Meeting of Repository Networks COAR
May 14 & 15, 2018 - Hamburg, Germany

Current network services: enhancement and curation, text and data mining, and heuristics

- **Lautaro Matas: Gerente Técnico. LA Referencia**
- **Washington Segundo. Coordenador do Laboratório de Medologias de Tratamento e Disseminacão da Informacão. IBICT Brasil.**
- **Alberto Cabezas B. Secretario Ejecutivo LA Referencia.**

http://www.lareferencia.info/

LA Referencia

- Mission:

To give visibility to the **publicly funded scientific production** in Latin America, through the cooperation and articulation of a **federated network** of institutional repositories, based on **regional agreements and national open access strategies.**

Associates

MINCYT - Ministerio de Ciencia, Tecnología e Innovación Productiva. Argentina

Ibict - Instituto Brasileiro de Informação em Ciência e Tecnologia. Brasil

Colciencias - Ministerio de Educación, Renata. Colombia

CONICYT - Comisión Nacional de Investigación Científica y Tecnológica. Chile

CONARE - Ministerio de Ciencia, Tecnología y Telecomunicaciones. Costa Rica

SENESCYT - Secretaria de Educación Superior, Ciencia y Tecnología. Ecuador

Ministerio de Educación Viceministerio de Ciencia y Tecnología. El Salvador

CONACYT - Consejo Nacional de Ciencia y Tecnología. México

CONCYTEC - Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica

RedCLARA - Cooperación Latinoamericana de Redes Avanzadas

Service

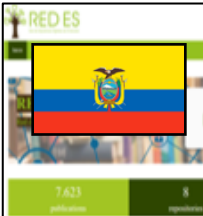| AGREEMENTS	GUIDELINES	TECHNOLOGY
Interoperability Regional Alliances Projects	OpenAIRE LR Application Profile Information quality Scientific Data Guidelines Training	Harvesters Transfer Pilots Developments Community

Transfer

Beta

3.2 version

- Reporting errors for IRs.
- Simplified Administration.
- LR version is the same as nodes.
- It supports current and future guidelines.
- End of Transfer: 2018,
- Potential data harvester.

Roadmap 2016: Common Guidelines
Challenges 2018

Challenges
- New Project OpenAIRE Advance 2018-2021
- Common statistics
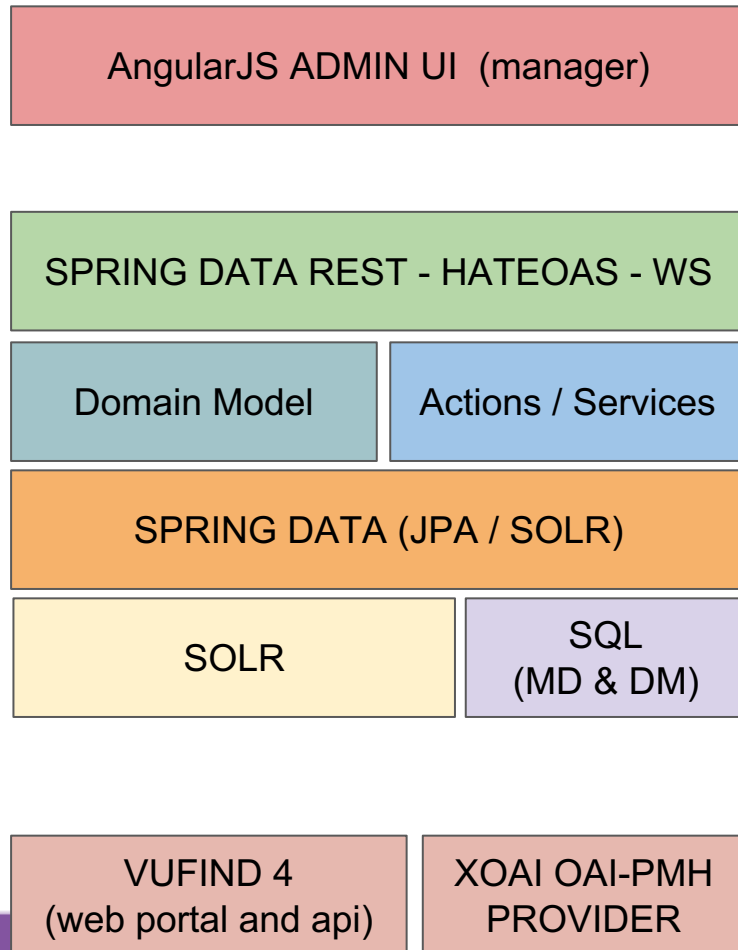- Notifications (broker)
- V. 3.0 to 4.0

