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 Respositories 2014
Helsinki, June 13th

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

New merged consortium in 2014

Consorci de Biblioteques Universitàries de Catalunya

CENTRE DE SUPERCOMPUTACIÓ DE CATALUNYA

CESCA

CSUC Consorci de Serveis Universitaris de Catalunya

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.

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

CSUC's DSpace repositories

TDX Tesis Doctorals en Xarxa

from 2001
www.tdx.cat

RECERCAT

from 2005
www.recercat.cat

mdx MATERIALS DOCENTS EN XARXA

from 2009
www.mdx.cat

CALAIX

from 2010
calaix.gencat.cat

FilmoTeca de Catalunya

from 2012
repositori.filmoteca.cat

ANELLA CULTURAL

Pilot on 2012

CIRAX
Col·laboratori interuniversitari de recursos d'aprenentatge en xarxa

from 2013
www.cirax.cat

MACBA MUSEU D'ART CONTEMPORANI DE BARCELONA

Coming soon on 2014

Generalitat de Catalunya
Departament de Salut

Coming soon on 2014

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

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 univeristy 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;

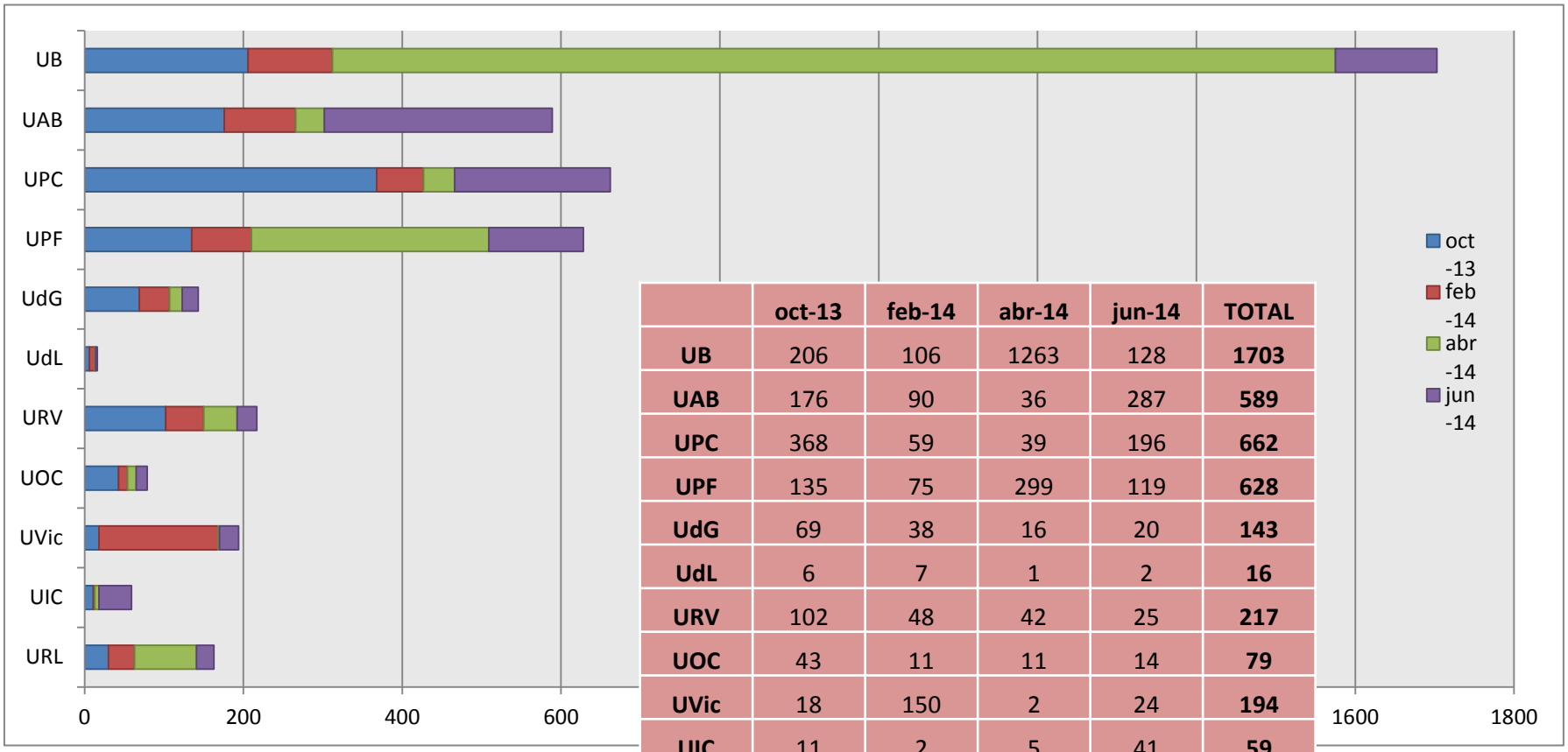
PRC building. Firsts decisions.

