The Heads of Cataloging Departments Interest Group
Midwinter ALA Denver

Monday, February 12th 2018

SHARE-VDE and BIBFRAME
Linked Data in real life

casalinilibri

3cult

Phase 1 and phase 2 introduction

SHARE Virtual Discovery Environment project

The main goal of this Research & Development library community driven project has been to experiment BIBFRAME with a critical mass of data (in heterogeneity and volumes).

Phase 1: analysis, enrichment, reconciliation, conversion in RDF, publication of two sets of bibliographic data for each participant library were foreseen (1985 and 2015 imprint titles). This phase also contained the release of Marc records enriched with URIs and BIBFRAME 1.0 datasets for each participant library.

A total of 2,249,387 bibliographical records and 3,601,327 authority records were converted into BIBFRAME 1.0 and published on the SHARE-VDE portal:

www.share-vde.org

Phase 1: from October 2016 to January 2017.

SHARE Virtual Discovery Environment project

Phase 2: data enrichment and conversion refinements and customization, enhanced data supply workflow experimentation, second release of the portal.

The library catalogue of each participant institution has been converted into BIBFRAME 2.0 and returned to each library (over 100 million records and consequent datasets are expected to be processed).

A relationship database that registers the relationships between entities (person, work, instances, subjects, publisher, ...) is established in order to assure a more precise identification rate of each entity to reach a higher quality of results without human intervention.

Refinements of data, e.g. for co-authors and editors, where there is a variety of ways in which they are identified in library records (Relator terms topic), is addressed.

Enrichment with URIs of the MARC records for each library's catalogue.

Phase 2: from March 2017 to February 2018.

Participants libraries (1)

Phase 1	Phase 2	(in Country/State order):
x	x	Stanford University
x	x	University California Berkeley
x	x	Yale University
x	x	Library of Congress
x	x	University of Chicago
x	x	University of Michigan Ann Arbor
x	x	Harvard University
x		Massachusetts Institute of Technology
	x	Duke University
x		Cornell University
x		Columbia University
x	x	University of Pennsylvania

Participants libraries (2)

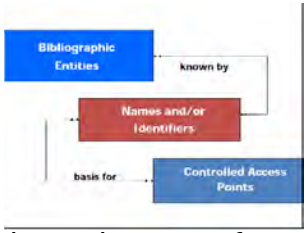
Phase 1	Phase 2	(in Country/State order):
	x	Pennsylvania State University
x	x	Texas A&M University
	x	University of Alberta
x		University of Toronto

SHARE-VDE is a Casalini Libri and @Cult partnership project, under the direction of Tiziana Possemato.

