kohacon17

FREE LIBRARY SYSTEM MAKATI, PHILIPPINES / June 19-23, 2017

# Publishing Library Linked Data via KOHA and ALIADA open source systems


***Cristina Gareta***

***SCANBIT, Technology and Services for Libraries***

***cgareta@scanbit.net***

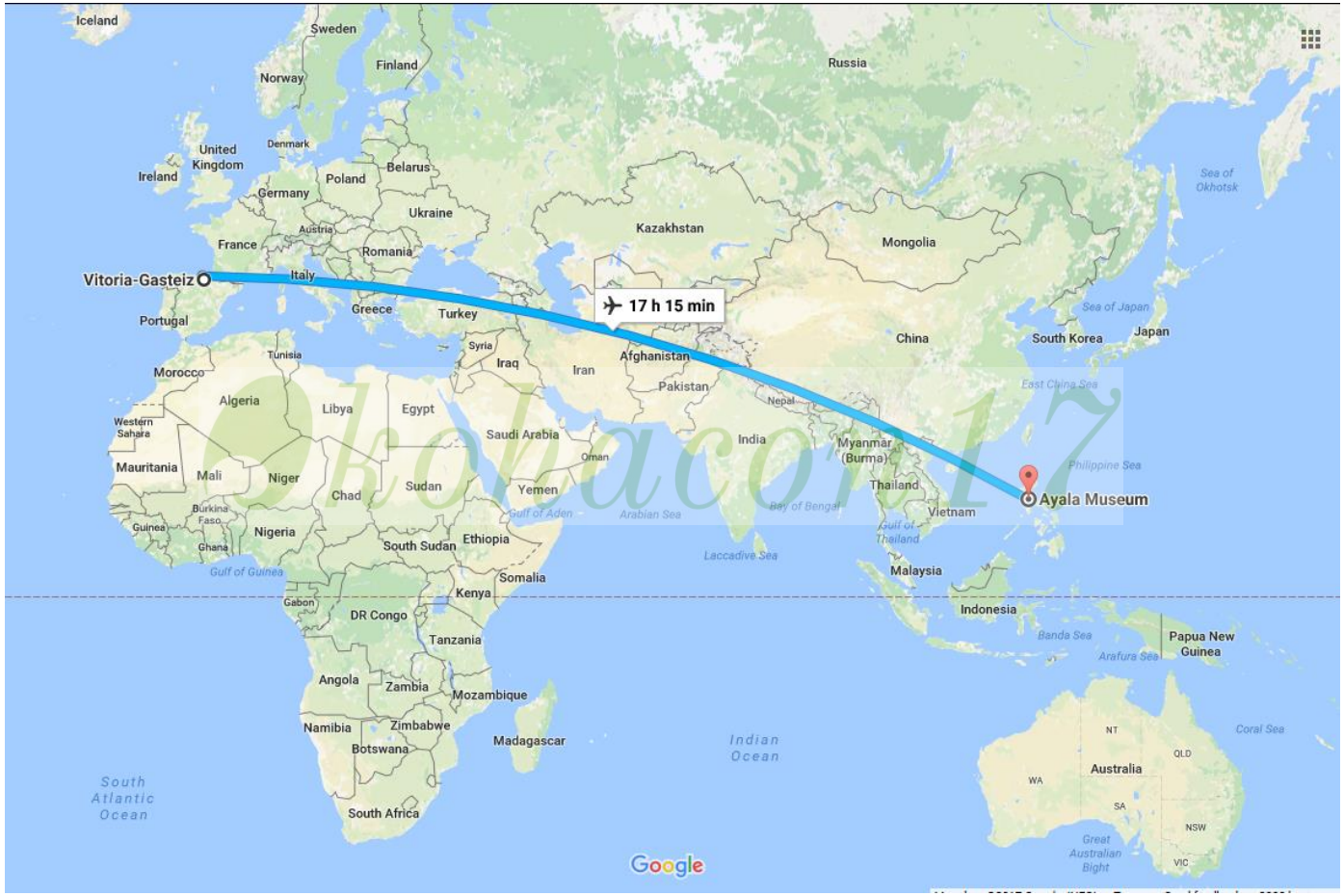

# A short presentation…

SCANBIT TECHNOLOGY & SERVICES

- **20-years** Spanish IT company specialized in technology and outsourcing services for **libraries (public or specialized), archives, museums**, universities and publishers
- Consultancy, development and implementation services for proprietary and open source library systems: **KOHA, VuFind, LibriSuite** (Dobis-Libis, AMICUS)**, Dspace**
- Projects in **Spain, Italy, UK, Germany or Mexico**
- Our team: **Software developers, Systems administrators and Librarians**
- **Innovation and ongoing training** (ALIADA)
- Presentations at **KOHAference** (Spain), **KOHA Italian Group Conference** and **KOHACon**
- **Some references**: Spanish Ministry of Culture, National Library of Spain, Universia-Santander Bank, Basque Government, SOAS University of London, Pontifical University of the Holy Cross,…

# A short presentation…

SCANBIT TECHNOLOGY & SERVICES






***Committed to the KOHA Community!***

# What is Linked Data?

ALIADA

- Tim Berners-Lee, director of the World Wide Web Consortium (W3C), coined the term in a design note about the Semantic Web project
- Method of publishing structured data so that it can be interlinked and become more useful
- Standard Web technologies: HTTP, RDF, URI, SPARQL, OWL…
- To share information in a way that can be read automatically by computers
- Enables data from different sources to be connected and queried

# What is Linked Data?

| | |
|---|---|
| ★ | Data published under open license |
| ★★ | Machine-readable data |
| ★★★ | Non-proprietary format |
| ★★★★ | RDF standards |
| ★★★★★ | Linked RDF |


http://www.w3.org/DesignIssues/LinkedData.html

"We're moving from a web of documents to a **web of data**"

# What is Linked Data?

ALIADA

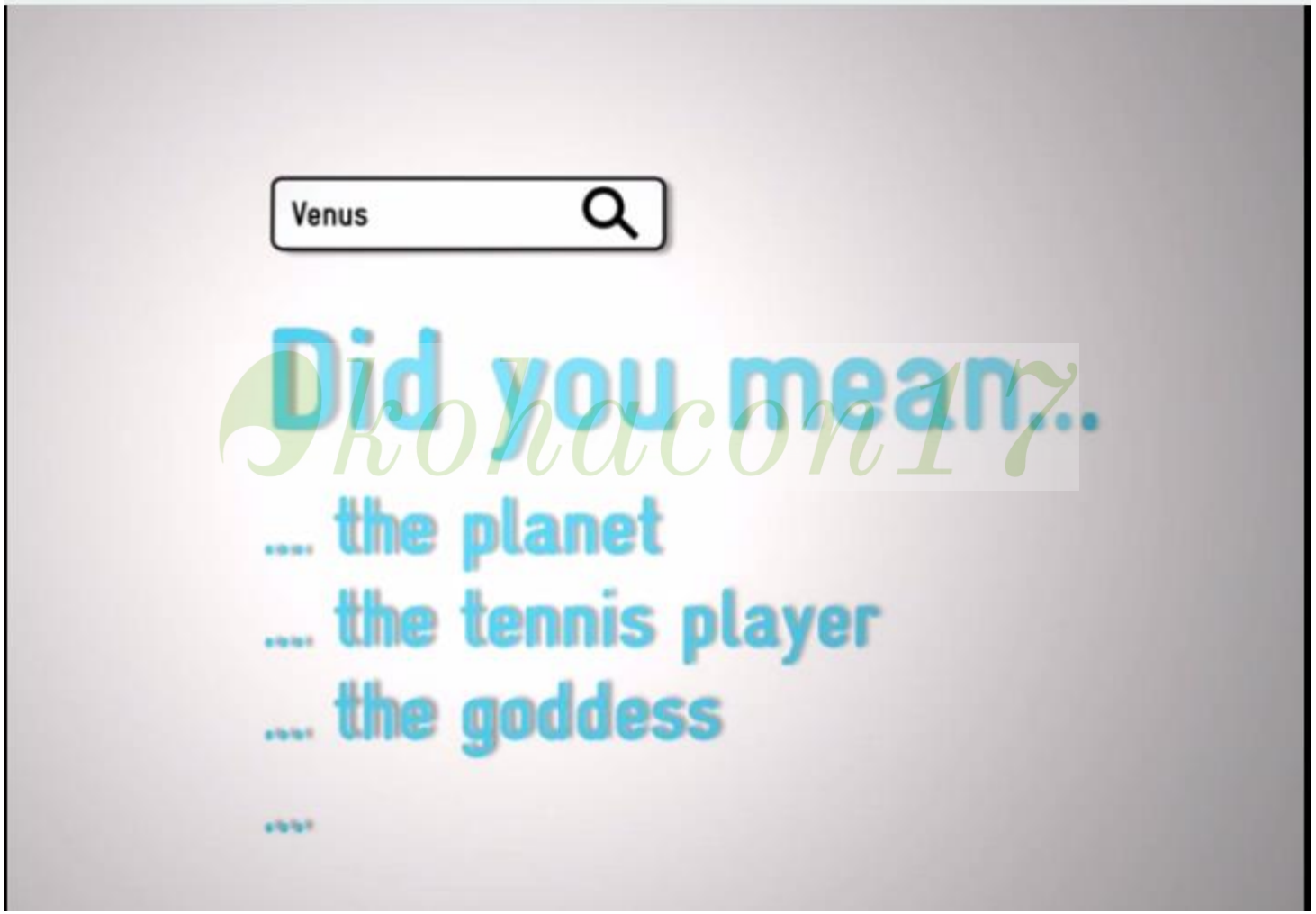
- Be a ‘thing’ (unique url)
- Use HTTP URIs (unique identifiers) to designate your data
- Provide useful information in your URIs
- Use standards (RDF*, SPARQL)
- Include links to other URIs to discover more things



