European BIBFRAME Workshop 2017

26th and 27th of September 2017
Deutsche Nationalbibliothek (DNB) in Frankfurt am Main, Germany

SHARE Virtual Discovery Environment in Linked Data

Tiziana Possemato
Director, @Cult
Chief Information Officer, Casalini Libri

Introduction
and Overall Project Goals

SHARE-VDE context

The development of this initiative started taking into account the following context:

- Libraries with different systems, habits and cataloguing rules and traditions;
- The emerging Linked Data paradigm;
- The broader GLAM (Gallery, Library, Archive and Museum) community expectations;
- The opportunity to provide more comprehensive access to researchers and students;
- The consciousness of cultural and language diversities, and the large differences among the needs of disciplines.

SHARE-VDE overall goals

The overall goals of the project are:

- To reach through short phases concrete results on which the library community can base further steps and decisions;

- To discuss, experiment and configure, together with the library community, the options for the future data creation, enhancement and sharing of all type of resources;

- To achieve a granularity and richness of relationships within library data and beyond, which is often hidden or unexpressed in a traditional catalogue;

- To prepare an environment which is useful for both library patrons - who can take advantage of more advanced discovery interfaces -, and librarians looking for cataloguing functions in native semantic web standards together with the integration of processes with the local systems and the implementation of tools in a shared collaborative environment.

SHARE-VDE approach

Among the guideline of the initiative:

- SHARE-VDE is a Research & Development project;
- The intiative is library community driven;
- The project aim is also to identify issues and problems related to new information management processes, and propose solutions;
- The components are designed to be available individually in as flexible as possible configuration for the library community's needs;
- Scalability will be tested up to the equivalent of over 100 million traditional bibliographic and authority records.

SHARE-VDE brief project overview

Three main goals:

- Experimenting with MARC records enriched with URIs through the reconciliation processes of varying forms of the same entity;

- Conversion into BIBFRAME 2.0 datasets and distribution procedures;

- Publication on a three-layered portal following the BIBFRAME data model (person/work -- instances (publication) -- item levels)

16 participant institutions from the US and Canada.

SHARE Virtual Discovery Environment project

The project is divided into three phases. Each participant decides whether or not to take part in the subsequent phase.

Phase 1 activities included: the analysis, enrichment, reconciliation and conversion into RDF of two sets of bibliographic data for each participating library (1985 and 2015 imprint titles); the publication of the first version of the platform applying the new data model; the release of test Marc records enriched with URIs and the BIBFRAME 1.0 datasets for each participating library.

A total of 2,249,387 bibliographical records and 3,601,327 authority records were converted into BIBFRAME 1.0 and published via the first release of the SHARE-VDE portal (www.share-vde.org).

Phase 1: from October 2016 to January 2017.

SHARE Virtual Discovery Environment project

Phase 2 activities includes:

The entire library catalogue of each participating institution will be converted into BIBFRAME 2.0 and returned to each single institution.

The second version of the platform will be released focusing also on the Linked Data oriented record-less approach.

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

Refinement 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).

SHARE Virtual Discovery Environment project

... cont. (**Phase 2** activities includes):

Export of data in Marc or RDF format filtering the library preferred URIs.

Inclusion of additional URI sources, e.g. specific sources for corporate bodies, subjects (LCSH, FAST, etc…) and RDA vocabularies.

Analysis for the creation of relationships among subject terms and strings in different languages.

Provenance declaration, update management and built-in instances will be addressed.

Phase 2: from March to September 2017.

Participating 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

Participating 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

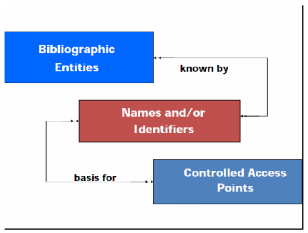
Phase 3 of SHARE-VDE, and its modular options for adoption, will also be driven by the library community

Example of possible components	Lib 1	Lib 2	Lib 3	Lib 4
Marc enrichment (URIs)	x	x		x
Dataset in RDF (BIBFRAME 2.0)			x	
Dataset in RDF (BIBFRAME 2.0) enriched with URIs	x			x
Ontologies-suite enrichment			x	x
Database of relationships	x			x
Knowledge base of clusters		x		x
URIs Registry		x		
Entity detection			x	
Data publication on SHARE-VDE portal	**x**	**x**		**x**
UC1: *Borrow Direct*	x		x	
UC2: *Borrow Direct in Franklin*		x		
UC3: *Collaborative selection tools*	x			x
UC4: *Data creation and management in RDF*			x	
UC5: *Community specific GUIs and functions*				x

Brief Theoretical Context

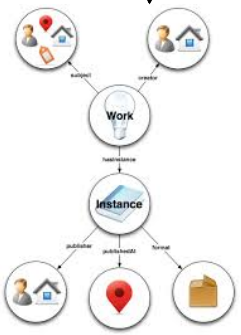
Brief theoretical context

Where we are going…

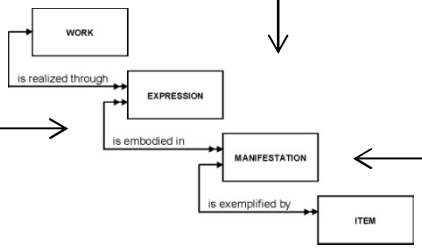


Functional Requirements for Authority Data

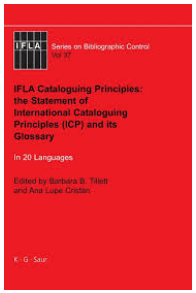
Resource Description and Access



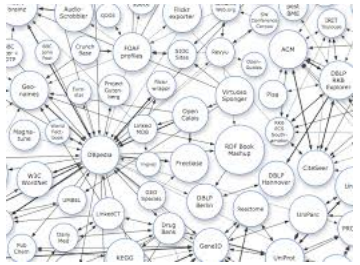
BIBFRAME



Functional Requirements for Bibliographic Records



International Cataloguing Principles



Semantic web/Linked data

Brief theoretical context

New standards, models and technologies as ways to approach entity **identification** and the **relationships** between entities, are recognized as the key elements in the construction of new entity detection and entity identification processes:

- **RDA – *Resource Description and Access,*** the new international guidelines to manage resources;

- **Linked Open Data** philosophy and technology;

- **BIBFRAME**, one of the more interesting models to convert and publish data. This model can be considered 'the core' ontology, completed with the ontologies for specific domains, that library, archive and museum communities will suggest.