LR Aggregation Model
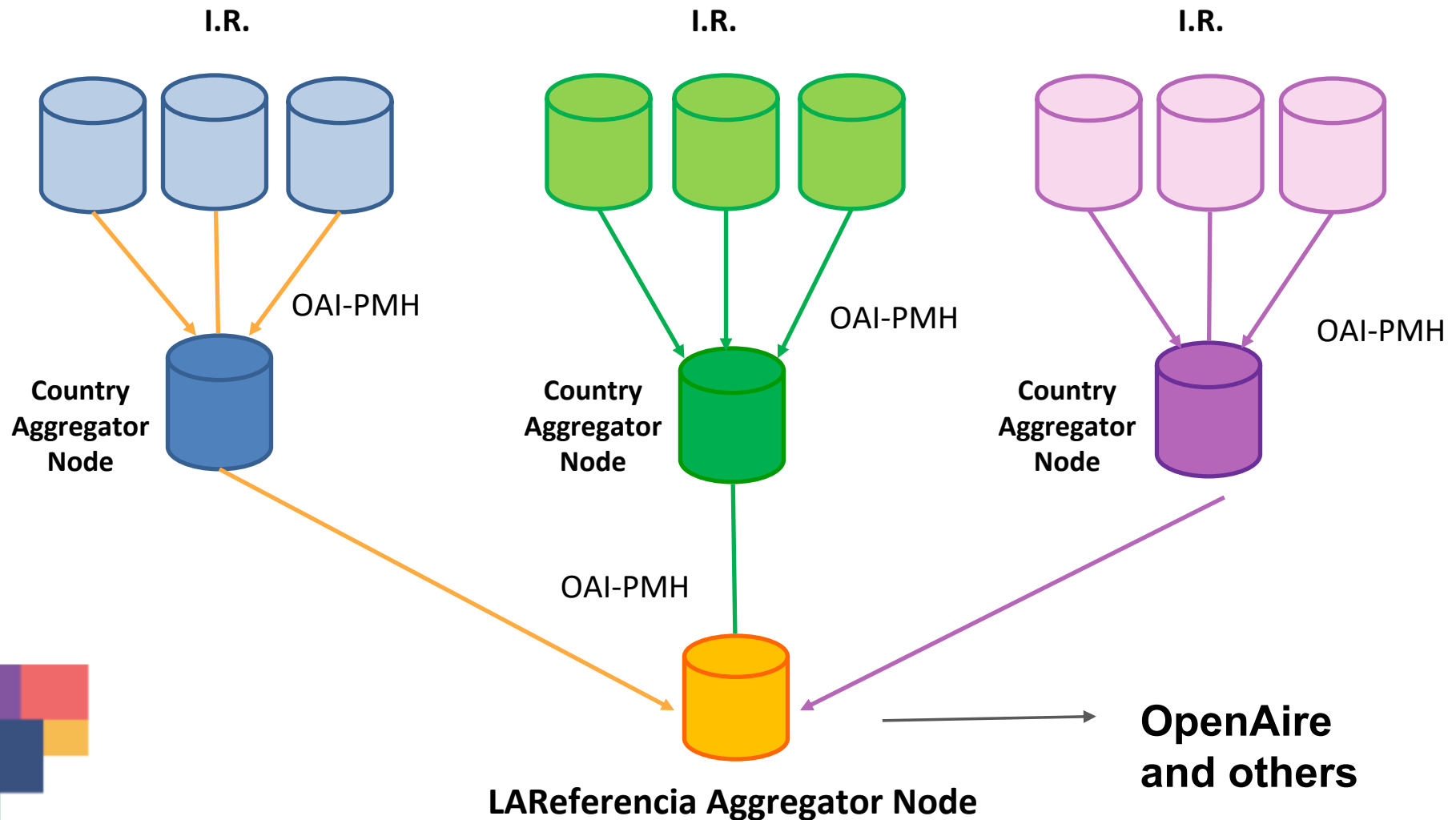
LA Referencia Harvester

- 5 years of development (2013-2018->)

- Easy to install / maintain

- Scalable: runs in low end laptops or across multiple servers in distributed mode