- 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



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

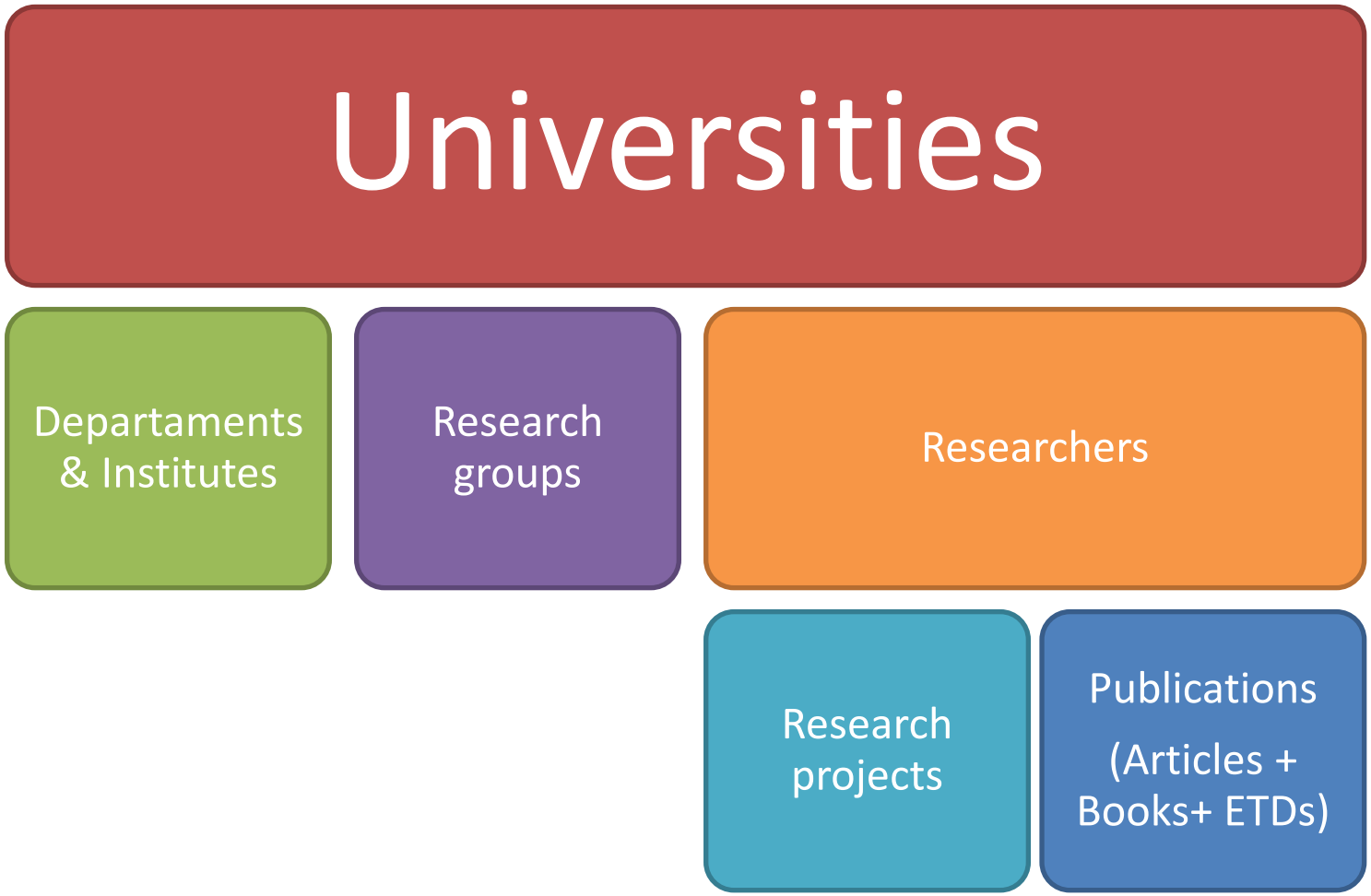
* Data provided by ORCID. Number of researchers registered with their university email.