Theoretical context

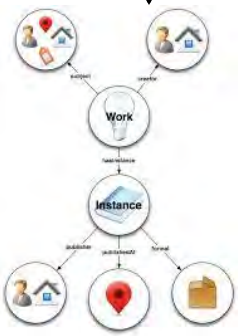
The theoretical context of the project

Where we are going…

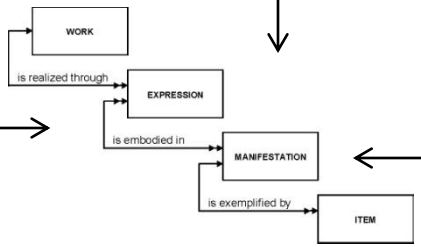


Functional Requirements for Authority Data

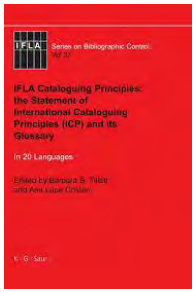
Resource Description and Access



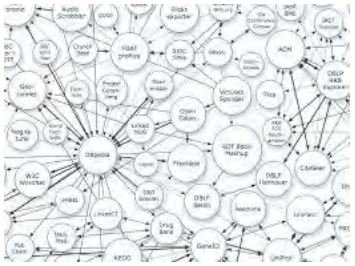
Bibframe and its extentions



Functional Requirements for Bibliographic Records
→ Library Reference Model



International Cataloguing Principles



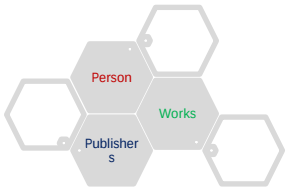
Semantic web/Linked data

SHARE-VDE process overview

The SHARE-VDE processes

OliSuite: manual process

Marc enriched/URIs



Database of relationships

Similarity's score

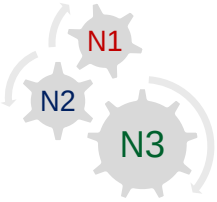
MARC
Bibliographic
Authority

XML
Schema 1
Schema 2

Entity detection

Enrichment

Reconciliation/Cluster



Knowledge base of clusters

Authify

Lodify

RDF/Bibframe dataset

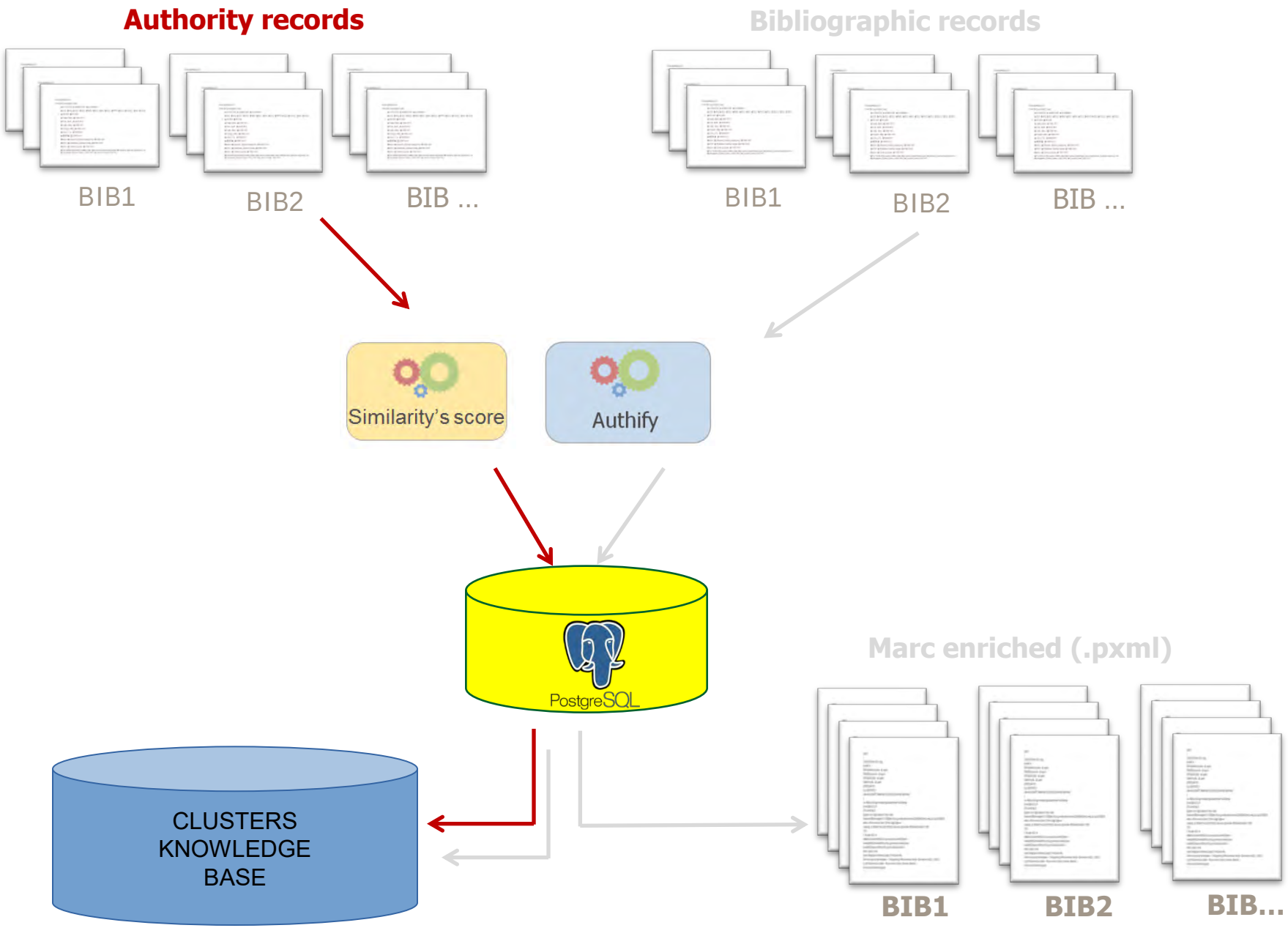
Dump db

APIs

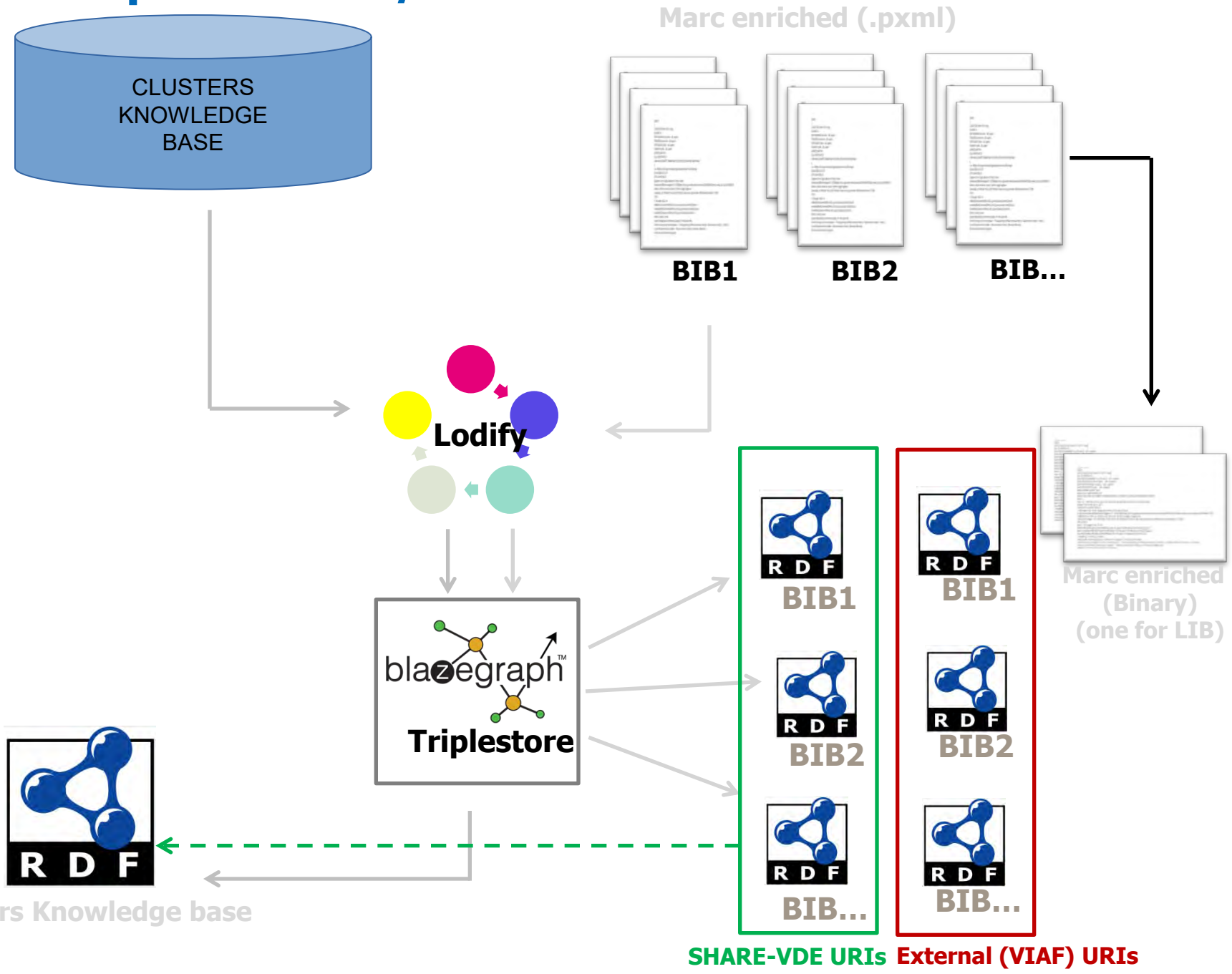
External sources

SHARE-VDE Portal

Focus on processes 1/2



Focus on processes 2/2



Entity identification

The new revolution: from record to entity

Full Record | MARC Tags

Where to Request

```
000  02384cam a2200445 a 4500
001  15789102
005  20100204124659.0
008  090624s2009 enka b 000 0 eng
010  __  |a  2009025024
015  __  |a GBA959347  |2 bnb
016  7_  |a 015291646  |2 Uk
020  __  |a 9780521519748 (hbk.)
020  __  |a 0521519748 (hbk.)
020  __  |a 9780521732505 (pbk.)
020  __  |a 0521732506 (pbk.)
035  __  |a (OCoLC)ocn416716652
040  __  |a DLC  |c DLC  |d YDXCP  |d BWKUK  |d BWK  |d FDA  |d UKM  |d BTCTA  |d CEF
                |d CDX  |d DLC
050  00  |a PR2803.A2  |b H35 2009
082  00  |a 822.3/3  |2 22
100  1_  |a Shakespeare, William,  |d 1564-1616.
245  10  |a As you like it /  |c edited by Michael Hattaway.
250  __  |a Updated ed.
260  __  |a Cambridge, UK ;  |a New York :  |b Cambridge University Press,  |c 2009.
300  __  |a xv, 240 p. :  |b ill. ;  |c 24 cm.
490  1_  |a New Cambridge Shakespeare
504  __  |a Includes bibliographical references (p. 240).
505  0_  |a Introduction. Journeys ; Plays within the play ; Theatrical genres ; Pastoral ;
         Counter-pastoral ; The condition of the country ; Politics ; 'Between you and the
         women the play must please' ; Gender ; Nuptials ; Sources ; Date and occassion ;
         Stage history ; Recent critical and stage interpretations -- Note on the text -- List of
         characters -- The play -- Textual analysis -- Appendixes : 1. An early court
         performance? ; 2. Extracts from Shakespeare's principal source, Lodge's
         Rosalind ; 3. The songs.
600  10  |a Shakespeare, William,  |d 1564-1616.  |t As you like it.
650  _0  |a Fathers and daughters  |v Drama.
650  _0  |a Exiles  |v Drama.
655  _7  |a Pastoral drama.  |2 gsafd
655  _7  |a Comedies.  |2 gsafd
700  1_  |a Hattaway, Michael.
```

As you like it [print]

As you like it [on-line]

