PCC Operations Committee Meeting (OpCo)
May 4-5, 2017
Library of Congress
Washington, DC

PCC
Program for Cooperative Cataloging
BIBCO / CONSER / NACO / SACO

SHARE-Virtual Discovery Environment and The Casalini experience and roadmap for supplying BIBFRAME data

Michele Casalini
Casalini Libri

Tiziana Possemato
Casalini Libri - @Cult

casalinilibri

@Cult

Index

- Introduction and Overall Project Goals
- The Theoretical Context
- SHARE-VDE Process Overview
- Entity Identification, Reconciliation and Data Enrichment
- Reconciliation & Enrichment – Automated Procedures
- Reconciliation & Enrichment – Manual Procedures
- Access Points and URIs
- Conversion in RDF/BIBFRAME
- Trust and Provenance
- Link to Examples in SHARE-VDE
- Triple Store Query Examples in Blazegraph

Introduction
and Overall Project Goals

Current activities and infrastructure

Casalini Libri produces, for publications from Romance language countries, more than 40,000 original bibliographic records in RDA as a member of the Program for Cooperative Cataloguing (PCC) with authority entries;

Bibliographic records are created using the @Cult OLISuite WeCat cataloguing modules;

@Cult, in addition to the LMS and Discovery tools field, is specialized in the development of software components and platforms to convert, enrich, reconcile and publish data of cultural institutions under the linked data paradigm.

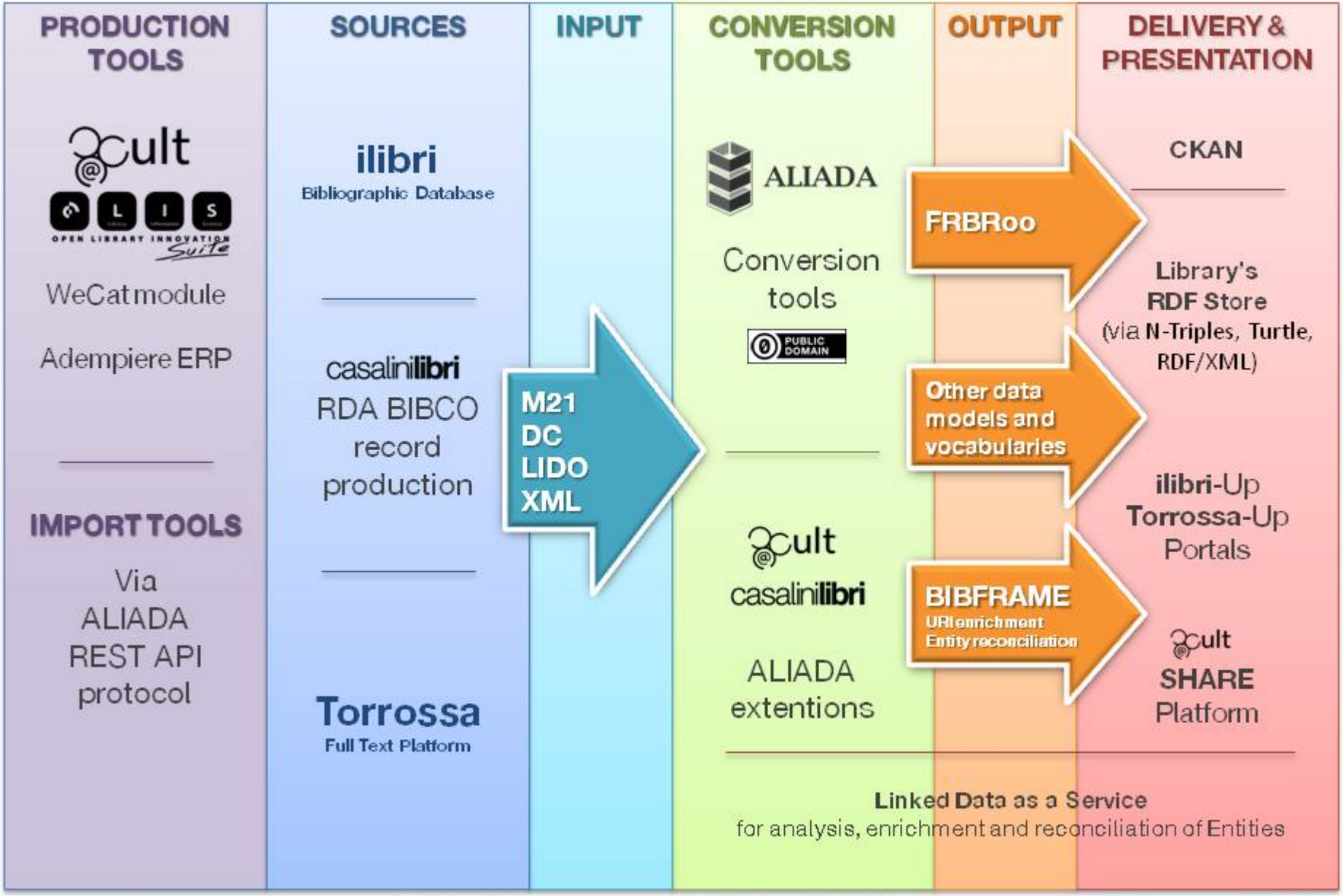
The three major areas of activity for the development of BIBFRAME/Linked Data

1. The enrichment of MARC records with URIs to simplify the BIBFRAME conversion;

2. The use of a framework to automate the conversion from MARC to RDF, using BIBFRAME vocabulary;

3. The creation of a BIBFRAME layered platform prototype starting from bibliographic and authority records, to test and demonstrate the advantages of the BIBFRAME data model.

Casalini's BIBFRAME Conversion, Distribution & Publication Options



November 2016

For further information please contact bibframe@casalini.it
@Cult is the technological partner of Casalini Libri

casalinilibri

SHARE-VDE overall goals

The main goals of Phase 1 and 2 Research & Development activities are:

- Reconciliation and clusterization of varying forms of the same entity;
- Enrichment of MARC records with URIs along with the development of detection procedures for entity identification, including relator terms;
- Conversion, supply and management of authority and bibliographical data in BIBFRAME, taking into account the complexity of the long transition time for both the library and the data producer;
- Publication of a BIBFRAME three layered platform prototype.

