

Publishing Library Linked Data via KOHA and ALIADA open source systems

> Cristina Gareta SCANBIT, Technology and Services for Libraries <u>cgareta@scanbit.net</u>



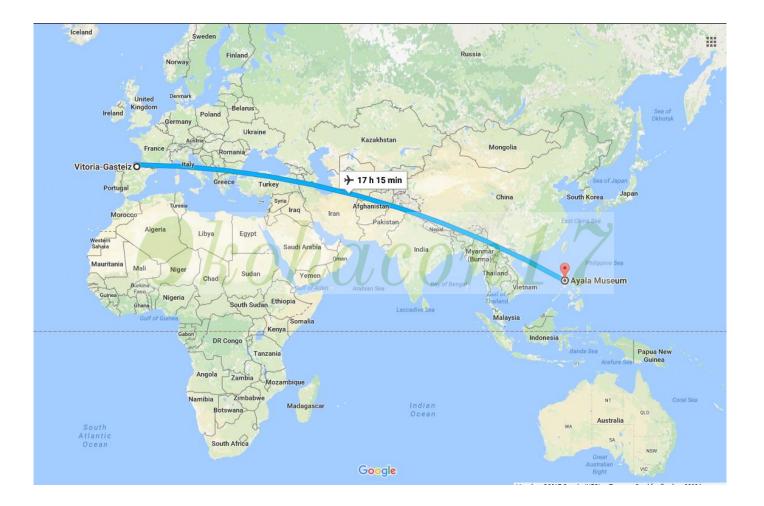
A short presentation...



- 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 <u>services</u> 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...





Committed to the KOHA Community!

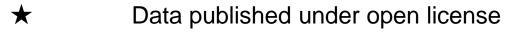
What is Linked Data?



- 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







- $\star \star$ Machine-readable data
- $\star \star \star$ 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?



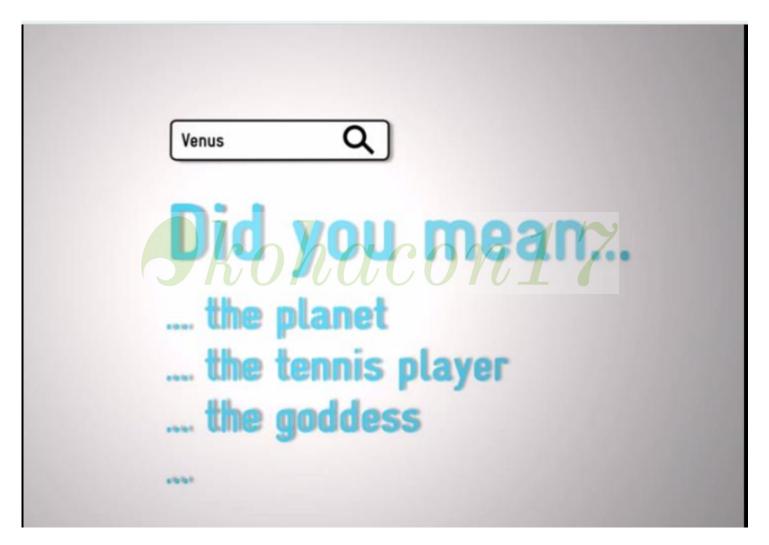
- Be a 'thing' (unique url)
- Use HTTP URIs (unique identifiers) to designate your data
- Provide useful information in your URIs
- Use standards (RDF*, SPARQL)
- Include links to other URIs to discover more things

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



- 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/Ild/wiki/Benefits



"(...) 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. <u>http://www.oclc.org/blog/main/getting-started-with-linked-data-3/</u>)

Why Library Linked Data



DATOS·BNE·es





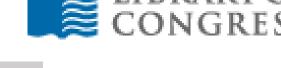












The KOHA Community's approach



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. <u>marc2rdf</u>)
 - $\checkmark\,$ 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 <u>http://bywatersolutions.com/2017/04/13/linked-data-at-the-marseille-hackfest/</u> How to start publishing as Linked Data?

