Lille 2017

What is the likely shape of the library of the future? And how do we build collections for it?

Enrichment, Reconciliation and Publication of Linked Data with the BIBFRAME model

Tiziana Possemato
Casalini Libri - @Cult

casalinilibri

@Cult

New cooperative scenarios

New context: new ways of cooperating between institutions and corporations, further removed from a complex *reductio ad unum* approach and physical merging.

The new generation of Authority control, Union catalogues and discovery tools: cross-institutional processes of integration and virtualization.

New data enrichment opportunities absolutely not possible in the past.

Focus on identifying entities and discovering their relationships with other entities.

Data entification, reconciliation, enrichment and publication

Bring together and make available data from different sources in a way that could be defined as *democratic* to better identify the entity in question.

Even wider reconciliation and enrichment processes form the basis of a number of projects that convert and publish bibliographic catalogues as linked open data, such as:

- **Share Catalogue**: www.catalogo.share-cat.unina.it (@Cult project)

- **Share VDE – *Share Virtual Discovery Environment***: www.share-vde.org (in partnership between Casalini Libri and @Cult)

SHARE Virtual Discovery Environment project

A prototype of a virtual discovery environment with a three BIBFRAME layer architecture (Person/Work, Instance, Item) will be established.

The project will create, in addition, a database of relationships that is open to the community and a common knowledge base of clusters that uses the paradigm of the semantic web.

The project will also identify issues and problems related to these new information management processes and propose solutions.

SHARE Virtual Discovery Environment project

The general project aim is to integrate the considerable knowledge base represented by the universities' authority and bibliographic catalogues to enrich it with the new and in-flux one generated by the web, creating an integrated information system to provide users with a single access tool for the various Libraries' OPAC.

SHARE-VDE brief project overview

Threefold goals:
-Conversion, supply and management of authority and bibliographical data in BIBFRAME taking into account the complexity of the long and heterogeneous transition time;

-Development of detection services for entity identification including relator terms, and creation of a common knowledge base of clusters of reconciliated results for names and works;

-Publication of a FRBR/BIBFRAME three layered platform with build-in instances techniques.

16 participant institutions

Phase 1 from October 2016 to January 2017

Phase 2 from March to September 2017

The theoretical context of the project

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

-the new international **RDA – *Resource Description and Access*** guidelines
-**Linked Open Data** philosophy and technology
-**BIBFRAME**: one of more interesting models to convert and publish data

RDA Toolkit: Identify and Link

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 Link

- 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

LINK

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).

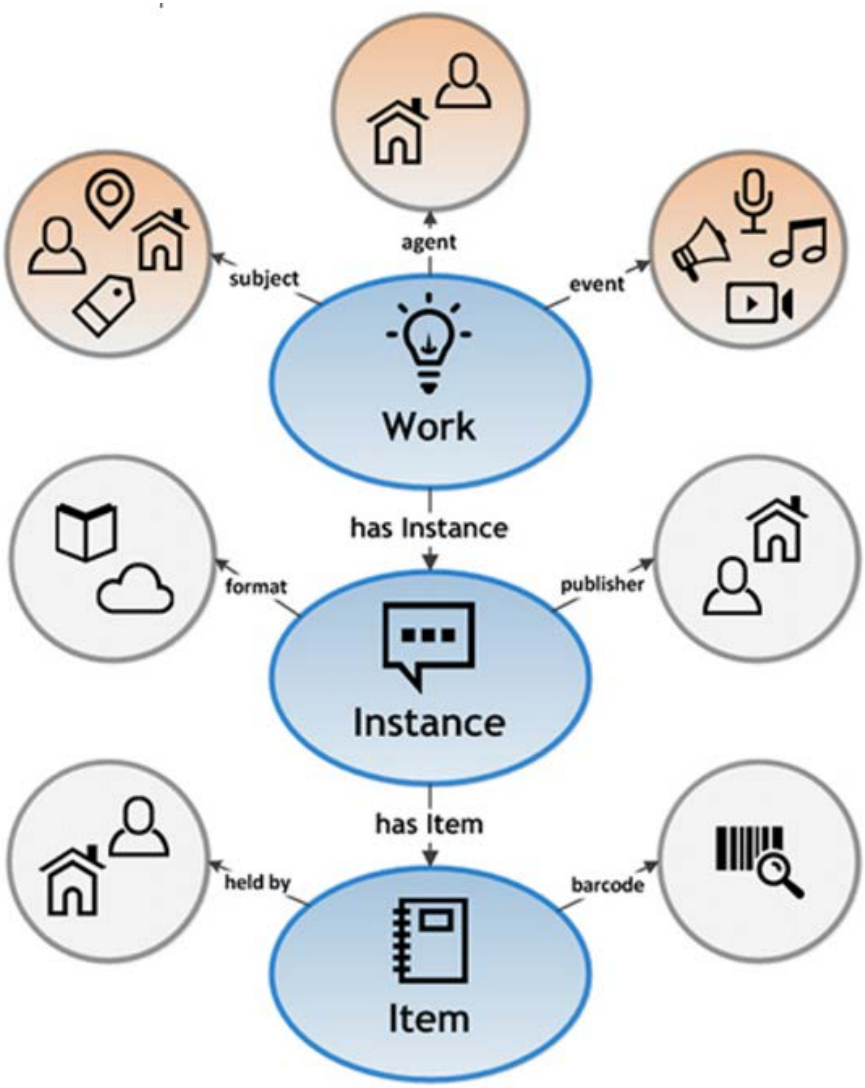
BIBFRAME – Bibliographic Framework Initiative