Cambridge University Press

Cambridge Press

Cambridge Univ. Press

Shakespeare, William, 1564-1616

Шекспир, У. 1564-1616 Уильям

Saixpēr, Gouilliam, 1564-1616

As you like it

Come ti piace

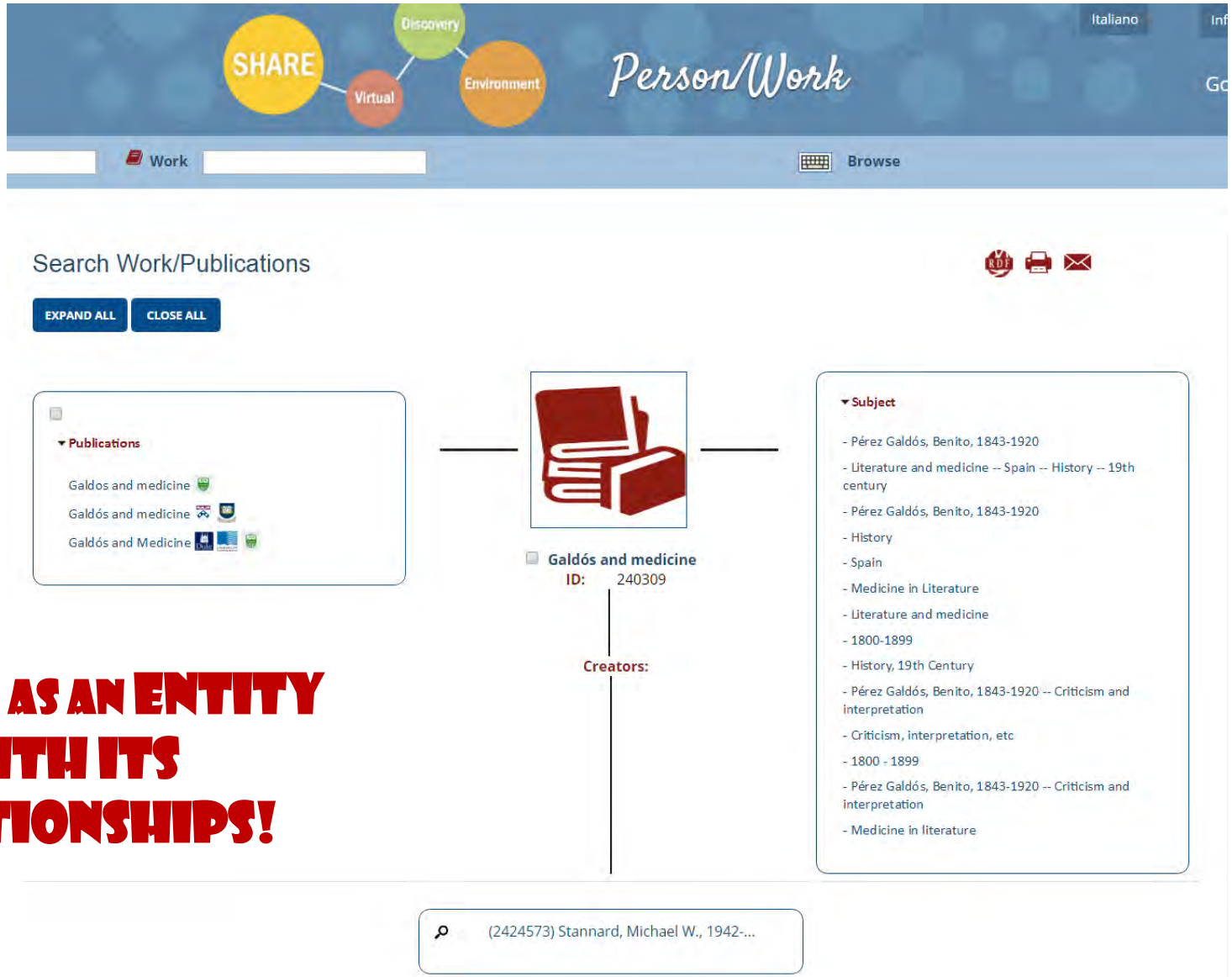
Comme il vous plaira

Fathers and daughters

Padri e figlie

Pères et filles

Different entities from the same MARC record



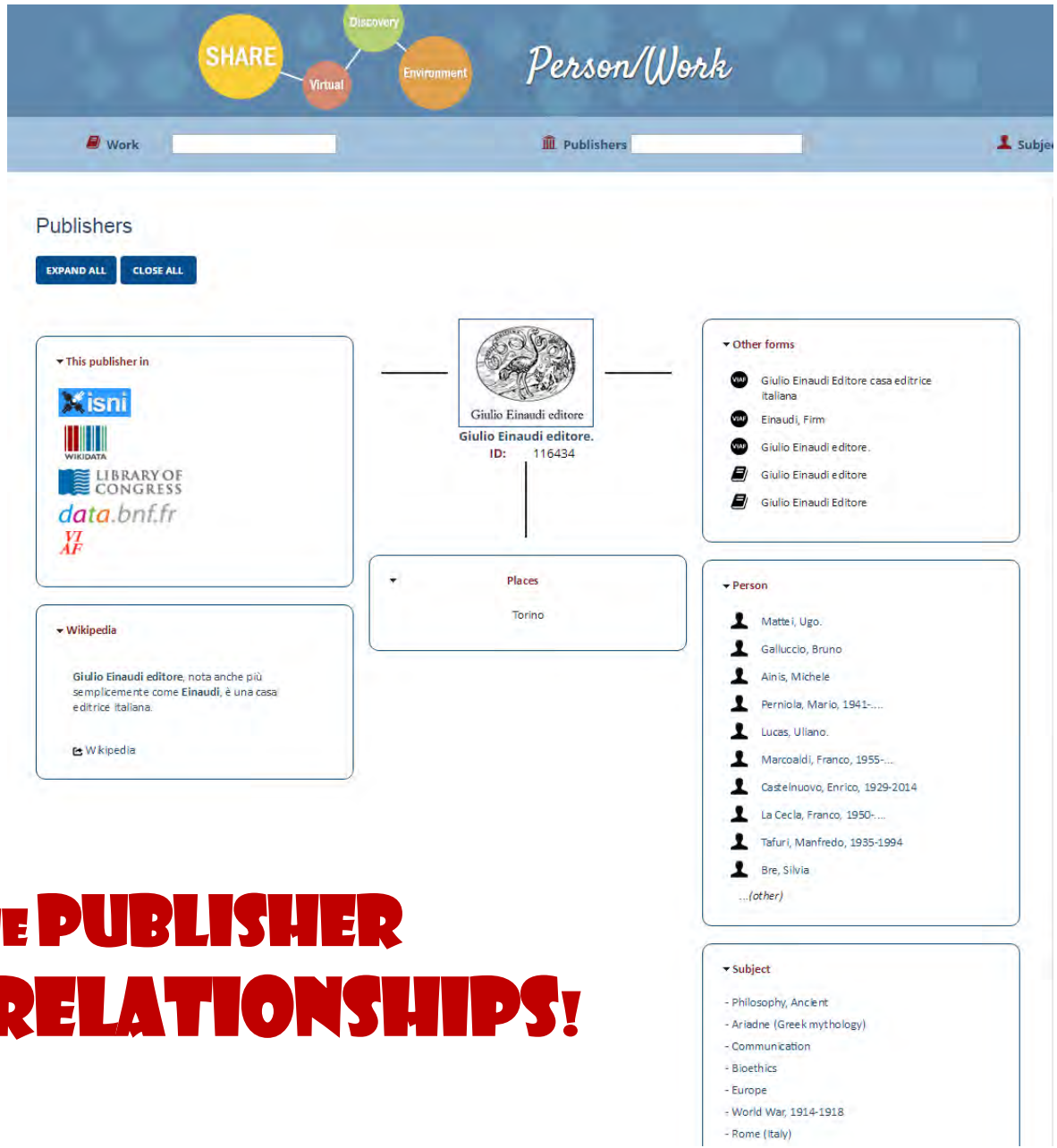
A WORK AS AN ENTITY WITH ITS RELATIONSHIPS!

http://share-vde.org/sharevde/searchTitles?t_cluster_id=240309&l=en

Different entities from the same MARC record

HERE THOMAS MANN IS THE SUBJECT OF A WORK!

Different entities from the same MARC record



THE PUBLISHER WITH ITS RELATIONSHIPS!

Entities in *cluster* : an example of collaboration and sharing

Vivaldi, Antonio, 1678-1741
ID: 37154

Questo autore in
WIKIDATA
LIBRARY OF CONGRESS
data.bnf.fr
VIAF

Altre forme del nome

- Vivaldi, Antonio, 1678-1741
- ויולד', אנטוניו, 1678-1741
- Vivaldi, Antonio, 1678-1741
- Vivaldi, Antonio
- Vivaldi, Antonio, sac., 1678-1741
- Вивальди, А. 1678-1741 Антонио
- Вивальди, Антонио, 1678-1741
- Vivaldi, Antonio, 1680-1741
- 1741-1678 ، انطونيو ، فيفالدي
- Antonio Vivaldi compositore e violinista italiano esponente di spicco del tardo barocco veneziano
- Vivaldi, Antonio, ca.1678-1741
- Vivaldi, Antonio (Italian composer and musician, 1678-1741)
- Prete rosso, 1678-1741
- Vivaldis, A., 1678-1741
- Vivaldi, A. (Antonio), 1678-1741
- Vivarudi, Antonio, 1678-1741
- Vivaldi, Antonio

...(altre forme)

The result of a reconciliation of the entity *Antonio Vivaldi* in the Share VDE project, with data from different sources and projects:

- the authorized form from a local authority file
- the variant forms originating from the references on the local authority records
- the variant forms originating from the VIAF
- the forms of the name used in the bibliographic records.

The cluster is completed and enriched with identifiers for the same entity, Antonio Vivaldi, from sources such as:

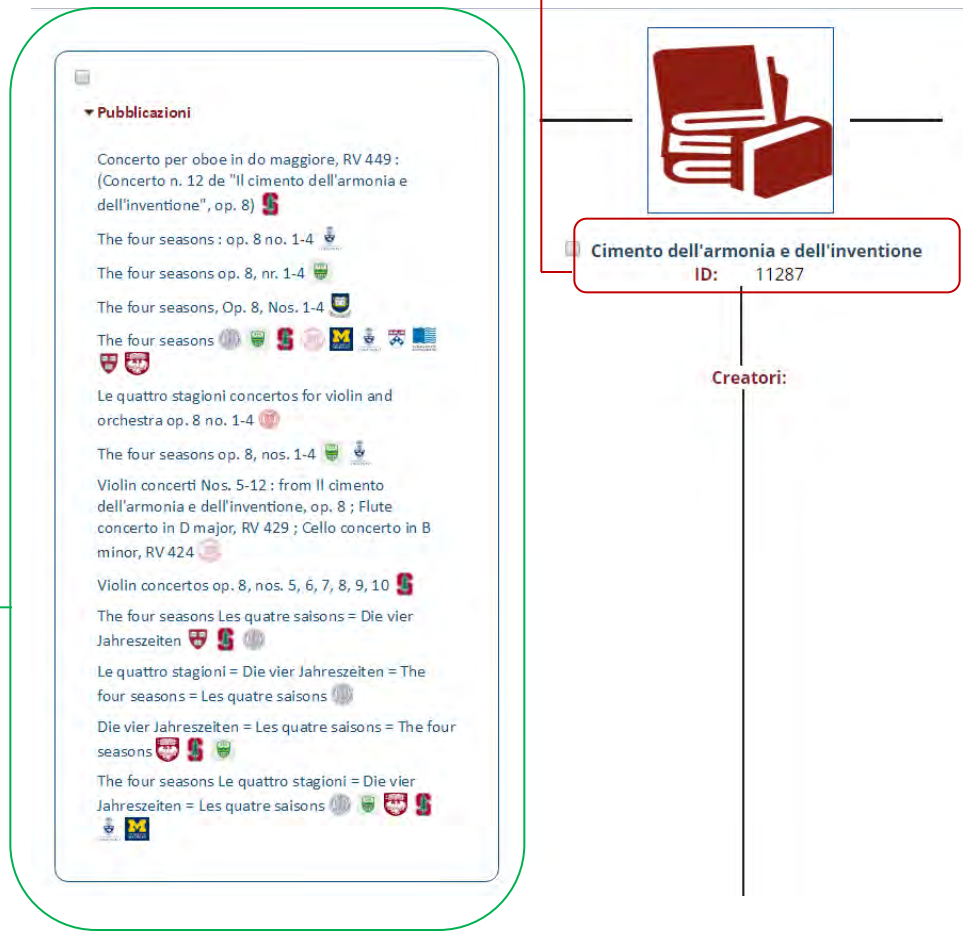
- Wikidata
- Library of Congress Name Authority File
- Data.bnf.fr
- VIAF

An example of Work/Instances reconciliation

Grouping under a single work title of the many publication titles in the catalogue for *Cimento dell'armonia e dell'inventione*

