OPEN REPOSITORIES
2019 | HAMBURG

LA Referencia
Red de repositorios de acceso abierto a la ciencia

ibict 65 anos
Instituto Brasileiro de Informação em Ciência e Tecnologia
DESDE DE 1954 NA VANGUARDA DA INFORMAÇÃO

Improving LA Referencia metadata by linking research profiles to repositories: the case of the Brazilian Digital Library of Thesis and Dissertations (BDTD) and the Lattes CV Platform

Lautaro J. Matas, LA Referencia, lmatas@gmail.com;
Washington L. R. de Carvalho-Segundo, IBICT, washingtonsegundo@ibict.br;
Thiago M. R. Dias, CEFET-MG, thiagomagela@gmail.com;

14th International Open Repositories Conference, June 10th-13th, Hamburg, Germany

Introduction

- This presentation shows a **collaborative regional effort** on **enrich theses metadata by linking repositories with a national CV system**:
  - Describe the “ecosystem” of **Brazil theses (BDTD)** and **CV (LATTES)** systems
  - Present the results a **pilot deduplication experience** using a basic algorithm
  - Show how this experience in being **integrated into LA Referencia LRHarvester software platform**.

BDTD Theses Records + Lattes CV Records = Delivery of Enriched metadata

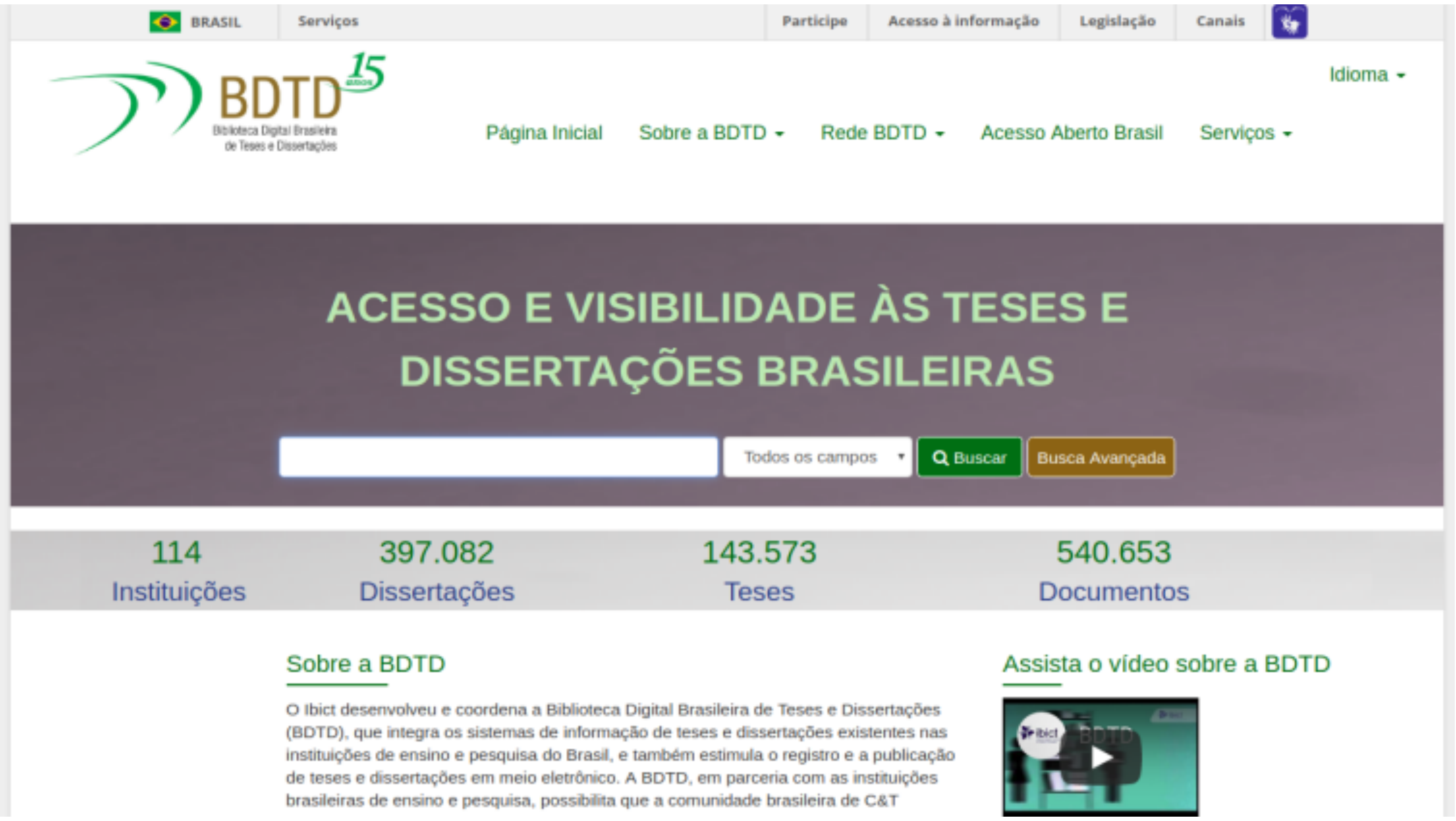
BDTD – BRAZILIAN DIGITAL LIBRARY OF THESES AND DISSERTATIONS

Created in 2002 by **IBICT – Brazilian Institute of Information in Science and Technology**

+**540K full-text** documents

114 Brazilian institutions

Is part of the **oasisbr – Brazilian Portal of Open Access Publications**



ACESSO E VISIBILIDADE ÀS TESES E DISSERTAÇÕES BRASILEIRAS

114	397.082	143.573	540.653
Instituições	Dissertações	Teses	Documentos

Sobre a BDTD

O Ibict desenvolveu e coordena a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), que integra os sistemas de informação de teses e dissertações existentes nas instituições de ensino e pesquisa do Brasil, e também estimula o registro e a publicação de teses e dissertações em meio eletrônico. A BDTD, em parceria com as instituições brasileiras de ensino e pesquisa, possibilita que a comunidade brasileira de C&T

An window to:

NDLTD (Network Digital Library of TDs)