Software

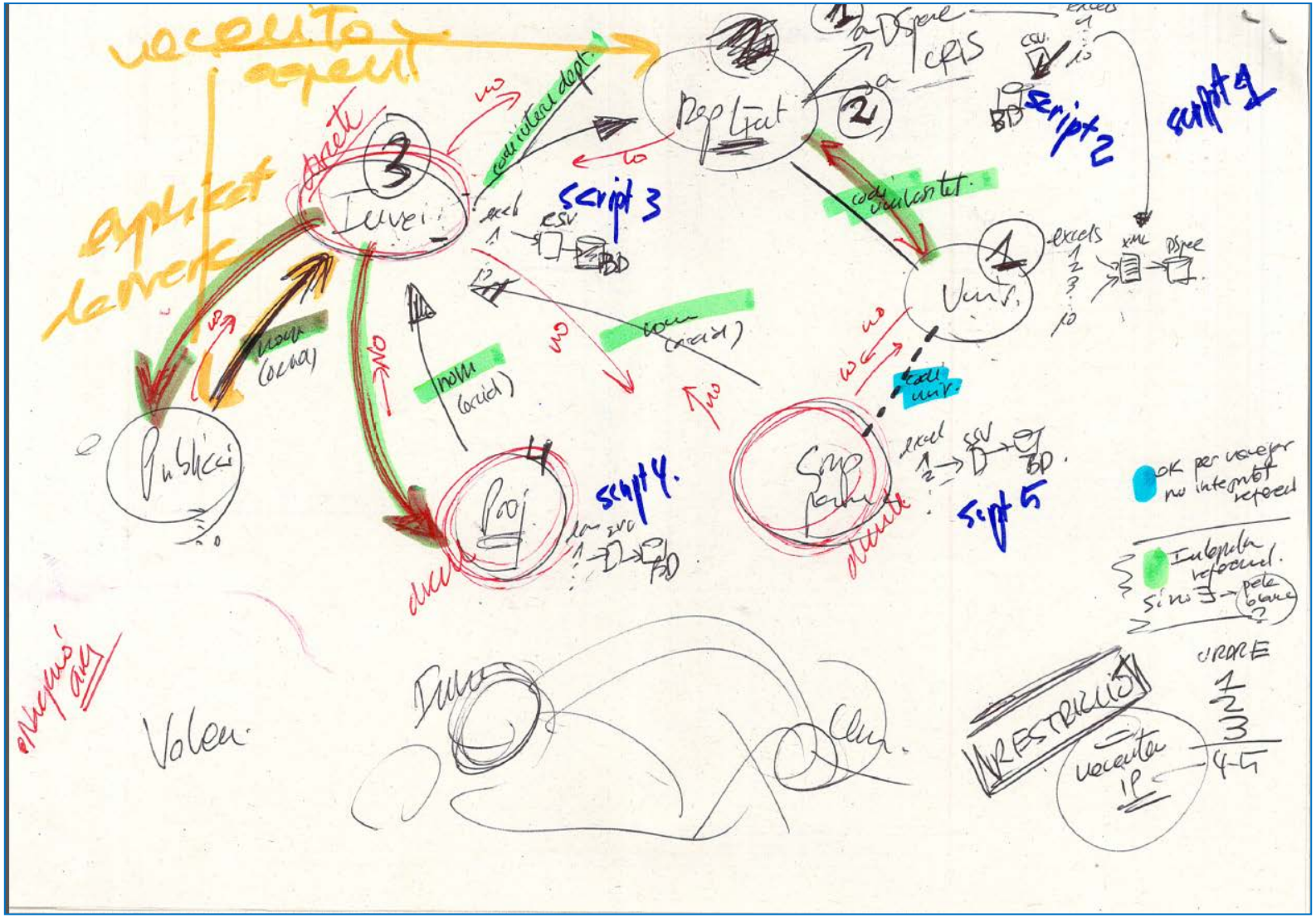
- Based on DSpace-CRIS of CINECA (like Hong Kong University)

- 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….)