- 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







"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)



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

ALIADA, the ally to publish LODLM

- 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

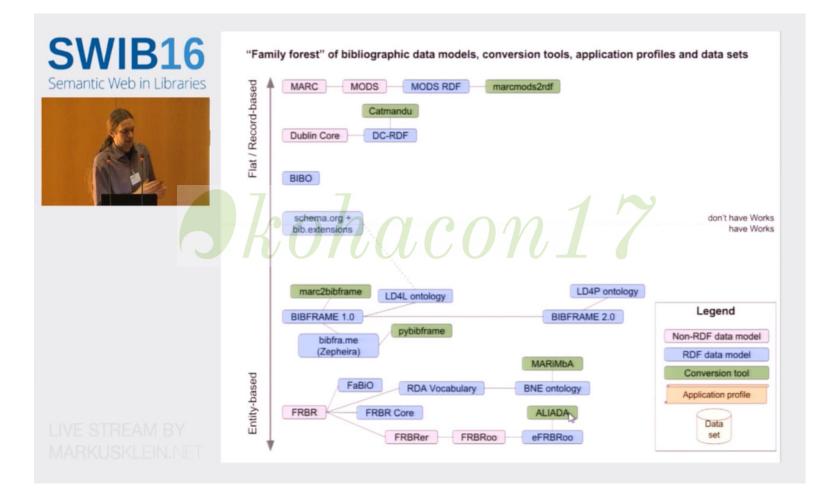


- Compliant with MARC XML, LIDO XML and Dublin Core formats
- Conversion to RDF triples (mapping) according to the ALIADA ontology, mainly based on FRBRoo, SKOS, WGS84 and FoaF ontologies
- Linking to other datasets, such as Europeana, British National Bibliography, Spanish National Library, Freebase 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



1st prototype (2014)

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

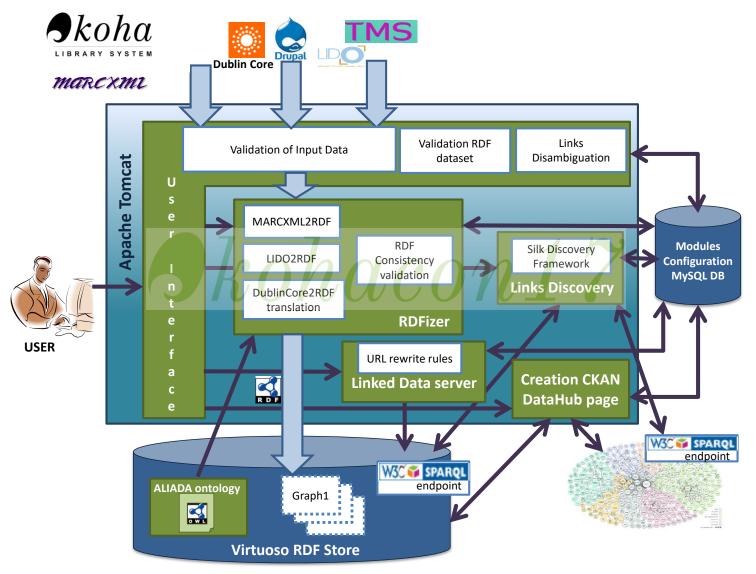


2nd prototype (2015)

- 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 Project Consortium



ALIADA Selection and Validation of Metadata

- MARCXML/Dublin Core
- From KOHA export files
- From KOHA OAI server

Sallada 🕈	artium Settings 📌 Uploads 🕑 Manu	al 🗓 admin 💟	English Hungarian Spanish Italian
Home > Import file > Conversion to RDF > Ac	Process as (select processing settings)	Choose a file from the next list:	Query RDF dataset
	MARC BIB Profiles Import xml file (Maximum size 1Gb): Browse artium_picasso.xml	File name Profile Status @ artium_picasso1437572375498.xml MARC BIB Image: Contemportal Status	
-	Import Error log		
			Next>>