LA Referencia

OpenAIRE

NDLTD

LA Referencia
Red de repositorios de acceso abierto a la ciencia

OpenAIRE

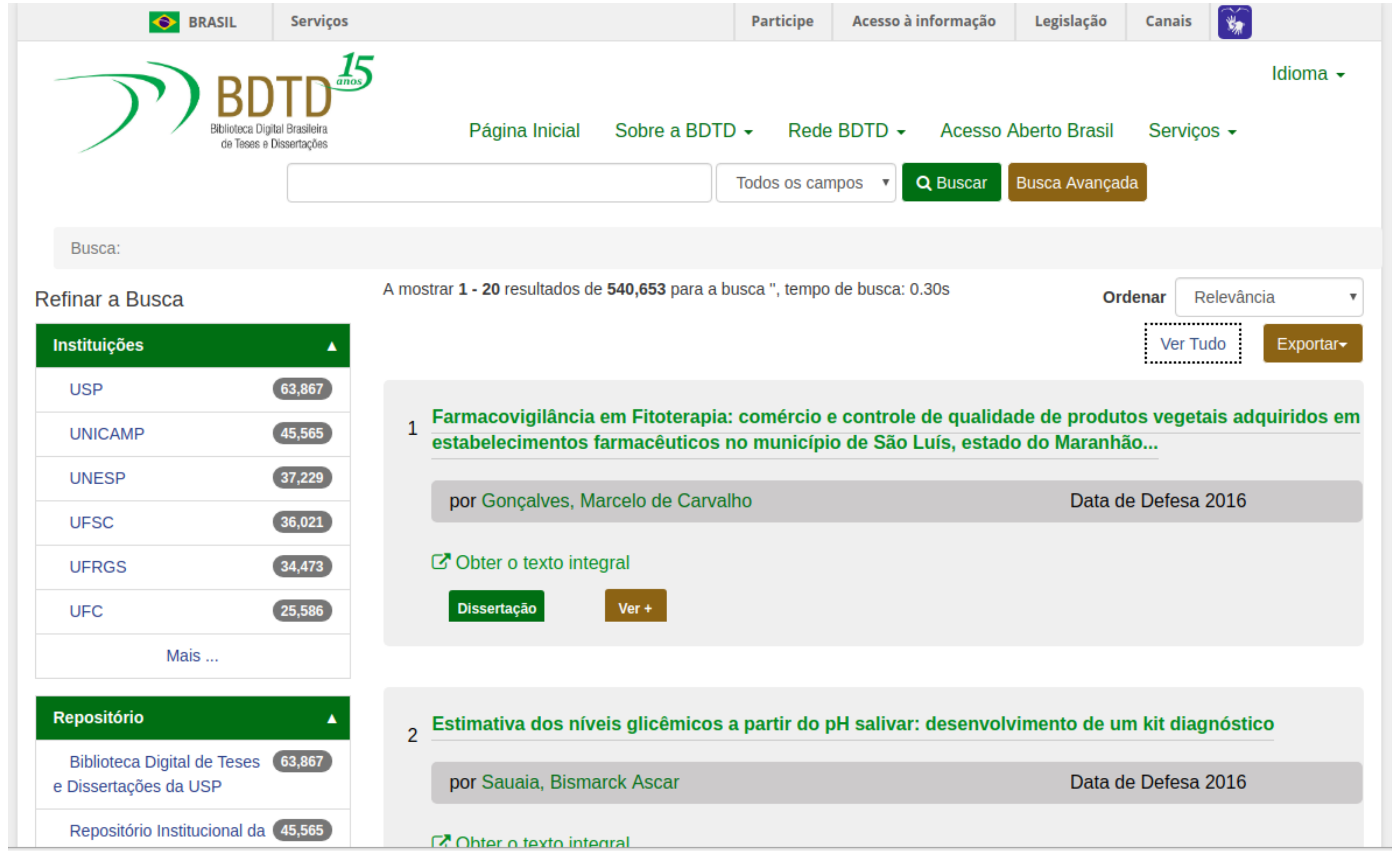
BDTD – BRAZILIAN DIGITAL LIBRARY OF THESES AND DISSERTATIONS

Public portal and metasearcher built over the LA Referencia software platform

- VuFind (Solr search engine)
- LRHarvester v 3.4
- OAI-PMH Provider

Local repositories are using different platforms:

- DSpace 4, 5 and 6
- A minor amount is using locally developed platform



Refinar a Busca

A mostrar **1 - 20** resultados de **540,653** para a busca '', tempo de busca: 0.30s

Instituições	
USP	63,867
UNICAMP	45,565
UNESP	37,229
UFSC	36,021
UFRGS	34,473
UFC	25,586
Mais ...	

Repositório	
Biblioteca Digital de Teses e Dissertações da USP	63,867
Repositório Institucional da	45,565

1 **Farmacovigilância em Fitoterapia: comércio e controle de qualidade de produtos vegetais adquiridos em estabelecimentos farmacêuticos no município de São Luís, estado do Maranhão...**

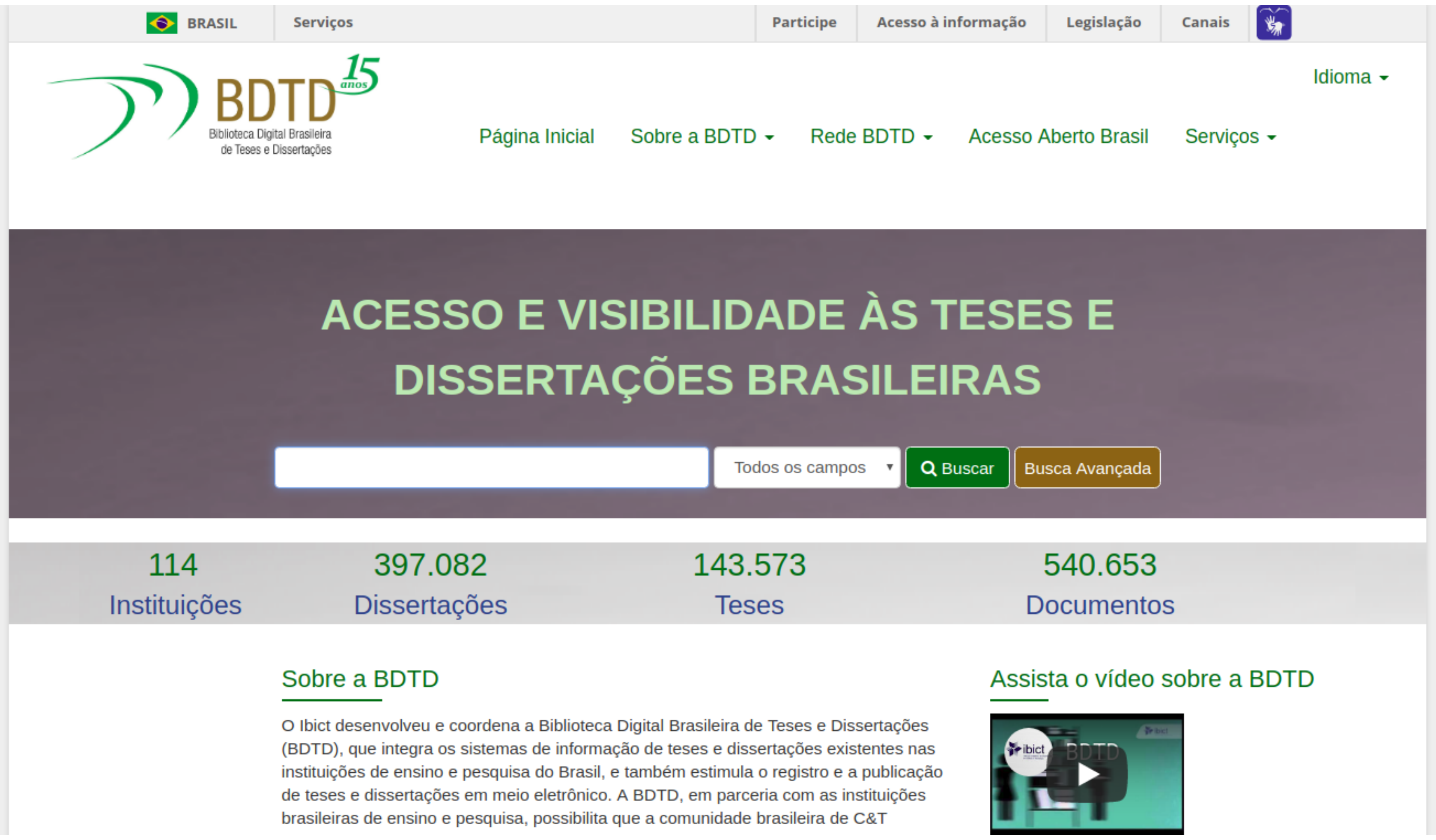
por Gonçalves, Marcelo de Carvalho — Data de Defesa 2016

