RuCRIS: a pilot CERIF-based system to aggregate heterogeneous data of Russian research projects

Guskov A.E. ¹, Zhizhimov O.L. ¹, Kikhtenko V.A. ², Skachkov D.M. ¹, Kosyakov D.V. ¹

¹Institute of computational technologies of the Siberian branch of Russian Academy of Sciences, ²Novosibirsk State University (Novosibirsk, Russia)
Introduction

There is a strong demand among scientific society and Russian government for information and analytics system for Russian research projects.

RuCRIS project, that was supported by Federal Program of Ministry of Education and Sciences of Russia is the effort to build such kind of system.

Project duration: 3 months*

*including 1 month for reporting stage
Goal

To develop technologies for aggregating and managing information on research projects and result outputs in a unified repository
Project tasks

1. Analysis of the existing methods for collecting, processing and distributing research deliverables

2. Design data gathering methods and procedures

3. Pilot implementation, which provides:
   - Research projects and results data harvesting.
   - Data classification, mapping and linking.
   - Storing the data in the central repository.
   - Visualizing data from the repository.
## Potential consumers

<table>
<thead>
<tr>
<th>Consumers</th>
<th>Purposes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>Research strategic management</td>
</tr>
<tr>
<td>Funders</td>
<td>Grants distribution</td>
</tr>
<tr>
<td>Research institutes and universities</td>
<td>Search for collaboration</td>
</tr>
<tr>
<td>Researchers</td>
<td>Search for related projects</td>
</tr>
<tr>
<td>Business, Industry</td>
<td>Search for innovations</td>
</tr>
</tbody>
</table>
## Data sources

<table>
<thead>
<tr>
<th>Repository</th>
<th>Access type</th>
<th>Project</th>
<th>Person</th>
<th>Organization</th>
<th>Result product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian Foundation for Basic Research</td>
<td>HTML</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian Foundation for Humanities Research</td>
<td>HTML</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Republican Research Scientific and Consulting Centre for Expertise</td>
<td>HTML</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Repository of Federal Programs of Russian Ministry of Education and Science</td>
<td>XML</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Center for Information Technologies and Systems for Executive Authorities</td>
<td>XML</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Projects repository of the Siberian Branch of the Russian Academy of Science</td>
<td>DB</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
## Problems

<table>
<thead>
<tr>
<th>Problem</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low collaboration level</td>
<td>Data owners aren’t engaged into integration efforts</td>
</tr>
</tbody>
</table>
Low collaboration problem
Data owners aren’t engaged into integration efforts

- Project catalogues on the website (name, annotation, executors)
- Some analytics in annual reports

- No definite results for almost all projects
- No API to access the data
- No custom analytics
## Problems

<table>
<thead>
<tr>
<th>Problem</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low collaboration level</td>
<td>Data owners aren’t engaged into integration efforts</td>
</tr>
<tr>
<td>Syntactic heterogeneity</td>
<td>There is no common way to collect the original data in preassigned systems</td>
</tr>
<tr>
<td>Semantic heterogeneity</td>
<td></td>
</tr>
<tr>
<td>Different data schemas</td>
<td>Different systems use their own data schemas and don’t provide any API.</td>
</tr>
<tr>
<td>Different classifiers</td>
<td>Similar classifiers contain very different classes (e.g. science fields)</td>
</tr>
<tr>
<td>Duplicate entities</td>
<td>The same entity exists in different sources.</td>
</tr>
</tbody>
</table>
Access data sources.
Plan A

Database 1
Adapter 1

Database 2
Adapter 2

Database 3
Adapter 3

RuCRIS repository

Continuous access to source DBs!
Access data sources.
Plan B

Database 1: XML exported
Converter 1
RuCRIS repository

Database 2: HTML crawled
Converter 2

Database 3: SQL DB exported
Converter 3

One-time or on schedule update
Problem: syntactic heterogeneity
no common way to collect the original data
Benefits of using CERIF

Research project management system have to be of national scale and even more – have strong connections with European CRIS systems.

Using CERIF for data storage and exchange is huge benefit in achieving of this goal.
Essential parts of CERIF
ruCRIS architecture

Client tier
- User interface

Data storage tier
- ruCRIS repository
- Data access

Data processing tier
- Formal checker
- Linking module
- Deduplication module
- Commit module

Data harvest tier
- Harvesting database
- Harvesting manager
- Adapters for external data sources

External data sources
Duplicates problem:
Is it the same John Smith or not?

<table>
<thead>
<tr>
<th>Entity</th>
<th>Preferable way</th>
<th>Alternative way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Check global IDs (e.g. ORCID)</td>
<td>Check name, degree, affiliation, email</td>
</tr>
<tr>
<td>Organization</td>
<td>Check global IDs (e.g. tax payer ID)</td>
<td>Check names, location</td>
</tr>
<tr>
<td>Classifier</td>
<td>Use unified classifiers</td>
<td>Generalization</td>
</tr>
</tbody>
</table>
RuCRIS statistics

- Projects: 22,130
- Persons: 9,300
- OrgUnits: 2,700

<table>
<thead>
<tr>
<th>Source</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFBR</td>
<td>9,587</td>
</tr>
<tr>
<td>RFHR</td>
<td>3,929</td>
</tr>
<tr>
<td>CIT</td>
<td>4,829</td>
</tr>
<tr>
<td>RRSCCR</td>
<td>2,247</td>
</tr>
<tr>
<td>FP</td>
<td>176</td>
</tr>
<tr>
<td>SBRAS</td>
<td>1,369</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>22,130</strong></td>
</tr>
</tbody>
</table>
1. Custom HR systems integration and research management workflows are handled on particular institution’s portal

2. Resulting data are replicated to the common CRIS Database

3. And published to the institution’s public-facing web site
CRIS/CERIF entities representation

All main CERIF entities are represented as corresponding web pages with unique and persistent URLs:

- OrgUnits
- Persons
- Projects
- Publications
- Et cetera
Publishing infrastructure

Based on Microsoft Sharepoint 2013 Cross-site Publishing features

- Sharepoint lists and libraries on different web-sites
- SQL Databases
- Sharepoint Search Databases
- Sharepoint Publishing Web Sites
Conclusion

• Mission completed: **pilot system works**
• Research open data in Russia: **still weak**
• Requirements for national CRIS:
  • global IDs
  • unified classification system
  • API for data access
RuCRIS: a pilot CERIF-based system to aggregate heterogeneous data of Russian research projects

Thank you for your attention!