SHARE - Virtual Discovery Environment project

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

Phase 1: analysis, enrichment, reconciliation, conversion into RDF and publication of two sets of bibliographic data for each participating library were planned (1985 and 2015 imprint titles). This phase also included the release of Marc records enriched with URIs and 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 SHARE-VDE portal www.share-vde.org.

Phase 1: from October 2016 to January 2017.

SHARE - Virtual Discovery Environment project

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

The library catalogue of each participating institution will be converted into BIBFRAME 2.0 and returned to each library (over 100 million records and subsequent datasets are expected to be processed).

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:

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

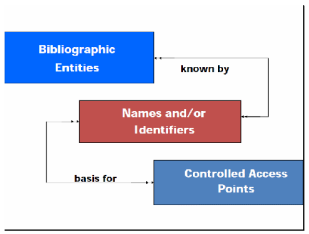
Phase 3 of SHARE-VDE, and its modular options for adoption, will also be driven by the library community

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
Knowledge base of clusters				x
URIs Registry		x		
Entity detection			x	
Data publication on SHARE-VDE portal		x		x
UC1: *Borrow Direct*	x		x	
UC2: *Borrow Direct in Franklin*		x		
UC3: *Collaborative selection tools*	x			x
UC4: *Community specific GUIs and functions*			x	
UC5: ……				

The Theoretical Context

The Theoretical Context

Where we are going…

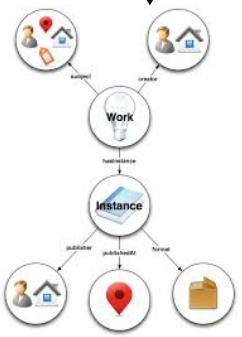


Functional Requirements for Authority Data

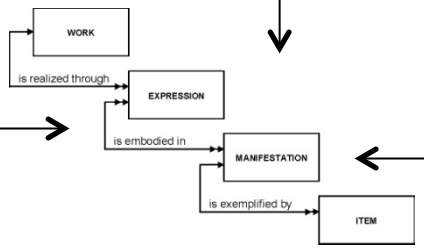
MARC 21

RDA
RESOURCE DESCRIPTION & ACCESS

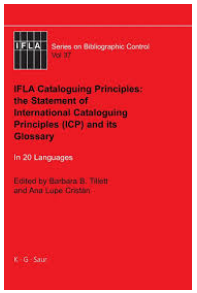
Resource Description and Access



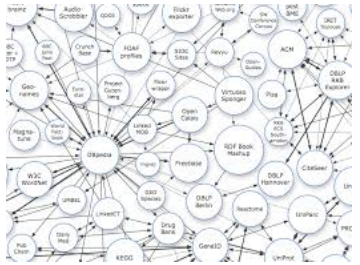
Bibframe



Functional Requirements for Bibliographic Records



International Cataloguing Principles



Semantic web/Linked data

The 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 is considered 'the core' ontology, completed with the ontologies for specific domains, that libraries will suggest