ALIADA Conversion to RDF

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

ALIADA	artium Settings	🖌 Uploads 🕏	Manual 🗓	admin 🔮			English Hungarian Spanish Italia
Home > Import file > Conve	rsion to RDF > Add links to external datasets	KOI	\mathcal{U}	CO	nl		
	Select graph to clean:			Prepare files t	o RDF-ize		
	http://datos.artium.org/id/co •	Input file		arti	um_picasso1437572375498.xml		
	Clean	Template			MARC BIB		
	If you want to add more data.	Dataset-Subsets	AI	iada Dataset	http://datos.artium.org/id/colle	ctions/library/bib	
If you v	Skip this step. You must clean the graph. When? vant to initialize the data stored in the dataset.	To convert Select the template, the dataset			RDF	ize	



ALIADA Conversion to RDF

Check of the triplestore (usability)





ALIADA Linking to External Datasets

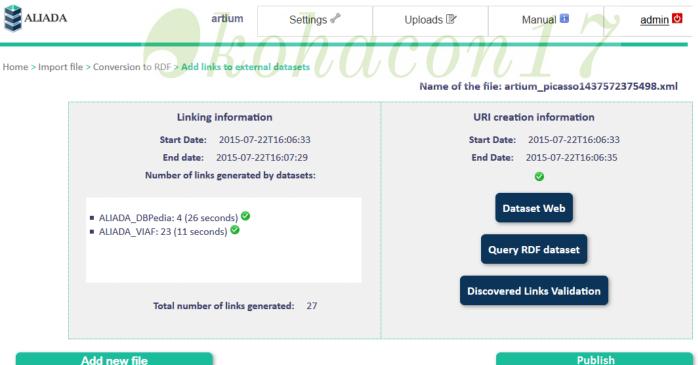
- Predefined list of datasets
- Selection of the datasets to discover the links
- Creation of links and URIs

Sallada 🕈	artium	Settings 🖋	Uploads 🖹 Manual	e <u>admin</u>	11	7	English Hungarian Spanish Itali
Home > Import file > Conv	rersion to RDF > Add links to external	datasets					
				le: artium_picasso1437572375498.xml :t of external datasets:			
	☑ DBPedia	GeoNames	🖾 Freebase	BNE BNE	BNB	🖾 Europeana	
	SZL NSZL	MARC	VIAF	🖾 Lobid: Libraries & rel. orgs	Lobid: Bibliographic Resources	Library of Congress SH	
	🖾 Open Library	Select all					
							Link



ALIADA Automatically Publishes the Dataset

- Dump of triples
- SPARQL end point
- Integration with Datahub





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 socalled "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)

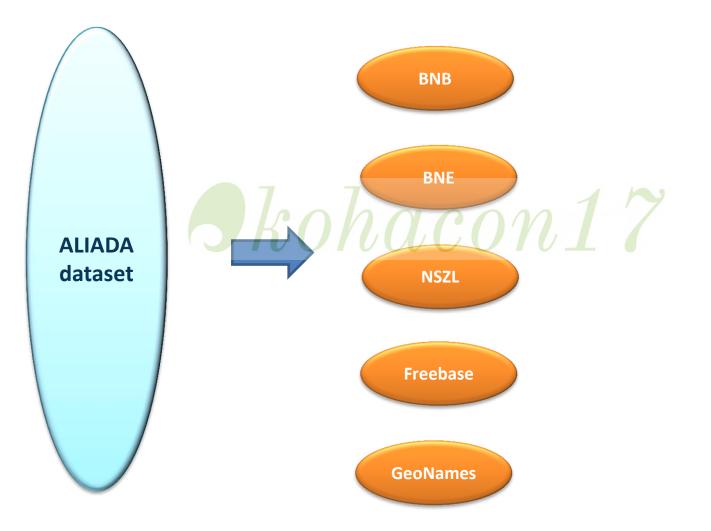


NER of free text fields.

