THE LA REFERENCIA SOFTWARE AND THE BRAZILIAN PORTAL OF SCIENTIFIC OPEN ACCESS PUBLICATIONS (OASISBR)

Washington L. R. de Carvalho-Segundo (IBICT)

Lautaro Matas (LA Referencia)

Alberto Cabezas (LA Referencia)

Bianca Amaro (IBICT)

Gabriel Gomes (IBICT)

June 28, 2017







Outline

Introduction

Problem definition

Adopted solution

Brazilian Institute of Information in Science and Technology (IBICT)



- National governmental organisation
- Created in 1954 as research institute of library science
- In the representative of Brazil node in the LA Referencia network

Brazilian Portal of Scientific OA Publications (oasisbr)



Almost 1,800,000 documents

711111030 1,000,000	accaments
Articles	46 %
Master thesis	28 %
Conference object	9 %
Doctoral thesis	9 %
Bachelor thesis	5 %
Book	1 %
Book chapter	1 %
Report	1 %





42 % Institutional Repositories



23 % Electronic Journals



13 % Electronic Thesis & Dissertations

Number of sources

- 75 Institutional Repositories
- 770 Electronic Journals
- 100 Electronic Thesis and Dissertations Databases
- 1 National Repository



Problem definition

- Multi-schema metadata: oai_dc, mtd-br, mtd2-br, mets, etd-ms, XOAI
- Non-compliance with international guidelines:
 - DRIVER
 - OpenAIRE
 - LA Referencia
- Connectivity:
 - lacktriangle Repositories o oasisbr

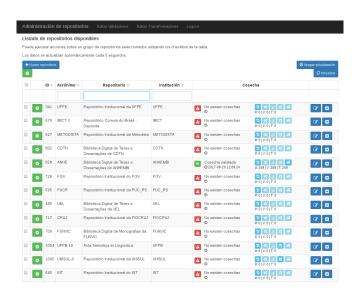
 - oasisbr → LA Referencia
 - Thesis & Dissertations → NDLTD



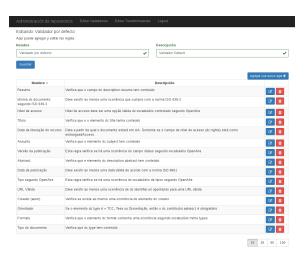
The LA Referencia Software

- Multi-schema metadata → XSLT transformation for each input schema to a common most complete description (XOAI Schema)
- ullet Non-compliance o metadata transformation / validation
- Connectivity: harvester and provider (XOAI interface) modules

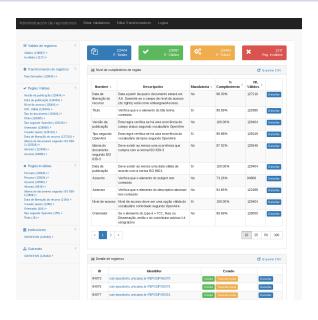
Administrator Interface



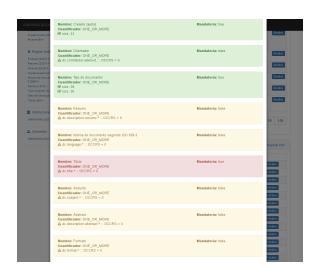
Transformation / Validation Rules



Interface of Metadata Validation



Interface of Metadata Validation



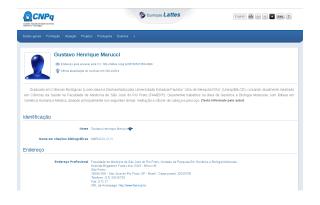
VuFind Meta-searcher integration



Research Profiles Linked

Access Level:	openAccess
Publication Date:	2011
Main Author:	Marucci, Gustavo Henrique 🕏
Advisor:	Pavarino-bertelli, érika Cristina 🦻
Co-advisor:	Goloni-bertollo, Eny Maria 6
Referee:	Machado, Ricardo Luiz Dantas 🦻, Cintra, Mariangela Torreglosa Ruiz 💁, Santos, Izaura dos 🚱
Format:	masterThesis
Language:	por
Published:	Faculdade de Medicina de São José do Rio Preto
Program:	Programa de Pós-Graduação em Ciências da Saúde
Department:	Medicina Interna; Medicina e Ciências Correlatas
Portuguese Subjects:	Sindrome de Down Trissomia do 21 Polimorfismo genético Homocistelina Ciências da Saúde
English Subjects:	Down Syndrome Triscory of 21 genetic polymorphism homocysten folate Health Sciences Polymorphism, Genetic Homocystate
Spanish Subjects:	Ciencias de la Salud
CNPQ Subjects:	CNPQ > CIENCIAS DA SAUDE
Online Access:	http://bdtd.famerp.br/handle/tede/108