RDA Toolkit: Identify and Relationship

The structure of the RDA Toolkit clearly expresses the importance given by the standard to concepts of identification and relationship:

- Section 1: Recording Attributes of Manifestations & Items

- Section 2: Recording Attributes of Works & Expressions

- Section 3: Recording Attributes of Agents

- Section 4: Recording Attributes of Concepts, Objects, Events & Places

IDENTIFY

RDA Toolkit: Identify and Relationship

- Section 5: Recording Primary Relationships between Works, Expressions, Manifestations & Items
- Section 6: Recording Relationships to Agents
- Section 7: Recording Relationships with Concepts, Objects, Events & Places
- Section 8: Recording Relationships between Works, Expressions, Manifestations & Items
- Section 9: Recording Relationships between Agents
- Section 10: Recording Relationships between Concepts, Objects, Events & Places

RELATIONSHIPS

The 4 rules for Linked Data creation

by *Sir Tim Berners-Lee*

1. Use URIs as names for things: give unique names to things;

2. Use HTTP URIs so that people can look up those names: the names assigned to things must also be machine readable;

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL): things must be self-explanatory (dereferencing);

4. Include links to other URIs so that they can discover more things: create links with other objects (any object can become the subject of a new statement).

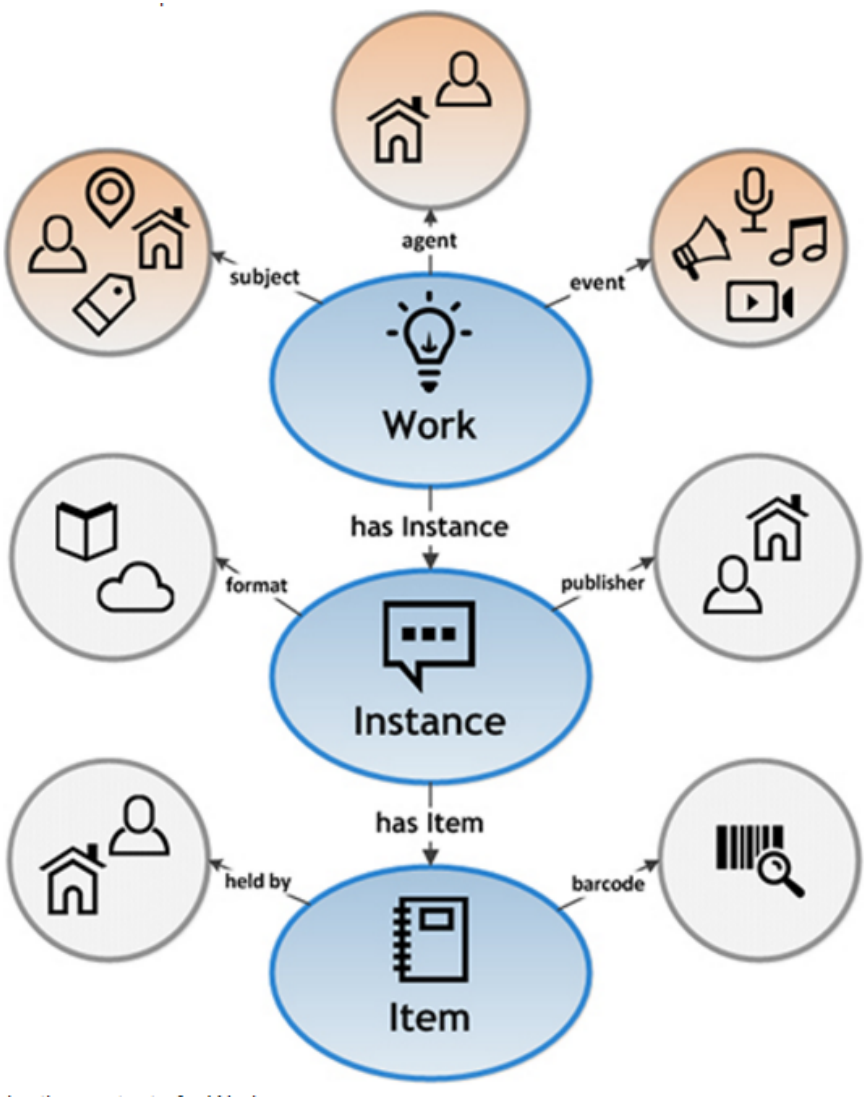
BIBFRAME – Bibliographic Framework Initiative

The **Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services** document published by the Library of Congress on November 21, 2012, sets out a new data model designed as an evolution, in linked open data, of the Marc21 format.

The reflections on the new cataloguing rules focus on some specific points, including:

- a greater level of identification and analysis of the data;
- greater attention to controlled vocabularies;
- more widespread use of terms instead of codes;
- emphasis on relationships;
- greater flexibility in controlled items.

BIBFRAME 2.0 data model



BIBFRAME 2.0 data model

"In translating the MARC 21 format to a Linked Data model it is important to deconstruct and then reconstruct the informational assets that comprise MARC". The BIBFRAME Model, version 2.0 (published on 2016, 21th of April) consists of the following core classes:

Work: The highest level of abstraction, a Work, in the BIBFRAME context, reflects the conceptual essence of the cataloged resource: authors, languages, and what it is about (subjects).

Instance: A Work may have one or more individual, material embodiments, for example, a particular published form. These are Instances of the Work. An Instance reflects information such as its publisher, place and date of publication, and format.

Item: An item is an actual copy (physical or electronic) of an Instance. It reflects information such as its location (physical or virtual), shelf mark, and barcode.

BIBFRAME 2.0 data model

BIBFRAME 2.0 further defines additional key concepts that have relationships to the core classes:

Agents: Agents are people, organizations, jurisdictions, etc., associated with a Work or Instance through roles such as author, editor, artist, photographer, composer, illustrator, etc.

Subjects: A Work might be "about" one or more concepts. Such a concept is said to be a "subject" of the Work. Concepts that may be subjects include topics, places, temporal expressions, events, works, instances, items, agents, etc.

Events: Occurrences, the recording of which may be the content of a Work.

SHARE-VDE Process Overview

The SHARE-VDE processes

Data creation
(RDF or MARC or .....)

Data enrichment (URIs)