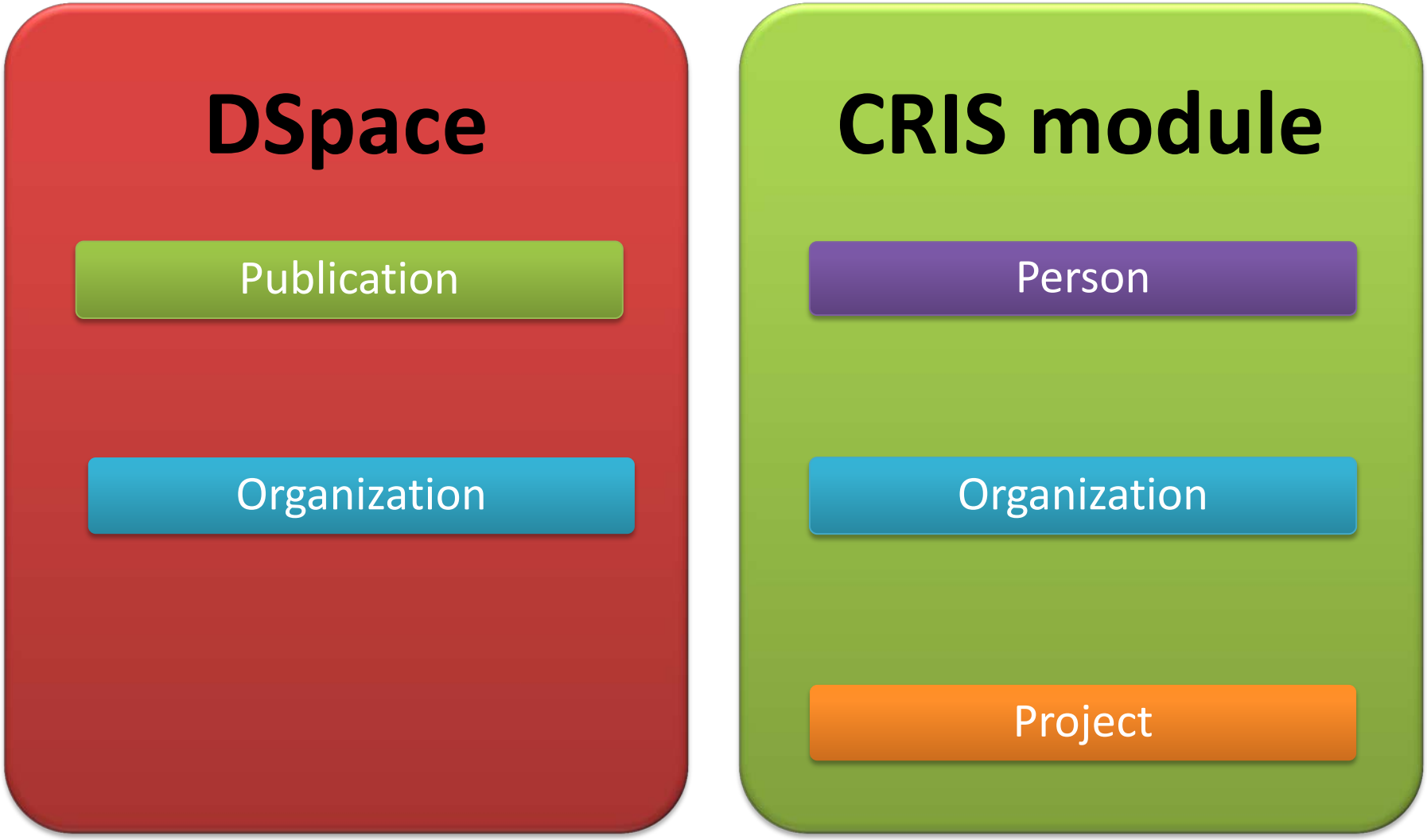
PRC entities



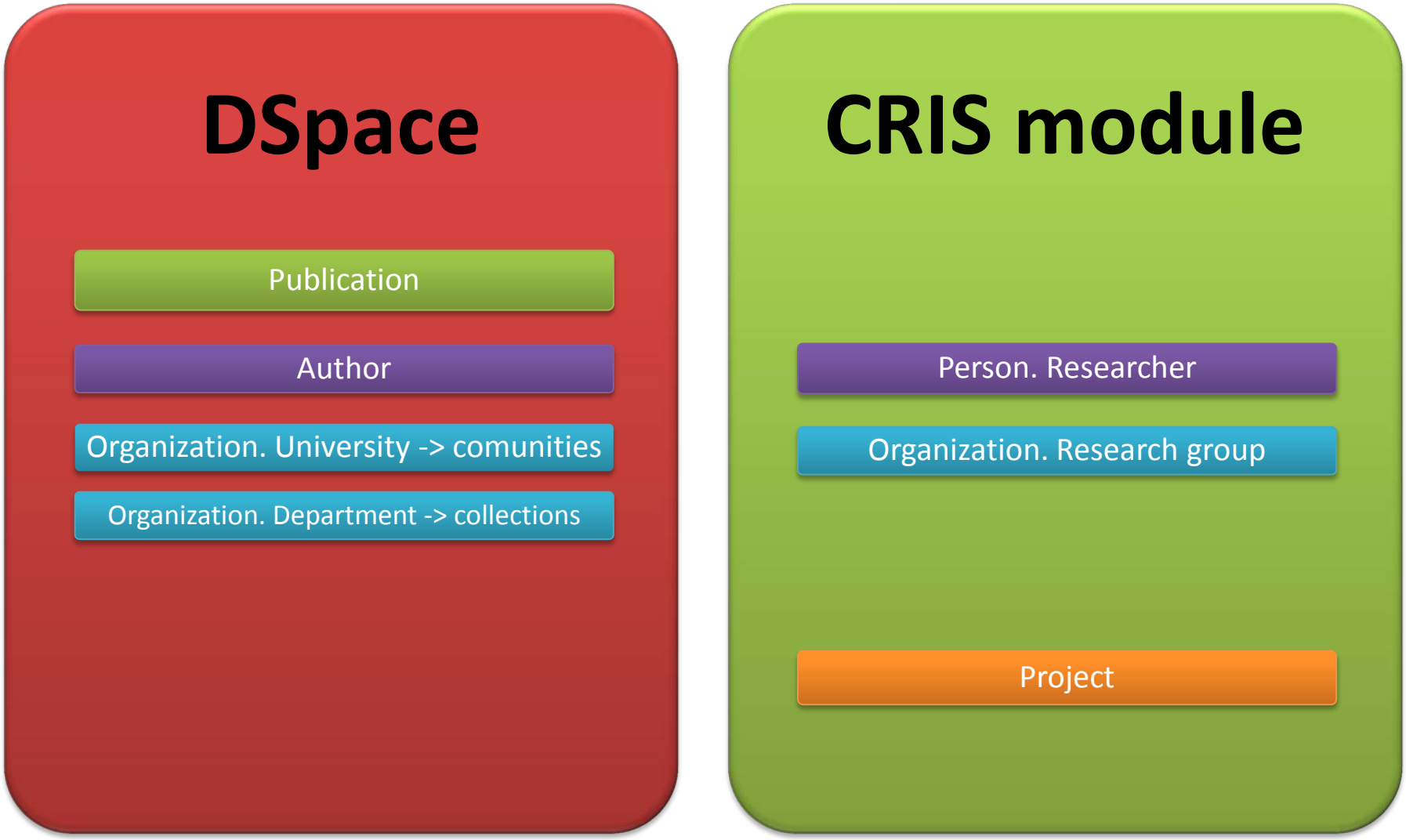
Lots of discussion on data mapping...