- GPL 3.0 License – Growing development community

- Currently supporting large repository networks (IBICT/BRASIL) harvesting/validation/transformation/indexing:
  - 200+ repositories
  - 1.5+ Million records
  - Multiple metadata schemas (standard o custom)

- OpenAire 3.0 (soon 4.0) compatible:
  - Medatada harvesting / validation / Transformation
  - Distributed statistics / OpenAire Broker integration (to be developed 2018 / 2019)

LA Referencia Software Modules

AngularJS ADMIN UI (manager)

SPRING DATA REST - HATEOAS - WS

Domain Model

Actions / Services

SPRING DATA (JPA / SOLR)

SOLR

SQL
(MD & DM)
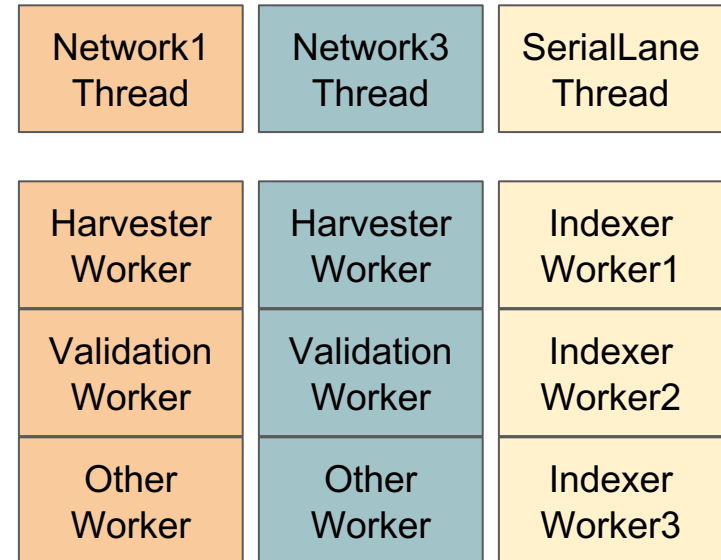
VUFIND 4
(web portal and api)

XOAI OAI-PMH
PROVIDER

- Java 8 Web Applications
- Spring 5 based
- Spring Data JPA / SOLR
- Hibernate5
- Spring Data REST - HATEOAS - HAL

- All components are integrated with dependency injection.
- Flexible architecture with component replacement/change philosophy.
- Public Web Services
- Private Web Service for admin tasks
- Vufind 4 vanilla biblio core indexing
- Vufind4 plugins for extra field display and export (tesis md fields)
- Vufind4 search and record API

Task Execution Model - Workers

- Multithreading scheduled task model based on Spring TaskScheduler
- Custom TaskManager implemented on top of Spring TS.
- A repository or national node is called **network**
- Networks can be processed using **actions**
- Actions can be launched on demand or scheduled
- An action is accomplished by one or more workers.
- Workers examples are: HarvesterWorker, ValidationWorker, IndexingWorker.
- Base classes are provided (record batch processing)
- Inside a network actions runs sequentially, but in parallel between networks
- Some actions can run in serial mode (SerialLane) to enforce sequential tasks.

Network1 Thread	Network3 Thread	SerialLane Thread
Harvester Worker	Harvester Worker	Indexer Worker1
Validation Worker	Validation Worker	Indexer Worker2
Other Worker	Other Worker	Indexer Worker3

Number of concurrent tasks is configurable. Allows to adapt the software to different hardware scenarios.

Runs in a 512mb RAM single core processor

XML Schema Transformation Service

- Records are transformed and stored in **XOAI format**.

- XOAI is a "neutral" format than can be easily transformed from/to present and future xml schemas. IE: Datacite/OpenAire4

- Transformations are made by a **configurable metadata schema transformation service**.

- Is easy to add new in/out schema format. (XSLT transformation and xml configuration) **No coding needed!**

- Customs schemas can be **harvested/validated/transformed/indexed and shared via OAI-PMH**