SHARE-VDE Process Overview

The SHARE-VDE processes

OliSuite: manual process

Marc enriched/URIs

Person
Works
Publishers

Database of relationships

Similarity's score

MARC
Bibliographic
Authority

XML
Schema 1
Schema 2

Entity detection
Enrichment
Reconciliation/Cluster

N1
N2
N3

Knowledge base of clusters

Authify

Dump db

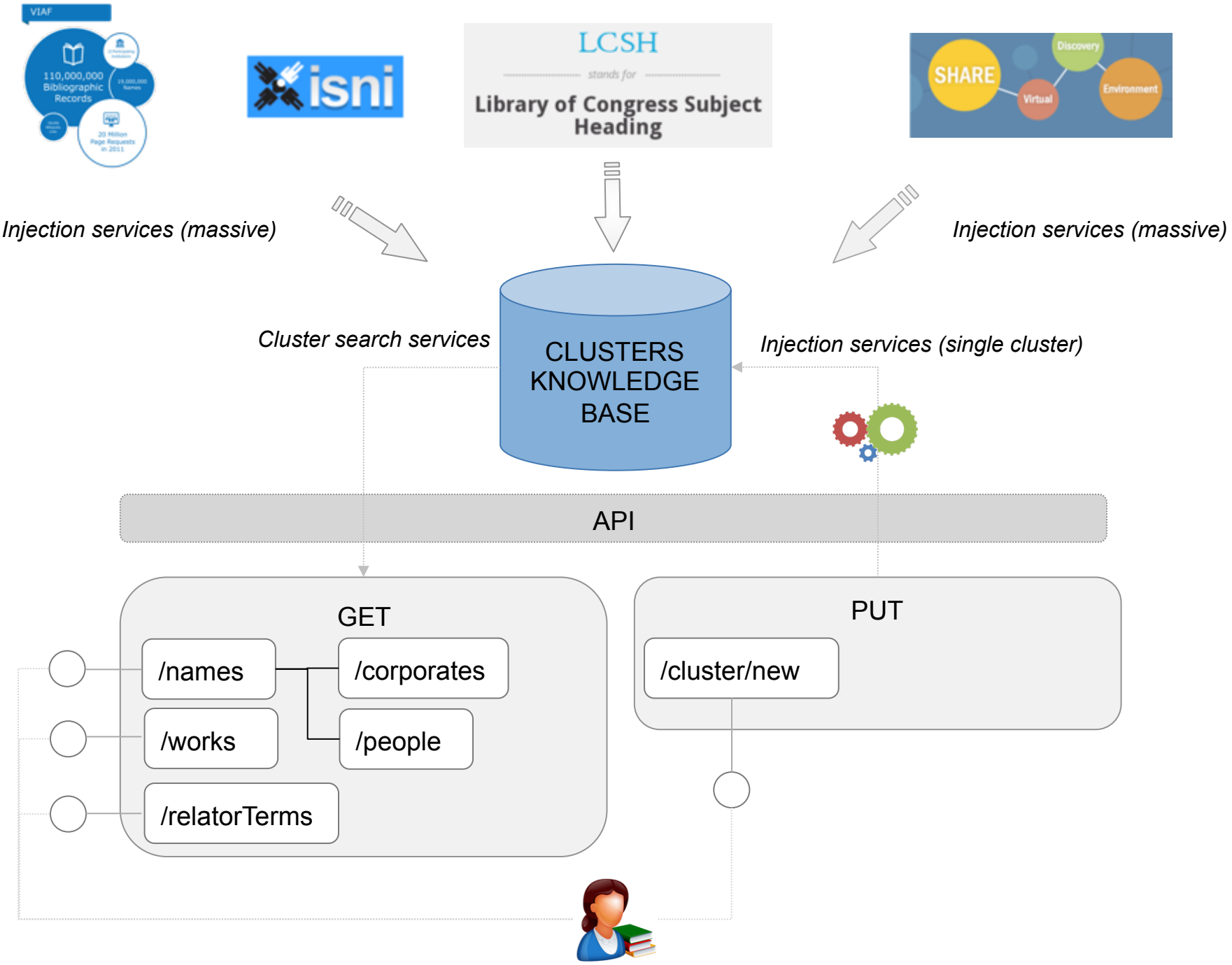
APIs

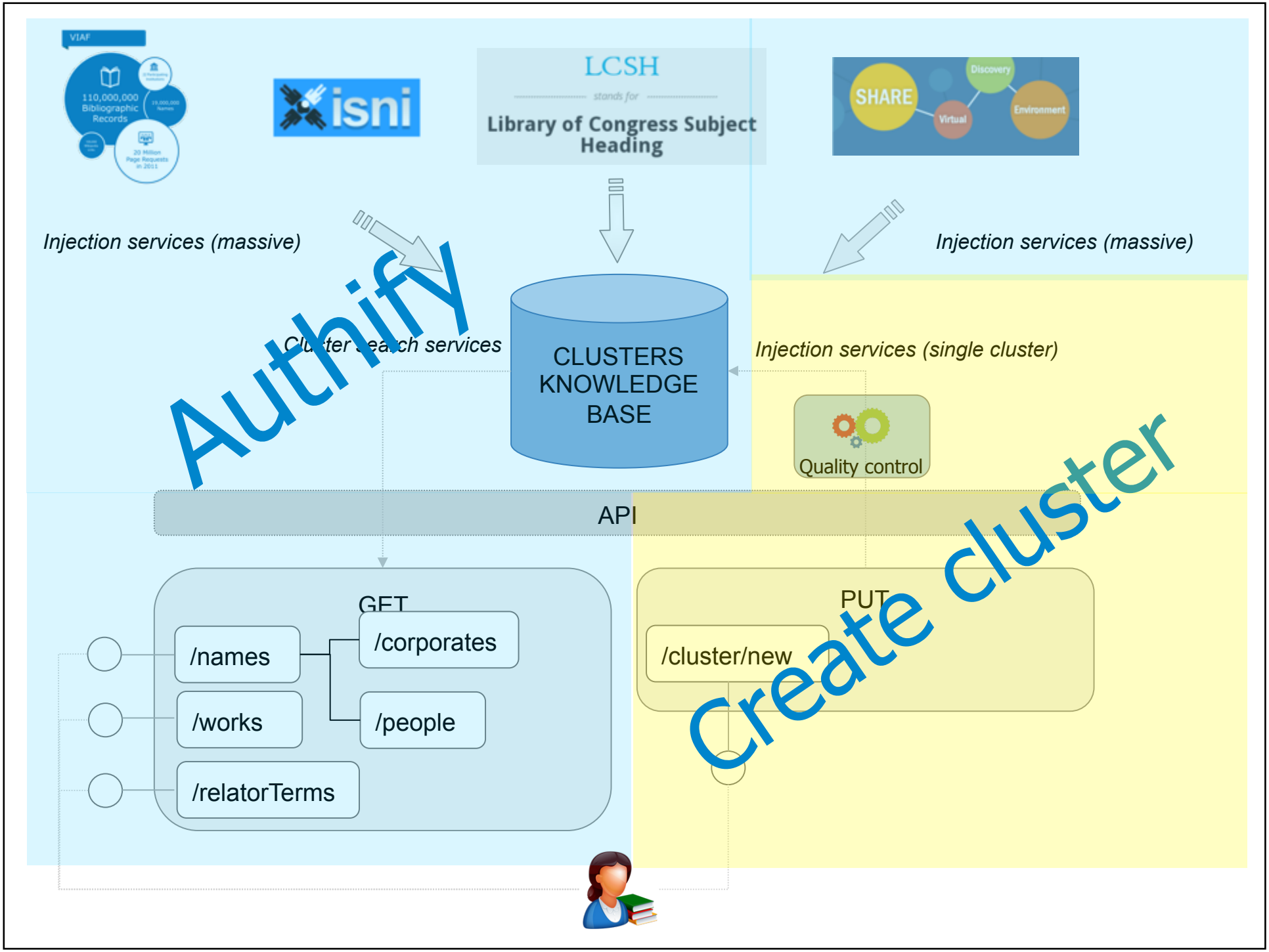
External sources

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


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)
Authify
Cluster search services
CLUSTERS KNOWLEDGE BASE
Injection services (single cluster)
Quality control
API
GET
/names
/corporates
/works
/people
/relatorTerms
PUT
/cluster/new
Create cluster

Technology Stack

LCSH
stands for
Library of Congress Subject Heading

isni

APACHE Spark™

spring batch by Pivotal

Scala

CLUSTER KNOWLEDGE BASE

PostgreSQL

Solr

blazegraph™

RESTFul API

spring by Pivotal

{ REST }

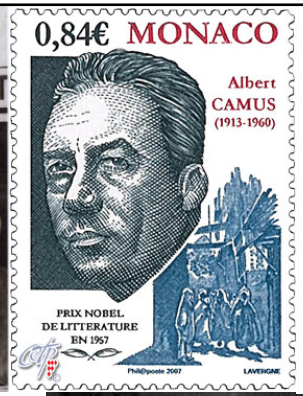
Java

Entity Identification, Reconciliation and Data Enrichment

Who's Who?