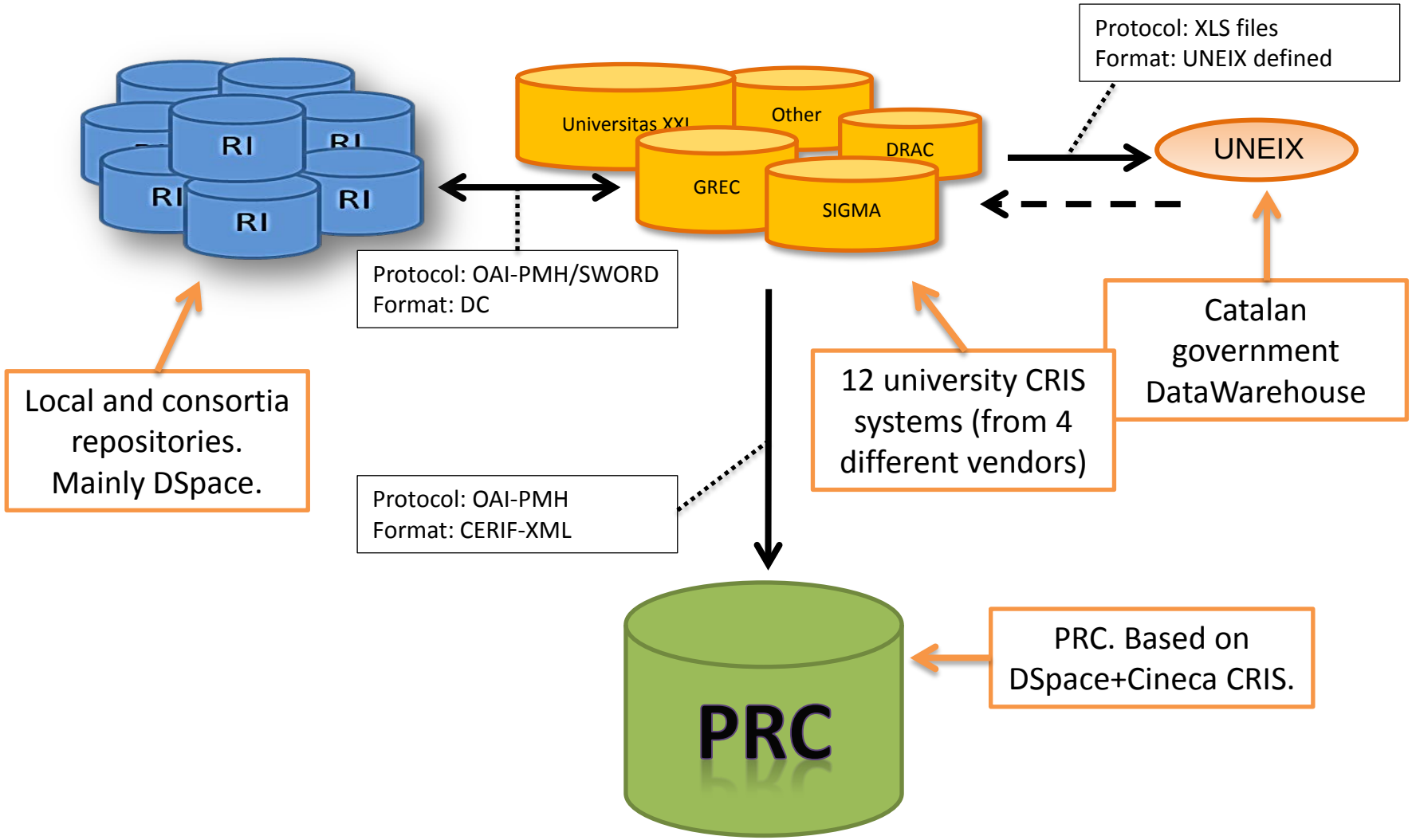
DSpace with the CRIS module. Main entities