http://viaf.org/viaf/110981647/#Kahlo,_Frida,_1907-1954
http://dbpedia.org/resource/Pablo_Picasso
http://vocab.getty.edu/ulan/500009666

“Use URIs as names for ***things*** (data)”

# What is Linked Data?




Europeana's video explaining LOD: https://vimeo.com/36752317

# Why Library Linked Data

ALIADA

- A **global pool of shared data** that can be re-used to describe resources will avoid the redundant effort of the current cataloging processes.
- The **use of the Web** and Web-based identifiers will make up-to-date resource descriptions directly citable by catalogers.
- Linked Data **is more durable and robust** than metadata formats that depend on a particular data structure.
- Developers will also **no longer have to work with library-specific data formats** (MARC, LIDO).
- With Linked Open Data, libraries can **increase their presence on the Web**, where most information seekers may be found.

http://www.**w3.org**/2005/Incubator/lld/wiki/Benefits

# Why Library Linked Data

*"(…) Because linked data can be easily understood by computers, resulting in opportunities for improved library workflows, enhanced user experiences, and discovery of library collections through a variety of popular sites and Web services, including Google, Wikipedia and social networks."*

(Tennant, Roy. "Getting started with linked data". *OCLC Next, 2016.* http://www.oclc.org/blog/main/getting-started-with-linked-data-3/)

# Why Library Linked Data

ALIADA

DATOS·BNE·es

The European Library
Connecting knowledge

DEUTSCHE NATIONAL BIBLIOTHEK

data.bnf.fr



BRITISH LIBRARY
THE BRITISH NATIONAL BIBLIOGRAPHY
as Linked Open Data

OCLC The world's libraries. Connected.™

LIBRARY OF CONGRESS

LODLAM
Linked Open Data in Libraries Archives and Museums

# The KOHA Community's approach

- https://wiki.koha-community.org/wiki/Linked_Data_RFC : thoughts on how KOHA can start to take advantage of these technologies
- Goals:
  - ✓ Keep MARC, for now
  - ✓ Explore and take advantage of RDF/Linked Data/Semantic Web technologies
  - ✓ Reuse stuff that others have made
  - ✓ Keep things flexible and light weight
  - ✓ Focus on making something that is useful in the OPAC
- First steps
  - ✓ Transform MARC to RDF (e.g. marc2rdf)
  - ✓ Save the resulting RDF in a triplestore
  - ✓ Create an interface in Koha for enriching the RDF in the triplestore with the knowledge of librarians and data from external sources.
  - ✓ Create an interface in the OPAC for exploring the data in the triplestore
  - ✓ Exploring links and relationships, compared to the free text search we are used to with MARC.

# The KOHA Community's approach

- KOHA records can be exported as RDF
- Other discussion topics:
  - Ontology based on BIBFRAME or FRBR?
- Linked Data at the Marseille Koha Hackfest
http://bywatersolutions.com/2017/04/13/linked-data-at-the-marseille-hackfest/

# How to start publishing as Linked Data?

ALIADA

- Cataloguing data according international conceptual models and standards (FRBR, BIBFRAME, CIDOC-CRM, …)
- Exporting records to standard metadata schemes (MARC, LIDO or Dublin Core) and formats (XML)
- Selecting an ontology and vocabularies
- Converting MARC/LIDO/DC metadata to RDF statements
- Linking the own dataset to other datasets (one domain or multidomain)
- Publishing Linked (Open) Data

# How to start publishing as Linked Data?

ALIADA

*"Librarians and curators are experts in cataloguing and making accessible their resources, but the don't know about Linked Data technology, so they need **an ally**"*

*(ALIADA Project Consortium, 2015)*

# ALIADA, the ally to publish LODLM

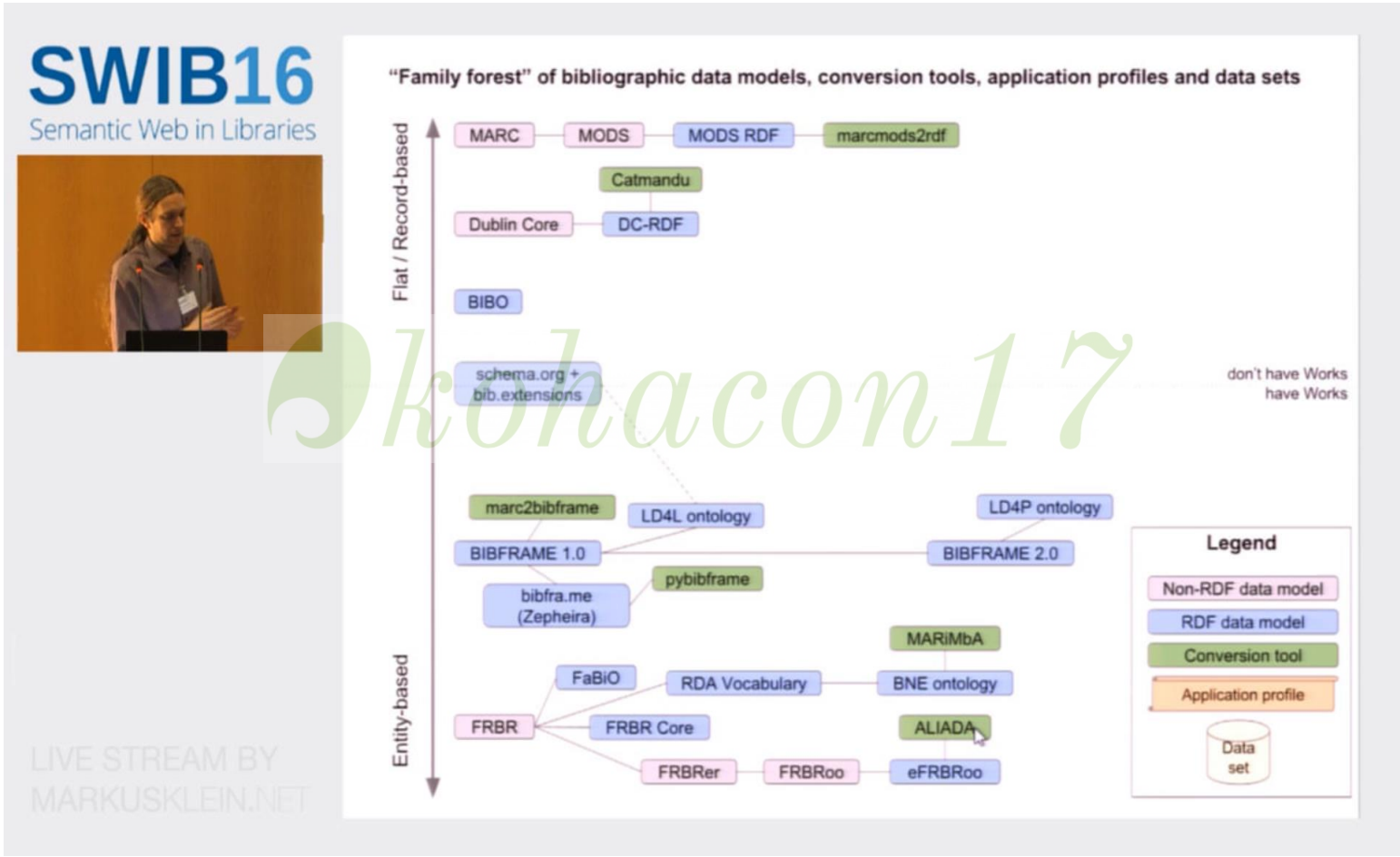
- **EU FP7**- ENV-2012 Collaborative project 2013-2015
- **Project website**: http://www.aliada-project.eu/
- **Partners**: Art museums, libraries, ILS vendors, experts on Semantic technologies
- **Final release**: October 2015
- **Open source** community: https://github.com/ALIADA/aliada-tool
- ALIADA is **free software**, you can redistribute it and/or modify it under the terms of the GNU GPL v3 (License)

