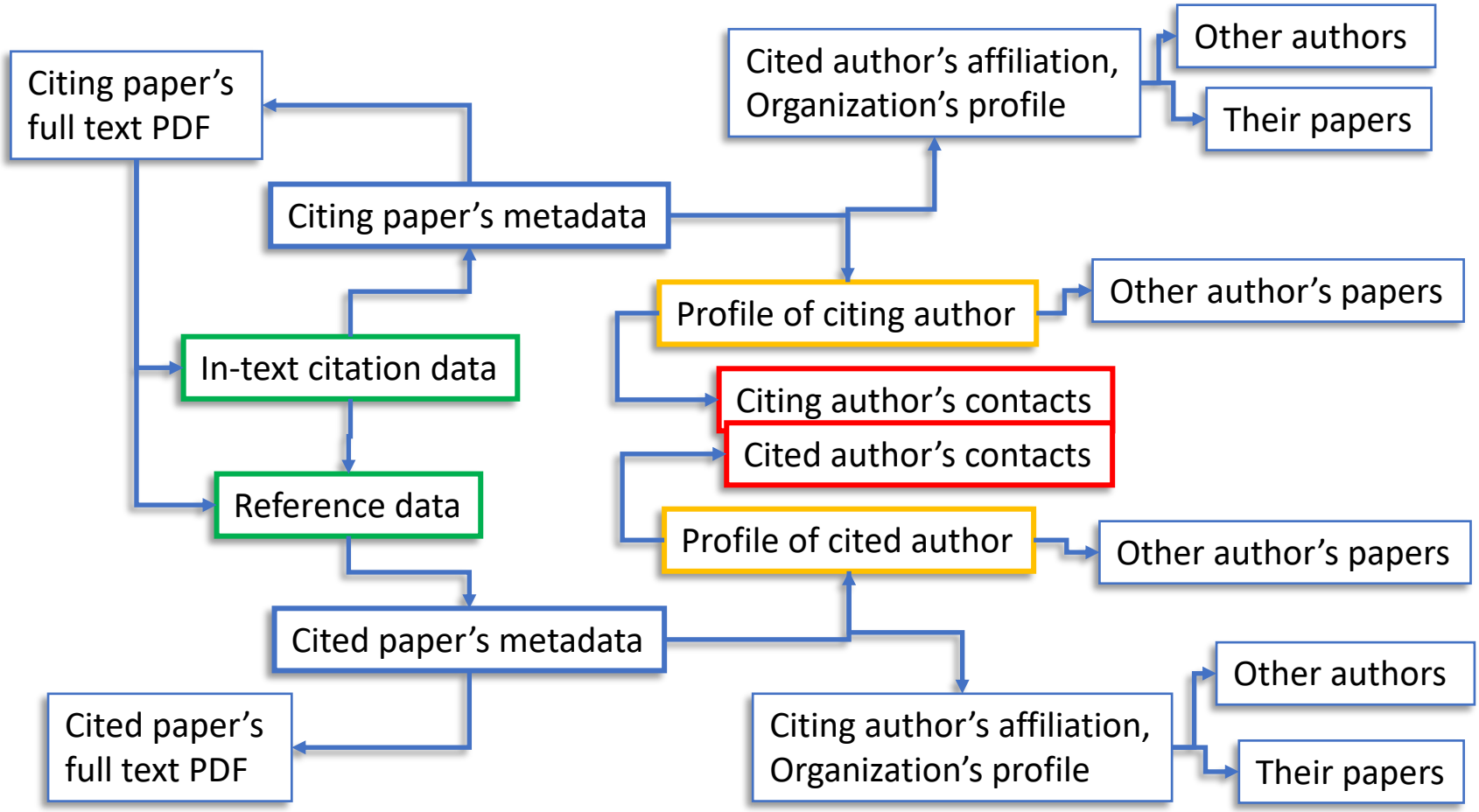
- Index of references that provides for each reference:
  - number/ID of papers where the reference occurs
  - number/ID of in-text citations for the reference (by papers)
  - citation contexts for the reference (by papers)
- Co-occurrence of references in papers
  - frequencies and list of references with common citation contexts
  - common citation contexts as characteristics of similarities between references
- Polarity of citation contexts (sentiment analysis)
- Word2vec and Doc2vec analysis of citation contexts (similarity analysis)