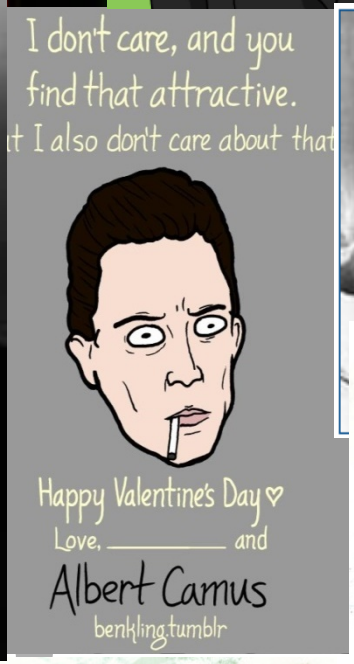
The question at hand:
how to identify an entity?

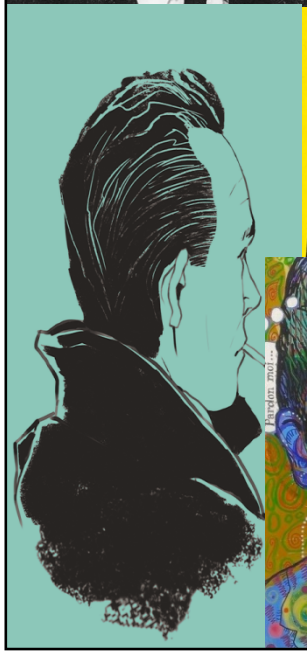

COMBAT
LE JOURNAL DE PARIS
ÉDITION SPÉCIALE
ALBERT CAMUS EST MORT
Une conscience contre le chaos

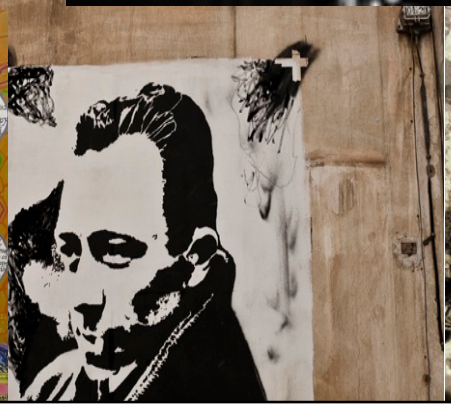

0,84€ MONACO
Albert
CAMUS
(1913-1960)
PRIX NOBEL
DE LITTERATURE
EN 1957
Phil@poste 2007
LAVERGNE

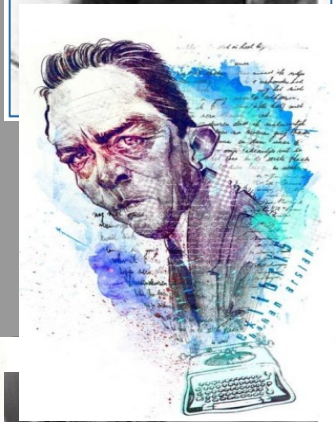

BACK TO YOU, B

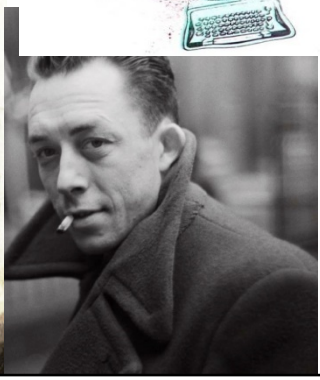

BIBLIOTHÈQUE
E.P.I.T.E.R


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





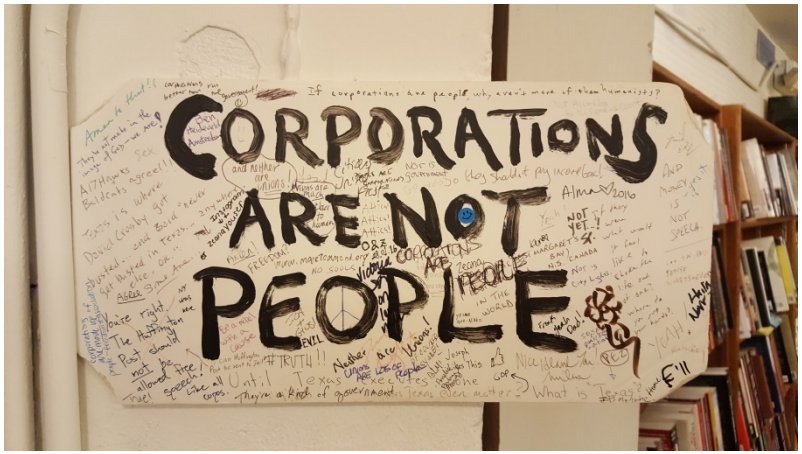




Albert Camus

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

The importance of identification (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.