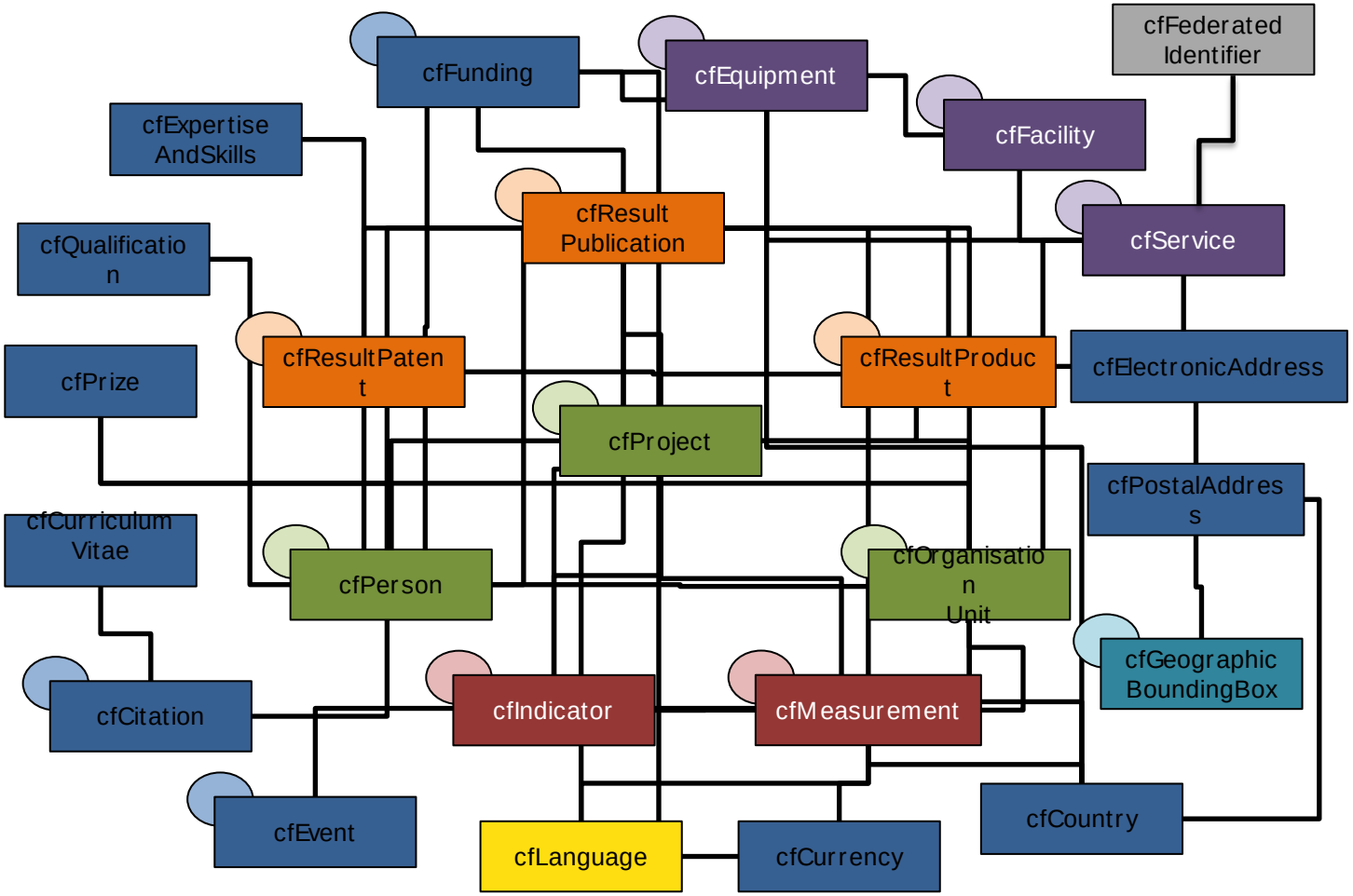
DSpace with the CRIS module. Detailed entities.