Validation / Transformation Rules

```
RULE (Java Object)
        ↓  ↑
JSON SERIALIZER /
DESERIALIZER
        ↓  ↑
Domain Model Rule
(JSON stored as text field)
        ↓  ↑
REST API
        ↓  ↑
ADMIN UI (View/Edit)
```
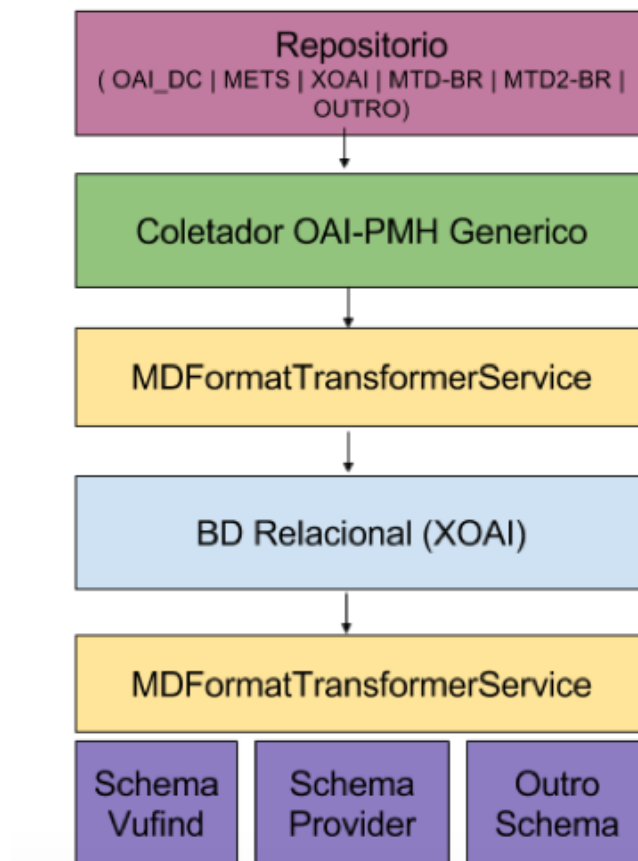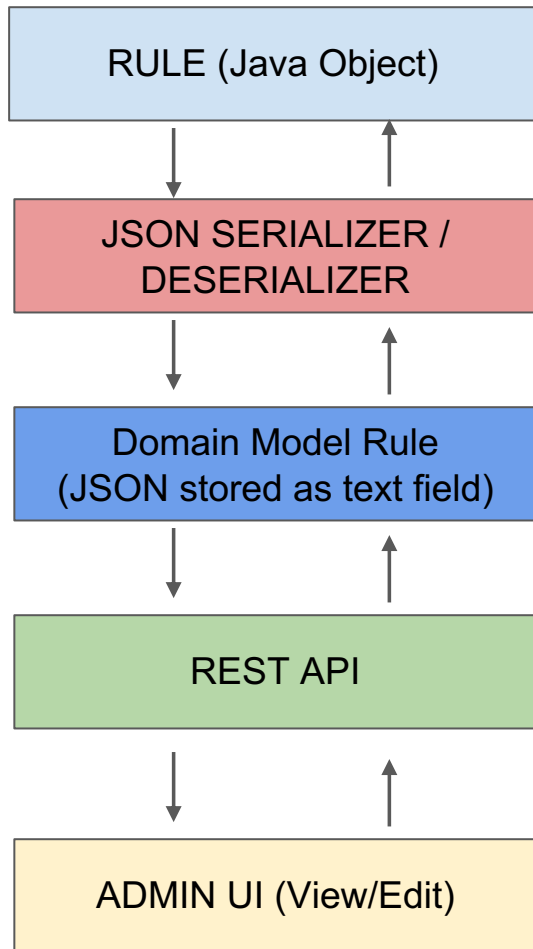
- Validation and Transformation are rule based
- Standard rules for validation/transformation are implemented: controlled vocabularies, regexp, search/replace, etc.
- A set of base classes is provided in order to encapsulate common tasks and give developers an easy way to implement more rules.
- The rules are stored as JSON Serialized java objects. It allows to view/edit rules from the web UI.
- Non-technical users can define the rules using a friendly interface.
- Rules are deserialized at validation/transformation runtime.

Validation Rule Editing

Transformation Rule Editing

Validation/Transformation result explorer

Validez de registros ‹

Inválidos (1510) +
Válidos (987) +

Transformación de registros ‹

Sin transformar (2124) +
Transformados (373) +

Reglas Válidas ‹

Filtros Argentina (2497) +
Filtros Brasil (2497) +
Filtros Perú (2497) +
Filtros México (2497) +
Filtros Costa Rica (2497) +
Driver:Status (2497) +
Identifier URL (2497) +
Title no vacío (2497) +
Validación Date (2497) +
Fecha de finalización de embargo (dc:date) (2496) +
Subject no vacío (2489) +
Type:Tesis (2409) +
Language iso-639-3 (2401) +
Relation no vacío (2236) +
Creator no vacío (2192) +
Nivel de acceso (dc:rights) (2134) +
Description no vacía (1970) +
Format (1851) +
Driver:Type (1397) +

| | 2497 R. Totales | | 987 R. Válidos | | 373 R. Transf. | | 1510 Reg. Inválidos |