Obter o texto integral

2 **Estimativa dos níveis glicêmicos a partir do pH salivar: desenvolvimento de um kit diagnóstico**

por Sauaia, Bismarck Ascar — Data de Defesa 2016

Obter o texto integral



ACESSO E VISIBILIDADE ÀS TESES E DISSERTAÇÕES BRASILEIRAS

114	397.082	143.573	540.653
Instituições	Dissertações	Teses	Documentos

Sobre a BDTD

O Ibict desenvolveu e coordena a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), que integra os sistemas de informação de teses e dissertações existentes nas instituições de ensino e pesquisa do Brasil, e também estimula o registro e a publicação de teses e dissertações em meio eletrônico. A BDTD, em parceria com as instituições brasileiras de ensino e pesquisa, possibilita que a comunidade brasileira de C&T

Assista o vídeo sobre a BDTD

Plataforma Lattes

LATTES RESEARCH PROFILE PLATFORM HTTP://LATTES.CNPQ.BR/

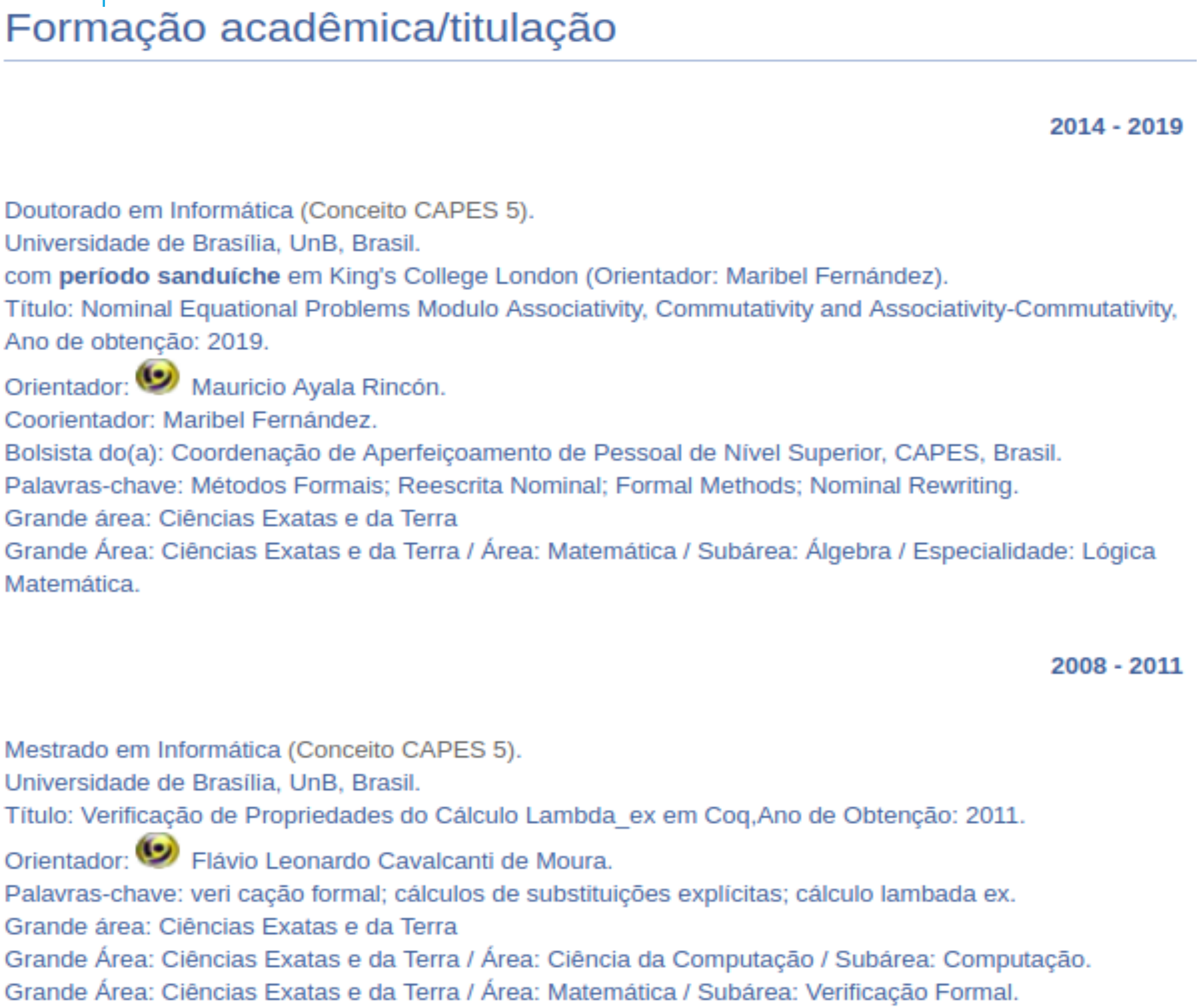
Created and supported from 1999 by the **National Council of Scientific and Technological Development (CNPq**)

+6m records => 99.9% of the researchers in Brazil have a profile in this platform

Academic history and researcher profile information

- Full name
- Research ID
- Affiliations
- Production
  - Theses and dissertations, articles, books, conferences
  - Projects
  - Founders