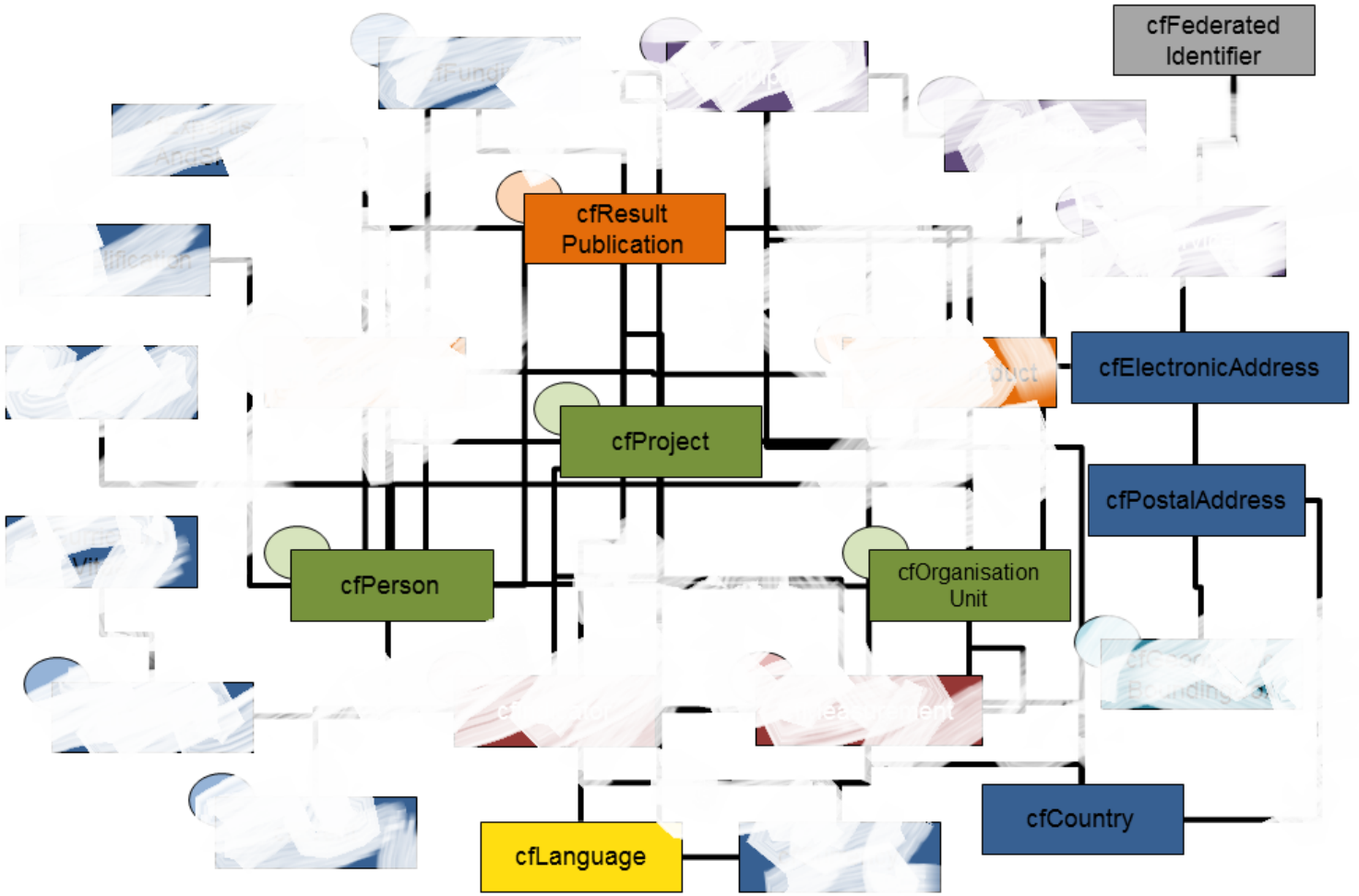
Data flow, protocols, sources and formats