- 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

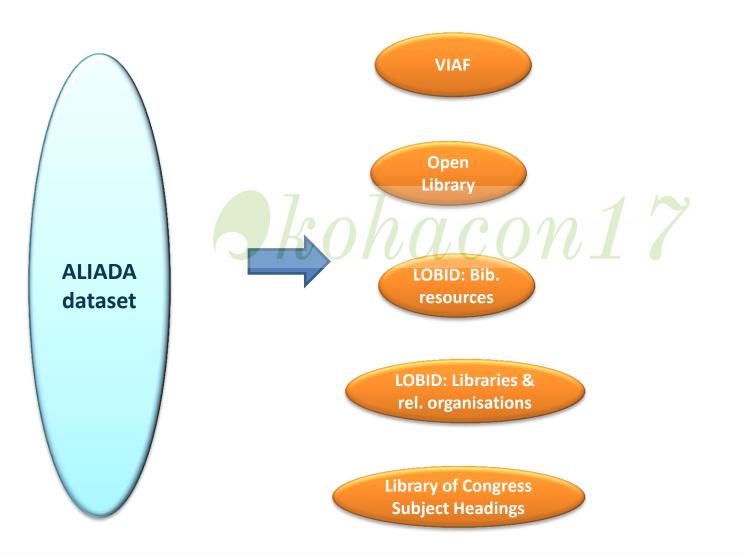


Linking to Datasets that provide SPARQL endpoint





Linking to Datasets that do not provide SPARQL endpoint





Links disambiguation

ALIADA	Α.	
	Ambiguous links (158)	All links (294)
	Ambiguou	s links
Show 100	▼ entries	Search:
	Aliada URI	External URI
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/00413d9a-6d92-39f1-ad0e-16fc8b380356	http://rdf.freebase.com/ns/m.0df65
http:/	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/00413d9a-6d92-39f1-ad0e-16fc8b380356	http://rdf.freebase.com/ns/en.egon_schiele
http:/	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/00413d9a-6d92-39f1-ad0e-16fc8b380356	http://data.nytimes.com/48774520370921546573
http:/	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/00413d9a-6d92-39f1-ad0e-16fc8b380356	http://data.nytimes.com/schiele_egon_per
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1b15cbf2-38bd-39b7-98c1-29e4c8385938	http://rdf.freebase.com/ns/m.01f7v_
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1b15cbf2-38bd-39b7-98c1-29e4c8385938	http://rdf.freebase.com/ns/en.wong_kar-wai
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1b15cbf2-38bd-39b7-98c1-29e4c8385938	http://data.nytimes.com/29647412799311593983
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1b15cbf2-38bd-39b7-98c1-29e4c8385938	http://data.nytimes.com/wong_kar_wai_per
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1d8731c8-7ba7-3c63-a3d4-7375fa04639a	http://rdf.freebase.com/ns/m.058w5
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1d8731c8-7ba7-3c63-a3d4-7375fa04639a	http://rdf.freebase.com/ns/en.michelangelo
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1d8731c8-7ba7-3c63-a3d4-7375fa04639a	http://data.nytimes.com/N12046271348993268513
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/1d8731c8-7ba7-3c63-a3d4-7375fa04639a	http://data.nytimes.com/michelangelo_buonarroti_per
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/2e97fc55-7977-3c20-ae82-70d47da96837	http://rdf.freebase.com/ns/m.060_7
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/2e97fc55-7977-3c20-ae82-70d47da96837	http://rdf.freebase.com/ns/en.pablo_picasso
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/2e97fc55-7977-3c20-ae82-70d47da96837	http://data.nytimes.com/N855344257183137093
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/2e97fc55-7977-3c20-ae82-70d47da96837	http://data.nytimes.com/picasso_pablo_per
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/2e97fc55-7977-3c20-ae82-70d47da96837	http://dbpedia.org/resource/Pablo_Picasso
<u>http://</u>	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/2e97fc55-7977-3c20-ae82-70d47da96837	http://dbpedia.org/resource/List_of_Picasso_artworks_1911%E2%80%9320
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/30f1387a-694e-33e6-a7fa-1c1c5957c783	http://rdf.freebase.com/ns/m.03j90
<u>http://</u>	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/30f1387a-694e-33e6-a7fa-1c1c5957c783	http://rdf.freebase.com/ns/en.hans_christian_andersen
<u>http://</u>	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/30f1387a-694e-33e6-a7fa-1c1c5957c783	http://data.nytimes.com/N48129769459947861233
<u>http://</u>	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/30f1387a-694e-33e6-a7fa-1c1c5957c783	http://data.nytimes.com/andersen_hans_christian_per
<u>http://</u>	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/34dbc6d1-41e8-3aff-b7a9-e57b69d73b79	http://data.nytimes.com/52199114612406425293
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/34dbc6d1-41e8-3aff-b7a9-e57b69d73b79	http://data.nytimes.com/viola_bill_per
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/34dbc6d1-41e8-3aff-b7a9-e57b69d73b79	http://rdf.freebase.com/ns/m.09lh9
http://	//aliada.scanbit.net:8891/id/collections/library/bib/E21_Person/34dbc6d1-41e8-3aff-b7a9-e57b69d73b79	http://rdf.freebase.com/ns/en.bill_viola