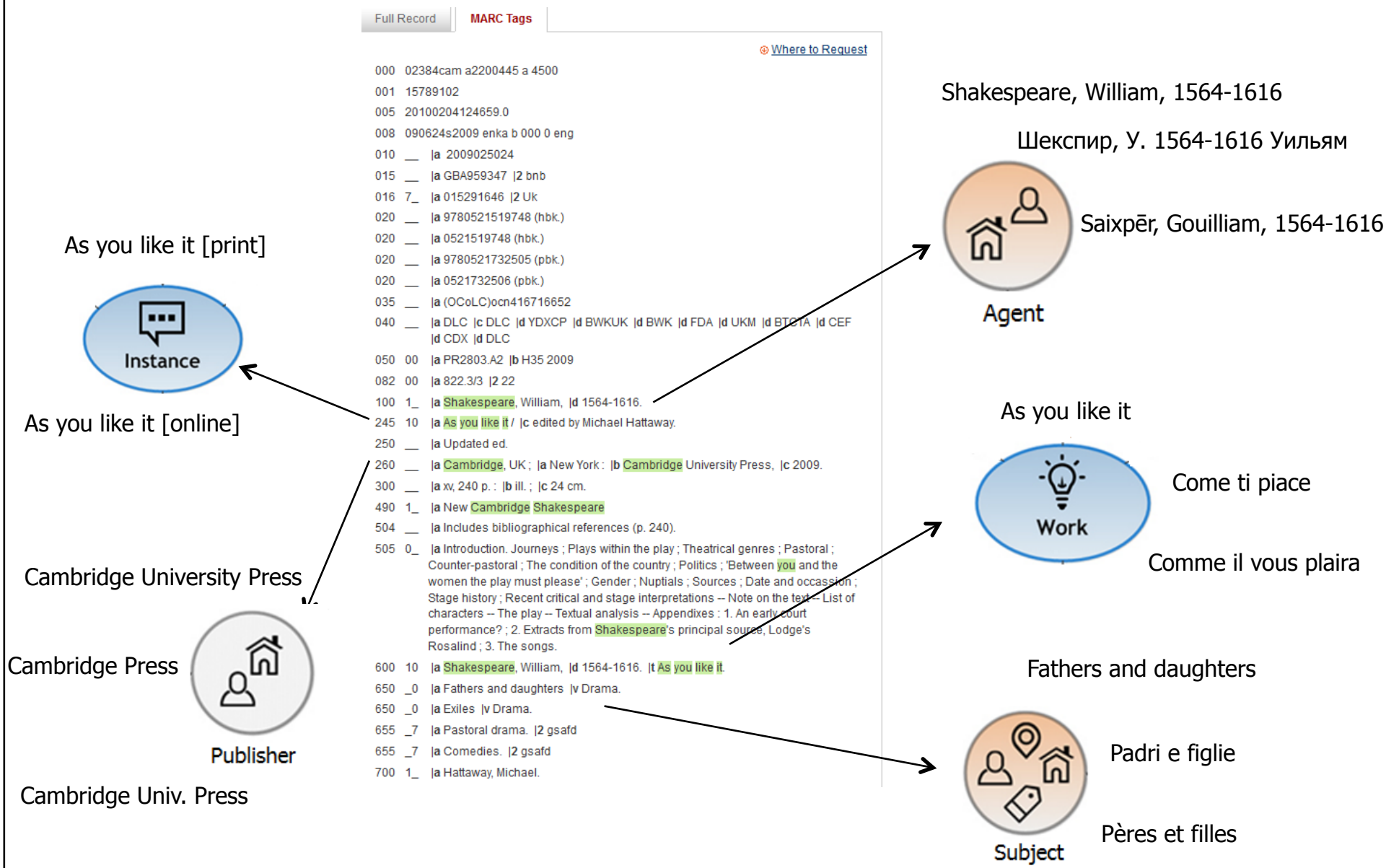
* *Both pictures are taken at the City Lights Bookstore, San Francisco*

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



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:

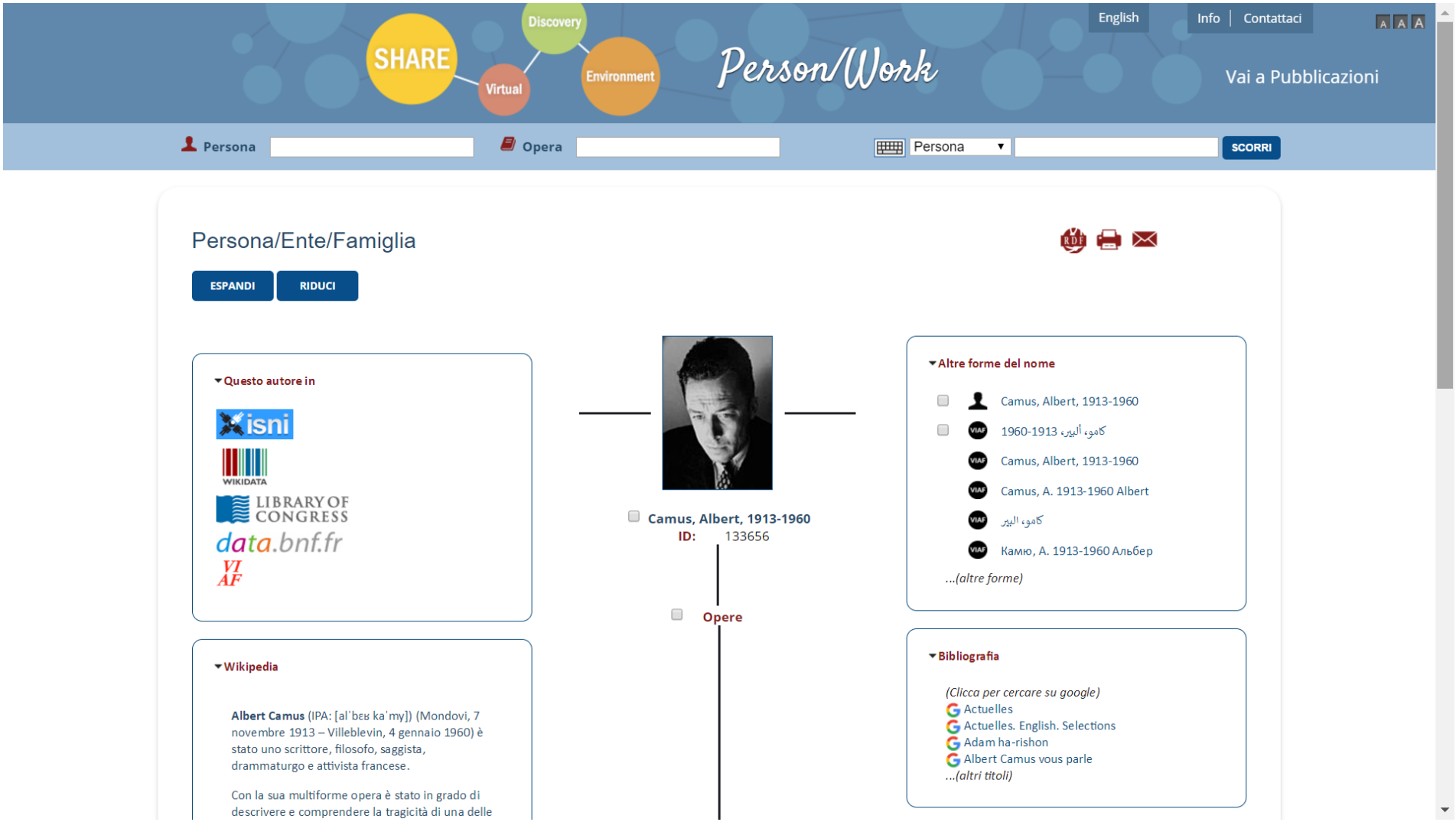
SHARE-VDE

SHARE Virtual Discovery Environment in Linked Data

www.share-vde.org

in partnership between Casalini Libri and @Cult.