Person
Works
Publishers

Database of relationships

MARC
Bibliographic
Authority

XML
Schema 1
Schema 2

Similarity's score

Entity detection
Authomatic enrichment
Reconciliation/Cluster

N1
N2
N3

Knowledge base of clusters

Authify

Dump db

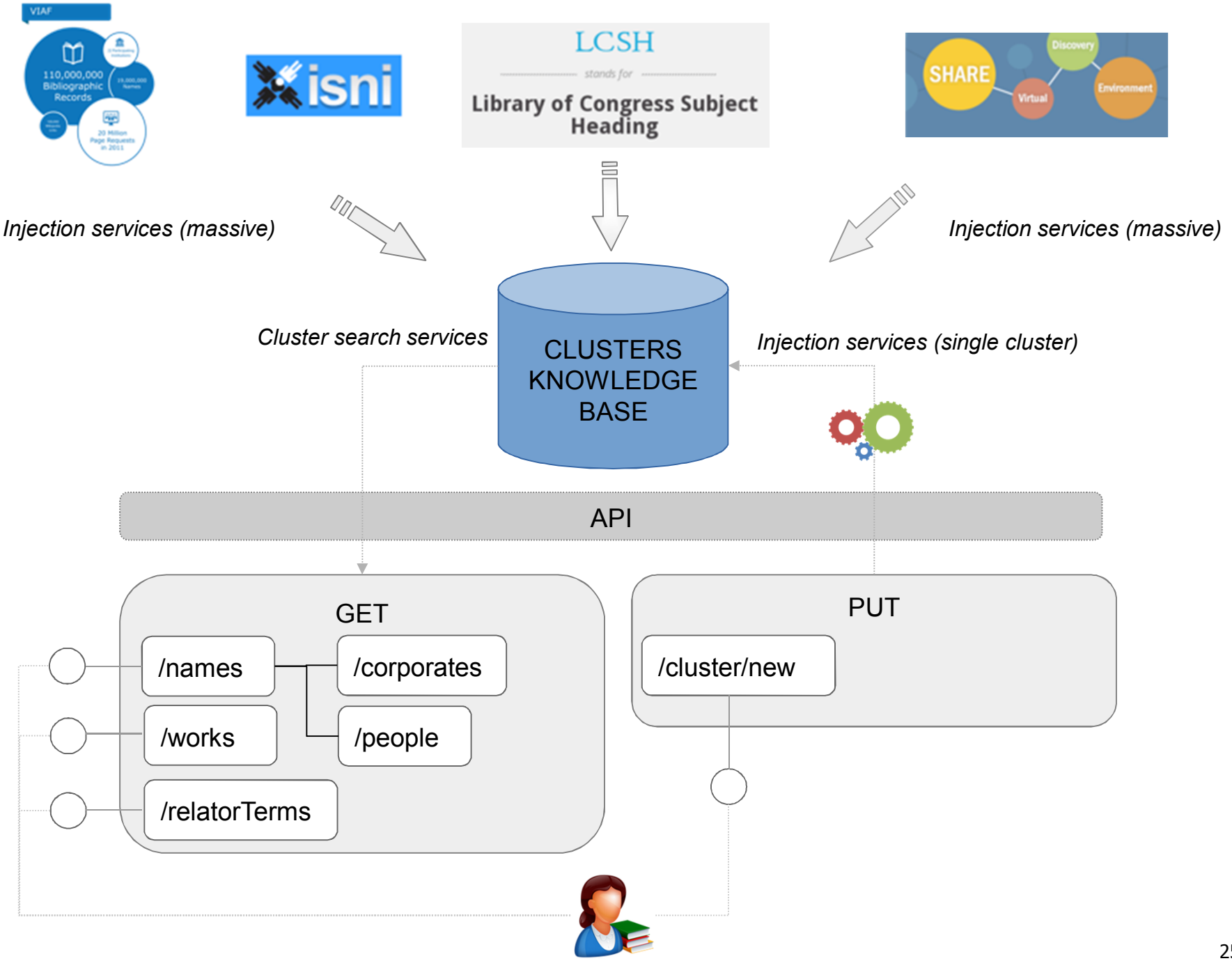
APIs

External sources

Lodify

RDF/Bibframe dataset

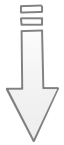
SHARE-VDE Portal


VIAF
110,000,000 Bibliographic Records
isni
LCSH
stands for
Library of Congress Subject Heading
SHARE
Discovery
Virtual
Environment
Injection services (massive)
Injection services (massive)
Cluster search services
CLUSTERS KNOWLEDGE BASE
Injection services (single cluster)
API
GET
/names
/corporates
/works
/people
/relatorTerms
PUT
/cluster/new

Technology stack

LCSH
stands for
Library of Congress Subject Heading

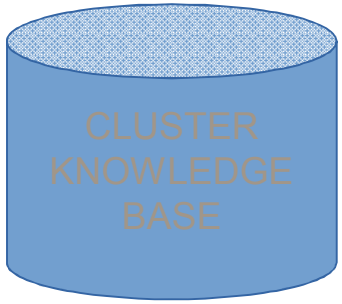
isni



APACHE Spark™

spring batch by Pivotal

Scala



PostgreSQL

Solr

blazegraph™

RESTFul API

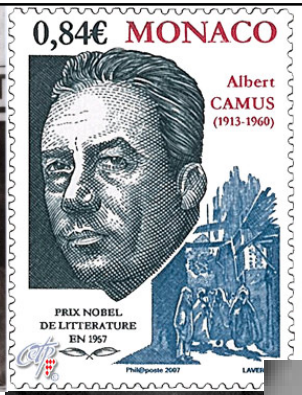
spring by Pivotal

{ REST }

Java

Entity Identification, Reconciliation and Data Publication on the SHARE-VDE Platform

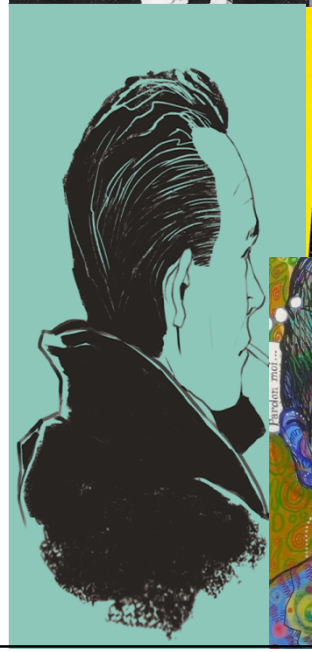
COMBAT
LE JOURNAL DE PARIS
ÉDITION SPÉCIALE
ALBERT CAMUS EST MORT
Une conscience contre le chaos

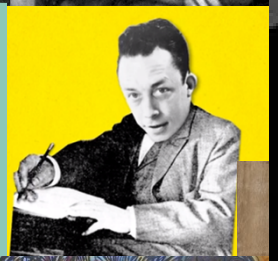

0,84€ MONACO
Albert
CAMUS
(1913-1960)
PRIX NOBEL
DE LITTERATURE
EN 1967

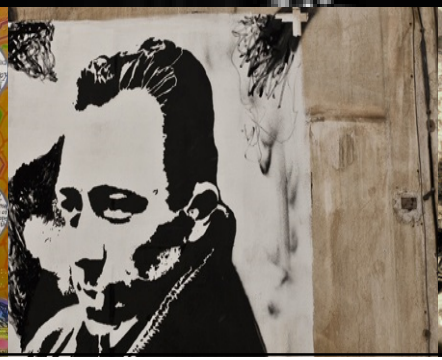
BACK TO Y

BIBLIOTHÈQUE
E.P.T.