# ALIADA, the ally to publish LODLM

ALIADA

- **Open source** Java application to **automatically publish** as Linked Data the metadata created by a library or museum management system
- Supported **metadata types** (types of datasets): bibliographic records, authority records, descriptions of museum objects and other information resources
- Compliant with **MARC XML, LIDO XML and Dublin Core** formats
- Conversion to RDF triples (mapping) according to the **ALIADA ontology**, mainly based on **FRBRoo**, SKOS, WGS84 and FoaF ontologies
- Linking to other datasets, such as Europeana, British National Bibliography, Spanish National Library, Freebase Visual Art, DBpedia, Hungarian National Library, Library of Congress Subject Headings, Lobid, MARC codes list, VIAF Virtual International Authority File or Open Library
- Automatic publication of dumps (URIs) and SPARQL Endpoint on DataHub

# ALIADA, the ally to publish LODLM




Suominen, Osma. "From MARC Silos to Linked Data Silos?". *SWIB16*. Bonn, 2016
http://swib.org/swib16/slides/suominen_silos.pdf

# ALIADA, the ally to publish LODLM

## 1st prototype (2014)

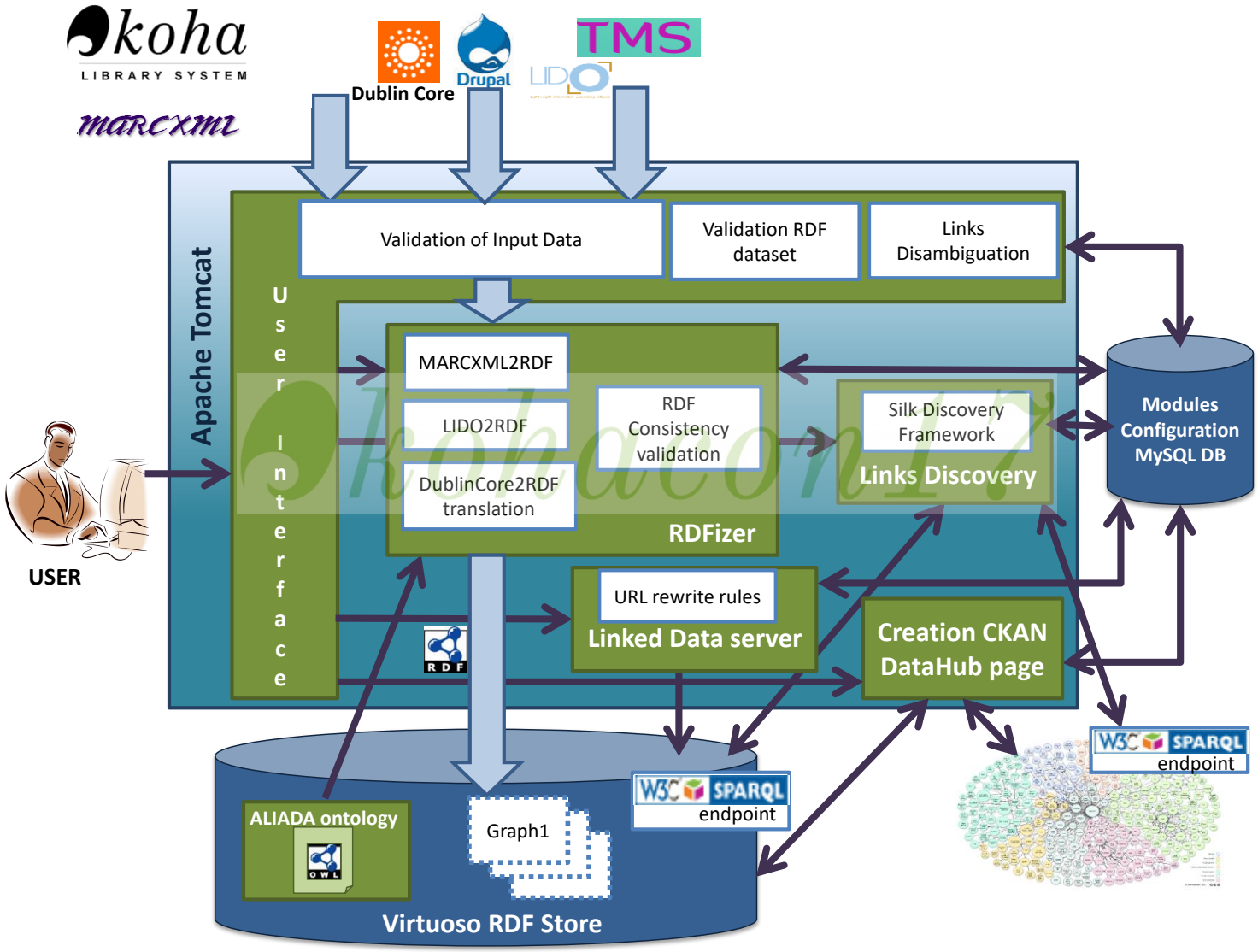
- User interface in Spanish and English.
- Validation of imported records (MARC Bibliographic and LIDO)
- Mapping templates (FRBRoo)
- RDF-izer (ALIADA ontology)
- Linking to some datasets: Europeana, BNB, BNE, Freebase, Dbpedia, NSZL, Geonames and MARC Code Lists (SPARQL end-point)
- Linked data server creation + SPARQL endpoint + URIs dereferencing .
- Linked dataset validation: through a number of SPARQL queries.