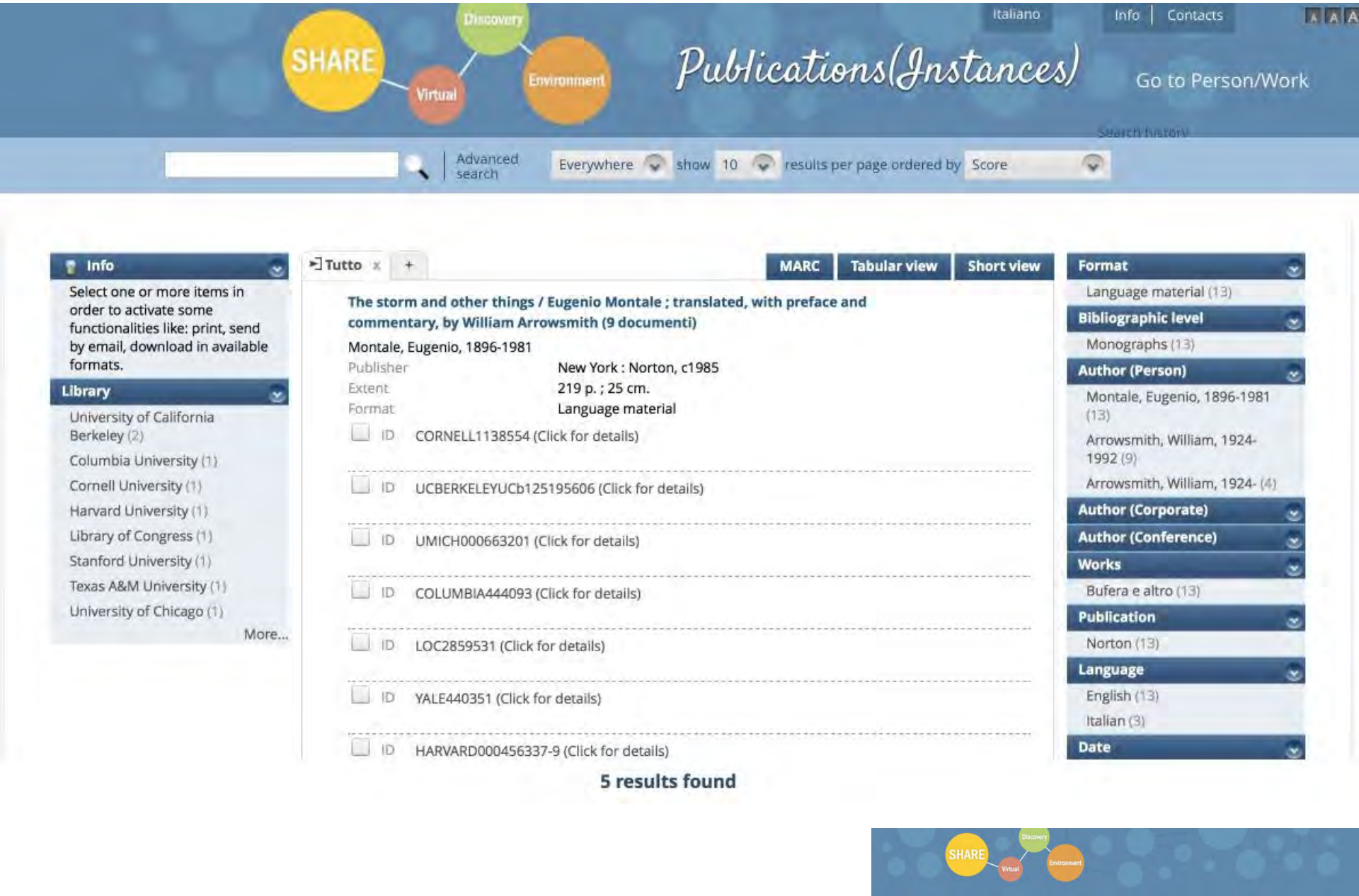
Single work title

Brings together different publications/resources present in different catalogues.



http://share-vde.org/sharevde/searchTitles?t_cluster_id=11287

Example of same Instances present in different libraries



Reconciliation & Enrichment
Automatic procedures

How reconciliation is obtained

Data reconciliation and enrichment is obtained by:

- *automated processes*
- *manual processes*

It is important to underline how the relationship between the reconciliation and validation of the results can differ profoundly between the automated and manual processes:

- automated processes: a high-level of reconciliation and clustering; a low-level of results validation;
- manual processes: a low-level of reconciliation and clustering; a high-level of results validation.

Authify – Cluster search services

Authify is a RESTFul module that offers several search and detection services, configurable also for those sources where the public Web API have limitations and it's necessary to download the whole dataset.

The Authify cluster search services provides, as the name suggest, a full-text search service among names and works clusters. The search Web API uses, behind the scenes, an "invisible queries" approach in order to (try to) find a match, as much precise as possible, within the managed clusters.

The invisible queries approach allows to make everything transparent to the caller: on top of a single search request, the system executes a chain of different search strategies with different priority, and the first match that produces a result will populate the response that will be returned.

For debugging purposes, the response will also include the matching strategy that produced the results.

Authify – Relator term detection

Another service which has been added to Authify is the so called "**Relator term detection**".

Starting from a MARC record (whatever is the specific dialect) the system analyses all (configured) tags that contain a name and, for each of them, tries to figure out (using the statements of responsibility of the input record) what is the corresponding role within the work represented by the given record.

Behind a simple token matching analysis, there is a more complicated logic that tries (using, among other things, the search services described in the previous point) to find the role of each found name using its variant forms or using a set of tokens that could identify such role (e.g. edited by, by, illustrated by).

A catch-all role is used when no valuable information can be desumed from the analysis.

Reconciliation & Enrichment
Manual procedures

PCC directives

PCC identifies and addresses policy issues on the use of identifiers in MARC:

- developing guidelines to include identifiers in MARC bibliographic and authority records
- the use of multiple identifiers for the same entity
- determining the entities for which identifiers should be provided in an initial implementation
- identifying automated methods for populating and maintaining new and existing records with identifiers

The importance of identification and detection in the Semantic Web

Key elements of the cataloguing workflow:

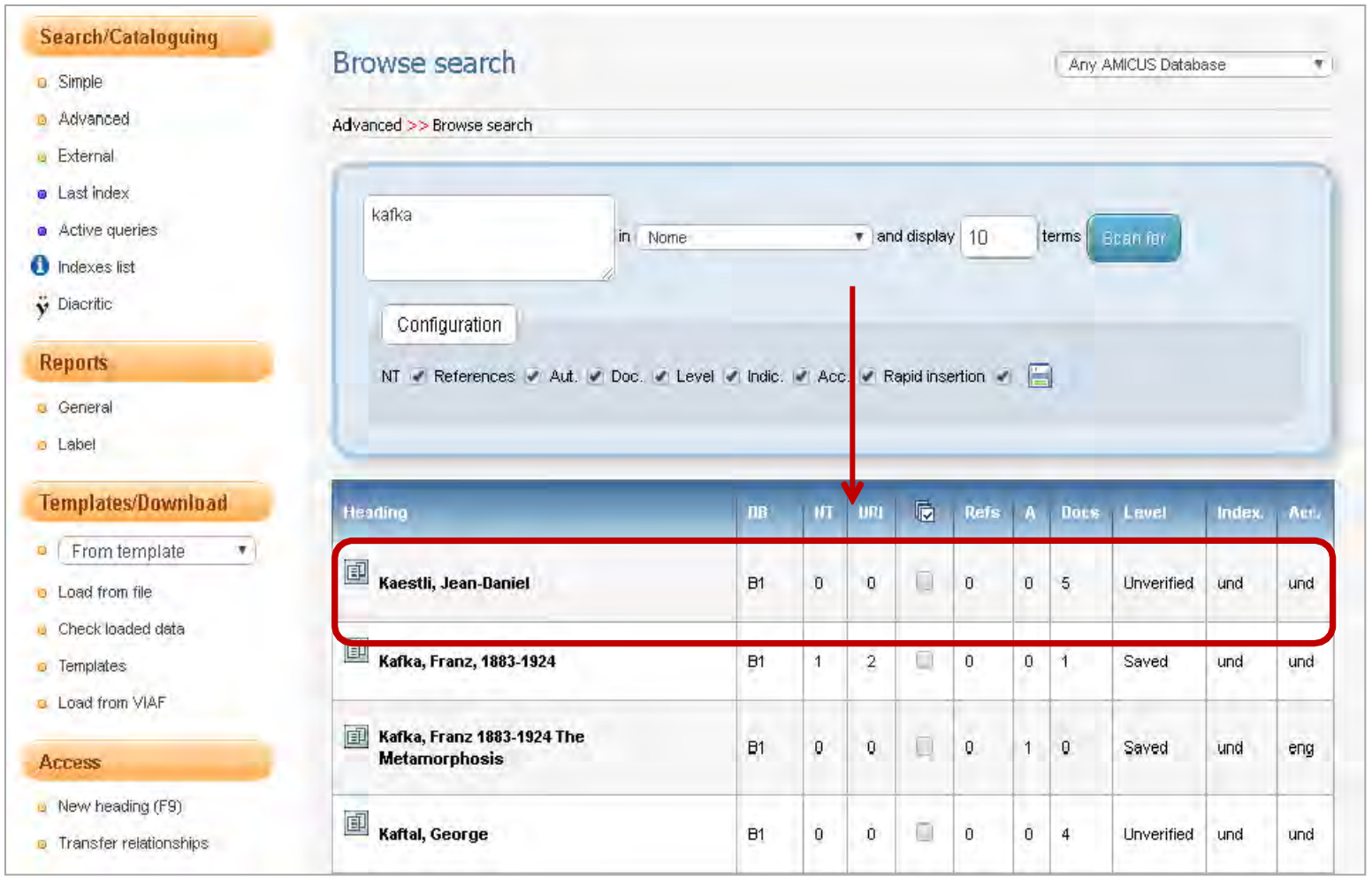
- entity identification
- reconciliation

To enrich a MARC record with URIs Casalini Libri uses the “URI Management System” (included in the OLISuite cataloguing module within WeCat developed by @Cult) that allows also the management of multiple identifiers for each access point or heading.