I don't care, and you
find that attractive.
t I also don't care about tha
Happy Valentine's Day♡
Love, _____ and
Albert Camus
benkling.tumblr



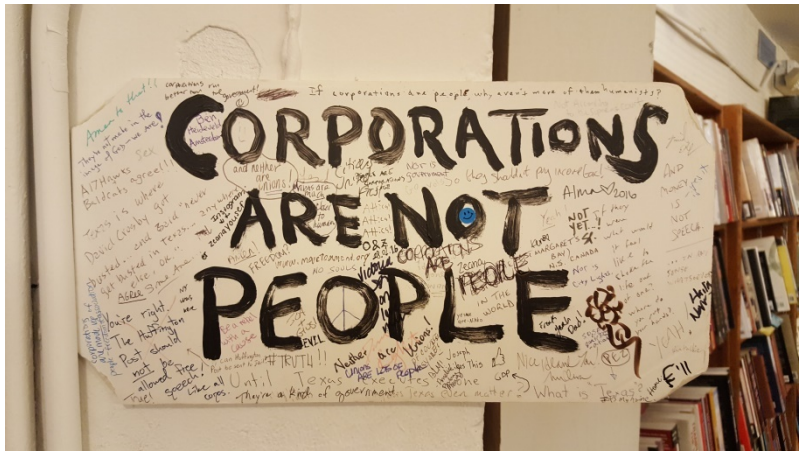




Albert Camus

http://share-vde.org/sharevde/searchNames?n_cluster_id=133656

The importance of identification (not only) in the catalographic tradition



Entity identification: it has traditionally been considered a highly important aspect of cataloguing.

But, the use of attributes to identify a person has not been widely used.

* *Both pictures are taken at the City Lights Bookstore, San Francisco*

Data reconciliation, enrichment and conversion

With the presence online of different catalogues and authority files available in various formats and, where possible, in open mode, the concepts of authority control and of catalogue unification have evolved into the grouping of an entity's identifying attributes from different sources.

The process is best known as ***reconciliation*** and consists of creating a cluster of data that all refer to the same entity.

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 occasion ; 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 [online]