LATTES RESEARCH PROFILE PLATFORM



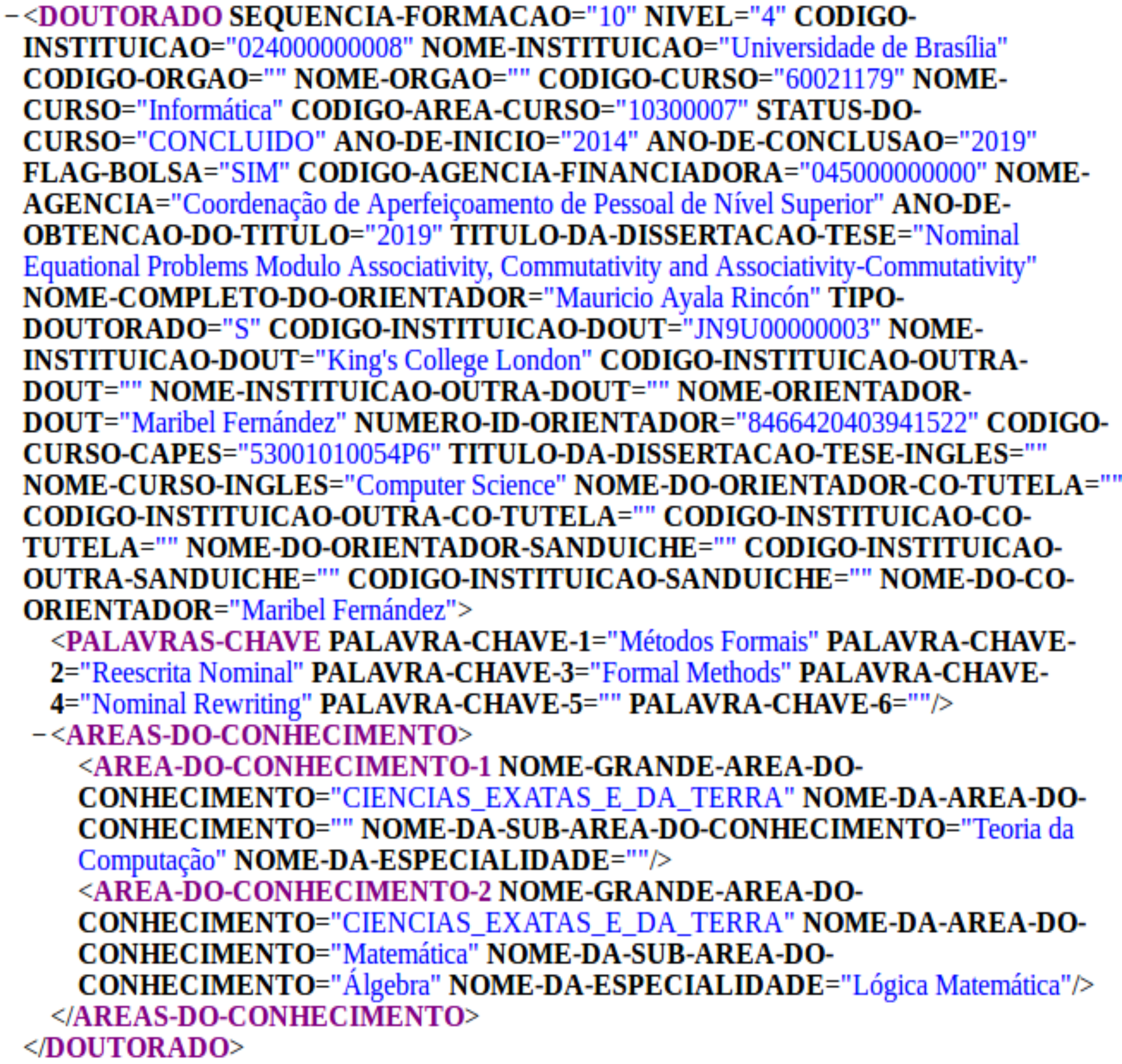
Formação acadêmica/titulação

2014 - 2019

Doutorado em Informática (Conceito CAPES 5).
Universidade de Brasília, UnB, Brasil.
com **período sanduíche** em King's College London (Orientador: Maribel Fernández).
Título: Nominal Equational Problems Modulo Associativity, Commutativity and Associativity-Commutativity,
Ano de obtenção: 2019.
Orientador: Mauricio Ayala Rincón.
Coorientador: Maribel Fernández.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.
Palavras-chave: Métodos Formais; Reescrita Nominal; Formal Methods; Nominal Rewriting.
Grande área: Ciências Exatas e da Terra
Grande Área: Ciências Exatas e da Terra / Área: Matemática / Subárea: Álgebra / Especialidade: Lógica Matemática.

2008 - 2011

Mestrado em Informática (Conceito CAPES 5).
Universidade de Brasília, UnB, Brasil.
Título: Verificação de Propriedades do Cálculo Lambda_ex em Coq,Ano de Obtenção: 2011.
Orientador: Flávio Leonardo Cavalcanti de Moura.
Palavras-chave: verificação formal; cálculos de substituições explícitas; cálculo lambada ex.
Grande área: Ciências Exatas e da Terra
Grande Área: Ciências Exatas e da Terra / Área: Ciência da Computação / Subárea: Computação.
Grande Área: Ciências Exatas e da Terra / Área: Matemática / Subárea: Verificação Formal.



```
-<DOUTORADO SEQUENCIA-FORMACAO="10" NIVEL="4" CODIGO-INSTITUICAO="024000000008" NOME-INSTITUICAO="Universidade de Brasília" CODIGO-ORGAO="" NOME-ORGAO="" CODIGO-CURSO="60021179" NOME-CURSO="Informática" CODIGO-AREA-CURSO="10300007" STATUS-DO-CURSO="CONCLUIDO" ANO-DE-INICIO="2014" ANO-DE-CONCLUSAO="2019" FLAG-BOLSA="SIM" CODIGO-AGENCIA-FINANCIADORA="045000000000" NOME-AGENCIA="Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" ANO-DE-OBTENCAO-DO-TITULO="2019" TITULO-DA-DISSERTACAO-TESE="Nominal Equational Problems Modulo Associativity, Commutativity and Associativity-Commutativity" NOME-COMPLETO-DO-ORIENTADOR="Mauricio Ayala Rincón" TIPO-DOUTORADO="S" CODIGO-INSTITUICAO-DOUT="JN9U00000003" NOME-INSTITUICAO-DOUT="King's College London" CODIGO-INSTITUICAO-OUTRA-DOUT="" NOME-INSTITUICAO-OUTRA-DOUT="" NOME-ORIENTADOR-DOUT="Maribel Fernández" NUMERO-ID-ORIENTADOR="8466420403941522" CODIGO-CURSO-CAPES="53001010054P6" TITULO-DA-DISSERTACAO-TESE-INGLES="" NOME-CURSO-INGLES="Computer Science" NOME-DO-ORIENTADOR-CO-TUTELA="" CODIGO-INSTITUICAO-OUTRA-CO-TUTELA="" CODIGO-INSTITUICAO-CO-TUTELA="" NOME-DO-ORIENTADOR-SANDUICHE="" CODIGO-INSTITUICAO-OUTRA-SANDUICHE="" CODIGO-INSTITUICAO-SANDUICHE="" NOME-DO-CO-ORIENTADOR="Maribel Fernández">
  <PALAVRAS-CHAVE PALAVRA-CHAVE-1="Métodos Formais" PALAVRA-CHAVE-2="Reescrita Nominal" PALAVRA-CHAVE-3="Formal Methods" PALAVRA-CHAVE-4="Nominal Rewriting" PALAVRA-CHAVE-5="" PALAVRA-CHAVE-6=""/>
 -<AREAS-DO-CONHECIMENTO>
    <AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA" NOME-DA-AREA-DO-CONHECIMENTO="" NOME-DA-SUB-AREA-DO-CONHECIMENTO="Teoria da Computação" NOME-DA-ESPECIALIDADE=""/>
    <AREA-DO-CONHECIMENTO-2 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA" NOME-DA-AREA-DO-CONHECIMENTO="Matemática" NOME-DA-SUB-AREA-DO-CONHECIMENTO="Álgebra" NOME-DA-ESPECIALIDADE="Lógica Matemática"/>
  </AREAS-DO-CONHECIMENTO>
</DOUTORADO>
```

BDTD COLLECTION AND CV LATTES LINKING PROOF OF CONCEPT

Trigram string similarity is a method of **identifying phrases that have a high probability of being variants of the same original phrase.** It is based on representing each phrase by a set of character trigrams. https://ii.nlm.nih.gov/MTI/Details/trigram.shtml

The initial strategy was to **calculate the distance for the title and author strings,** hypothesis was that the **joint probability of two records having high coefficients in the two fields and not being the same record is extremely low.**

The strategy proved to be **very accurate,** but with **impractical computational cost,** at least for our infrastructure.

We implemented a Elasticsearch trigram indexing and “More Like This Query” heuristic, given the **title of a record, to obtain a small list of possible candidates to be compared with the trigram-based strategy.**

As a result, now the method can be used to compare **two arbitrary collections of millions of records in a few hours.**

BDTD COLLECTION AND CV LATTES LINKING RESULTS.