Advanced URI de-referencing

URI regulation : <u>http://www.w3.org/TR/cooluris/</u>

- 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



• URI name convention of ALIADA: URI structure

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

data.szepmuveszeti.hu/id/collections/museum/E18_Physical_Thing/sz epmuveszeti.hu_object_29

acon



• 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



Dataset Default Web Page created by ALIADA

Art Jum Aliada datase	
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: <u>http://aliada.scanbit.net:8891/doc/collections/library/bib</u> Library Authority data: <u>http://aliada.scanbit.net:8891/doc/collections/library/auth</u> Museum data: <u>http://aliada.scanbit.net:8891/doc/collections/museum</u> DC data: <u>http://aliada.scanbit.net:8891/doc/collections/dc</u>



Publication in CKAN Datahub

aliada-scanbit-net	👍 Conjunto de datos 🗿 Flujo d	de Actividad 🛛 🔚 Relacionados	📌 Ed
Seguidores 0	aliada-scanbit-net		
• Seguir	Open linked data from the Library and	Museum of ARTIUM	
U Sogui	Datos y Recursos		
🗒 Organización	Sparql Endpoint SPARQL endpoint for aliada-scant	Más información	Ir al recurs
Art	Dataset dump of Library Bibl Dataset dump in N-Triples format.	iographic data in Más información	Ir al recurs
ium	Dataset dump of links to exte	mul defeasts of	
ARTIUM	Dataset dump in N-Triples format.	Más información	Ir al recurs
Basque Museum-Center of	aliadascanbitnet8891.ttl		
Contemporary Art leer más	Void file describing the main featu	res of the aliada-scanbit-n	Ir al recurs
C Social	Example resource of LIDO	Más información	Ir al recurs
Social Google+	Example resource of LIDO in RDF/	XML format	Lo ir ai recurs
Twitter	Example resource of DC Example resource of DC in RDF/XI	ML format	Ir al recurs
Facebook	Example resource of MARC Example resource of MARC BIB in		Ir al recurs
Licencia Creative Commons CCZero	Example resource of MARC		Ir al recurs
	RDF Schema RDF Schema	Más información	C Ir al recurs
	format-foaf format-frbroo format-wgs84 license-metadat provenance-metadata publicat		o cab-mappings
	Campo	Valor	
	Fuente	http://biblioteca.artium.org	
	Autor	Aliada Consortium	

