Let's do data research work: the creation of a portal with research information from Catalan Universities

Ramon Ros i Gorné
also Lluís M. Anglada i de Ferrer, Sandra Reoyo i Tudó and Ricard de la Vega i Sivera
(CSUC)

Open Repositories 2014
Helsinki, June 13th
1. Who we are
2. What we have (DSpace repositories)
3. The PRC project and firsts decisions
   - Identifiers
   - Software
   - Data mapping
   - Data flow
   - Data exchange format
4. Current status
5. Work to be done
New merged consortium in 2014

for catalan universities

with more services and projects

- The current CBUC ones
- The current CESCA ones
- Join purchases (electricity, printing, cleaning, facilities, etc.)
- Common data center
- Portal for the research output (PRC)
- Electronic administrative procedures.
- Etc.
1. Who we are
2. What we have (DSpace repositories)
3. The PRC project and firsts decisions
   - Identifiers
   - Software
   - Data mapping
   - Data flow
   - Data exchange format
4. Current status
5. Work to be done
CSUC’s DSpace repositories

TDX Tesis Doctorals en Xarxa
from 2001
www.tdx.cat

Mdx Materiales Docents en Xarxa
from 2009
www.mdx.cat

FilmoTeca de Catalunya
from 2012
repositori.filmoteca.cat

CALAIX
from 2010
calaix.gencat.cat

CIRAX
from 2013
www.cirax.cat

ANELLA CULTURAL
Pilot on 2012

MACBA Museu d’Art Contemporani de Barcelona
Coming soon on 2014

Coming soon on 2014

www.tdx.cat
www.re cercat.cat
www.recercat.cat
www.cirax.cat
www.cirax.cat
www.cirax.cat
www.cirax.cat
1. Who we are
2. What we have (DSpace repositories)
3. The PRC project and firsts decisions
   - Identifiers
   - Software
   - Data mapping
   - Data flow
   - Data exchange format
4. Current status
5. Work to be done
Situation in 2012

- CBUC promotes IR since 1999
- Some universities (UPC & UPF) already have research portals
- There are new standards and protocols that help interoperability between IR and CRIS
- Research output is becoming more important for the university managers.
Decision in 2012

What
• To create a portal to find the research outputs of the Catalan research system

Why
• To increase the visibility of the research done in Catalonia
• To foster OA
• To increase interoperability between data

How
• Taking advantage of the leverage work previously done
  – In IR, CRIS and statistical data (Uneix)
• The central idea: the works done for the portal will improve local IR and CRIS
• Following international best practices
  – Narcis / Holland; HKU Scholars Hub / Hong Kong;
PRC building. Firsts decisions.

- Identifiers -> ORCID
- Software -> DSpace + CINECA CRIS
- Data mapping
- Data flow -> from local CRIS systems
- Data exchange format -> CERIF XML
ORCID as researcher identifier

1. Selection of identifiers
   - Decision based in a CBUC report: Sistemes d’identificació unívoca d’investigadors / Àngel Borrego

2. Technical work
   - Modify all the local CRIS in order to allow to load the ORCID identifier
   - Promotion of ORCID id in other working groups: repositories, CCUC, Mendeley...

3. ORCID diffusion
   - We studied the ORCID API to create ORCID id automatically, but we decided not to use it
   - Merchandising, translations, videos, ‘good practices’ document ...
   - UB (the biggest university) have a mandate for an ORCID id in some process related with research assessment
Evoloution of ORCID registered researchers

* Data provided by ORCID. Number of researchers registered with their university email.