Nivel de cumplimiento de reglas   ⎘ Exportar CSV

Nombre ⇕	Descripción	Mandatoria ⇕	% Cumplimiento ⇕	#R. Válidos ⇕	
Type:Tesis	Si el elemento dc.type es = a tesis entonces dc.contributor es obligatorio.	No	96.48%	2409	Detalles
Fecha de finalización de embargo (dc:date)	Fecha partir a partir de la cual el OD estará en AA. Solo si el campo Nivel de Accesibilidad (dc:rights) informa OD bajo período de embargo (info:eu-repo/semantics/embargoedAccess)	No	99.96%	2496	Detalles
Driver:Type	Esta regla evalúa que haya al menos una ocurrencia del vocaburio de tipo driver	Si	55.95%	1397	Detalles
Driver:Status	Esta regla evalúa que haya al menos una ocurrencia del vocaburio de status driver	Si	100.00%	2497	Detalles
Identifier URL	Debe existir al menos una ocurrencia de dc:identifer apuntando a URL válida	Si	100.00%	2497	Detalles
Title no vacío	Verifica que exista dc:title	Si	100.00%	2497	Detalles
Creator no vacío	Verifica que exista dc:creator no vacío.	Si	87.79%	2192	Detalles
Validación Date	Debe existir al menos una fecha válida	Si	100.00%	2497	Detalles

Validation details by record

Transformed XML record display

Detalles de validación: oai:http://redicces.org.sv:10972/1291

registro válido

Validación | XML del registro transformado

```xml
<metadata
    xmlns="http://www.lyncode.com/xoai"
    xmlns:fn="http://www.w3.org/2005/xpath-functions"
    xmlns:fo="http://www.w3.org/1999/XSL/Format"
    xmlns:xdt="http://www.w3.org/2005/xpath-datatypes"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.lyncode.com/xoai http://www.lyncode.com/xsd/xoai.xsd">
<element name="dc">
    <element name="title">
        <element name="none">
            <field name="value">La moral del sermón de la montaña de Mateo</field>
        </element>
    </element>
    <element name="creator">
        <element name="none">
            <field name="value">Alegre, Xavier, 1940-</field>
        </element>
    </element>
    <element name="subject">
        <element name="none">
            <field name="value">Moral bíblica</field>
            <field name="value">Teología</field>
            <field name="value">Biblia.-N.T.-San Mateo-Comentarios</field>
            <field name="value">Sermón de la montaña-Crítica, interpretación, etc.</field>
            <field name="value">Bienaventuranzas</field>
        </element>
    </element>
    <element name="publisher">
        <element name="none">
            <field name="value">Universidad Centroamericana "José Simeón Cañas"</field>
        </element>
    </element>
```

Administració

Language iso-639-
Type:Tesis (88) +
Subject no vacío (8
Fecha de finalización
(1) +

🏛 Instituciones

UNKNOWN (2497)

⛓ Subredes

UNKNOWN (2497)

Exportar CSV

LRSoftware 3.2 Highlights
(2017 - actual version)

- Incremental:

  - Harvesting
  - Validation
  - Transformation
  - Indexing (SOLR, VUFIND, XOAI Provider)

- Full text download and indexing

- Vufind 4.0 Indexing and search

- Author ID and CV metadata linking (when available)

- JSON Search/Result API

- Custom metadata fields (IE: BDTD Thesis specific fields)

- CV Data links

LRHarvester 3.3

(now in test phase)

— **LA Referencia - IBICT (Brasil) / RCAPP (Portugal) source code integration. Bitbucket code repository / Common issue reporting / Testing and Documentation**

— GPL 3.0 Software License / Open Source Community

- Custom repository attributes profiles:

    - Examples: geographic coordinates, URL, repository manager data, notes, tags

    - Custom fields and profiles can be added without via XML configurations

    - Customizable Admin UI with tabs and controls

    - Repository data is indexed in a separated solr core

- Harvesting historic data store: historic harvesting information now is stored in a separated solr core *

- Improvements in fulltext indexing and search

- OpenAire4 Harvesting (alpha): preliminary openaire4/datacite harvesting profile. Metadata transformation, validation and store. *