The **Elasticsearch+Trigram-based strategy** was applied to the **BDTD collection (543,161 metadata** records) and compared to a **Lattes CV Platform declared thesis collection (1.364.279 records).**

Additionally, a **subset of BDTD collection (87.341 records)** that have the **ID Lattes assigned** was used as a **control set and for error calculation**.

The was executed using .60 as threshold trigram cosine distance for author and title comparisons, and the candidates were selected considering titles with 55% of coincident trigrams for "More Like This" queries.

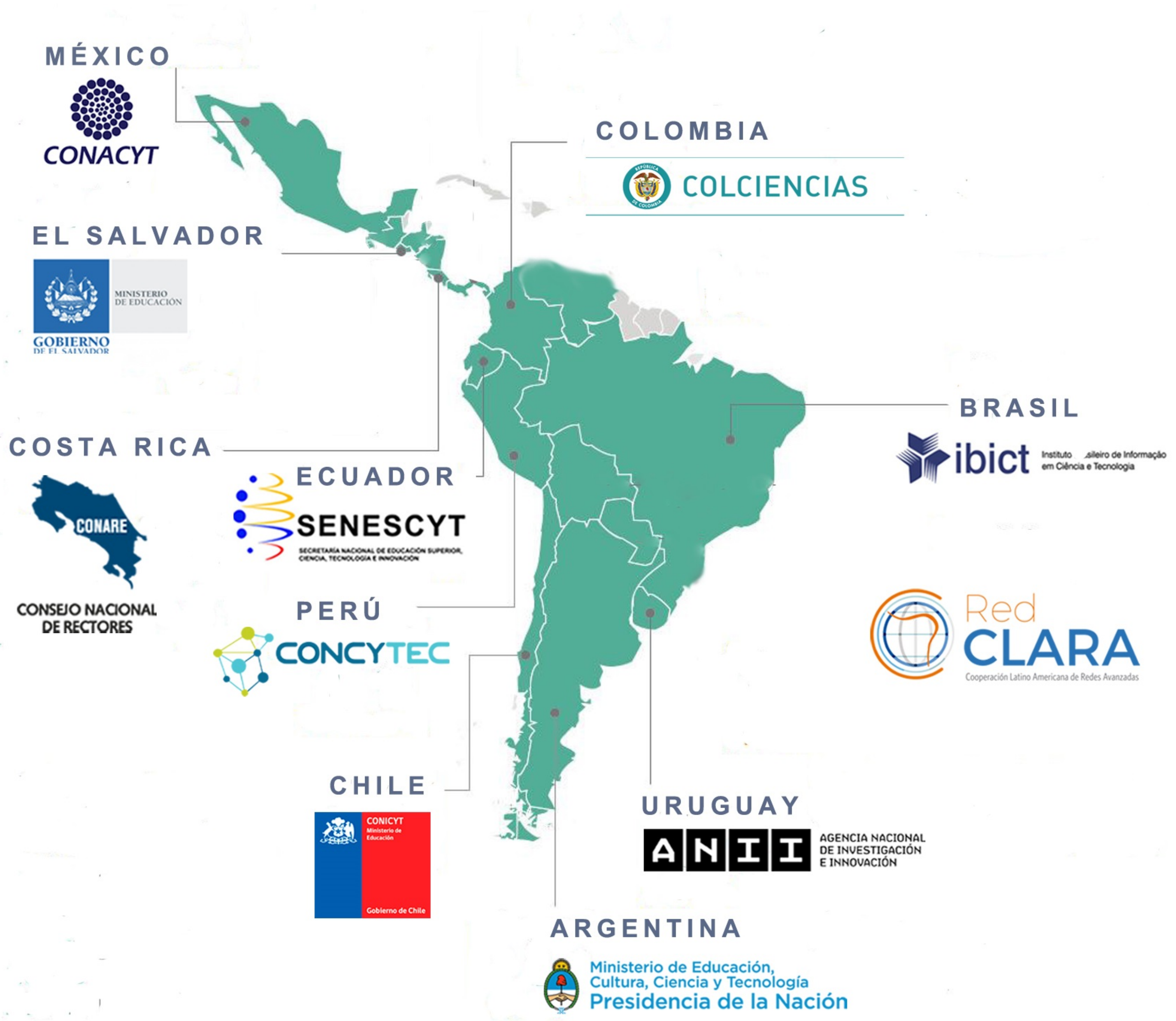
As result **401.723 BDTD records (73,96%)** were **identified in the Lattes CV Platform.**

Regarding the **control subset, 65.981 (75,54%) were matched with 6 (0,01%)** wrong matches (**Error Type I**: positive match for different IDs Lattes).

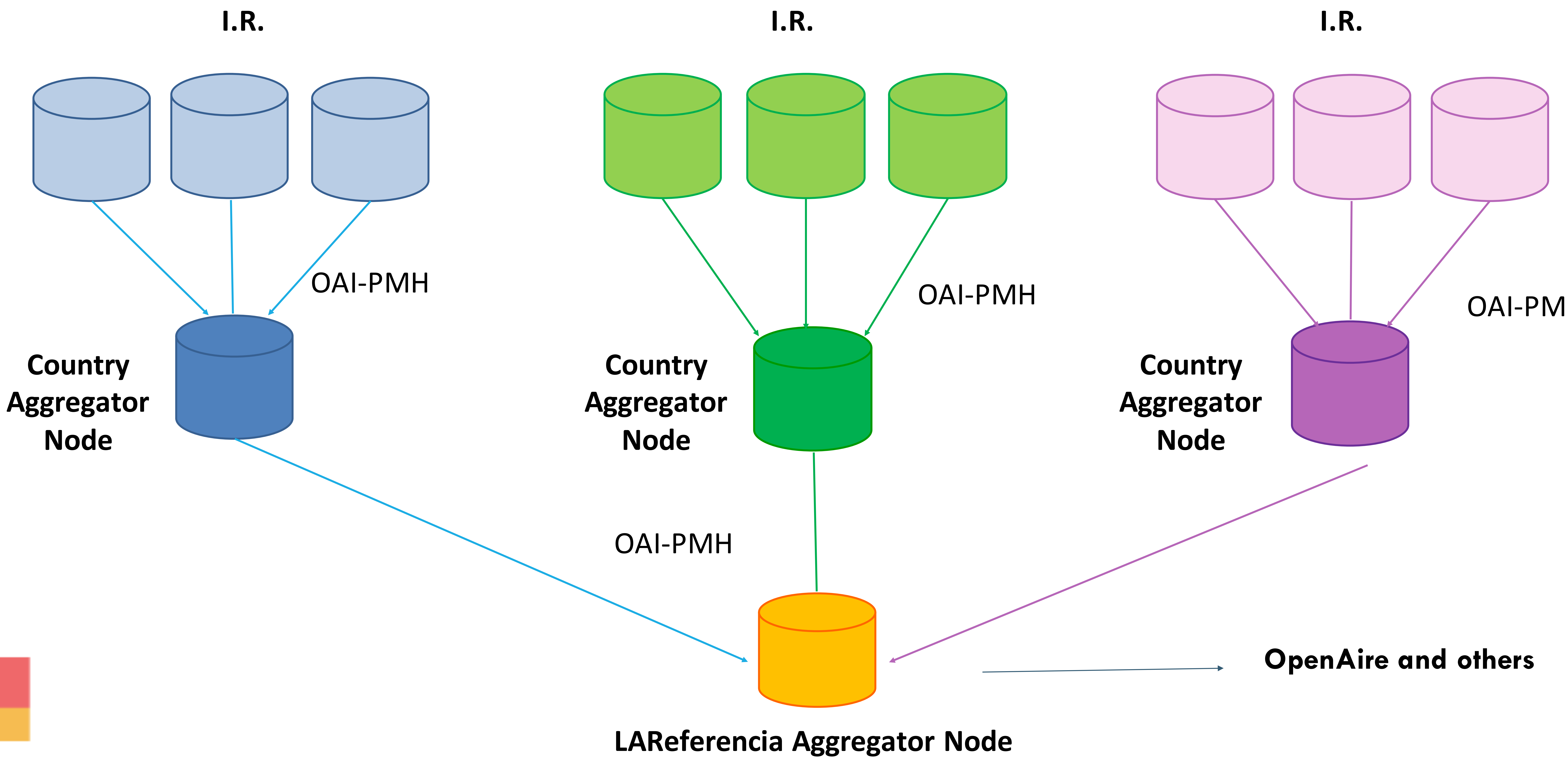
Regarding **Error Type II** (not matched with same Lattes ID) **17.085 records (19,56%) were missed.**