# Future Cirtec: ambitious aims

- Transformation of the in-text citations into interactive elements:
  - to make channels for scholarly communication and research cooperation
- Using these channels:
  - the cited authors know who used what of their outputs
  - the cited authors can inform the citing authors about upgrades with cited outputs
  - the citing authors can send requests to cited authors on needed development of cited outputs
- As a result,
  - the research community has wider, than now, scholarly cooperation
  - scholars have better individual research performance

# Research Information System (RIS)

If we integrate citation data into RIS with a rich semantic layer, we can enrich the data by many additional attributes, like citing/cited authors contact data, etc.



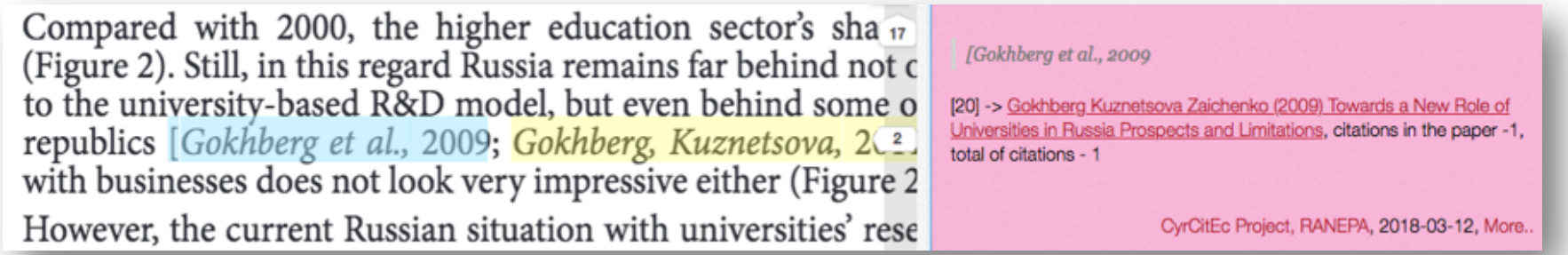
# Interactive in-text citations: first experiments

- PDF.js module to convert PDF to JSON
- Hypothes.is annotation tool within Socionet
- formatting citation data by the Web Annotation Data Model

Computer-generated annotations for the in-text citations


```
{"target":·[¶
··{"source":·"·https://foresight-journal.hse.ru/data/…",·¶
···"selector":·[¶
  →  {"type":·"TextPositionSelector","end":·8781,·"start":·8757},¶
  →  {"exact":·"[Gokhberg·et·al.,·2009"}¶
  →  ]¶
··"tags":·[some·tags],·¶
··"text":·"some·information·and·statistics·about·cited·reference"¶
}]}¶
```


A fragment of paper's PDF with annotated in-text citations




source: https://goo.gl/bZJwzZ

# Taxonomy of cited author's reactions

| VALUES | FOR CITING AUTHORS | FOR CITED AUTHORS | FOR READERS |
|---|---|---|---|
| agree with this citation, comment | √ | √ | √ |
| disagree with this citation, comment | √ | √ | √ |
| ready to improve my paper | | √ | |
| ready to help with taking better effect from using my paper | | √ | |
| propose making a joint paper | | √ | |
| propose a joint development of my results | | √ | |
| misunderstanding of my paper | | √ | |
| protest against style of this citation | | √ | |

# Contacts

- Web: http://cirtec.ranepa.ru/
- Oxana Medvedeva, Cirtec project head, oxana.medvedeva.1984@gmail.com
- Sergey Parinov, Cirtec development group leader, sparinov@gmail.com