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

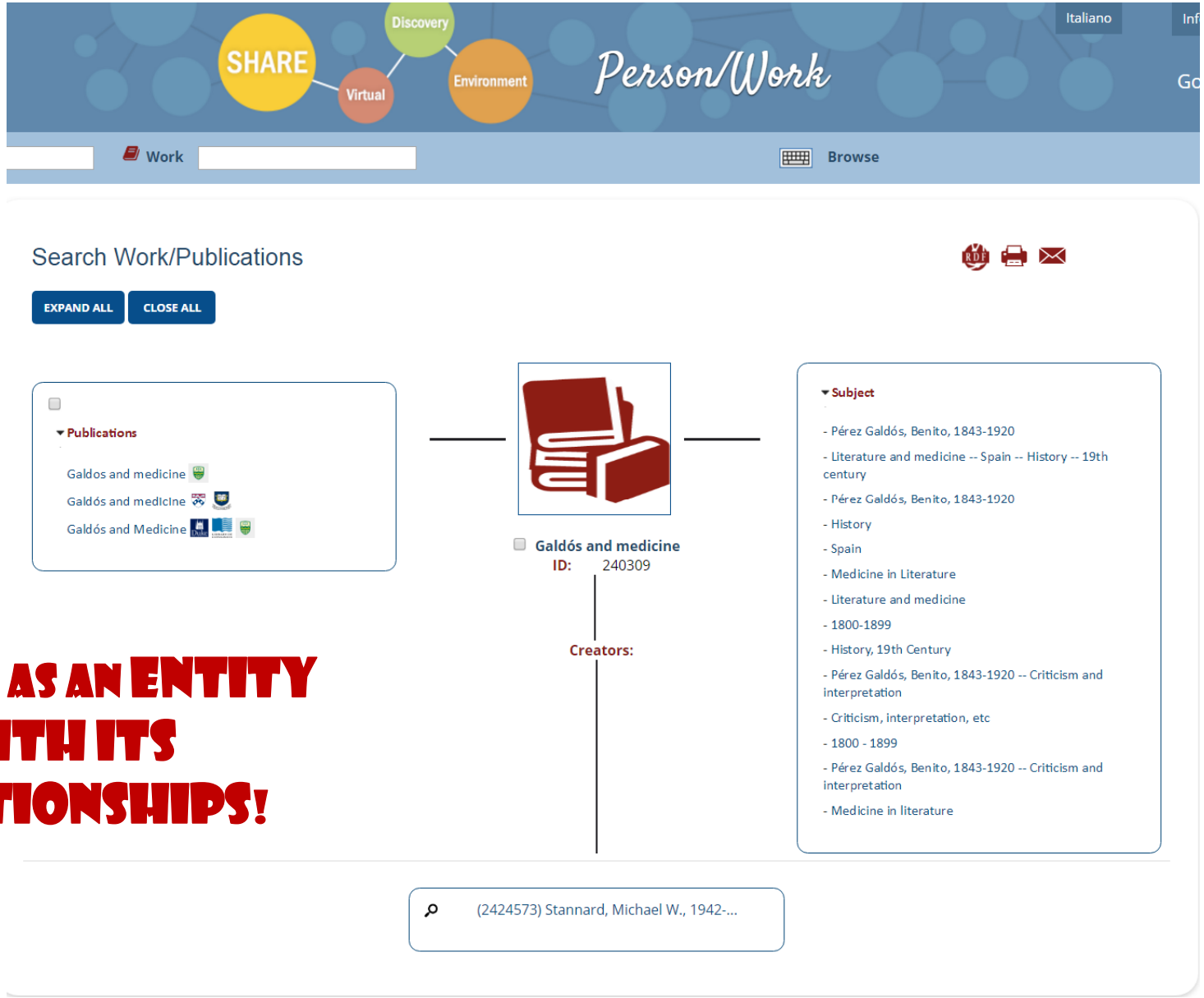
Padri e figlie

Pères et filles

Albert Camus on the SHARE-VDE platform

http://share-vde.org/sharevde/searchNames?n_cluster_id=133656

An entity as Work on the SHARE-VDE platform



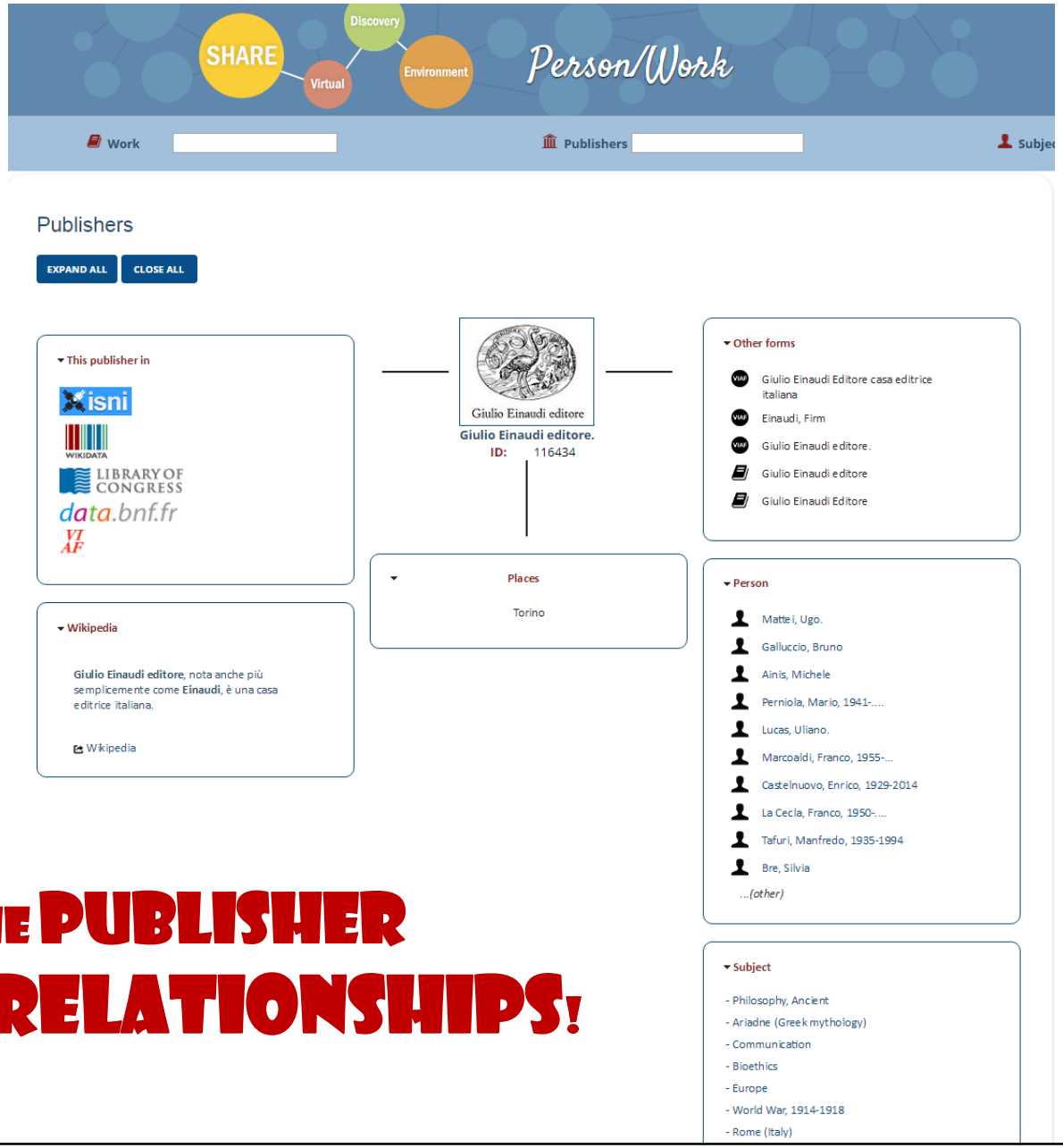
A WORK AS AN ENTITY WITH ITS RELATIONSHIPS!

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
- ויולדי, 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
- فيفالدي، انطونيو، 1678-1741
- 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:

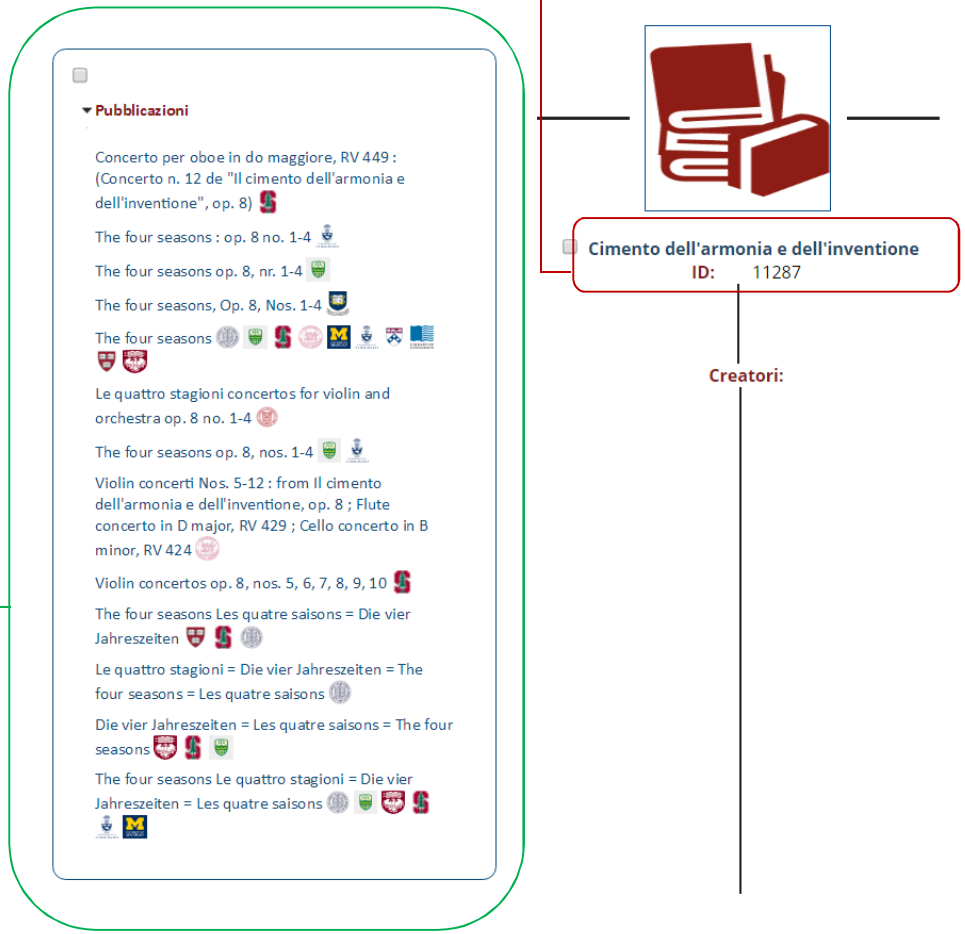
- Wikidata
- Library of Congress Name Authority File
- Data.bnf.fr
- VIAF