Research Profiles Linked





ASSOCIATES



MINCYT - Ministerio de Ciencia, Tecnología e Innovación Productiva



SENESCYT - Secretaria de Educación Superior, Ciencia y Tecnología





Ibict - Instituto Brasileiro de Informação em Ciência e Tecnologia



Ministerio de Educación Viceministerio de Ciencia y Tecnología



Colciencias - Ministerio de Educación, Renata



CONACYT - Consejo Nacional de Ciencia y Tecnología



CONICYT - Comisión Nacional de Investigación Científica y Tecnológica



CONCYTEC - Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica Observer Country



CONARE - Ministerio de Ciencia, Tecnología y Telecomunicaciones.

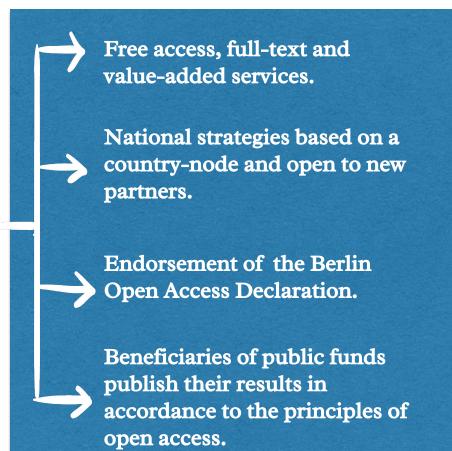


2010-2013 MILESTONES

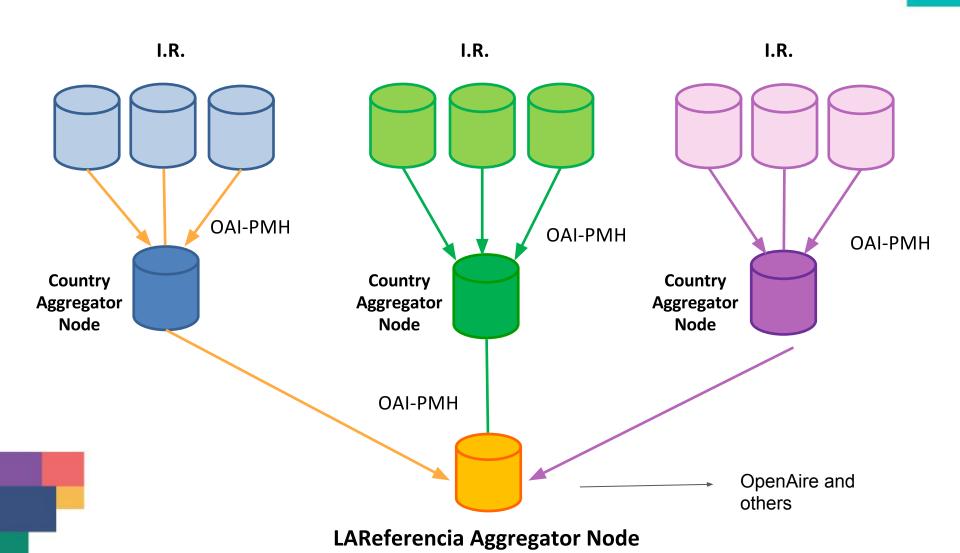
Project Funded by IDB / 2010 - 2013. Regional Public Good

Consensus strategy; political and interoperability agreements; Pilot project.

Agreement of cooperation between high authorities of Science, Technology and Innovation of Latin America, Buenos Aires, November 29, 2012.