# ALIADA, the ally to publish LODLM

## 2nd prototype (2015)

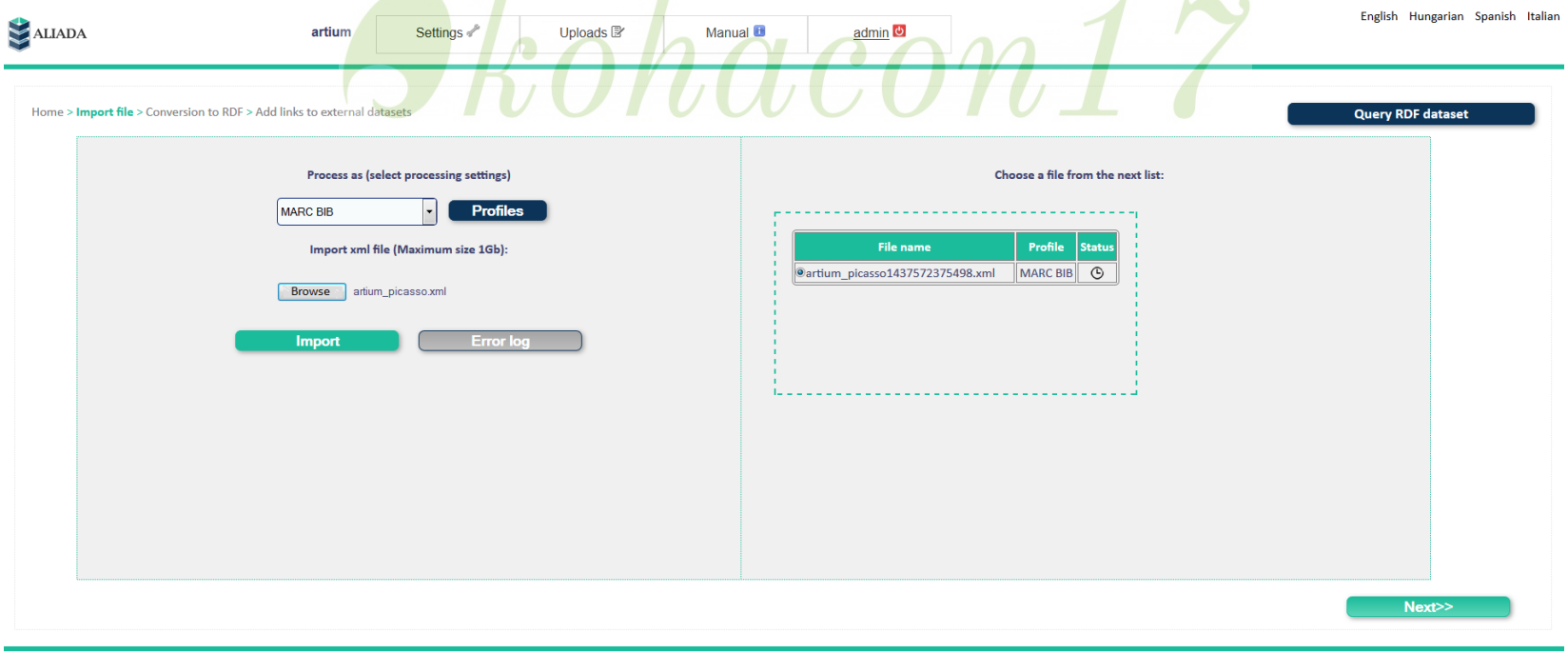
- Translation of DublinCore XML and MARC XML Authorities to ALIADA ontology.
- Validation of RDF dataset consistency.
- NER (Named Entity Recognition) for some text free elements
- Ad-hoc linking to some of the listed external datasets , which do not provide a SPARQL endpoint such as VIAF, LOBID, Open Library and Library of Congress Subject Headings.
- Links disambiguation: the system offers to the user a set of possible ambiguous links: so the user can decide which links are correct and which ones should be rejected.
- Advanced URI de-referencing and creation of a web page for the generated dataset.
- Publication in CKAN of the created linked dataset + DataHub LOD validator.
- Translation of the user interface to Italian and Hungarian.
- ALIADA offers REST services in order to be integrated with other systems: ILS and CMS.

# ALIADA, architecture



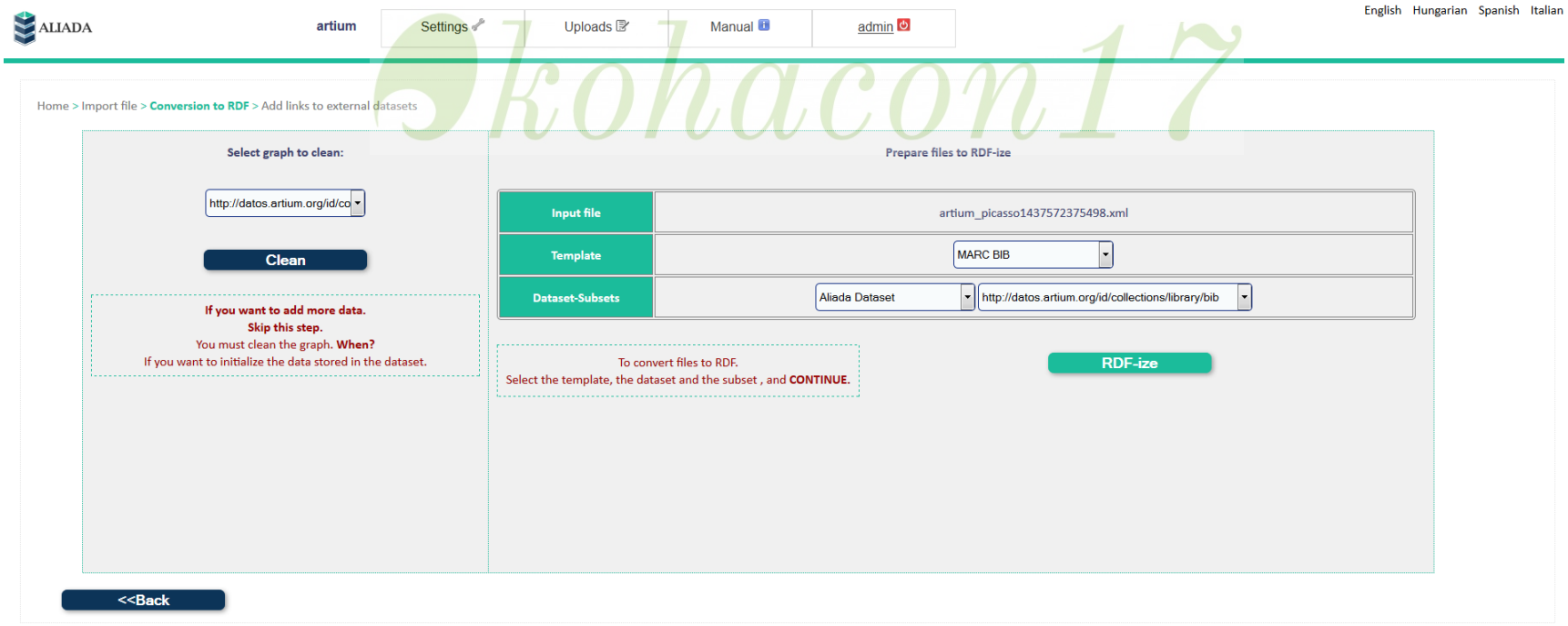
# ALIADA, modules

- **ALIADA Selection and Validation of Metadata**
  - MARCXML/Dublin Core
  - From KOHA export files
  - From KOHA OAI server



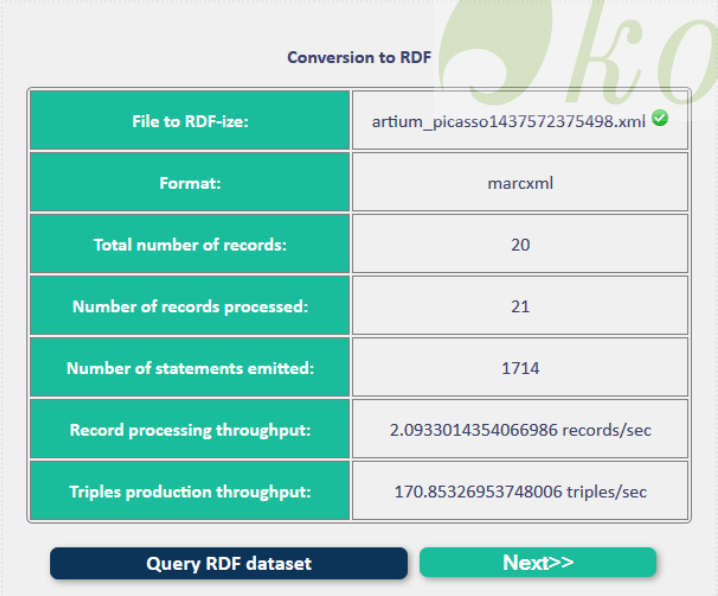
# ALIADA, modules

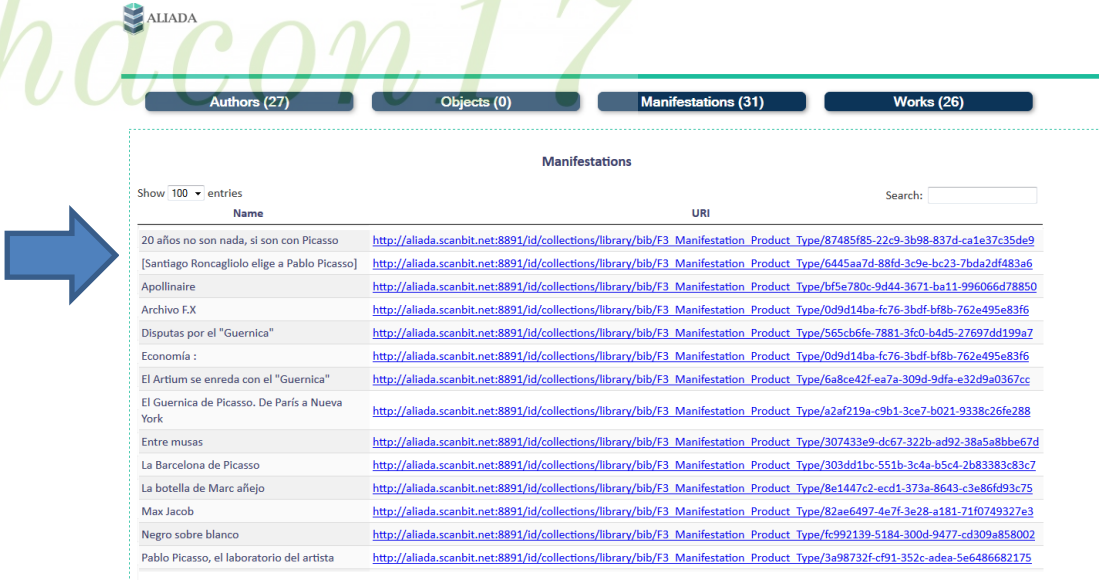
- **ALIADA Conversion to RDF**
  - According to ALIADA Ontology (FRBRoo, SKOS, FOAF…)
  - Predefined mapping templates
  - Datasets for Bib and Aut records