A careful analysis of the data showed that most cases of Error Type II correspond to titles declared in different languages (English versus Portuguese).

LA REFERENCIA NETWORK – 10 COUNTRIES AND GROWING


MÉXICO
CONACYT
COLOMBIA
COLCIENCIAS
EL SALVADOR
MINISTERIO DE EDUCACIÓN
GOBIERNO DE EL SALVADOR
BRASIL
ibict
Instituto Brasileiro de Informação em Ciência e Tecnologia
COSTA RICA
CONARE
CONSEJO NACIONAL DE RECTORES
ECUADOR
SENESCYT
SECRETARÍA NACIONAL DE EDUCACIÓN SUPERIOR, CIENCIA, TECNOLOGÍA E INNOVACIÓN
PERÚ
CONCYTEC
Red CLARA
Cooperación Latino Americana de Redes Avanzadas
CHILE
CONICYT
Ministerio de Educación
Gobierno de Chile
URUGUAY
ANII
AGENCIA NACIONAL DE INVESTIGACIÓN E INNOVACIÓN
ARGENTINA
Ministerio de Educación, Cultura, Ciencia y Tecnología
Presidencia de la Nación

LA REFERENCIA AGGREGATION MODEL



LA REFERENCIA LRHARVESTER SOFTWARE

6 years of development (2013-2019->), 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:
- repositories
- 1.5+ Million records
- Multiple metadata schemas (standard o custom
- OpenAire 3.0 (4.0 work in progress) compatible
- Metatata harvesting / validation / transformation

OpenAIRE Distributed usage statistics and broker as a service integration (work in progress 2019)

Data repositories aggregator pilot (end of 2019)

Validez de registros
nválidos (1510) +
álidos (987) +

Transformación de registros
in transformar (2124) +
ransformados (373) +

Reglas Válidas
iltros Argentina (2497) +
iltros Brasil (2497) +
iltros Perú (2497) +
iltros México (2497) +
iltros Costa Rica (24
river:Status (24
entifier URL (2497) +
itle no vacío (2497) +
alidación Date (2497) +
echa de finalización de embargo (dc:date) 2496) +
ubject no vacío (2489) +
ype:Tesis (2409) +
anguage iso-639-3 (2401) +
elation no vacío (2236) +
reator no vacío (2192) +
ivel de acceso (dc:rights) (2134) +
escription no vacía (1970) +
ormat (1851) +
river:Type (1397) +

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

Nivel de cumplimiento de reglas

Exportar CSV

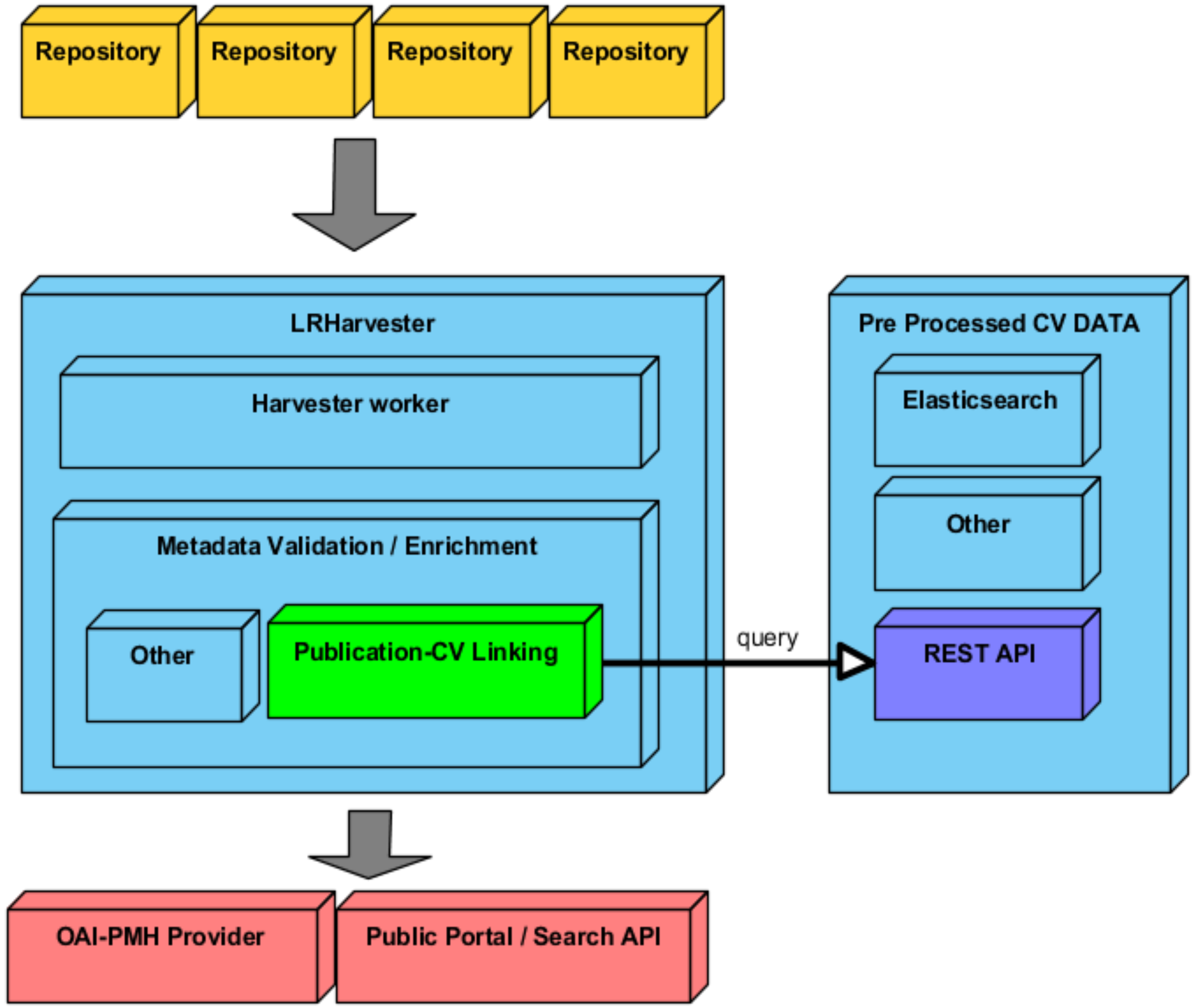
Nombre	Descripción	Mandatoria	% Cumplimiento	#R. Válidos	
Type:Tesis	Si el elemento dc.type es = a tesis entonces dc.contributor es obl	No	96.48%	2409	Detalles
Fecha de finalización de embargo	Fecha partir a partir de AA. Solo si el campo Nivel de Accesibilidad (dc:rights) informa OD bajo periodo de embargo info:eu-repo/semantics/embargoedAccess)				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:identifier 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