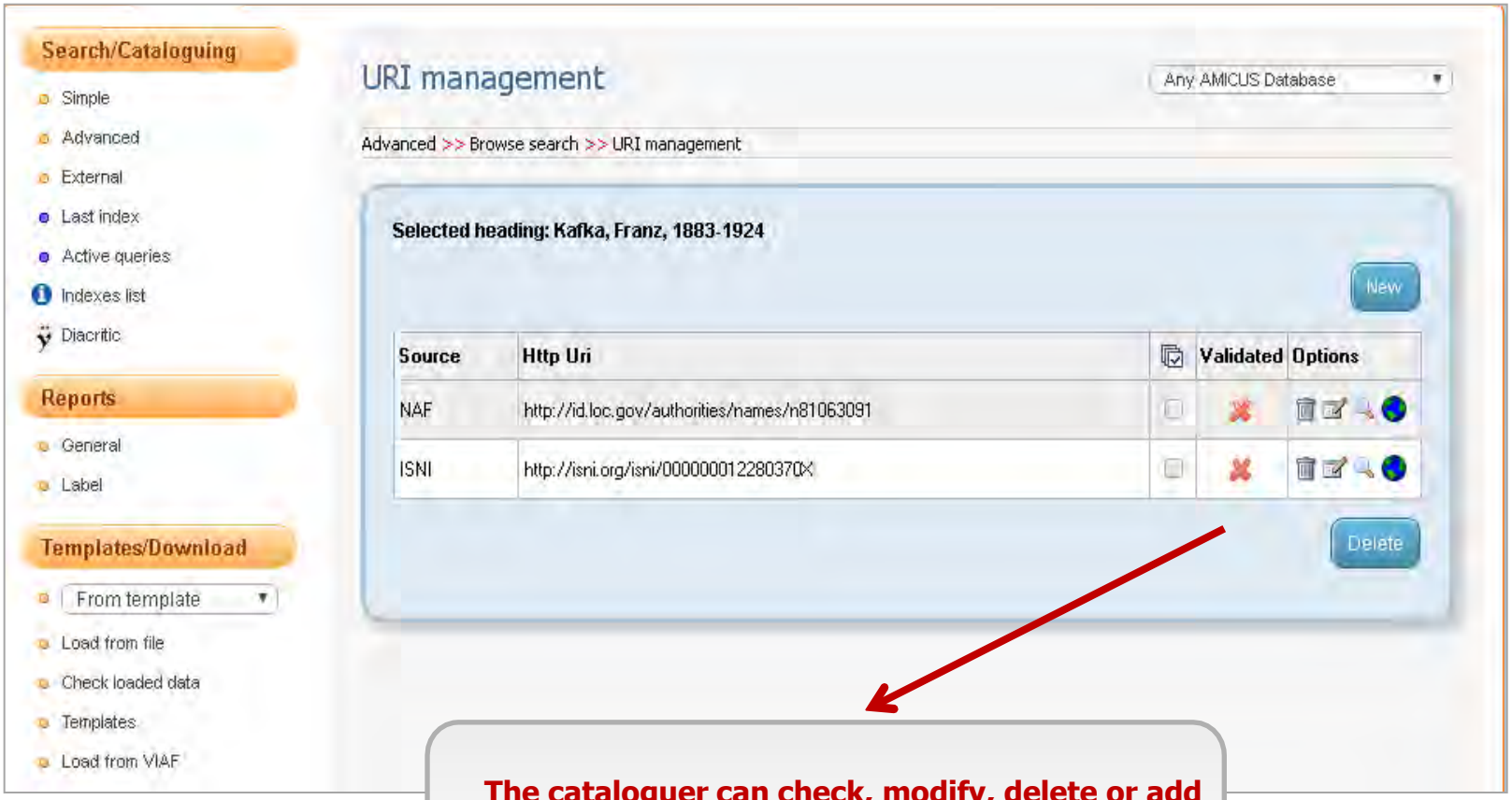
This simplifies the reconciliation of varying forms of the same entity with the development of detection procedures for entity identification and the conversion to BIBFRAME.

URI Management System in OLISuite

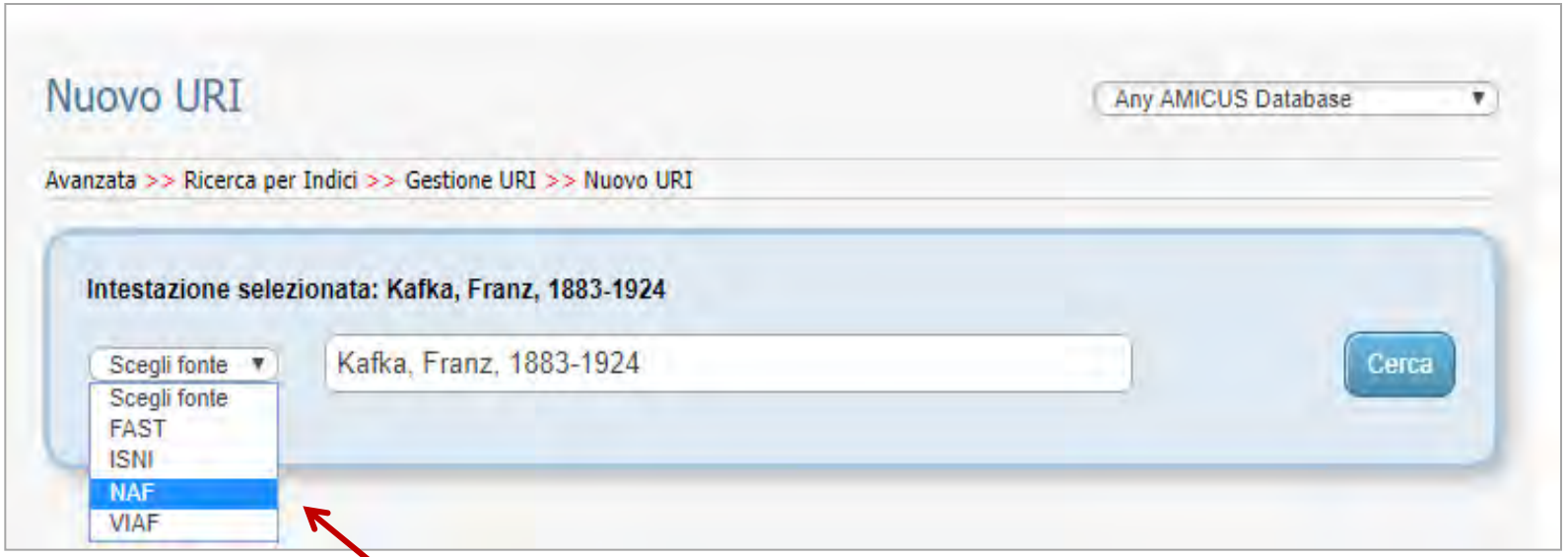
This component is integrated in Casalini Libri's original cataloguing workflow in order to have the option of supplying MARC enriched records and the equivalent BIBFRAME datasets.