<table>
<thead>
<tr>
<th>University</th>
<th>oct-13</th>
<th>feb-14</th>
<th>abr-14</th>
<th>jun-14</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>UB</td>
<td>206</td>
<td>106</td>
<td>1263</td>
<td>128</td>
<td>1703</td>
</tr>
<tr>
<td>UAB</td>
<td>176</td>
<td>90</td>
<td>36</td>
<td>287</td>
<td>589</td>
</tr>
<tr>
<td>UPC</td>
<td>368</td>
<td>59</td>
<td>39</td>
<td>196</td>
<td>662</td>
</tr>
<tr>
<td>UPF</td>
<td>135</td>
<td>75</td>
<td>299</td>
<td>119</td>
<td>628</td>
</tr>
<tr>
<td>UdG</td>
<td>69</td>
<td>38</td>
<td>16</td>
<td>20</td>
<td>143</td>
</tr>
<tr>
<td>UdL</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>URV</td>
<td>102</td>
<td>48</td>
<td>42</td>
<td>25</td>
<td>217</td>
</tr>
<tr>
<td>UOC</td>
<td>43</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>79</td>
</tr>
<tr>
<td>UVic</td>
<td>18</td>
<td>150</td>
<td>2</td>
<td>24</td>
<td>194</td>
</tr>
<tr>
<td>UIC</td>
<td>11</td>
<td>2</td>
<td>5</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>URL</td>
<td>30</td>
<td>33</td>
<td>78</td>
<td>22</td>
<td>163</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1164</td>
<td>619</td>
<td>1792</td>
<td>878</td>
<td>4453</td>
</tr>
</tbody>
</table>

* Data provided by ORCID. Number of researchers registered with their university email.
Software

• Based on DSpace-CRIS of CINECA (like Hong Kong University)

• Main challenges (to adapt/develop)
  – From one institution to multi-institution
  – From submit contents to harvest from local CRIS instances
  – Massive import mechanisms are needed (XML-CERIF....)
PRC entities

Universities

Departaments & Institutes
Research groups
Researchers
Research projects
Publications (Articles + Books+ ETDs)
Lots of discussion on data mapping...
DSpace with the CRIS module.
Main entities
DSpace with the CRIS module.
Detailed entities.

**DSpace**
- Publication
- Author
- Organization. University -> communities
- Organization. Department -> collections

**CRIS module**
- Person. Researcher
- Organization. Research group
- Project
Data flow, protocols, sources and formats

Local and consortia repositories. Mainly DSpace.

Protocol: OAI-PMH/SWORD
Format: DC

12 university CRIS systems (from 4 different vendors)

Protocol: OAI-PMH
Format: CERIF-XML

PRC. Based on DSpace+Cineca CRIS.

Pronto: XLS files
Format: UNEIX defined
CERIF model
Simplification of CERIF for PRC
Simplified CERIF subset for PRC
Outline

1. Who we are
2. What we have (DSpace repositories)
3. The PRC project and firsts decisions
   - Identifiers
   - Software
   - Data mapping
   - Data flow
   - Data exchange format
4. Current status
5. Work to be done
Main achievements

• Good working team
  • People from ≠ universities and ≠ services
• Agreement: to use ORCID for researchers
• Already done
  – We succeed to export 20 complete data records from 11 universities (using 5 different CRIS)
  – All the CRIS systems already have a field for ORCID
  – A good program selected
    • Adopted by EUROCRIS as repository because CERIF compliance
Step 1: prototype
Sample data
Manual entry

Step 2: first batch load
Data sample from all universities.
CSV/XLS format

Step 3: full batch load
All data from all universities.
CSV/XLS format

Step 4: CERIF-XML ingest
First manual CERIF-XML ingest

Step 5: OAI-PMH automatic ingest.
Full synchronization with local CRIS systems.
Outline

1. Who we are
2. What we have (DSpace repositories)
3. The PRC project and firsts decisions
   - Identifiers
   - Software
   - Data mapping
   - Data flow
   - Data exchange format
4. Current status
5. Work to be done
Work to be done & challenges

- Organizational:
  - More meetings with expert group
  - ORCID ids implementation
  - MoU for personal data

- External adaptation
  - Local CRIS system to adapt XML-CERIF wrapping (export).

- Portal implementation
  - Ingest the full data of all institutions
  - Design and build the user interfaces
  - Develop the CERIF-XML import mechanisms
  - Think about depuration & deduplication data mechanisms
Thanks!

Ramon Ros i Gorné
(CSUC)
ramon.ros@csuc.cat
http://www.csuc.cat