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*

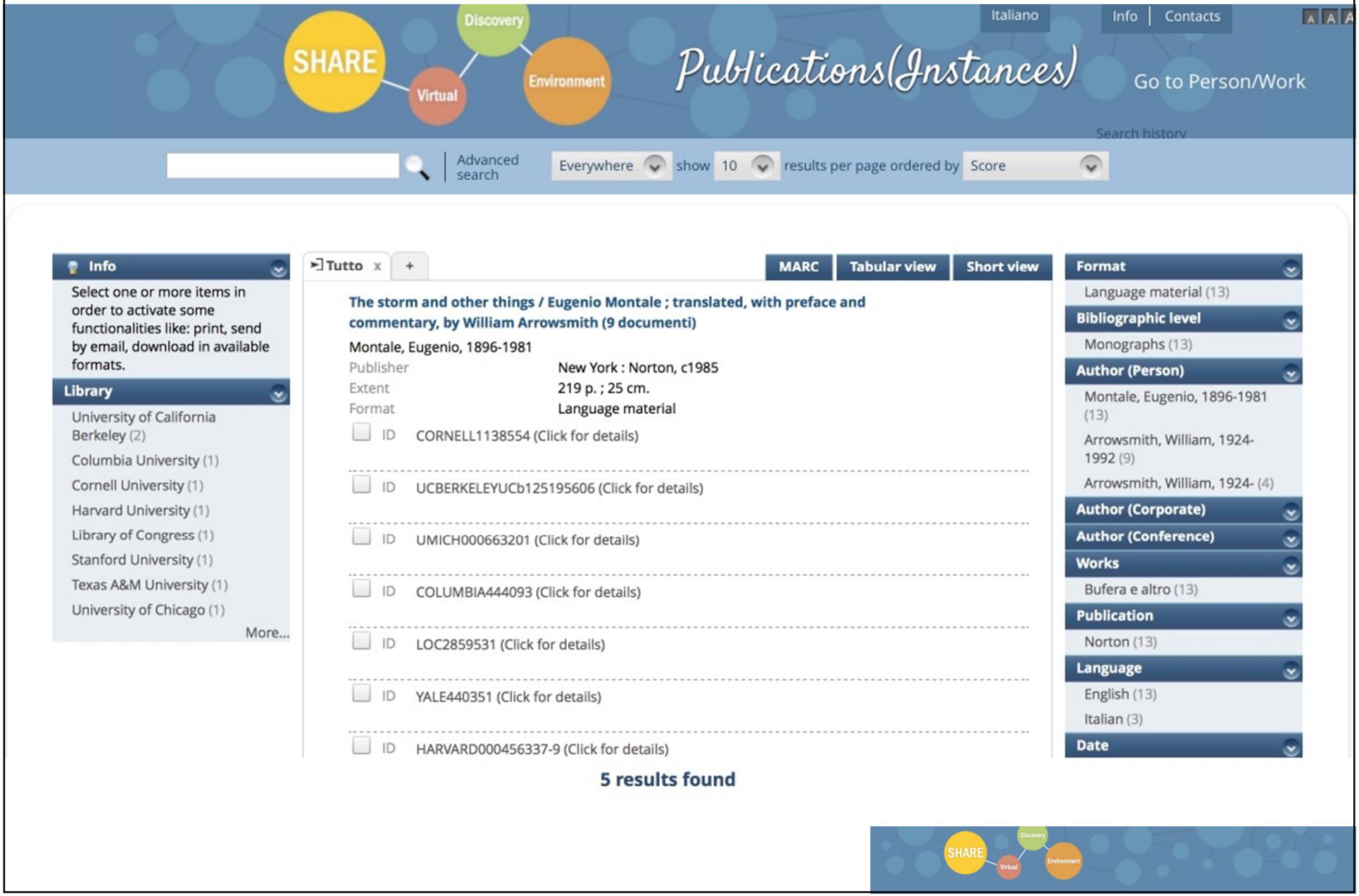
Single work title

Brings together different publications/resources present in different catalogues.



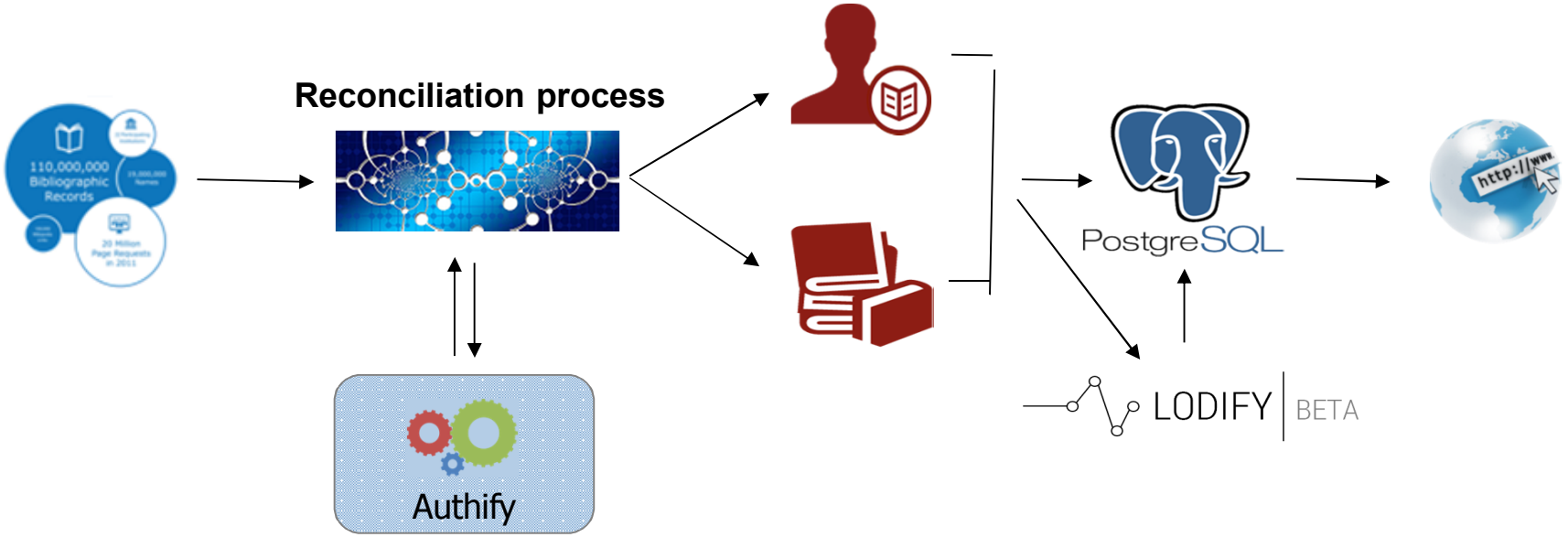
http://share-vde.org/sharevde/searchTitles?t_cluster_id=11287