URI Management System in OLISuite

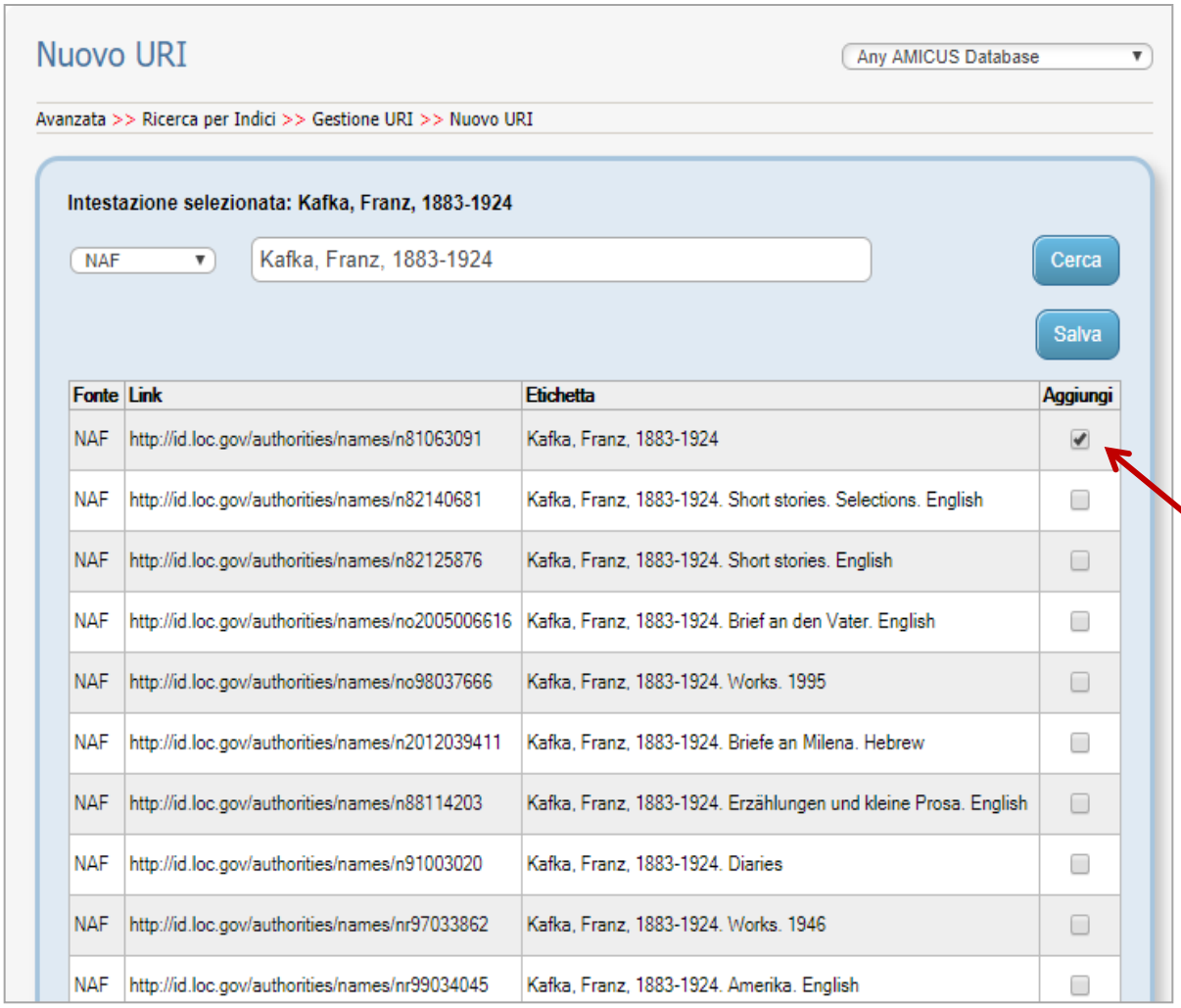


Adding new URIs to a heading in OLISuite



From this drop down menu the cataloguer can choose the desired source and start the URI search

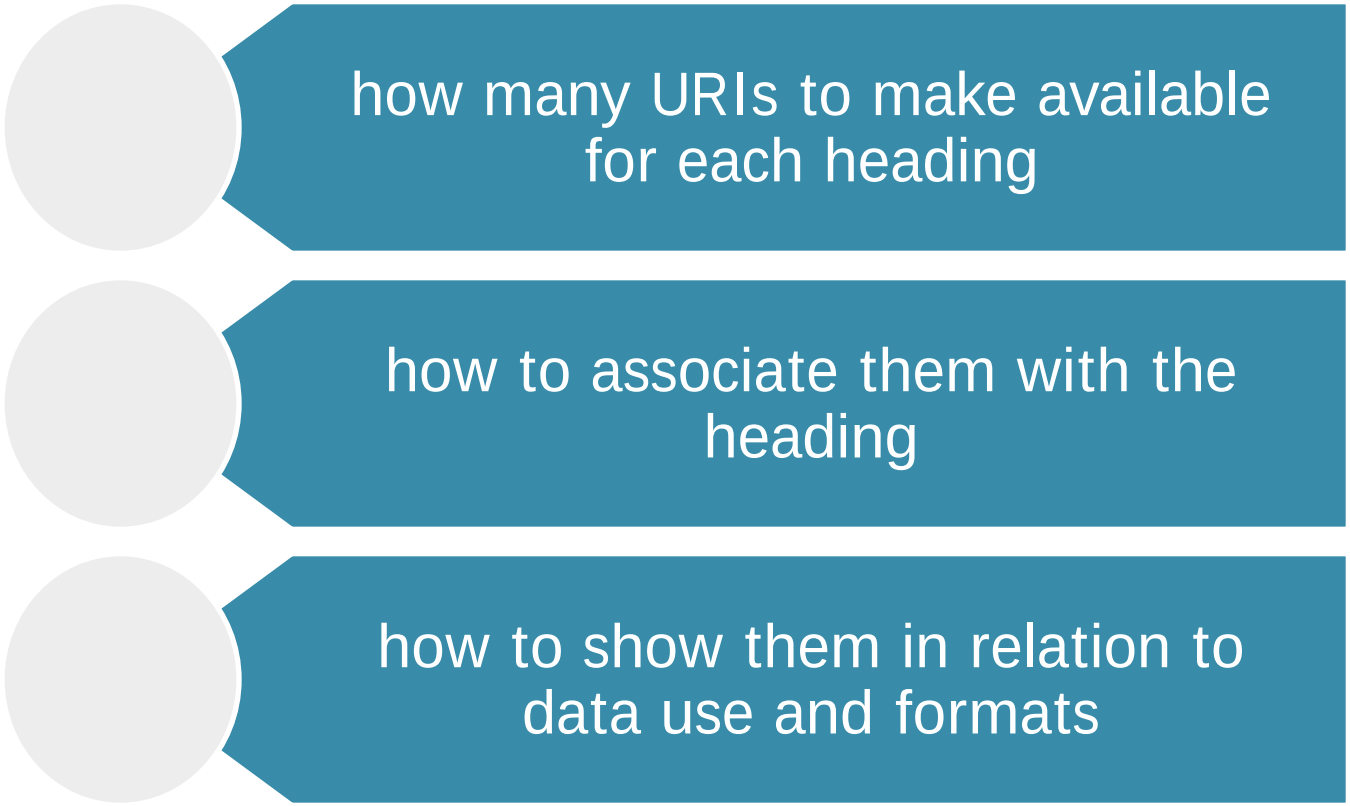
Adding new URI to a heading OLISuite



From the search result window choose the desired URI and SAVE

Access points and URIs

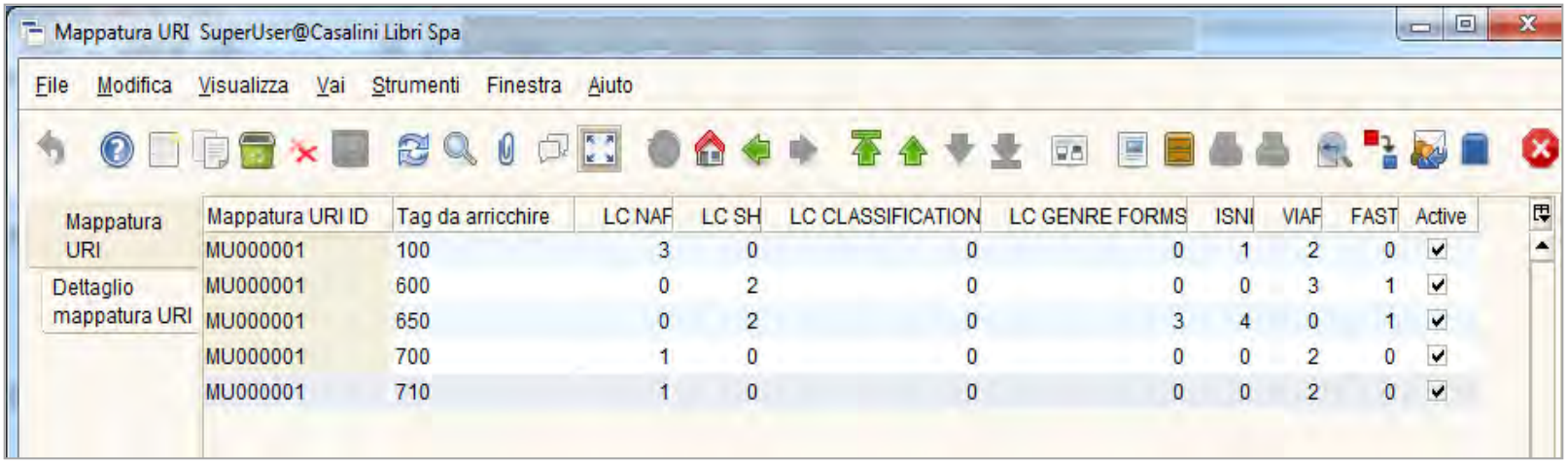
Saving the different URIs in a database table, allow them to be used in various ways, selected during the data export/conversion:



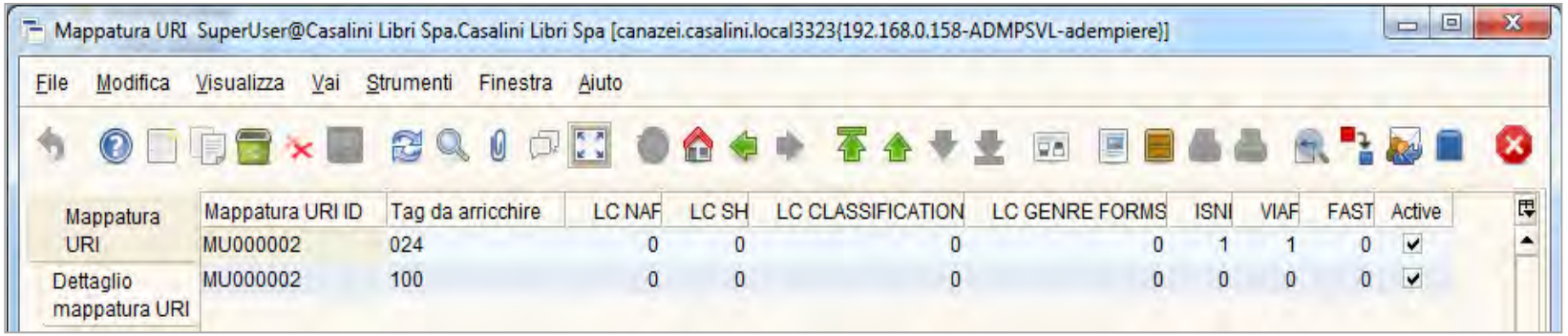
Different library profiles that were previously defined in the ERP Adempiere are considered.