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
- ויוולדי, 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 ،فيفالدي، انطونيو
- 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 present in different catalogues.

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

Reconciliation & Enrichment – Automated Procedures

How reconciliation is obtained

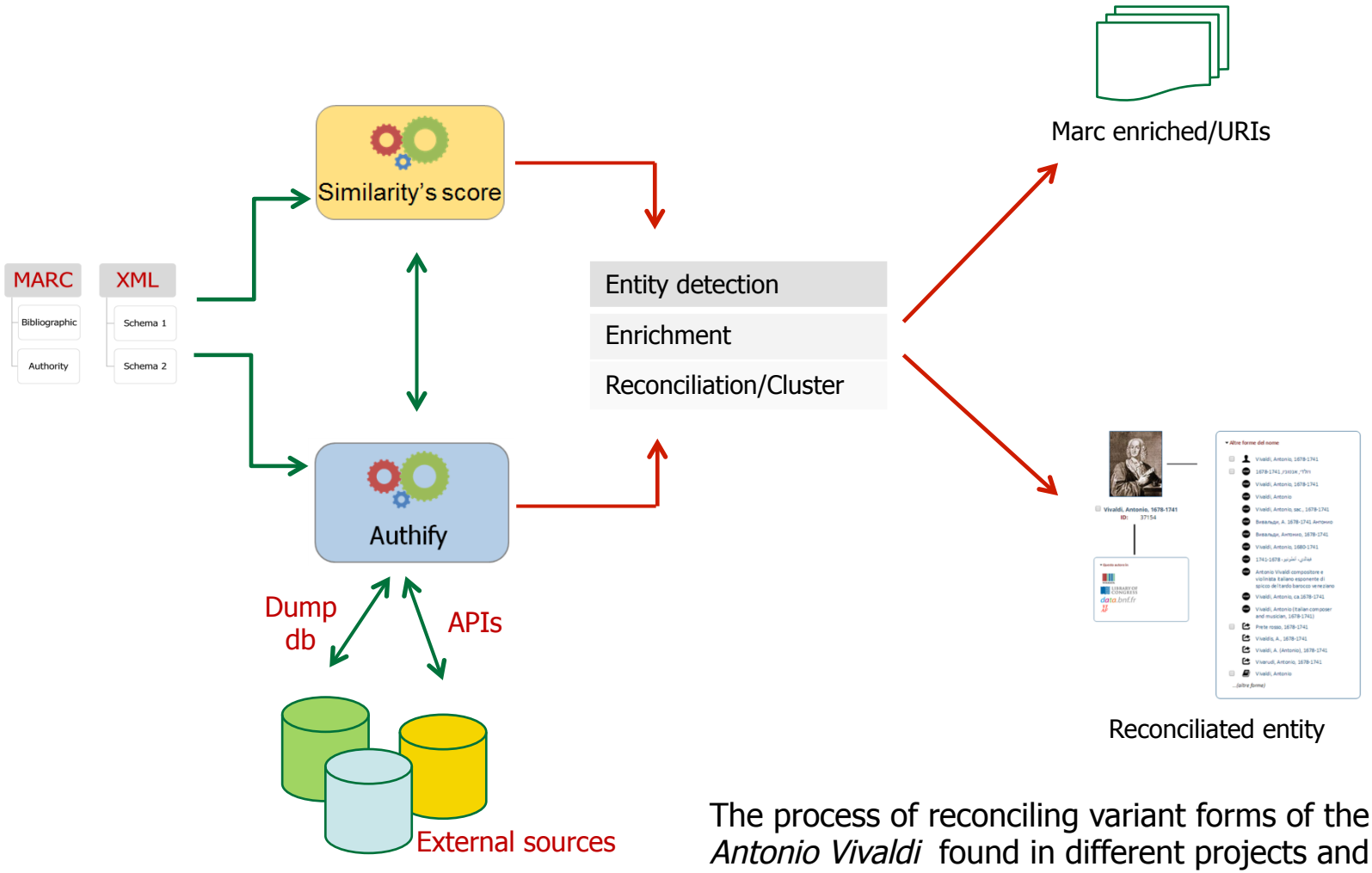
Data reconciliation and enrichment is obtained by:

- *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.

Automated reconciliation and enrichment



The process of reconciling variant forms of the entity *Antonio Vivaldi* found in different projects and catalogues.

Authify – General description of the tool

Authify is a RESTFul module that offers several search and detection services. The project was started, initially, to overcome some of the limitations of the public VIAF Web API.

VIAF, being a public project, does not allow a massive invocation of its API: for those cases where such this is necessary, the project provides a download of the whole dataset.

This was the main reason for the implementation of Authify: to index and store the dataset of VIAF clusters and provide, in addition, powerful full-text and bibliographic search services.

It's possible to add to Authify other dump databases, from external projects that make them available.

Authify – Cluster search services

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 as precise a match as possible within the managed clusters.

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

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

Authify – Cluster search services

The system has been built with extensibility in mind, so the mentioned chain is fully configurable; for instance, here is a brief description of the current configuration when searching names clusters:

- **Subfields matching**: the query language allows the caller to specify the source tag / subfields that compose the heading (which is the actual input query string).
- **Input heading exact match**: the system tries to find an exact match with the provided query string.
- **Full text search**: if exact match is not possible, then a regular full text search is executed, with things like proximity search for names (e.g. Bertrand Meyer = Meyer Bertrand), special detection for a particular entity (e.g. birth and death dates).
- As last resort, the system executes a **search by "initials"**, in order to find a valid match in those cases when the input string (or the indexed heading) contains the name in its shortened form. As in the previous point, this could lead to a response with reduced precision.