LR Aggregation Model





2014-2016 SERVICES

Sustainability / membership fees since the end of 2013

Consolidation of Board of Directors and Technical Groups

From pilot to service platform. From 3 countries with harvests to the entire network. From 320 thousand records to 1.43 million.

Quality criteria, generation of projects and definition of priorities.

From definition of metadata guidelines to the implementation at national nodes.

9	1 431 703 DOCUMENTOS	799 411 ARTÍCULOS	187 350 TESIS DE DOCTORADO
NODOS NACIONALES		7 107 REPORTES	437 835 TESIS DE MAESTRÍA



AGREEMENTS

GUIDELINES

TECHNOLOGY



Interoperability
Regional
Alliances
Projects



OpenAIRE
LR Application Profile
Information quality
Scientific Data
(Guidelines)
Training



Harvesters
Transfer
Pilots
Developments
Community

LaReferencia Software Modules

ANGULAR ADMIN UI (node manager)

SPRING DATA REST - HATEOAS - WS

Domain Model

Actions / Services

SPRING DATA (JPA / SOLR)

SOLR

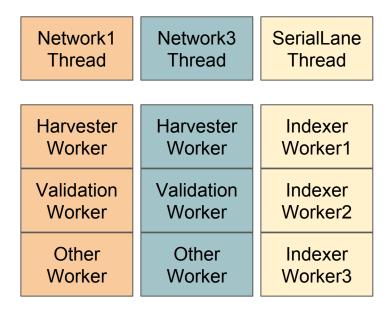
SQL (MD & DM)

VUFIND 4 (web portal and api)

XOAI OAI-PMH PROVIDER

- Java 8 Webapps
- Spring 4 based
- Spring Data JPA / SOLR
- Hibernate4
- Spring Data REST HATEOAS HAL
- All components are integrated with dependency injection.
- Interface based 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



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

Runs in a 512mb RAM single core processor

- 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.

Listado de repositorios disponibles

Puede ejecutar acciones sobre un grupo de repositorios seleccionados utilizando los checkbox de la tabla.

Los datos se actualizan automáticamente cada 5 segundos.

0		ID ÷	Acrónimo ‡	Repositorio ‡	Institución ‡	Co	secha	
0	ø	1	AR	Argentina	MINCYT	No existen cosechas	*: 0 V: 0 T: 0	(2)
0	•	8	PE	Perú	CONCYTEC	No existen cosechas	#: 0 V: 0 T: 0	☑ 🙃
0	•	10	CR	Costa Rica	Costa Rica	No existen cosechas	#: 0 V: 0 T: 0	☑ 🙃
0	٠	5	EC	Ecuador	SENACYT	No existen cosechas	#: 0 V: 0 T: 0	2 0
)	•	2	BR	Brasil	IBICT	No existen cosechas	#: 0 V: 0 T: 0	2 0
)	•	6	SV	El Salvador	El Salvador	Cosecha validada Ø 2017-05-08 05:41:15	#: 2497 V: 987 T: 373	☑ •
)	•	12	motri	Revista Motricidade	Motricidade	Cosecha validada Ø 2017-05-11 06:56:22	#: 429 V: 0 T: 0	€ 0
)	٠	13	arca	Arca	IGC	Cosecha validada Ø 2017-05-11 07:28:04	#: 512 V: 512 T: 0	€ 0
)	•	11	rcomum	Repositório Comum	Projeto RCAAP	Cosecha validada Ø 2017-05-11 07:10:42	#: 14705 V: 14705 T:	☑ •



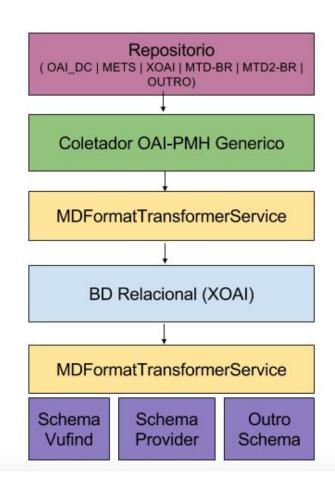
		ID ÷	Acrónimo ‡		Repositorio ÷	Institució
	•	1	AR	Argentin	na	MINCYT
0	Ejecutar acciones configuradas					CONCYTEC
0	Acciones individuales Cosecha Validación/Transformación Indexación Vufind Exportar OAI-PMH Borrado índice Vufind Borrado índice OAI-PMH Limpiar Snapshots Antiguos Eliminar la red				a	Costa Rica
0				ón		SENACYT
0						IBICT
0					or	El Salvador
					lotricidade	Motricidade
0	Detener todas las tareas Re-Agendar tareas programadas					IGC
0	٠	11	rcomum	Reposit	ório Comum	Projeto RCAA