ALIADA, integration with the ILS

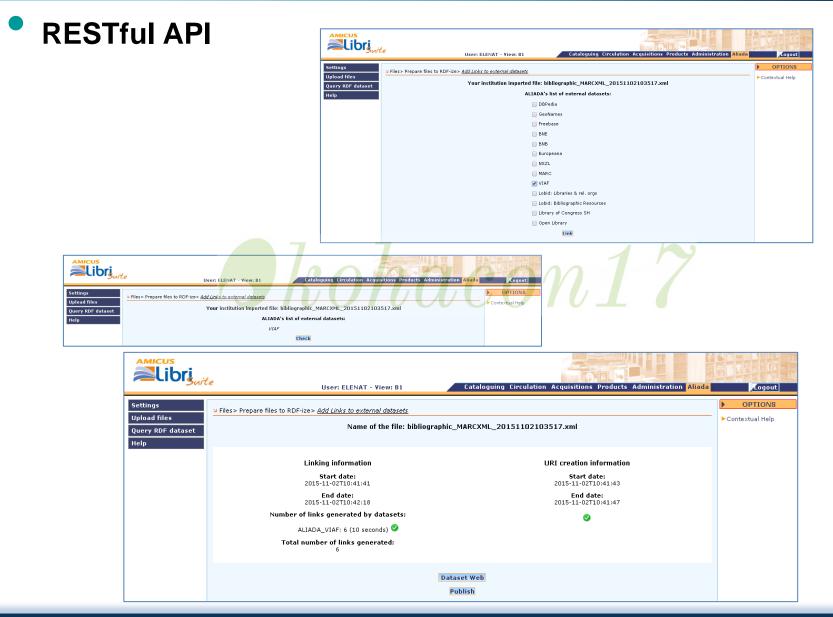


RESTful API

User: ELENAT - V	e iew: B1		Cataloguing Circulation Acquisitions Products Administration Alia	da Cogour		
Authority Bibliographic Search > Create new record > Bibliographic Models Register book	AW m	ch results (Bibliogra ialagon 10 • results Sort malagon	Search	Create bibliog Create b	raphy raphy 	
Serials Active queries Reports Help	1-5/		1 1	Settings Upload files Upload files Query RDF dataset Help	User: ELENAT - View: B1 Lataloguing Linculation Acquisitions Products Kaministration Audoa	Acgout OPTIONS itextual Help
		Physical description: Publisher:	pasemos al siguiente estado de ánimo, un creciente y cada vez más generalizado cabreo ante la realidad que, debido a algunos, tenemos muchos. (Fuente: Planeta de Libros) BCD4 134 p. : il. ; 20 cm Barcelona : Grupo Planeta, 2013 Amicus LibriSuite 2.2.8.2	Settings Upload files Query RDF dataset Help	J Files> Conversion to RDF	Cogout OPTIONS Itextual Help
				Settings Upload files Query RDF dataset Help	Le User: ELENAT - View: B1 Cataloguing Circulation Acquisitions Products Administration Aliada	Conout OPTIONS ontextual Help

ALIADA, integration with the ILS







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



ALIADA open source community



ALIAD.	A		0	
Home About Publications	Community Do	ownload News Contact N	Aember Area	FAQ
🗱 datahub	. Datase	ets Organizations About Blog Help	Search	Q.
Inking Open Data Cloud This group catalogs data sets that are available on the Web as Linked Data and contain data links pointing at other Linked Data sets. The descriptions of the data sets	Datasets A Datasets Add Dataset Soarch datasets 212 datasets VIAF: The Virtual Internation VIAF (Virtual Internation	the Linking O Cloud using A found or international Authority File	data in pen Dat LIADA der by: Reisvarce	17
ALIADA: Automatic publication und ALIADA (ally in Spanish, female genre) will au by different Library or Collection Managemer ALIADA will support the whole life cycle of re libraries involved in the consortium, providing and linking of datasets in the Linked Data Clo institutions that own datasets managed by lib	tomatize the publicat It Software. use of multilingual op g a usable and open s ud by the ALIADA use	tion in the Linked Open Data cloud of dat pen data from public bodies, initially the source tool that automatize the seleccion ers: IT staff, documentalist, curators and	tasets hosted museums and n, publication	Search ALIADA's website: Get Involved Username * Password *

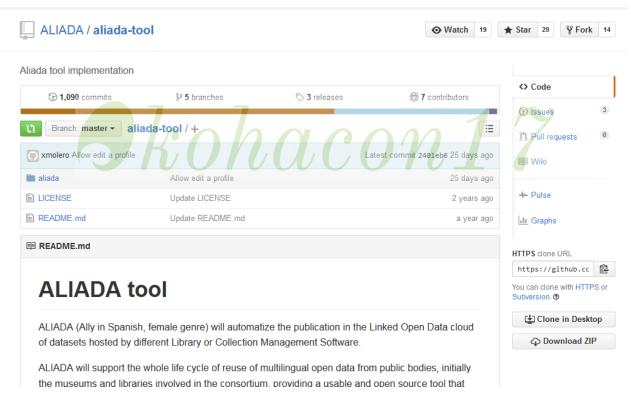
ALIADA will be an open source plugin for the library or collection management software, initially for the ones developed

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

Create new account



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/



What's next?

- Add BIBFRAME
- Add schema.org
- 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