LA REFERENCIA LRHARVESTER ARCHITECTURE 4.0

Build, enrich and store an entity-relation (cerif like) model based on the different metadata sources (literature repositories, aggregators, API´s, CV´s, founders metadata)

Use the entity-relation model to curate and enrich metadata. Interoperate with original sources (and actors) to provide feedback.

Use the entity-relation model to feed a service API and interoperate with other system and services in S&T ecosystem (CV´s, CRIS)



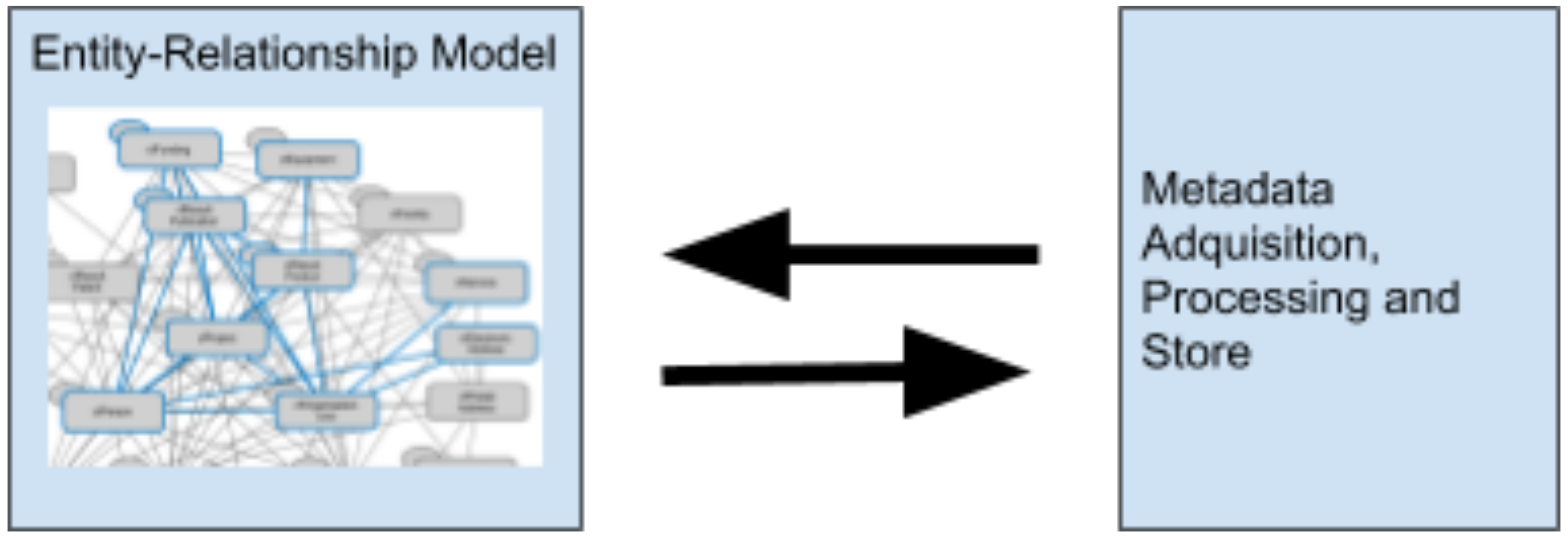
Repository
Repository
Repository
Repository

LRHarvester
Harvester worker
Metadata Validation / Enrichment
Other
Publication-CV Linking