# ALIADA, modules

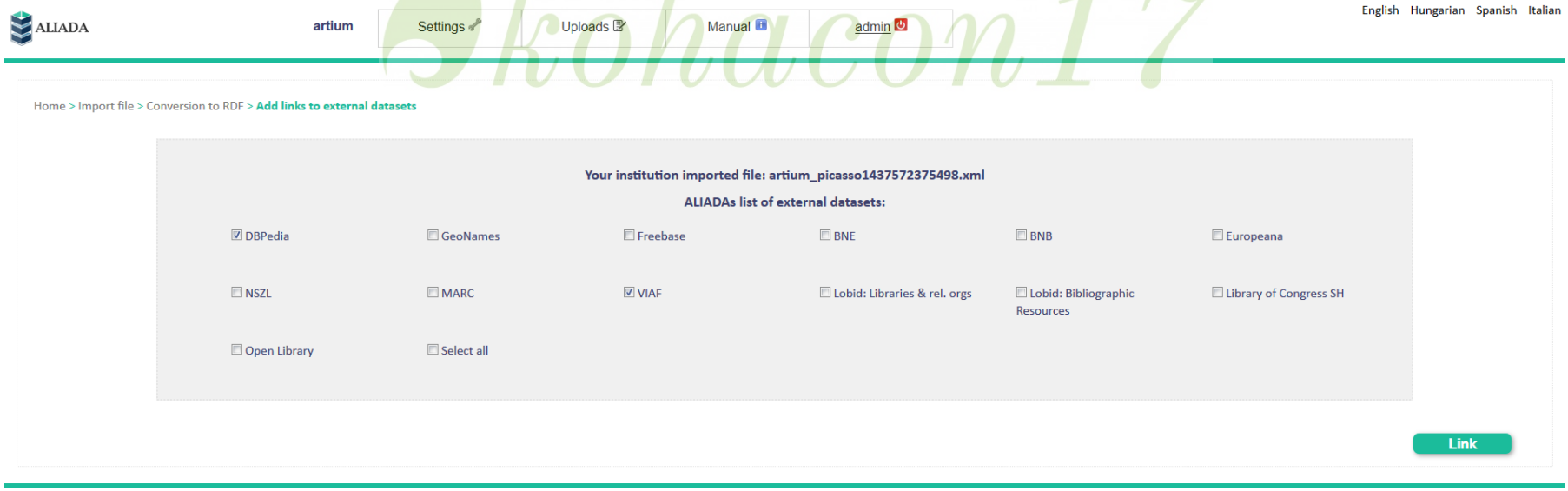
- **ALIADA Conversion to RDF**
  - Check of the triplestore (usability)



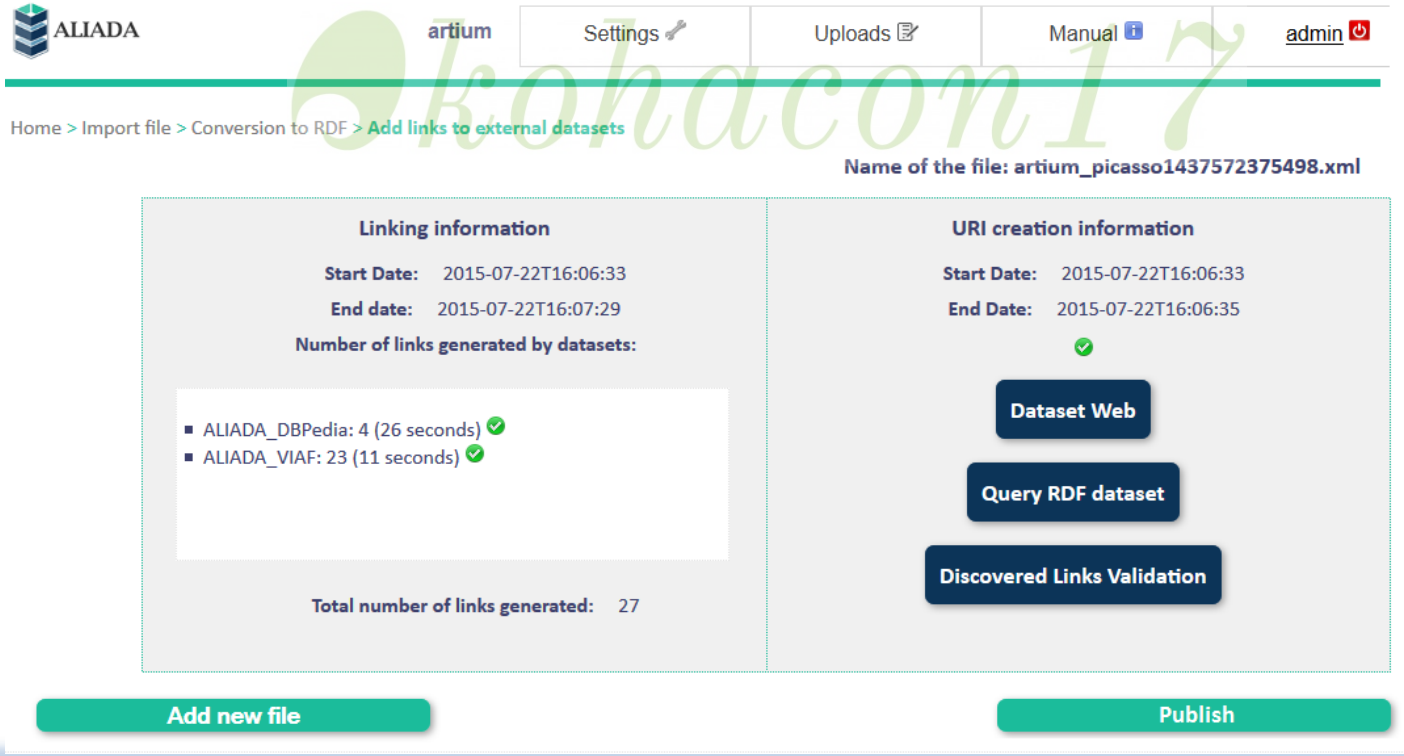
# ALIADA, modules

- **ALIADA Linking to External Datasets**
  - Predefined list of datasets
  - Selection of the datasets to discover the links
  - Creation of links and URIs



# ALIADA, modules

- **ALIADA Automatically Publishes the Dataset**
  - Dump of triples
  - SPARQL end point
  - Integration with Datahub