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 Marc 21 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 – Data model v. 2.0



BIBFRAME – Data model v. 2.0

“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 – Data model v. 2.0

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.

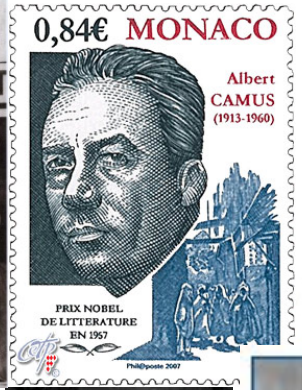
Events: Occurrences, the recording of which may be the content of a Work

Who’s Who?

The question at hand:

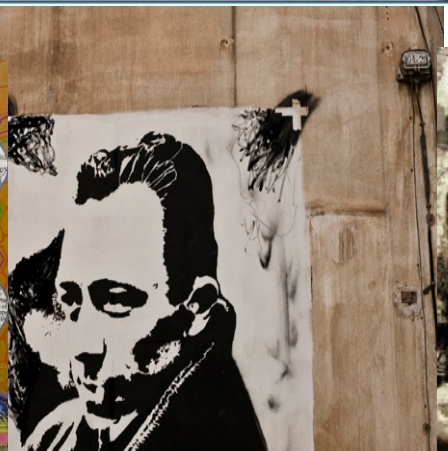
how to identify an entity?


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 1957


BACK TO YOU, BEAUVOIR.


I don't care, and you
find that attractive.
But I also don't care about that.
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 in the catalographic tradition

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

But, the use of attributes to identify an entity has not been widely used

Data reconciliation, enrichment and conversion

With the on-line presence of different catalogues and authority files available in various formats and, where possible, in open mode, also the concept of authority control and of union catalogue has evolved into the grouping of an entity's identifying attributes from different sources.