URIs/tag mapping

URIs/tag mapping for Library A



URIs/tag mapping for Library B



URI history => URI Registry

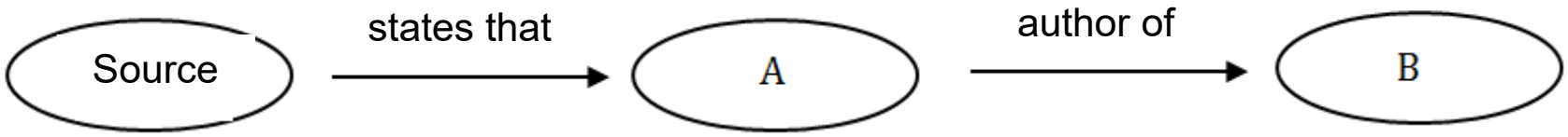
The reorganisation of a cluster can modify its original content, so we need to save the relevant cluster updates in a URI Registry.

The URI Registry could keep information such as (but not limited to):

- the resources added to the cluster, but also modified or removed from it
- the date of the update
- the particular operation performed
- the status of an URI (for instance valid or invalid)
- the URI Aliases

Guarantee of authority and quality in the new LOD environment

The source or *provenance*, which, in turn, must be constructed with reference to specific ontologies, providing the classes, properties and restrictions needed for identifying it, becomes the *fourth element* added to every triple (assertion) to certify its validity, transforming the triple into a quadruple.

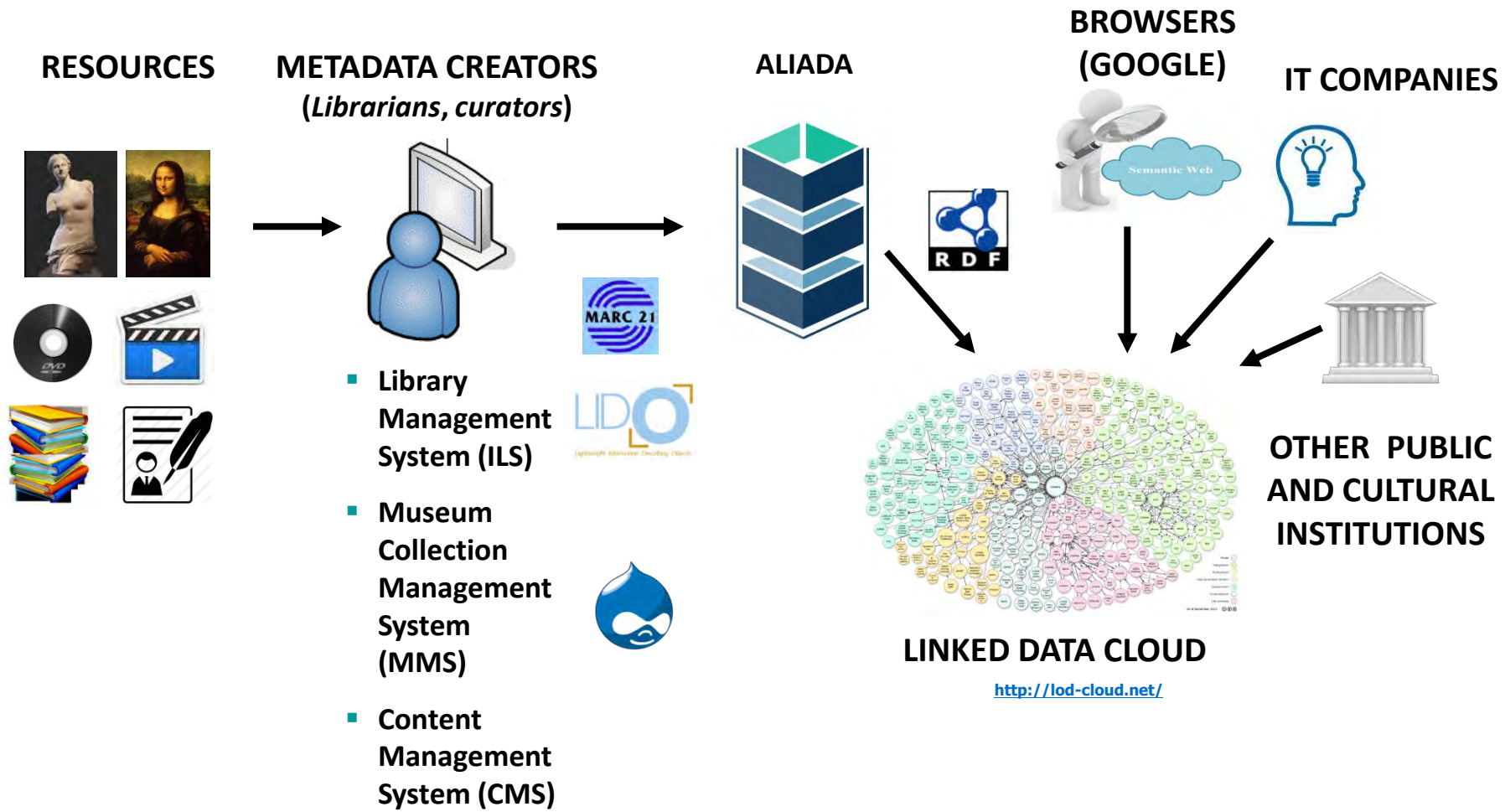


Stating the *provenance* of a piece of information is an essential element for increasing the trust that can be placed in data, and facilitating its use and sharing by end users or by the institutions choosing to cooperate in this way.

Conversion in RDF/BIBFRAME

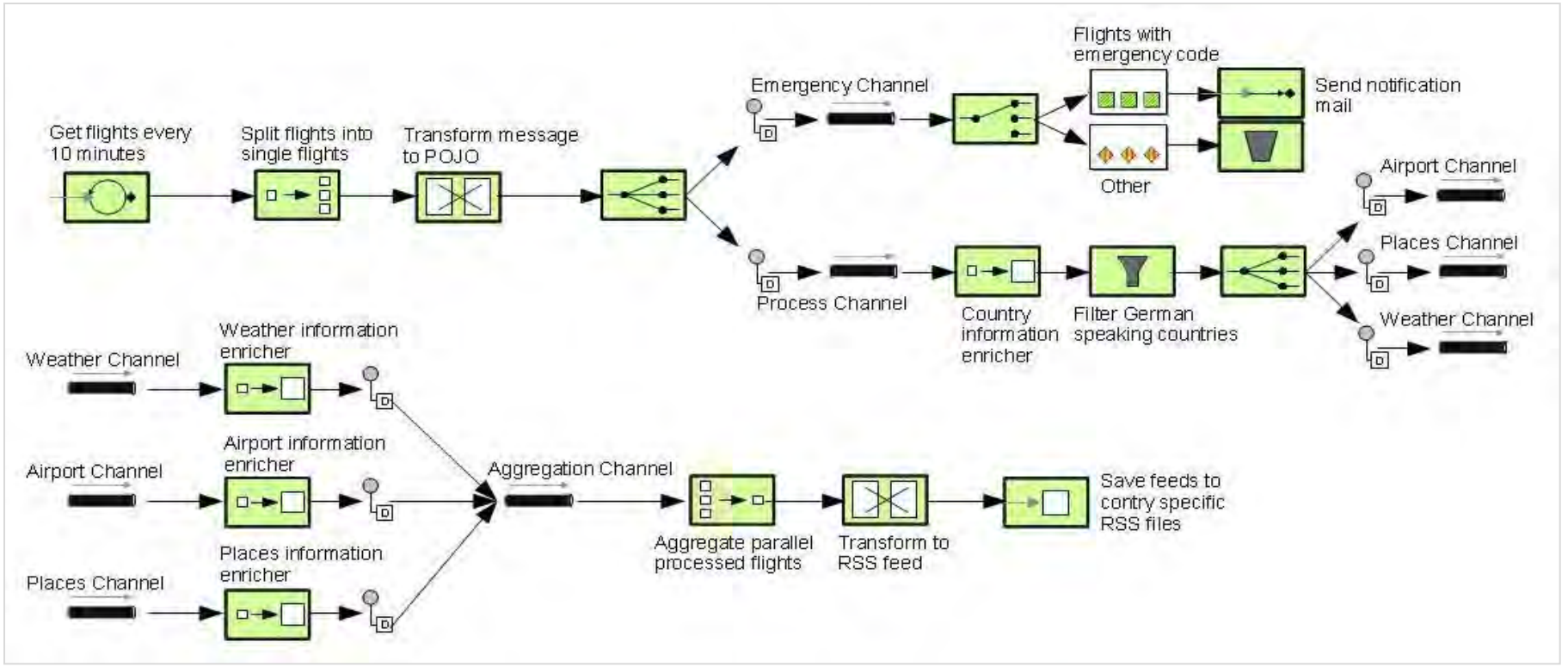
Lodify: the evolution of Aliada for BIBFRAME conversion