Harvester Worker

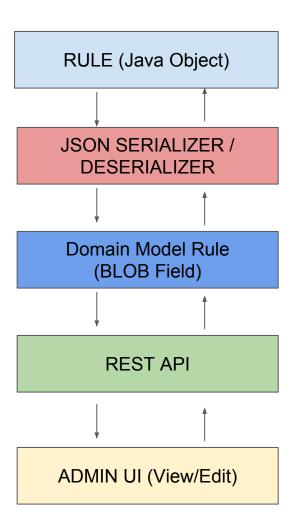
- Based on improvement of OCLCHarvester2 OAI-PMH harvester
- HTTPS Harvesting PATCH
- Bad character encoding transformer/filter to avoid harvesting errors in xml processing.
- Retry model: wait and retry model with incremental wating times.
- Event model that captures responses, stores metadata records o inform errors.
- Incremental harvesting model with states, that permits later incremental validation, transformation and indexing.

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.
- Transformations are made by a configurable metadata schema transformation service.
- To add a new in/out schema format an xlst transformation and xml configuration file must be provided. No coding needed!
- New schema can be defined to harvest or feed index workers and oai-pmh provider.

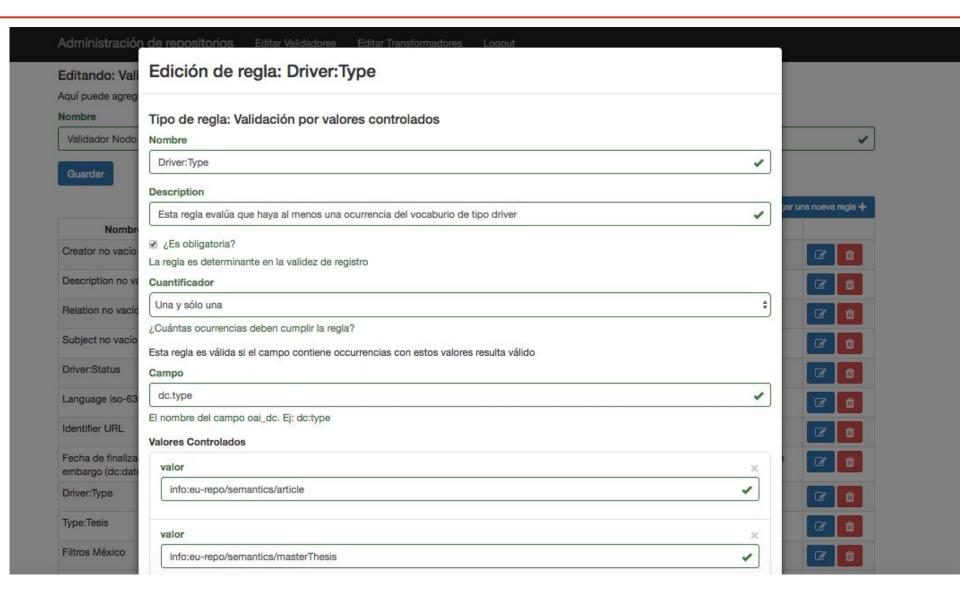


Validation / Transformation Rules

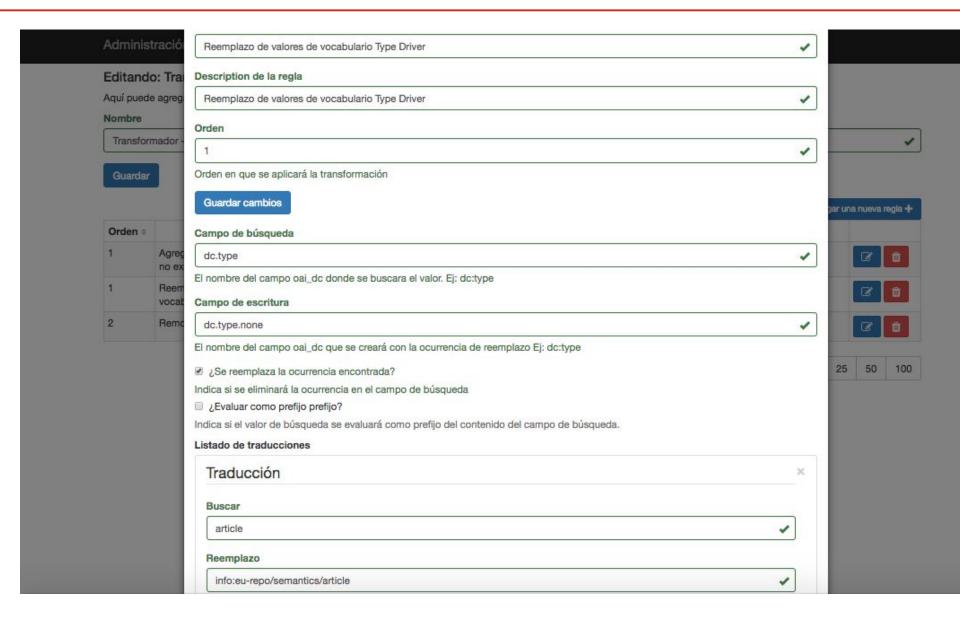


- Validation and Transformation are rule based
- Standard rules for validation/transformation are implemented: controlled vocabularies, regexp, search/replace, etc
- A set of base classes is provided ir order to encapsulate common tasks and give developers an easy way to implement more rules.
- The rules 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.