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

Reconciliation => *Re- conciliare*

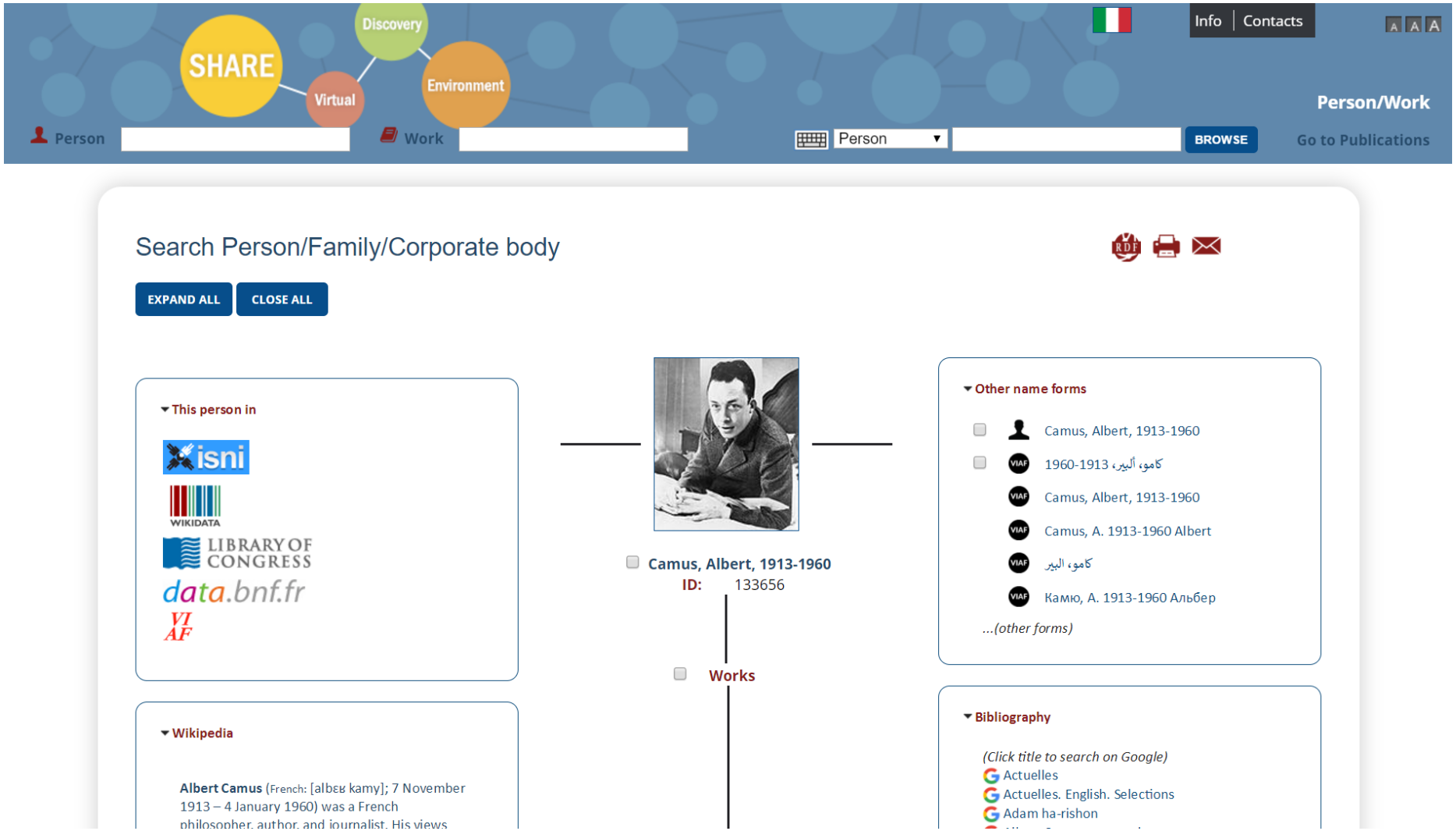
The term *reconcile*, comes from the Latin *reconciliare*, made up of *re-* and *conciliare* “to bring together, conciliate”: bring together the different name variants referring to the same entity.

Reconciliare: “To bring to agreement, restore to peace and harmony”.

Reconciliation

begins with the assumption that an entity may be known by different names deriving from cultural differences, different cataloguing rules, linguistic differences, simple typographical errors, and accepting this variety, making it into a value.

An example of Reconciliation: Albert Camus in Share-VDE project



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

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:

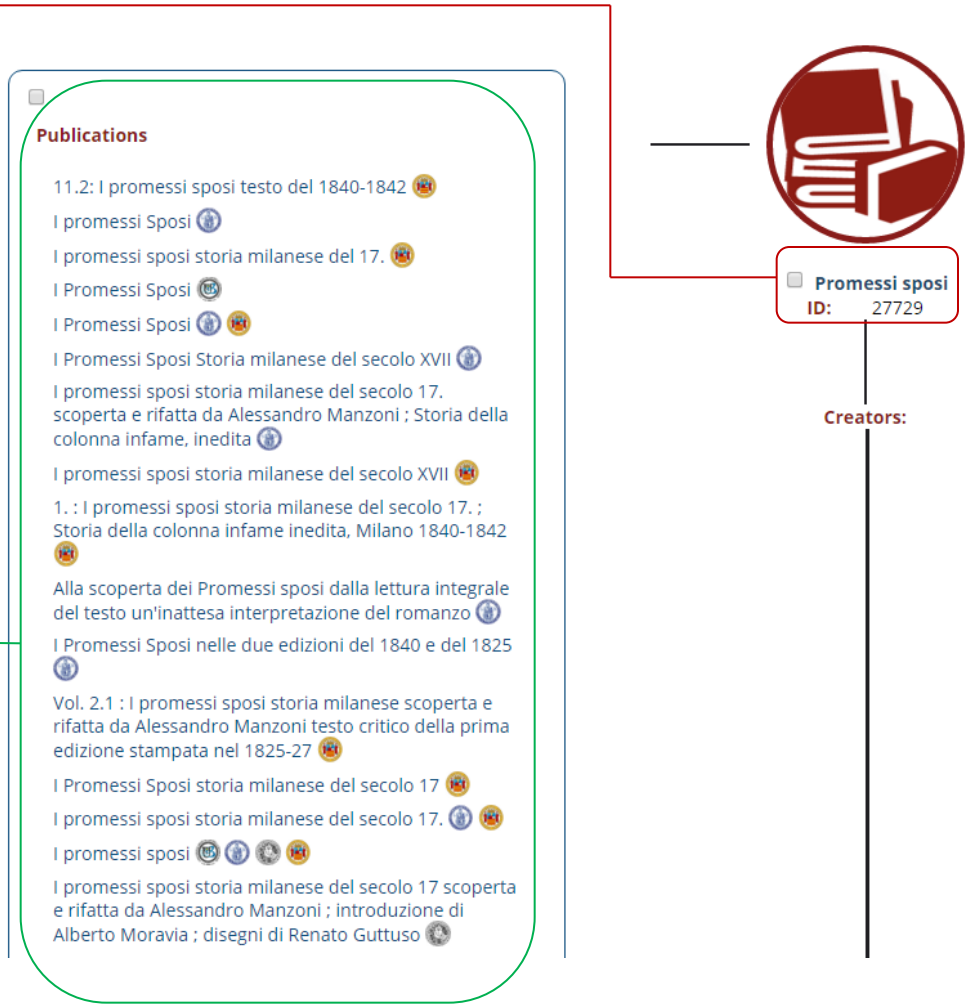
- Wikidata
- Library of Congress Name Authority File
- Data.bnf.fr
- VIAF

An example of Title reconciliation

Grouping under a single work title of the many publication titles in the catalogue for *Promessi sposi*.

One work title

Brings together more than 70 different publications present in different catalogues.



Injection services (massive)

Injection services (massive)

Cluster search services

CLUSTERS KNOWLEDGE BASE

Injection services (single cluster)