Authify – Cluster search services – Response

The query interface: http://labs.atcult.it/authify/names?q=bertrand_Meyer : the system will provide a response like this:


```
{
 "responseHeader" : {
   "QTime" : 3,
   "matching-strategy" : "name::headings-exact-match",
   "status" : 0
 },
 "response" : {
   "docs" : [ {
     "id" : "51714577",
     "type" : "Personal",
     "uri" : "http://viaf.org/viaf/51714577/",
     "headings" : [
                              "Meyer, Bertrand, 1950-....",
                              "Bertrand Meyer",
                              "Meyer, Bertrand" ],
     "sources" : [
                              "BNF|12079479",
                              "DNB|112127843",
                              "ISNI|0000000109003927",
                              "LC|n  86061235",
                              "LNB|LNC10-000142119",
                              "NDL|00471567",
                              "NKC|skuk0004073",
                              "NLA|000035194108",
```

Authify – Relator term detection

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

Starting from a MARC record (regardless of the 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.

So for instance, on top of the following input (the example shows only the relevant tags):

245 10$aFondamenti di teoria dei circuiti /$cCharles A. Desoer, Ernest S. Kuh ; prefazione all'edizione italiana di G. Biorci

100 1 $aDesoer, Charles A.

700 1 $aBiorci, Giuseppe

700 1 $aKuh, Ernest S.

Authify – Relator term detection

The system will answer with a response like this:


```
{
 "id": "LE02614324",
 "statements": [
  "245 10$aFondamenti di teoria dei circuiti /$cCharles A. Desoer, Ernest S. Kuh ; prefazione all'edizione italiana di G. Biorci"
 ],
 "names": [
  "100 1 $aDesoer, Charles A.",
  "700 1 $aBiorci, Giuseppe",
  "700 1 $aKuh, Ernest S."
 ],
 "responsibilities": {
  "content": {
   "http://id.loc.gov/vocabulary/relators/oth": {
    "headings": [
     {
      "name": "Biorci, Giuseppe"
     }
    ],
    "relatorTermCode": "oth",
    "relatorTermText": "Other"
   },
   "http://id.loc.gov/vocabulary/relators/aut": {
    "headings": [
     {
      "name": "Kuh, Ernest S."
     },
     {
      "name": "Desoer, Charles A."
     }
    ],
    "relatorTermCode": "aut",
    "relatorTermText": "Author"
   }
```

Authify – Relator term detection

In these examples you can see that two main roles have been detected:

- **authors**
- **other** (unclassified role).

The "other" role is a catch-all role used when no valuable information can be obtained from the analysis.

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 by using a set of tokens that could identify such role (e.g. edited by, by, illustrated by).

Entity detection (example 1)

=LDR 00833nam a2200217 i 4500
=001 LE02519084
=005 20020503192020.0
=008 970703s1990\\\\uk\\\\\\\\\\\\\|||\|\eng\\
=020 \\$a0415030889
=040 \\$aFac. Economia$bita
=082 0\$a820.9
=100 1\$aStephens, John
=245 10$aLiterature, language and change **:$bfrom Chaucer to the present /$cJohn Stephens and Ruth Waterhouse**
=260 \\$bRoutledge,$cc1990
=300 \\$aix, 293 p. ;$c20 cm.
=650 \4$aLetteratura inglese$xStoria e critica
=650 \4$aLingua inglese
=700 1\$aWaterhouse, Ruth

Entity detection - Authify/Detect response (1)

Response Body servizio authify/detect:


```
{
 "id": "LE02519084",
 "statements": [
  "245 10$aLiterature, language and change :$bfrom Chaucer to the present /$cJohn Stephens and Ruth Waterhouse"
 ],
 "names": [
  "100 1 $aStephens, John",
  "700 1 $aWaterhouse, Ruth"
 ],
 "responsibilities": {
  "content": {
   "http://id.loc.gov/vocabulary/relators/aut": {
    "headings": [
     {
      "name": "Stephens, John"
     },
     {
      "name": "Waterhouse, Ruth"
     }
    ],
    "relatorTermCode": "aut",
    "relatorTermText": "Author"
   }
  }
 }
}
```

Entity detection (example 2)

=LDR 01127pam a2200325 a 4500
=001 7486885
=005 20150720142401.0
=008 090901t20152015mauab\\\\b\\\\001\0\eng\\
=010 \\$a 2009036444
=020 \\$a9781566567879$qpaperback
=020 \\$a1566567874$qpaperback
=024 \\$a99963025763
=035 \\$a(OCoLC)908588988
=035 \\$a(OCoLC)ocn908588988
=035 \\$a(NNC)7486885
=040 \\$aDLC$beng$cDLC$dBTCTA$dBDX$dOCLCF$dOCLCO$dMNM$dNhCcYBP
=043 \\$aa-is---$aawba---
=050 00$aDS109.93$b.J48 2015
=082 00$a956.94/4205$222
=245 00$aJerusalem interrupted :$bmodernity and colonial transformation 1917-present /$cedited and introduced by Lena Jayyusi.
=260 \\$aNorthampton, Mass. :$bOlive Branch Press,$c2015.
=300 \\$axxii, 499 p. :$bill., maps ;$c24 cm.
=504 \\$aIncludes bibliographical references and index.
=651 \0$aJerusalem$xHistory$y20th century.
=651 \0$aJerusalem$xHistory$y21st century.
=651 \0$aJerusalem$xInternational status.
=650 \0$aArab-Israeli conflict.
=700 1\$aJayyusi, Lena.