# ALIADA, main features

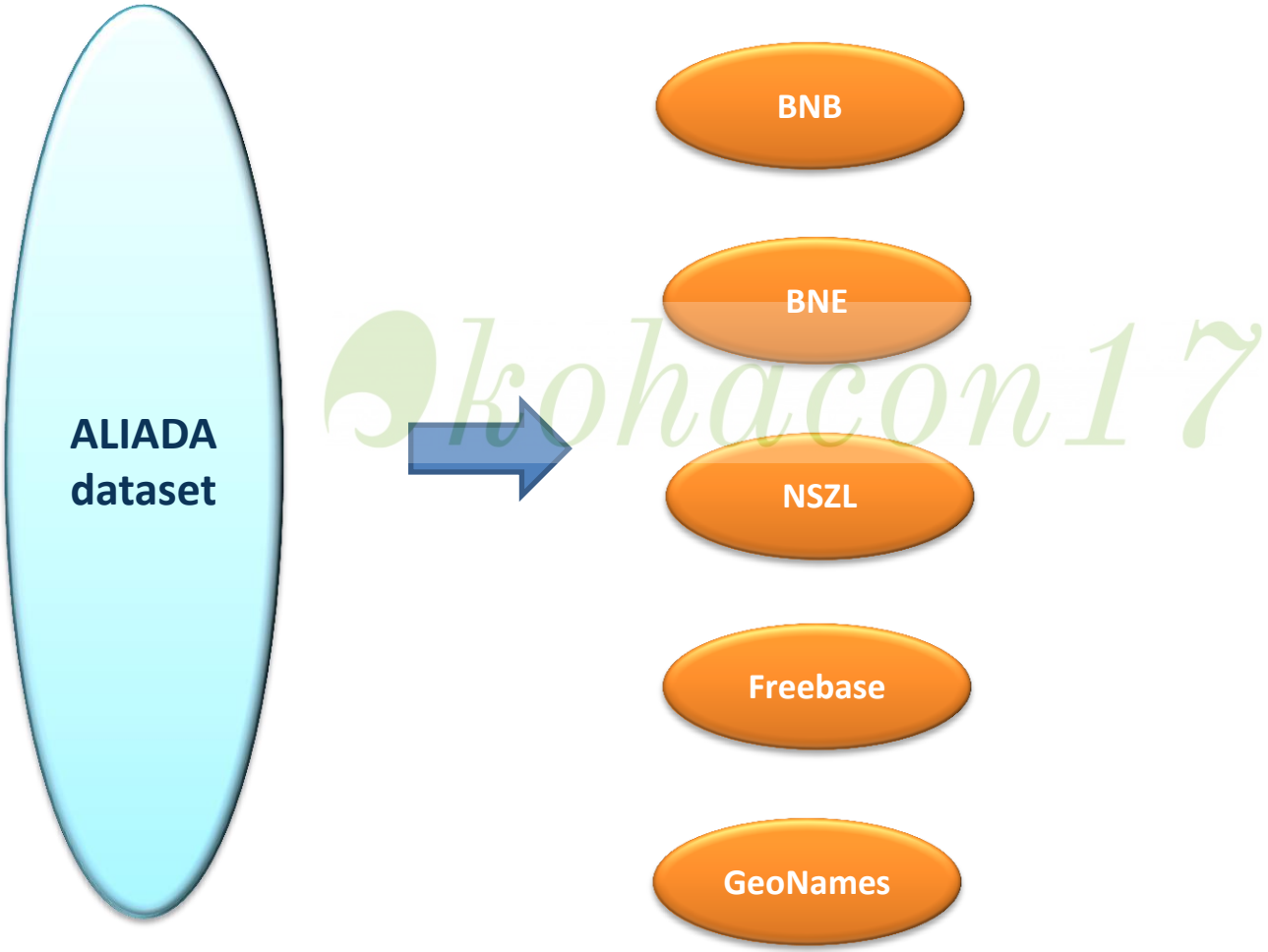
- **ALIADA RDFizer**
  - Scalability (RESTful application, Apache Camel asynchronous channels, JEE web application)
  - Modularity (Conversion templates are configurable and extensible)
  - Reusability (Standalone installation of the RDFizer)
  - Easy to use/maintain (Conversion job is controlled by the so-called “conversion templates”, which are runtime-interpreted scripts very easy to maintains)
  - Easy to extend (The library is free to create their conversion template, producing an arbitrary output format, with another ontology)
  - Validation before the conversion (Jena OWL Micro Reasoner)

# ALIADA, main features

- **NER of free text fields.**
  - A dedicated component for doing NLP (Natural Language Processing). Recognizes sequences of words in a text which are the names of things, such as person, company names and places.
  - In LIDO it is applied to <lido:descriptiveNoteValue xml:lang="en" lido:type ="physical-description">Sculpture of Mozart</lido:descriptiveNoteValue >
  - In MARC it is applied to
    - “522” ”a”: Geographic Coverage Note.
    - ”525” ”a”: Supplement Note.
    - ”520” ”a”: Summary, etc.
    - ”520” ”b”: Expansion of summary note.
  - The NER results are stored as RDF triples that enrich the owning records