API

GET

/names

/corporates

/works

/people

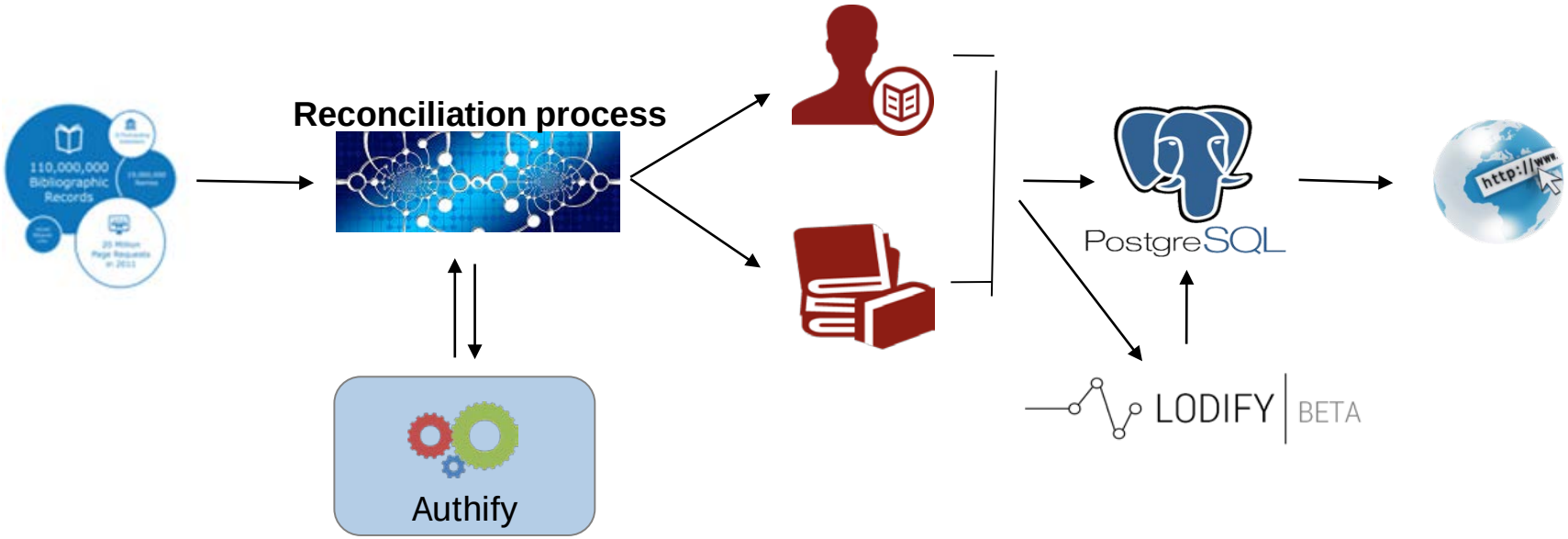
/relatorTerms

PUT

/cluster/new

Massive clusters process

- 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

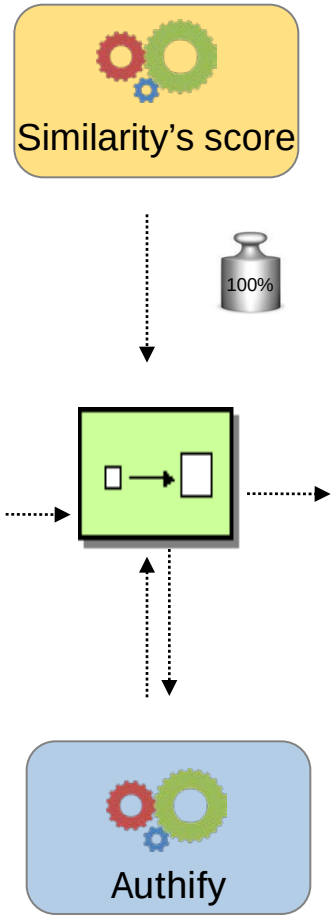
Authority form:

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:

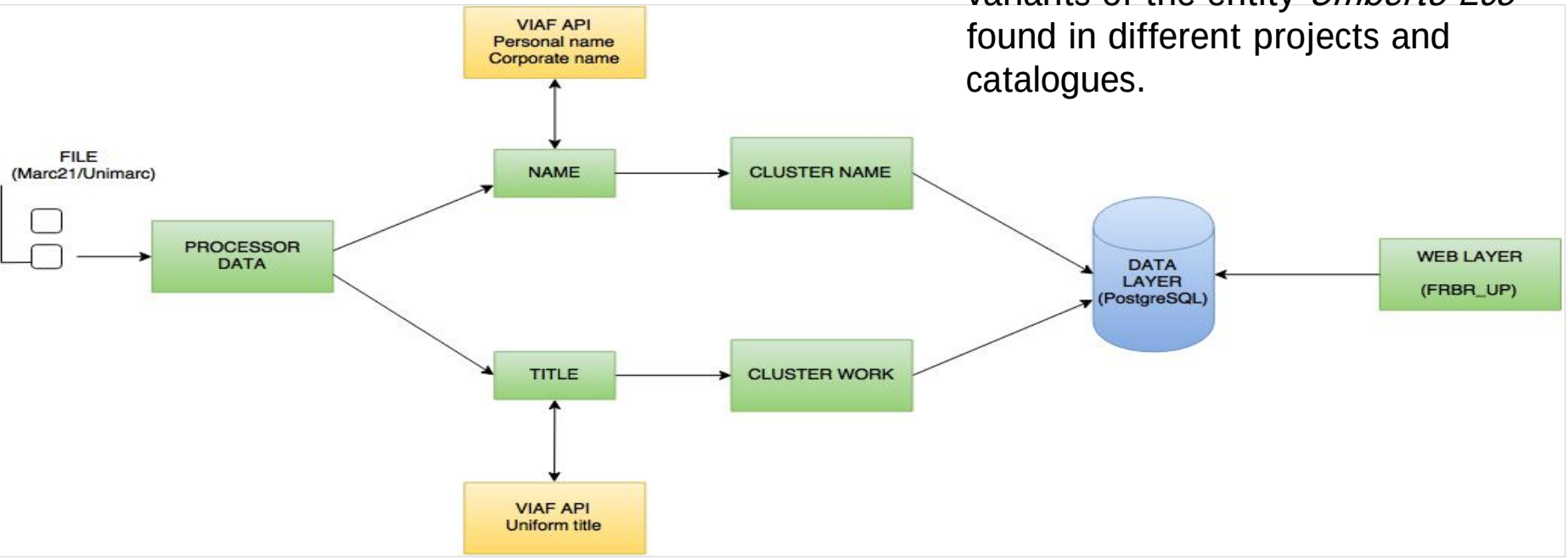
- *automated processes*
- *manual processes*

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.