Example of same Instances present in different libraries

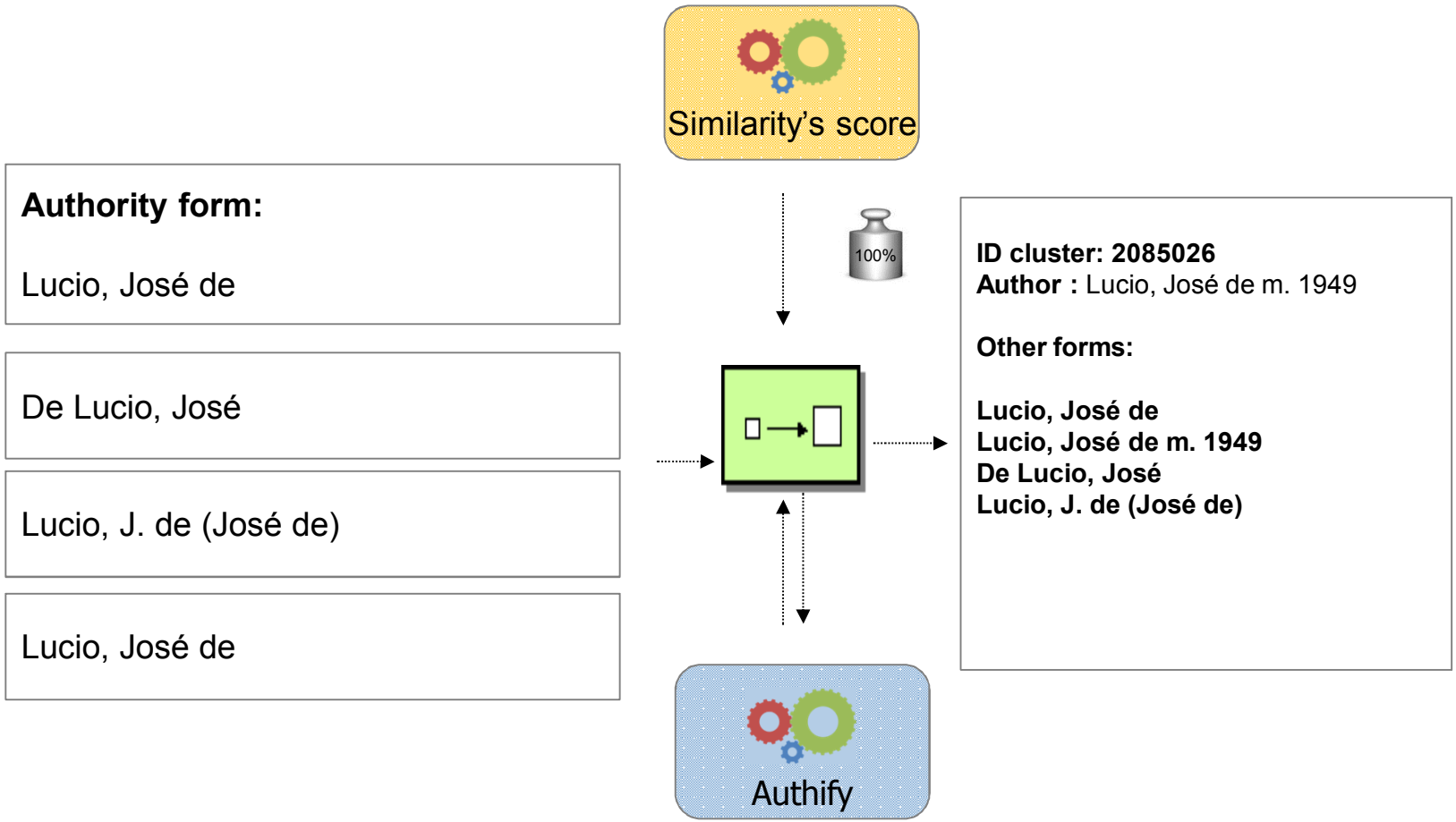


Massive clusters processes

- Authority headings analysis and process in PostgresSql;
- Data enrichment with external sources
- Marc bibliographic process
- Entity detection (authors and co-authors identification process)
- Name heading-to-Authority names association (through a comparison algorithm weights)
- Name heading-to-Variant names association
- Cluster check (it exists = add, it doesn't exist = create new)



Name cluster process



Authority form:

Lucio, José de

De Lucio, José

Lucio, J. de (José de)

Lucio, José de

ID cluster: 2085026
Author : Lucio, José de m. 1949

Other forms:

Lucio, José de
Lucio, José de m. 1949
De Lucio, José
Lucio, J. de (José de)

How reconciliation is obtained

Data reconciliation and enrichment is obtained by:

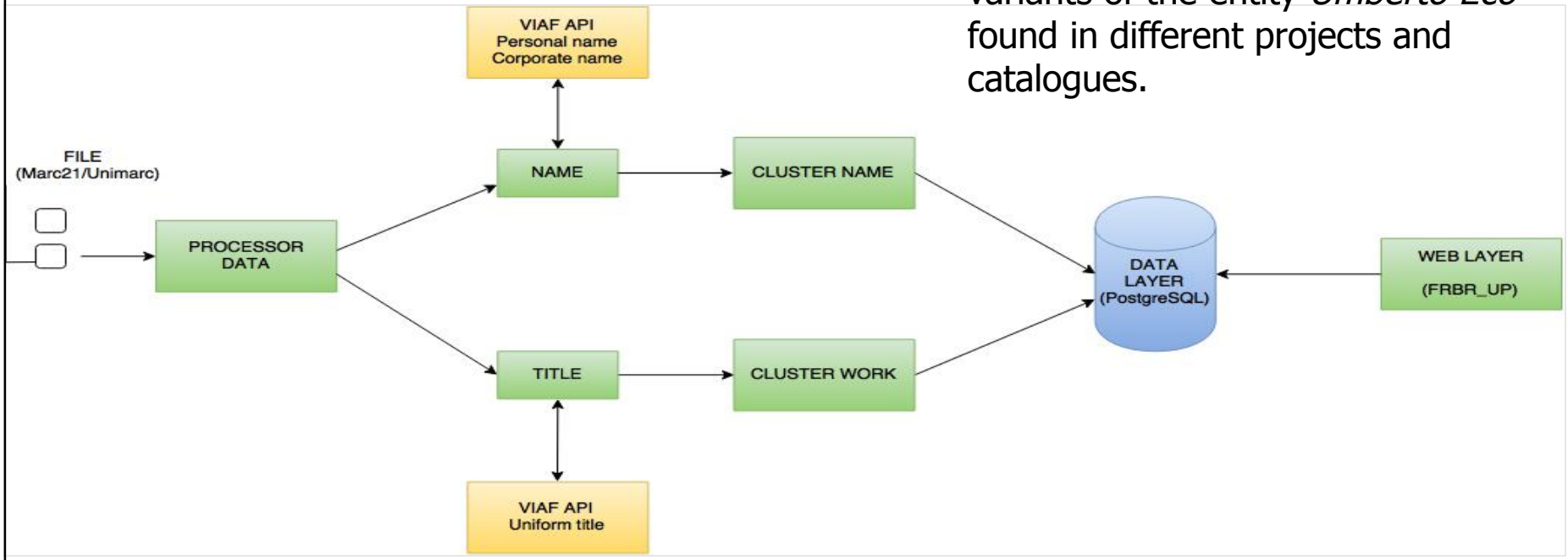
- *automated processes*
- *manual processes*

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.

Automated reconciliation

The process of reconciling name variants of the entity *Umberto Eco* found in different projects and catalogues.