  * Developed by RCAAP Team / Ibict Team

LRHarvester 3.4 / 3.5
Preliminary Roadmap – 2018/2019

- Distributed statistics (OpenAire Advance Project)
- OpenAire 4.0 Harvesting (final version)
- Data Repostories Harvesting/Integration testing (end 2018)
- Author ID and Funding data link and curation
- OpenAIRE Broker integration (notifications)
- AngularJS to Angular UI migration

Addenda

- Case of Portugal
- Curation

Demonstração
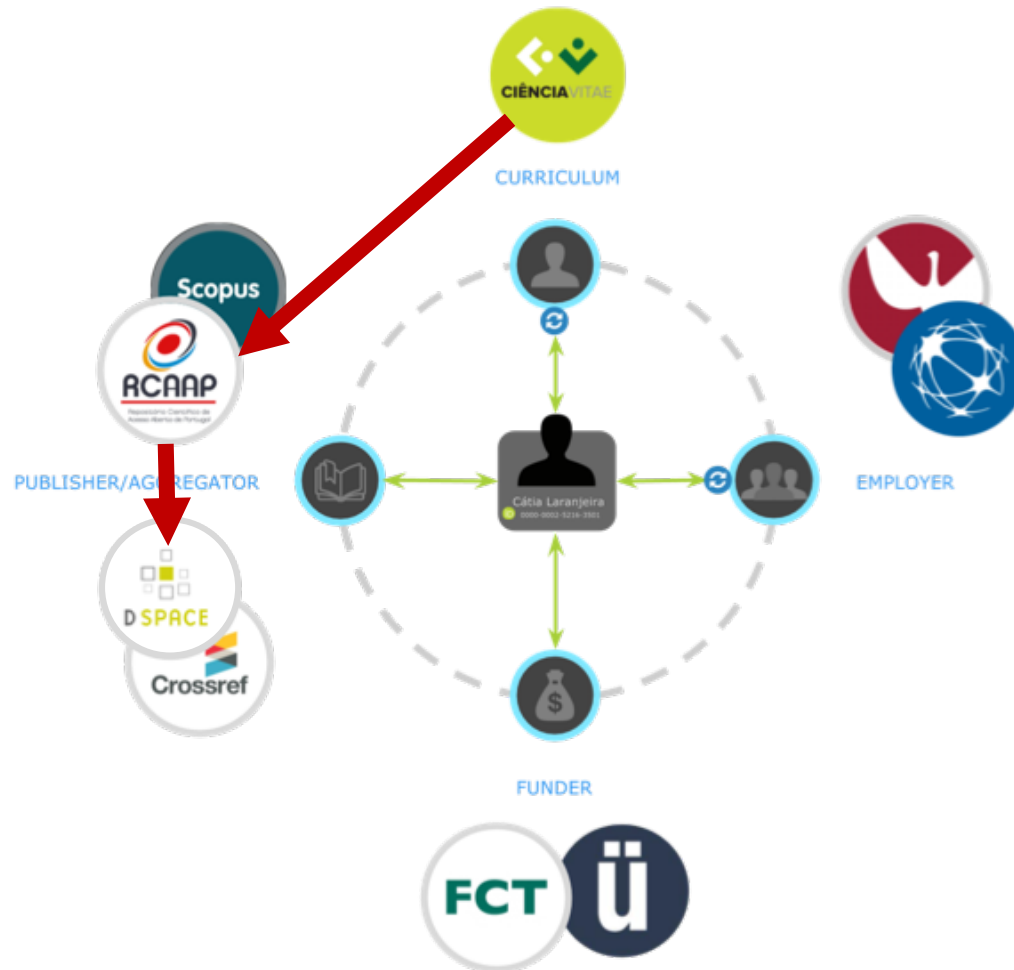
F4 – Curadoria de autor

F5 – Atualização de registo

Relatório de Validação da Agregação

- Validação dos Registos

- Identificação de Melhorias

Listagem de Melhorias Identificadas

Informação Pormenorizada

Atualização no Repositório

- **Atualmente**: Registo Manual
- **Futuro**: Registo Automático com aprovação do gestor do repositório

Detalhes da relação: 4bf507c7-ee45-4b06-88d0-950584fcc2f2

O registo: Dissertação da Cátia   tem uma relação no Ciência Vitae com o autor: Laranjeira, Cátia
Identificadores do Autor:
ORCID: 0000-0002-5216-3501

Cerrar

Após Alteração no Repositório

Caso de Uso para 2 Cenários