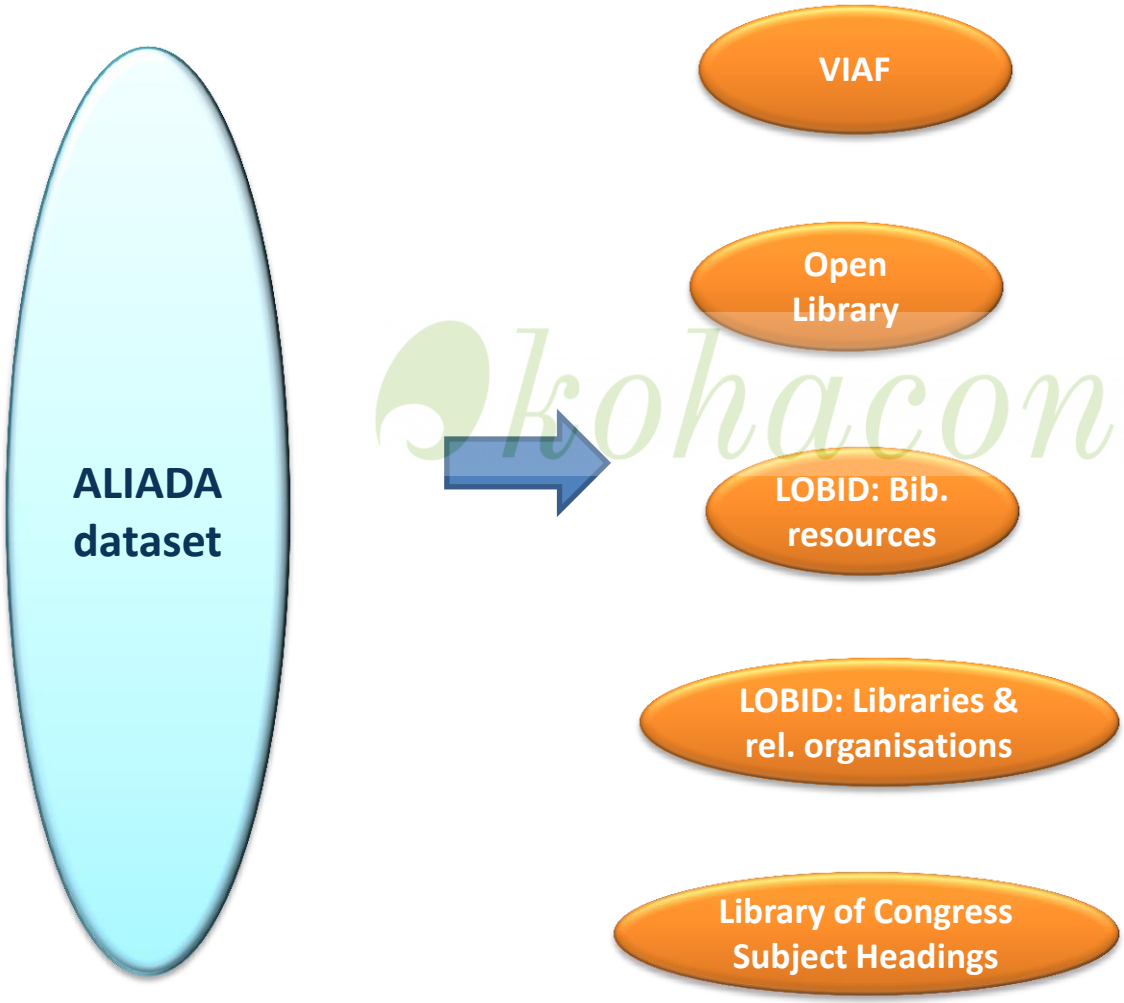
# ALIADA, main features

- **Linking to Datasets that provide SPARQL endpoint**



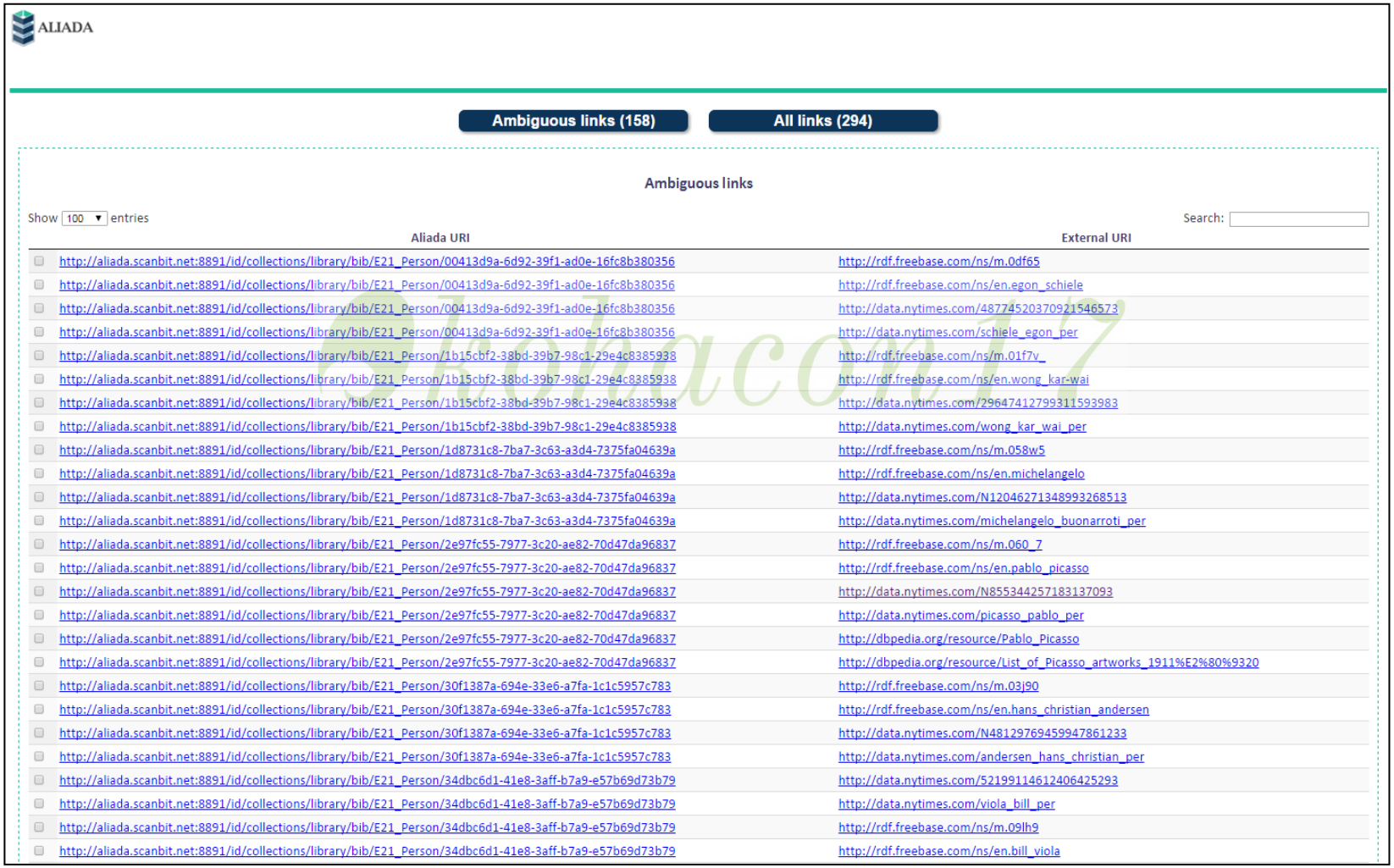
# ALIADA, main features

- **Linking to Datasets that do not provide SPARQL endpoint**



# ALIADA, main features

- **Links disambiguation**



# ALIADA, main features

ALIADA

- **Advanced URI de-referencing**

  URI regulation : http://www.w3.org/TR/cooluris/

  - Be on the web
    - Machines and people should be able to retrieve a description about the resource identified by the URI from the Web.
    - Machines should get RDF data and humans should get a readable representation, such as HTML
  - Cool URIs
    - Simplisity: Short and mnemonic
    - Stability: Does not change as long as possible
    - Managabilty: Keep all URIs in a dedicated subdomain

# ALIADA, main features

- **URI name convention of ALIADA: URI structure**

**{domain}/{type}/{concept}/{class}/{reference}.{format}**

*data.szepmuveszeti.hu/id/collections/museum/E18_Physical_Thing/szepmuveszeti.hu_object_29*

# ALIADA, main features

- **URI name convention of ALIADA: URI structure**

| Extension | Media type | Common name |
|---|---|---|
| .html | text/html | HTML |
| .json | application/rdf+json | JSON |
| .jsonld | application/ld+json | JSON-LD |
| .nt | text/plain | N-Triples |
| .opac | OPAC of the institution (if restful) | OPAC |
| .rdf | application/rdf+xml | RDF/XML |
| .ttl | text/rdf+n3 | N3/Turtle |

# ALIADA, main features

- **Dataset Default Web Page created by ALIADA**

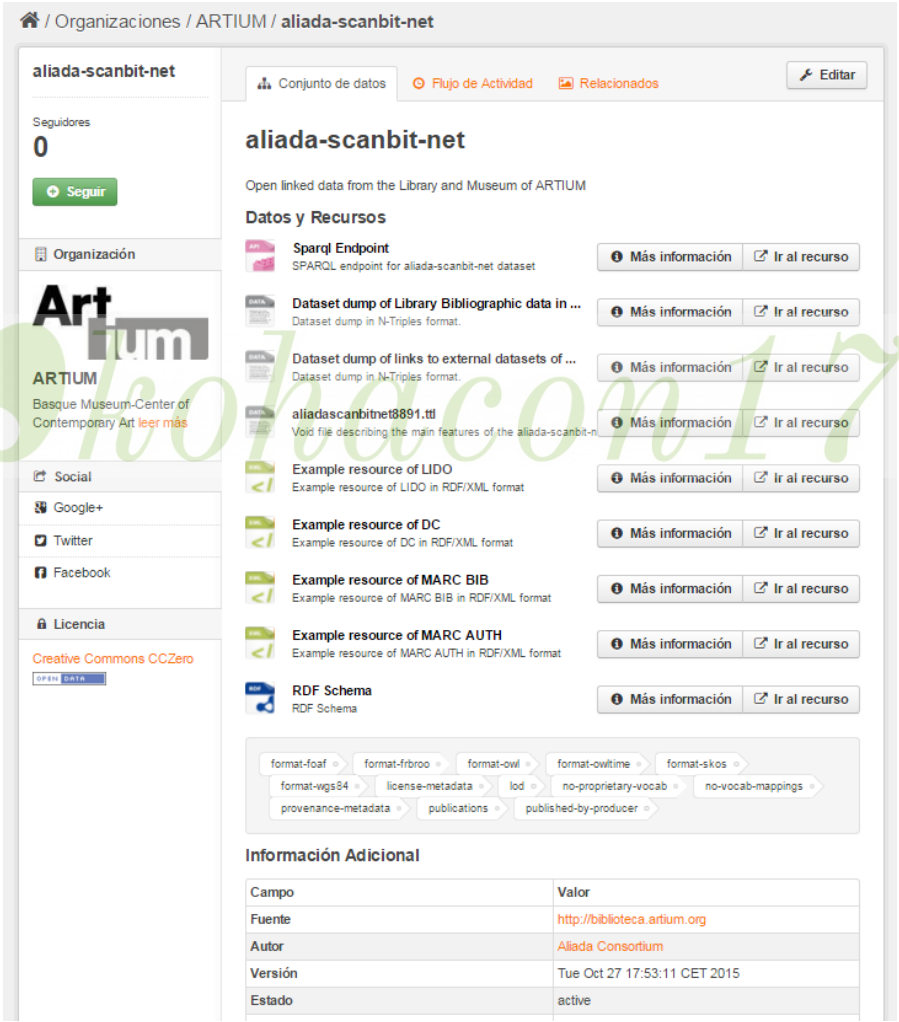
Art ium

## Aliada dataset

| | |
|---|---|
| description | Open linked data from the Library and Museum of ARTIUM |
| publisher | ARTIUM |
| source | http://biblioteca.artium.org |
| created | 2015-10-05 |
| contributor | Aliada Consortium |
| license | http://creativecommons.org/publicdomain/zero/1.0/ |
| SPARQL endpoint | http://aliada.scanbit.net:8891/sparql |
| vocabulary | http://aliada-project.eu/2014/aliada-ontology# |
| number of triples | 11566 |
| list of resources | http://aliada.scanbit.net:8891/doc/collections |
| subsets | Library Bibliographic data: http://aliada.scanbit.net:8891/doc/collections/library/bib<br>Library Authority data: http://aliada.scanbit.net:8891/doc/collections/library/auth<br>Museum data: http://aliada.scanbit.net:8891/doc/collections/museum<br>DC data: http://aliada.scanbit.net:8891/doc/collections/dc |