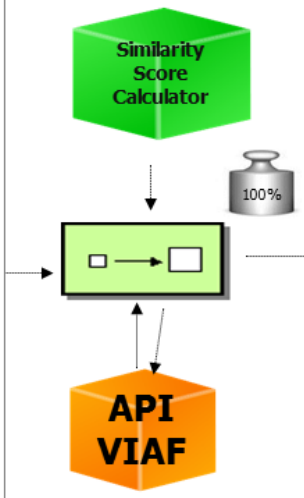



```
001 00001
200 \1  $aEco,$bUmberto$f<1932- >.
400  \0 $aDedalus
997  \\ $aAUTHORITY
```

```
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
```

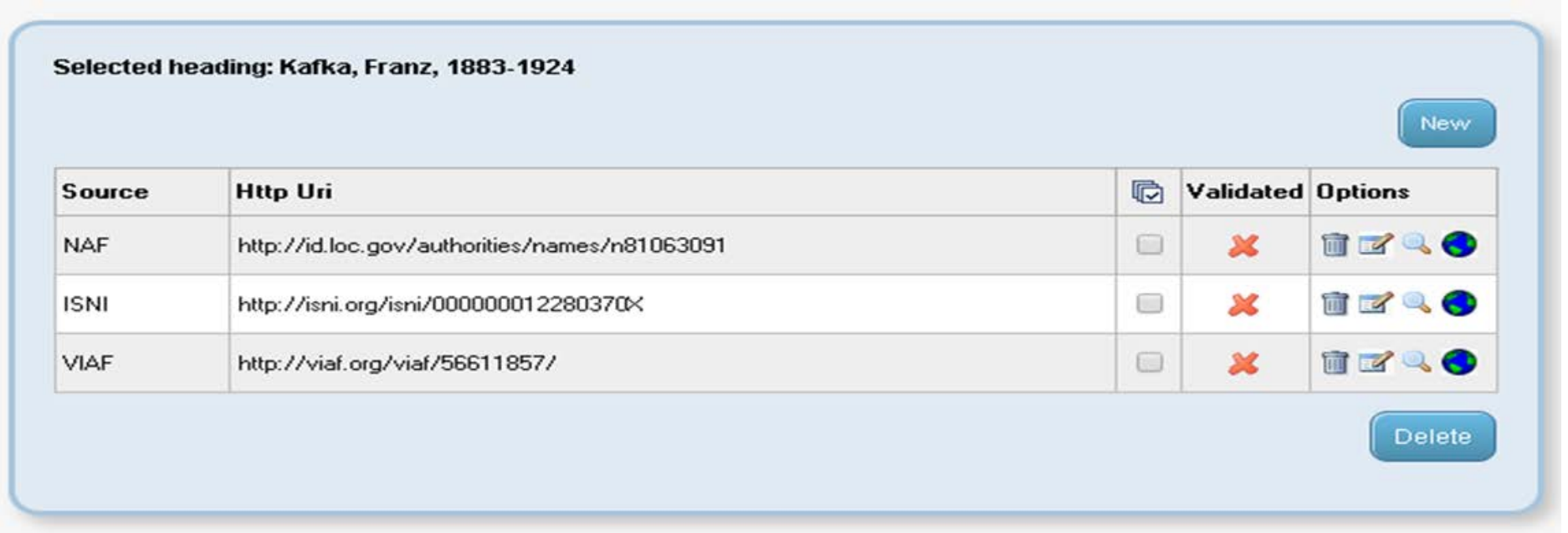


ID cluster: 1281829
Author : Eco, Umberto, 1932-...
VIAF ID: 108299403

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,אומברטו ,אקו-

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

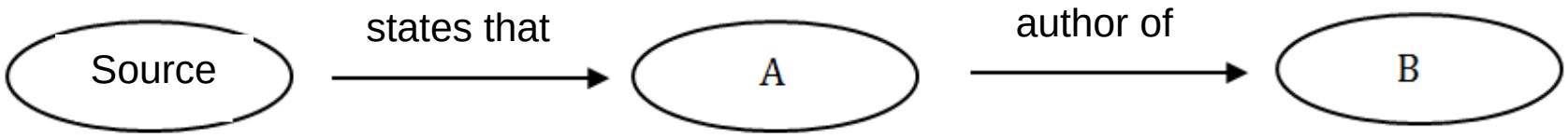
Need to guarantee the accurancy of this information

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.

Thank you.

The project is driven by the library community input and we will be very grateful for any feedback, proposals and suggestions.

Tiziana Possemato
tiziana.possemato@casalini.it
tiziana.possemato@atcult.it