query

Pre Processed CV DATA
Elasticsearch
Other
REST API

OAI-PMH Provider
Public Portal / Search API

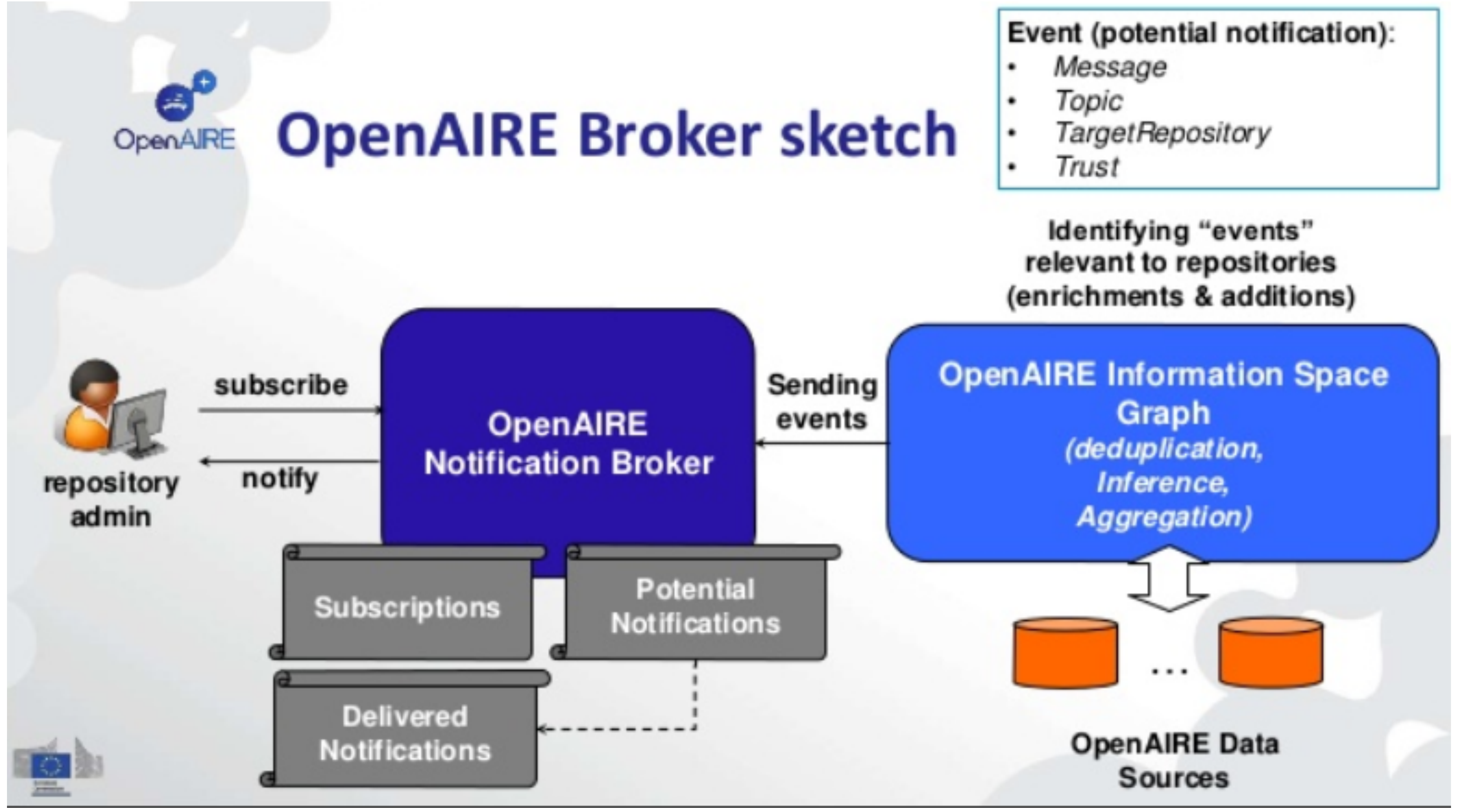


Entity-Relationship Model

Metadata Adquisition, Processing and Store

LA REFERENCIA – OPENAIRE INTEGRATION

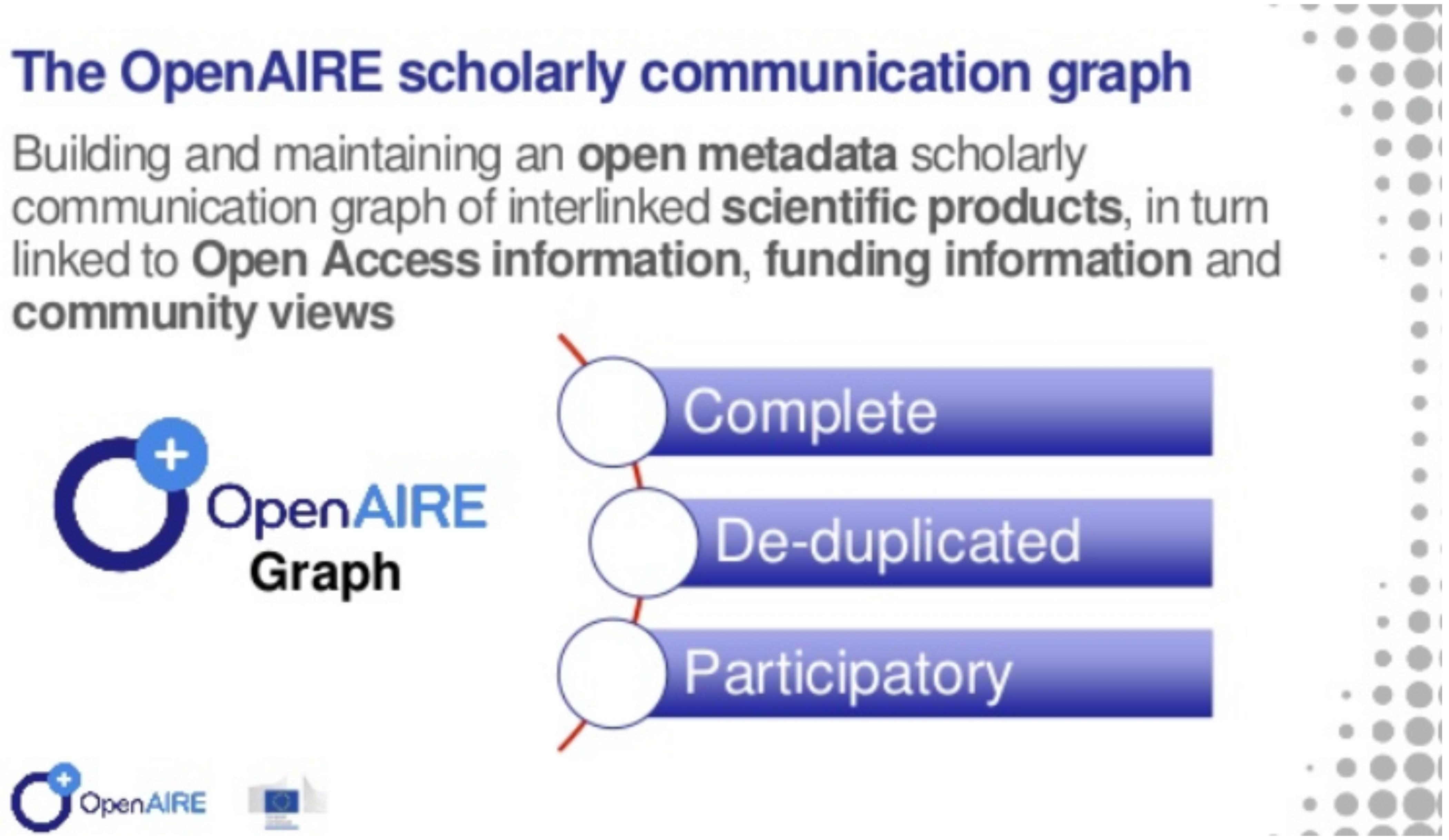
Broker as a service integration – Consume events / Integration into national network dashboard (to be developed during 2019/2020)



OpenAIRE Broker sketch

Event (potential notification):
- *Message*
- *Topic*
- *TargetRepository*
- *Trust*

OpenAIRE
Identifying "events" relevant to repositories (enrichments & additions)
subscribe
notify
repository admin
OpenAIRE Notification Broker
Sending events
OpenAIRE Information Space Graph
(deduplication, Inference, Aggregation)
Subscriptions
Potential Notifications
Delivered Notifications
...
OpenAIRE Data Sources



The OpenAIRE scholarly communication graph

Building and maintaining an **open metadata** scholarly communication graph of interlinked **scientific products**, in turn linked to **Open Access information**, **funding information** and **community views**

OpenAIRE Graph
Complete
De-duplicated
Participatory
OpenAIRE

* Slides by Paolo Manghi at DI4R2018

LRHARVESTER 3.5/4.0 2019/2020 ROADMAP

Configurable Entity-Relation Meta Model (DB stored) – BETA

Entity-Relation Model - 2019

- instantiation (OpenAire4/CRIS) - WIP
- feeding the model from metadata – Harvester Workers
- CRUD Rest API / HATEOAS (content)
- Public REST API
- Indexing (SOLR / ELASTICSEARCH) / REST API (search)
- Model enrichment and entity deduplication

Repository administrator dashboard 2019 (validation results, broker notifications, statistics)

Metadata acquisition (other sources than OAI-PMH) - 2020

- CRIS Sources
- Implement Resource Sync
- API´s (ORCID, NATIONAL DATA)
- Large DUMPs loading (OpenAIRE Graph, ORCID)

LA REFERENCIA LRHARVESTER 4.0 POTENTIAL SERVICES

LA Referencia

Red de repositorios de acceso abierto a la ciencia

Better integration with CV and CRIS Systems (CVLattes, VIVO, DSPACE CRIS, ORCID)

OpenAIRE Broker event delivery to national networks and repositories

OpenAIRE Graph: dump loading and integration into entity model

Metadata enrichment for building indicators and decision making tools.

Metadata curation and enrichment at repository level

OpenAIRE 3.0 to 4.0 migration services for repositories

Regional and national usage statistics aggregator portals

Thank you !!

Lautaro J. Matas, LA Referencia, lmatas@gmail.com;
Washington L. R. de Carvalho-Segundo, IBICT, washingtonsegundo@ibict.br