Analytics & Reporting at different levels for a CRIS based on DSpace

the use case of the Peruvian National Platform
Funded by the World Bank, #PeruCRIS is the Peruvian project for setting up and operating the National Information Network on Science, Technology and Technological Innovation. It is based on open source software and open standards, specifically as an extension of the already existing network of Peruvian open access repositories, and it is led by Concytec, the Peruvian Council for Science, Technology and Innovation.

The vision for #PeruCRIS Project is:

**Vision**
To have an Information ecosystem in science, technology and technological innovation in Peru to provide value, accessibility and development for all people.

In order to accomplish the goal for #PeruCRIS Project, Concytec implements the following strategy:

**Main strategy**

1. Articulate, operate and regulate the National Information Network on STI.

2. Strengthen the National Network of Open Access Repositories (RENARE).

3. Launch the #PeruCRIS platform and interoperate with institutional research management systems.

4. Provide value-added services for accessing and managing information on STI in the country.
EXPECTED BENEFITS

- Timely statistics and reports on national R&D activities.
- Better monitoring and evaluation for public funding and national open access policy.
- Disseminate research results, analyze trends and impact.
- Sharing and discovery for innovative technologies and ideas, new markets, competitors and partners.
- Better decision making at different levels: national, local, institutional, private sector and general population.
Sustainability:
- Open-Source Software
- Agile Methodology

Interoperability:
- Open standards
- Best practices

**Project pillars**

<table>
<thead>
<tr>
<th>Stories</th>
<th>To Do</th>
<th>In Progress</th>
<th>Testing</th>
<th>Done</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Examples of interoperability tools and guidelines:**
- CERIF (Common European Research Information Format) CERIF XML & CERIF API
- OpenAIRE Guidelines for CRIS Managers
- OpenAIRE Guidelines for Literature Repositories
- OAI-PMH Harvesting Protocol
- Persistent Identifiers ORCID, ROR, DOI, Handle
- #PeruCRIS Guidelines (application profile, semantics, syntax and scope of data)
Main Directories will be populated through the following flows:

1. A subsystem allows for institutions to have private entities collections under their own private communities, where they are able to bulk import or directly edit information to be submitted to the main public directories.
2. OAI-PMH harvesting with PeruCRIS custom CERIF XML profile, designed to be the preferred way for data ingesting.
3. Custom connectors for national interoperability (personal national records - RENIEC, national degrees database - SUNEDU, national institutions tax database - SUNAT) and other bibliographic sources (Scopus, Web of Science, and Latin American SciELO collections).
4. A new version of the national profile CV system is being developed as its own DSpace community.
5. PeruCRIS Platform goes beyond mere aggregation, supporting data normalization, enrichment and curation for collected data, and direct editing in the Directorios.

The design and development of the platform is based on DSpace-CRIS and carried out by 4Science, [https://www.4science.it/](https://www.4science.it/).
Analytics components

OpenSearch is a community-driven, **open source search and analytics suite** derived from Apache 2.0 licensed Elasticsearch 7.10.2 & Kibana 7.10.2

Dremio is a **data lake engine** that creates a semantic layer and supports interactive queries.

Apache Superset is a modern **data exploration and visualization** platform
DSpace CRIS \rightarrow OpenSearch "former" dremio

Open Distro

Apache Superset

OpenSearch

RDBMS

Sparql

CSV
Different users & project scales require different solutions

"Built-in" powered by OpenSearch

Analytics add-on powered by OpenSearch

Data lake powered by Dremio & Superset
DSpace-CRIS supports a flexible search engine that allows data analysis and exploration. Results can be exported in configurable formats including CSV, XLS, PDF. Aggregation can be used to narrow the analysis (faceted browsing) and provide basic visualization.
It is possible to configure the graphical visualization (bar charts, pie charts, lines) with respect to any dimension of aggregation for searchable or predetermined lists (projects of a department, of a researcher, of a certain area, etc). The visualization is interactive, and it is possible to progressively shift the focus using both graphical and numerical facets, and search queries.
Data can be extracted according to user privileges.
4Science engineered and generalized the solution used in the PeruCRIS project and offers it to institutions for a share of the initial cost of design. 4Science engineers will support the institution to configure the add-on module to meet the institution specific needs, extracting the most from data. It provides self-service analytics capabilities, rows and columns level security, automatic periodic generation of reports. During the ingest, data are de-structured to provide easy analysis from different perspectives. Data can be accessed via JDBC / ODBC as they are in a traditional database ANSI SQL, allowing further reuse in external tools such as PowerBI, Qlik Sense, Tableau or even Excel.
# Dashboards

**Search:**

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>OrgUnits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications - single contribution view</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications - current affiliation view</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Rows per page:**

- **20**

---

**Dashboards**

**Search:**

- **Publications**

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications - single contribution view</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications - current affiliation view</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Rows per page:**

- **1**

---

**Create dashboard**
Projects dashboard

Projects count

1,558

Projects by Start Year

Projects by Dept

Projects Dept Heat Map

Institute AAAA

Institute AAAAA

Institute BBBB

Institute CCCC

Institute DDDD

Institute EEEE

Institute HHHH

Institute IIII

Institute JJJJ

Economic values (in K euro)
Publications dashboard
Self-service capabilities
Dremio provides the bridge from the "known institutional" data available in DSpace-CRIS via the Analytics add-on, and other external sources (database, SPARQL endpoints, xls files, S3, etc.). The data lake can be visualized and explored via the integrated Apache Superset webapp or external tool, accessing it as a virtual traditional SQL database.
Superset provides a self-service environment with almost the same features as OpenSearch, but independent from a specific data source. In our case we connect it to the "DREMIO" virtual database.

Data in the national DSpace-CRIS installation are dynamically joined with external demographic sources on Dremio to provide a graphical visualization over a map of research activities.
Patents dashboard
Thanks for your attention!

Andrea Bollini, 4Science
andrea.bollini@4science.it
Praxistreffen 2022