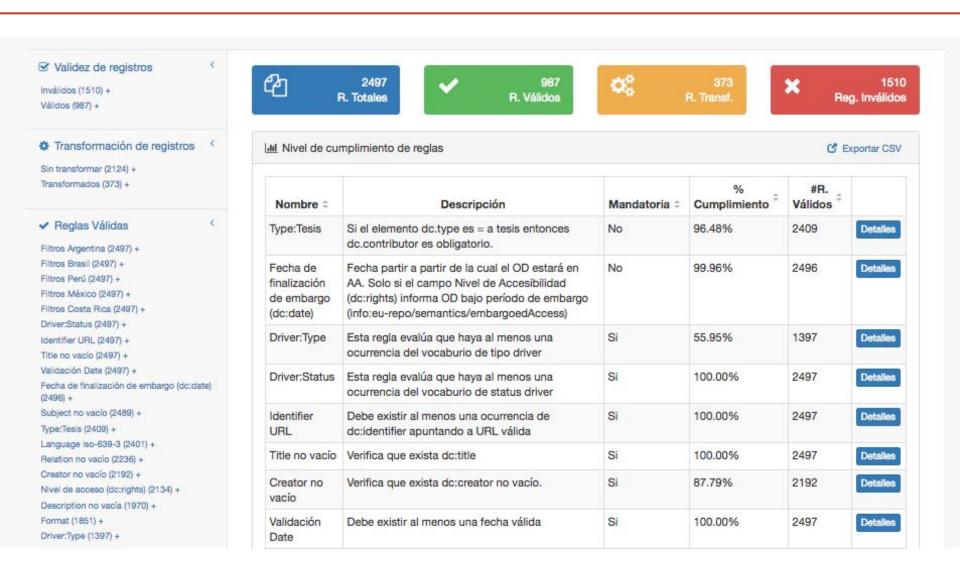
Validation Rule Editing



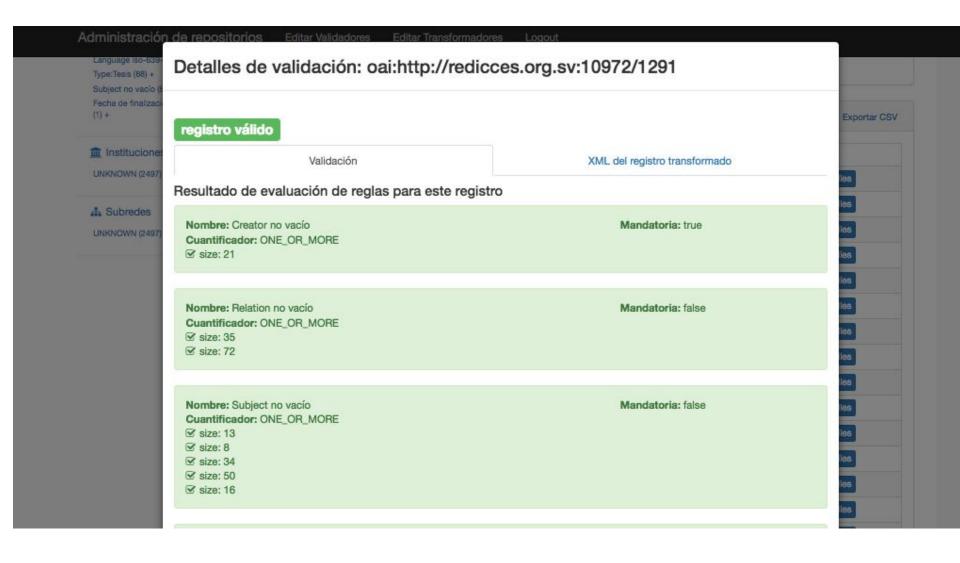
Transformation Rule Editing



Validation results explorer

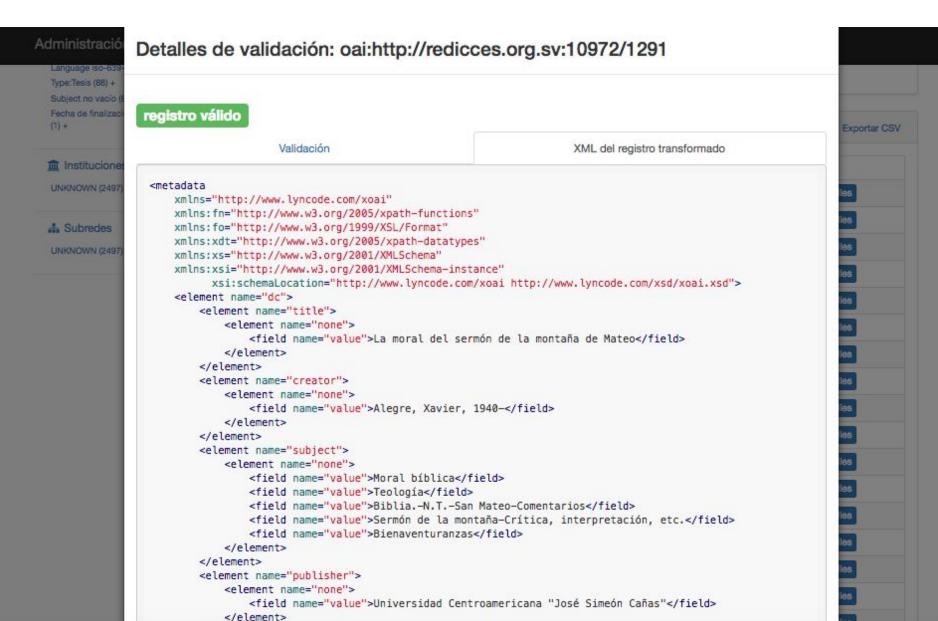


Validation details by record



Transformed XML record display

</element>



LRSoftware 3.2 (IBICT) Highlights / Future

- Incremental Model
- Digital object download
- Full Text indexing
- Vufind 4.0
- JSON Search/Result API
- Especial fields (Thesis)
- CV Data links

Next version challenges:

- Get a better name for the software platform :)
- Distributed statistics OpenAire
- OpenAIRE Broker integration (notifications)
- Dspace 7 Integration via API ??

THANK YOU!

WASHINGTON CARVALHO-SEGUNDO (WASHINGTONSEGUNDO@IBICT.BR)

LAUTARO MATAS (LMATAS@GMAIL.COM)

ANY QUESTIONS?

OASISBR (HTTP://OASISBR.IBICT.BR)

LA REFERENCIA (HTTP://www.lareferencia.info)