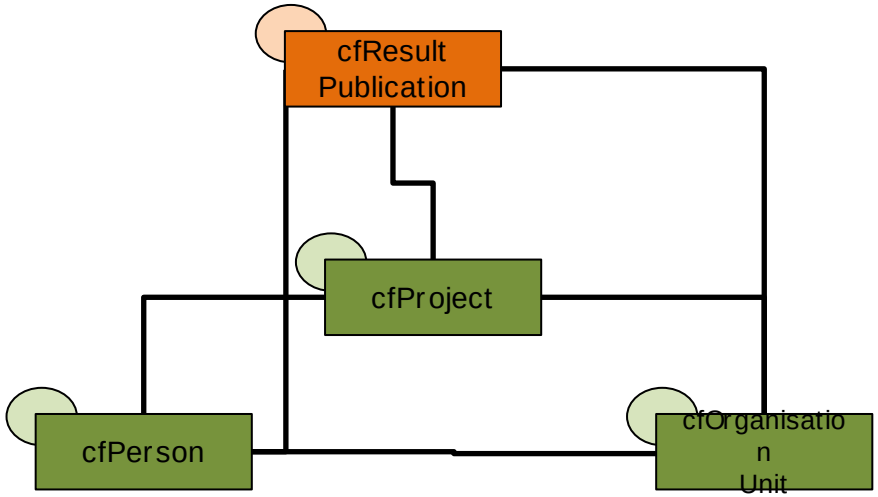
CERIF model


cfFunding
cfEquipment
cfFederated Identifier
cfExpertise AndSkills
cfFacility
cfResult Publication
cfQualificatio n
cfService
cfResultPaten t
cfResultProduc t
cfElectronicAddress
cfPrize
cfProject
cfPostalAddres s
cfCurriculum Vitae
cfPerson
cfOrganisatio n Unit
cfGeographic BoundingBox
cfCitation
cfIndicator
cfMeasurement
cfEvent
cfLanguage
cfCurrency
cfCountry

Simplification of CERIF for PRC



Simplified CERIF subset for PRC



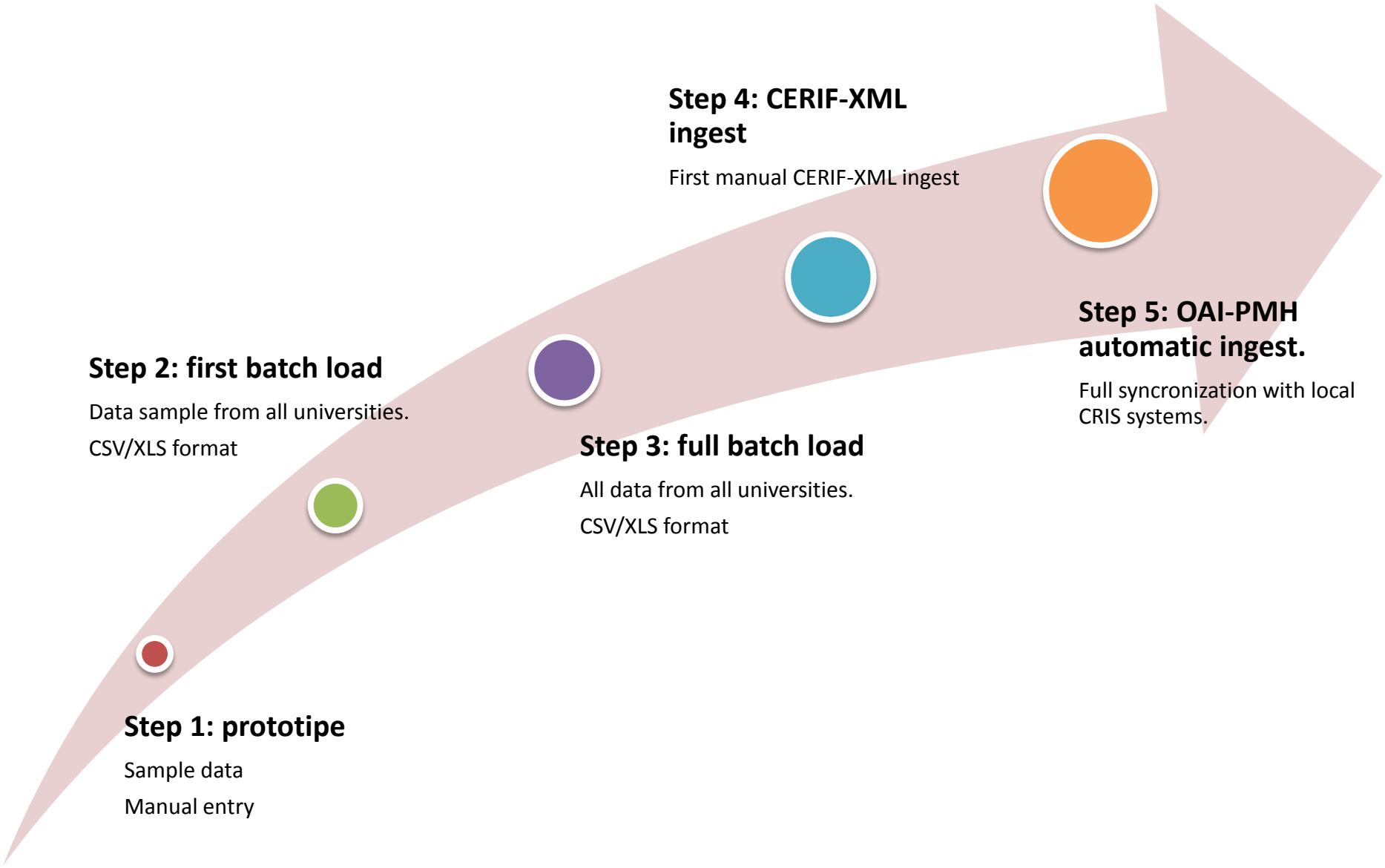
Outline

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

Implementation steps



Step 1: prototipe

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 syncronization 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