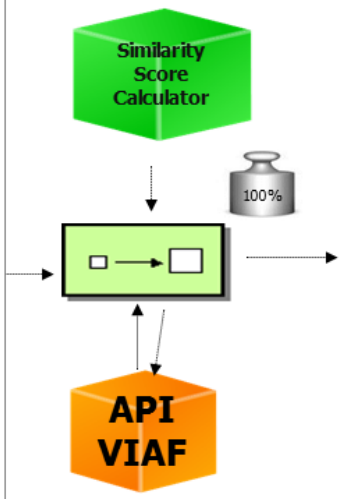



```
001 00001
200 \1  $aEco,$bUmberto$f<1932- >.
400  \0 $aDedalus
997  \\ $aAUTHORITY
```

```
001 27283
700 \1  $aEco,$bUmberto$f<1932- >.
997 \\  $aUNINA
```

```
001 7258
700 \1  $aECO,$bUmberto
997 \\  $aUNISA
```

```
001 7258
700 \1  $aEco,$bUmberto
997 \\  $aUNIBAS
```

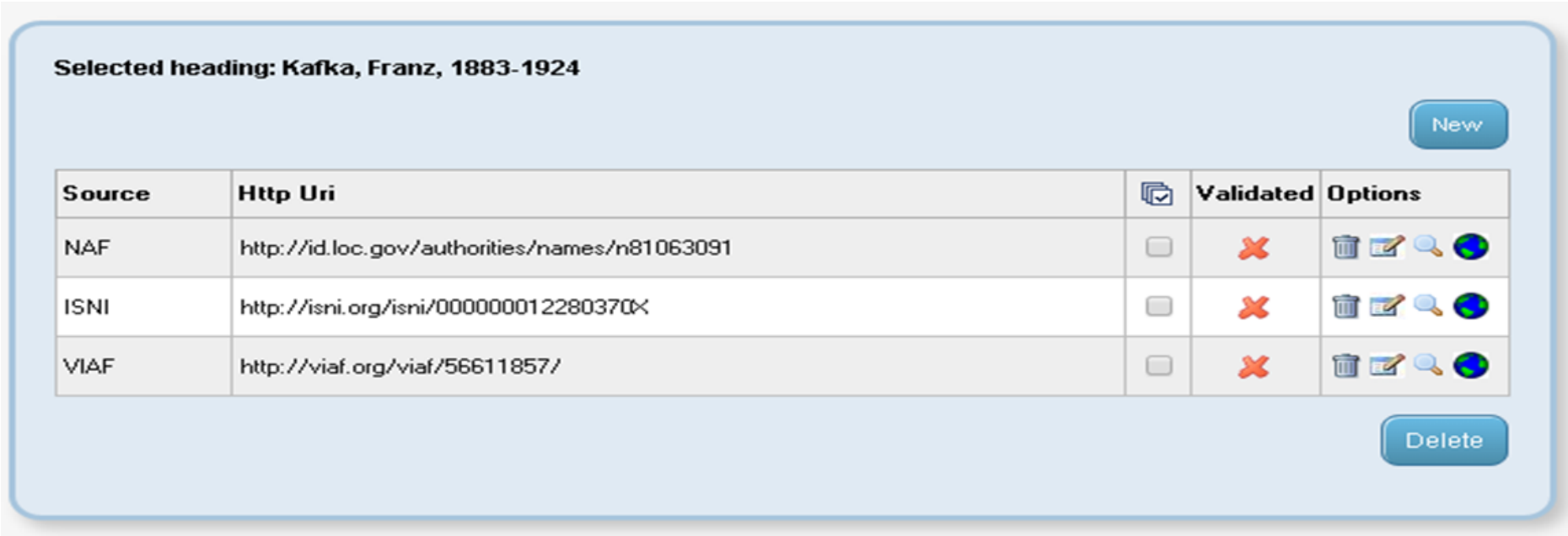


ID cluster: 1281829
Author : Eco, Umberto, 1932-...
VIAF ID: 108299403

Other forms:

ECO, Umberto
Eco, Umberto
Dedalus
1932- اكو، أمبيرتو
إكو، أمبرتو
Eco, Umberto (Italian writer,
architect, and critic, born 1932)
Umberto Eco semiologo,
filosofo e scrittore italiano
1932 ,אקו, אומברטו-

Manual reconciliation



The same result of entity enrichment, but carried out, in the cataloguing workflow, using manual processes, which enable a more precise verification of the results: the availability of API and web services allows the use of external sources (in this example, NAF, ISNI and VIAF) and the association of the "Franz Kafka" cluster with the URIs that identify it in each of the projects. This is the starting point for the automatic processes of cluster creation by means of aggregating the multiple name variants.

Guarantee of authority for the new virtual authority files

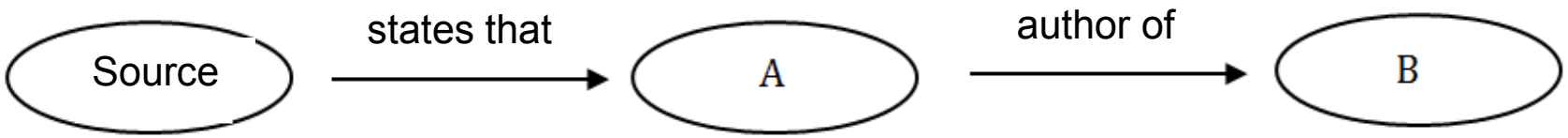
Need to guarantee the accurancy of this information

Knowing the *provenance* of a piece of information – *its origin*, authorship or matrix – is a key factor in determining *the extent to which it can be trusted*.

The information source has become the guarantor of quality: creating a link between information and its source has become essential for the purpose of guaranteeing the authority of the information itself.

Guarantee of authority for the new virtual authority files

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.

Conclusions: the sharing and reuse of information resources

All of the efforts made to facilitate the sharing and reuse of assets, and tools produced by libraries, museums and other institutions, and to guarantee their availability to a wider public, enriching the World Wide Web with valuable information that would otherwise remain mostly hidden in archives, collections and catalogues, promote a culture of open access to knowledge, with numerous advantages for each link in the information chain.

Libraries, archives and museums all benefit from the possibility of more comprehensive and well-structured tools which provide end users with a vast wealth of information, and create new cooperative tools for sector professionals.

In line with this new open philosophy of data sharing and reuse, even traditional authority controls, union catalogues and discovery systems are evolving.

Links to Live Examples in SHARE-VDE

Some examples on the SHARE-VDE platform

www.share-vde.org

Emily Bronte:
http://share-vde.org/sharevde/searchNames?n_cluster_id=318705

and this Work Wuthering Heights:
http://share-vde.org/sharevde/resource?uri=LOC18843460&v=l&dcnr=1

Frankenstein:
http://share-vde.org/sharevde/resource?uri=LOC18789412&v=l&dcnr=8

Eugenio Montale:
http://share-vde.org/sharevde/searchNames?n_cluster_id=166369

and his Works:
http://share-vde.org/sharevde/resource?uri=UCBERKELEYUCb232697760&dir=1&v=l

Instances reconciliation:
http://share-vde.org/sharevde/search?q=Android+studio+essentials&v=ll&h=any_bc&s=10&o=score

European BIBFRAME Workshop 2017

Thank you!

Tiziana Possemato

tiziana.possemato@atcult.it
tiziana.possemato@casalini.it