# ALIADA, main features

- **Publication in CKAN Datahub**

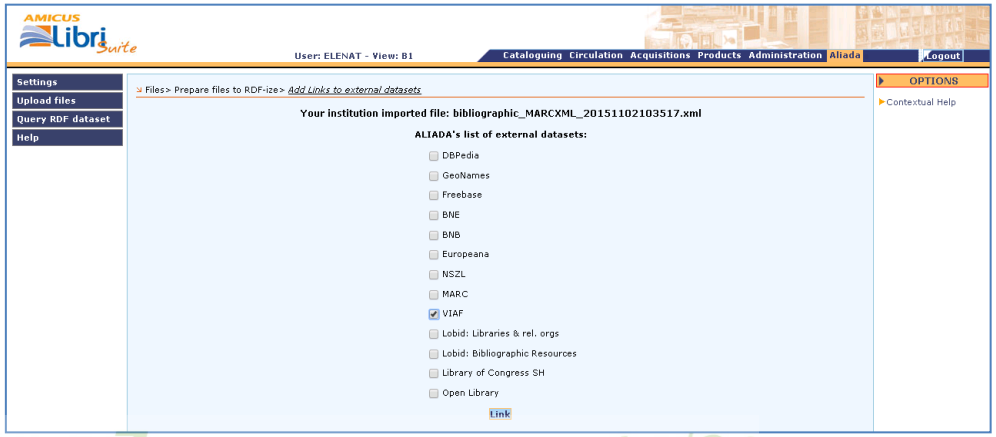


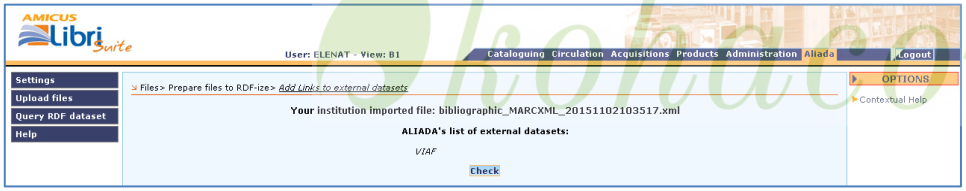
# ALIADA, integration with the ILS

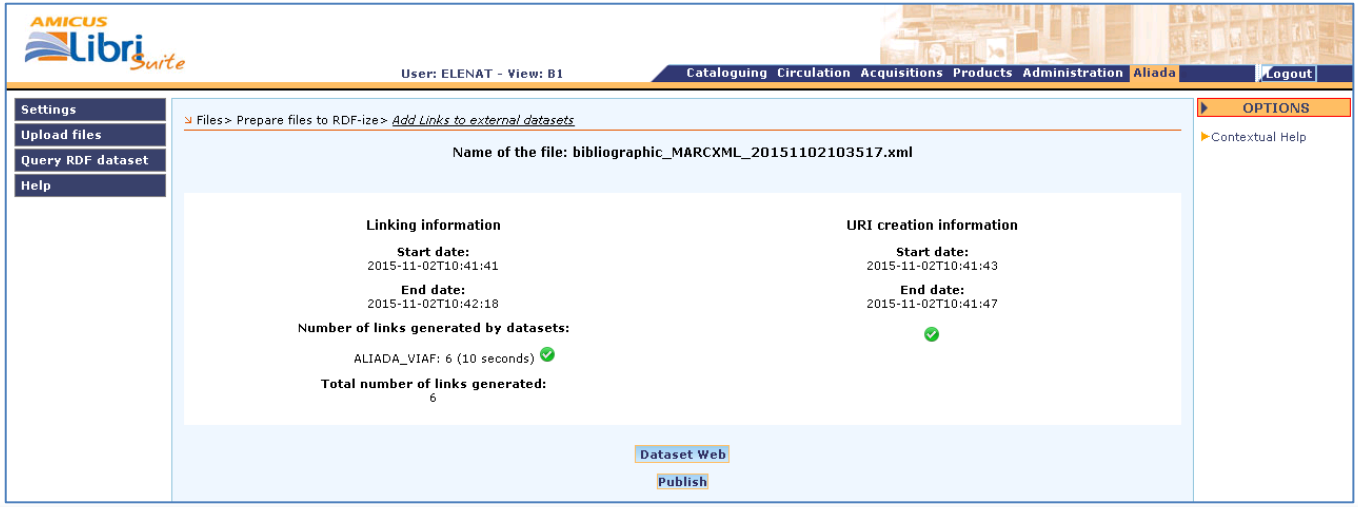
- **RESTful API**

# ALIADA, integration with the ILS

- **RESTful API**



# ALIADA, integration with the ILS

*New KOHA Linked Data Module based on ALIADA open source Library Linked Data Publisher?*

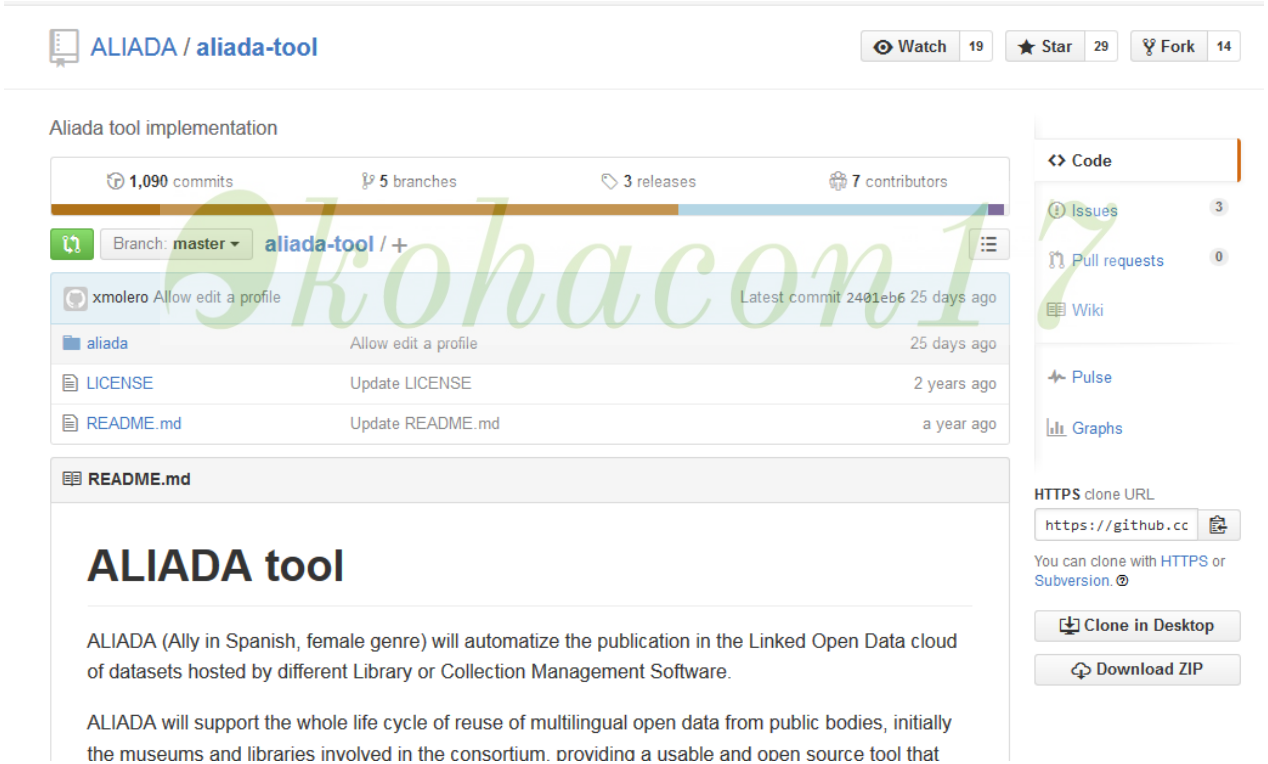


# ALIADA open source community


http://www.aliada-project.eu/

# ALIADA open source community

The source code of the second release has been uploaded to the GitHub's ALIADA repository and the technical documentation, the user manual and the wiki have also been updated.




https://github.com/ALIADA/aliada-tool/

# ALIADA open source community

*What's next?*

- *Add BIBFRAME*
- *Add schema.org*
- *Customize the RDF Conversion mapping*
- *Data Quality Assurance*
- *User interface for searching datasets*
- *…*

**https://github.com/ALIADA/aliada-tool/**

# THANK YOU FOR YOUR ATTENTION!
# ANY QUESTIONS?



www.scanbit.net