The conversion process from any format to RDF



Lodify - The asynchronous pipeline

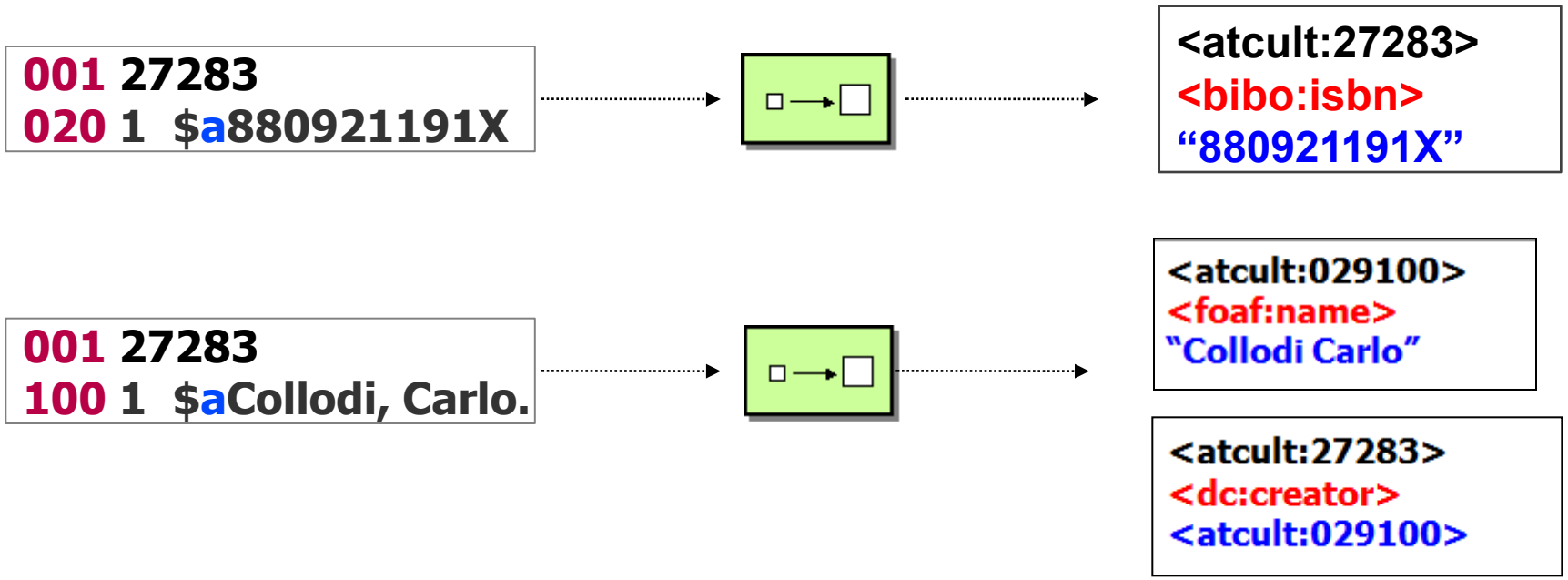
Lodify building block, realized through Apache Camel. The process is split into atomic pieces (processors), each of these responsible for a small part of the overall task. Each processor can act as a splitter or aggregator and can achieve content manipulation on the incoming message.



Lodify - Conversion templates

Lodify converts each incoming record by means of Conversion templates.
Each template associates:

- a MARC record belonging to the incoming data-stream
- with a set of (conversion) rules associated with one or more ontologies.



BF LC vocabulary extensions: the state-of-art

Class List

AppliesTo	DemographicGroup	Relation
CreatorCharacteristic	EncodingLevel	Relationship

Property List

applicableInstitution	name10MatchKey	title00MarcKey
appliesTo	name11MarcKey	title00MatchKey
creatorCharacteristic	name11MatchKey	title10MarcKey
demographicGroup	primaryContributorName00MatchKey	title10MatchKey
encodingLevel	primaryContributorName10MatchKey	title11MarcKey
metadataLicensor	primaryContributorName11MatchKey	title11MatchKey
name00MarcKey	projectedProvisionDate	title30MarcKey
name00MatchKey	relation	title30MatchKey
name10MarcKey	relationship	title40MarcKey
		title40MatchKey
		titleSortKey

URI table for external sources

SHARE-VDE — URI_table_for_external_sources_2017-12-22.xls — 2017-12-22

CODE	SOURCE	INSTITUTION	LINK TO PROJECT PAGE	URI	URI EXAMPLE	MARC21 TAG	API/WS
NAF (Names)	Library of Congress Names	Library of Congress	http://id.loc.gov/authorities/names.ht	http://id.loc.gov/authorities/nam	http://id.loc.gov/authorities/nam	100/110/111 $0	S/M
NAF (NameTitle + Titles)	Library of Con						
LCSH	Library of Con Headings						
LC Classification	Library of Con						
LC Identifiers	Library of Con Identifiers Sch						
Geographic Areas	MARC List for						
Relators	MARC Code Li						
Languages	MARC List for						
Countries	MARC List for						
ISNI	ISNI - Internat Identifier						
VIAF (Names)	VIAF - Virtual File						

SHARE-VDE — URI_table_for_external_sources_2017-12-22.xls

SPARQL endpoint	RDF dump	LICENSE TERMS	URL SEARCH	NOTES	AUTHIFY (Yes for search URL specified)	SUBFIELD $0/$1
	Yes		http://id.loc.gov/search/?q=Shakespeare+William&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fauthorities%2Fnames&format=atom&start=1		http://${host}:${port}/authify/lcnaf/persons?q=Eco, U	$0
	Yes		http://id.loc.gov/search/?q=bible+old+testament&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fauthorities%2Fnames OR http://id.loc.gov/search/?q=niccolo+macchiavelli+arte+della+guerra&q=cs:http://id.loc.gov/authorities/names		http://${ tle=Bible http://${ niccolo% guerra	
	Yes		http://id.loc.gov/search/?q=poetry&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fauthorities%2Fsubjects&format=atom&start=1		http://${ q=Diarie	
	NO		http://id.loc.gov/search/?q=poetry&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fauthorities%2Fclassification&format=atom&start=1		NO	
	NO		http://id.loc.gov/search/?q=viaf&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fvocabulary%2Fidentifiers&format=atom&start=1			
			http://id.loc.gov/search/?q=italy&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fvocabulary%2FgeographicAreas&format=atom&start=1	It can be added as a fix value "http://id.loc.gov/vocabulary/geographicAreas/" + $a. For example: http://id.loc.gov/vocabulary/geographicAreas/a-ko	NO	
			http://id.loc.gov/search/?q=creator&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fvocabulary%2Frelators&format=atom&start=1	Addition of URI in export process: creation of a table that searches the corresponding URI from the $e field. - The /detect service exist, it retrieves the relator code from the responsability expressed in the 245 $c tag.	/detect	
			http://id.loc.gov/search/?q=italian&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fvocabulary%2Flanguages&format=atom&start=1		NO	
			http://id.loc.gov/search/?q=italy&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fvocabulary%2Fcountries&format=atom&start=1		NO	
			http://isni.oclc.nl/sru/?query=pica.na%3D%22Shakespeare%2C%20William%22&operation=searchRetrieve&recordSchema=isni-b&maximumRecords=3&recordPacking=xml	URI implementation with the available API. Massive usage with authentication.	NO	
	Yes		http://viaf.org/viaf/search?query=local.personalNames=+%22Shakespeare%20William%22&sortKeys=holdingscount&maximumRecords=2&httpAccept=application/xml		http://${ =manzor alessand	

SHARE-VDE

URI
http://id.loc.gov/authorities/names/n79021285
http://id.loc.gov/authorities/names/n80106714 http://id.loc.gov/authorities/names/nr00019197
http://id.loc.gov/authorities/subjects/sh85037601
http://id.loc.gov/vocabulary/geographicAreas/a-ko
http://id.loc.gov/vocabulary/relators/aut
http://isni.org/isni/0000000120981297
http://viaf.org/viaf/14356

SHARE-VDE phase 2 deliverables

Phase 2 deliverables overview

Linked

D1 the catalogue of each library converted into BIBFRAME 2.0 format

Entities are reconciled in the dataset and linked to SHARE-VDE project URIs of D2 for identification.

D2 the SHARE-VDE project Knowledge base of clusters in RDF format

Is common for all institutions as it includes data from all of the participants. Entities in D2 are enriched with URIs from external sources. All variant forms are included.

D3 the dataset converted in BIBFRAME 2.0 with external URIs included

This dataset includes a certain number of relationships already present in the knowledge base. Works autonomously from D2.

D4 the MARC21 version of D3

It includes all of the institution's records enriched with URIs.

The Heads of Cataloging Departments Interest Group
Midwinter ALA Denver

Monday, February 12th, 2018

Thank you!

Any questions or feedback are greatly appreciated.

Michele Casalini, Managing Director
michele@casalini.it

casalinilibri

@cult