Entity detection - Authify/Detect response (2)


```
{
  "id": "7486885",
  "statements": [
    "245 00$aJerusalem interrupted :$bmodernity and colonial transformation 1917-present /$cedited and
     introduced by Lena Jayyusi."
  ],
  "names": [
    "700 1 $aJayyusi, Lena."
  ],
  "responsibilities": {
    "content": {
      "http://id.loc.gov/vocabulary/relators/edt": {
        "headings": [
          {
            "name": "Jayyusi, Lena."
          }
        ],
        "relatorTermCode": "edt",
        "relatorTermText": "Editor"
      }
    }
  }
}
```

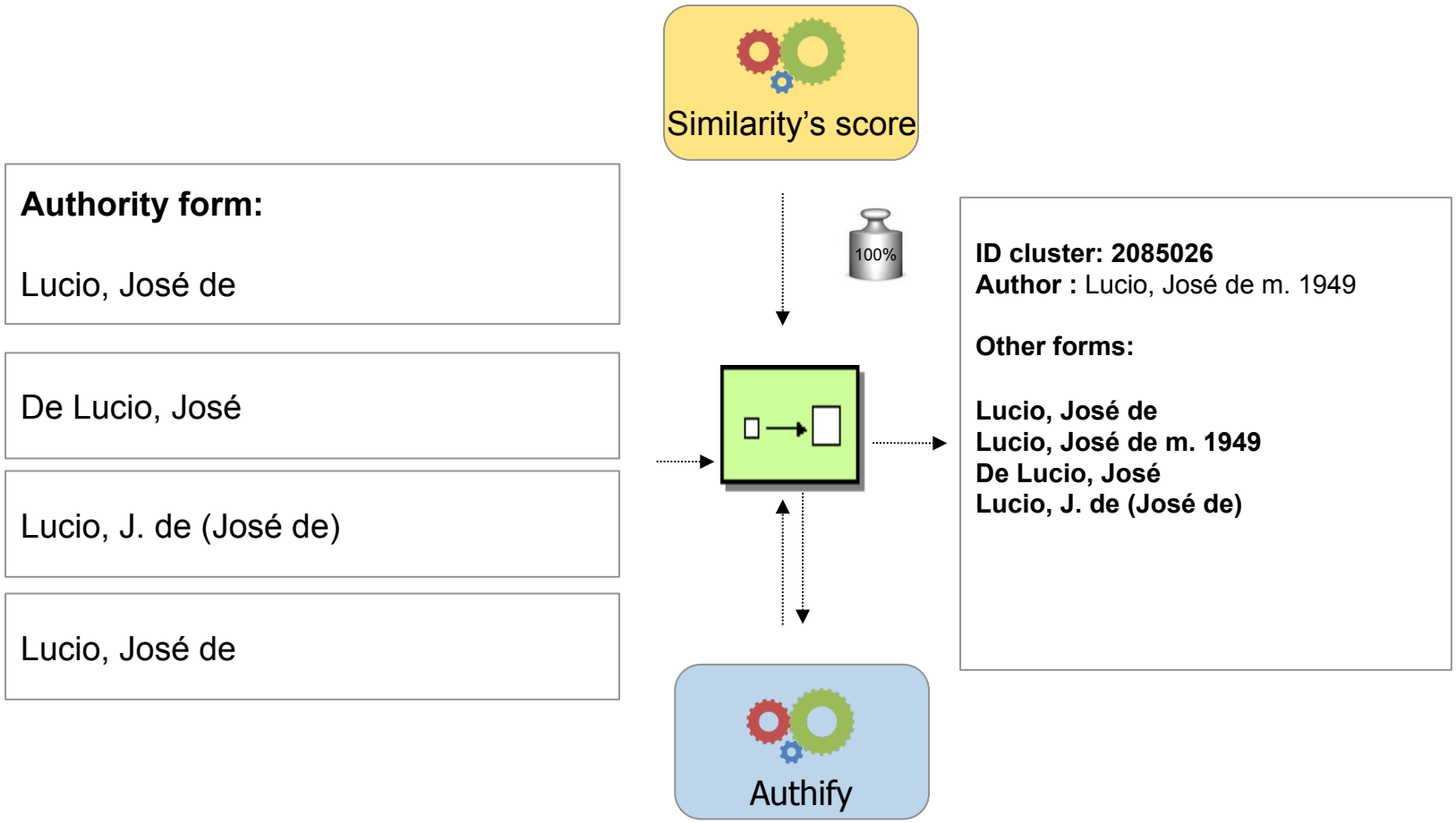
Entity detection (example 3) – Critical case

=LDR 01145nam a2200241 i 4500
=001 LE01988135
=005 20020503105244.0
=008 010702s1999\\\\it\\\\\\\\\\\\000\0\lat\\
=020 \\$a882092868X
=040 \\$aDip.to Beni Arti e Storia$bita
=082 0\$a264.024
=245 00$aBreviarium Romanum :$beditio princeps, 1568 /$cedizione anastatica, introduzione e appendice a cura di Manlio Sodi, Achille Maria Triacca ; con la collaborazione di Maria Gabriella Foti ; presentazione di Virgilio Noè
=260 \\$aCittà del Vaticano :$bLibreria editrice Vaticana,$c1999
=300 \\$aXXII, 1056 p. ;$c25 cm
=440 \0$aMonumenta liturgica concilii tridentini$v3
=700 1\$aSodi, Manlio
=700 1\$aTriacca, Achille Maria
=700 1\$aFoti, Maria Gabriella
=700 1\$aNoè, Virgilio
=907 \\$a.b10000914$b02-04-14$c29-05-02

Entity detection - Authify/Detect response (3)


```
{
 "id": "LE01988135",
 "statements": [
  "245 00$aBreviarium Romanum :$beditio princeps, 1568 /$cedizione anastatica, introduzione e appendice a cura di Manlio Sodi, Achille Maria
      Triacca ; con la collaborazione di Maria Gabriella Foti ; presentazione di Virgilio Noè"
 ],
 "names": [
  "700 1 $aFoti, Maria Gabriella",
  "700 1 $aNoè, Virgilio",
  "700 1 $aSodi, Manlio",
  "700 1 $aTriacca, Achille Maria"
 ],
 "responsibilities": {
  "content": {
   "http://id.loc.gov/vocabulary/relators/oth": {
    "headings": [
     {
      "name": "Sodi, Manlio"
     },
     {
      "name": "Triacca, Achille Maria"
     },
     {
      "name": "Foti, Maria Gabriella"
     },
     {
      "name": "Noè, Virgilio"
     }
    ],
    "relatorTermCode": "oth",
    "relatorTermText": "Other"
   }
  }
 }
```

Name cluster process



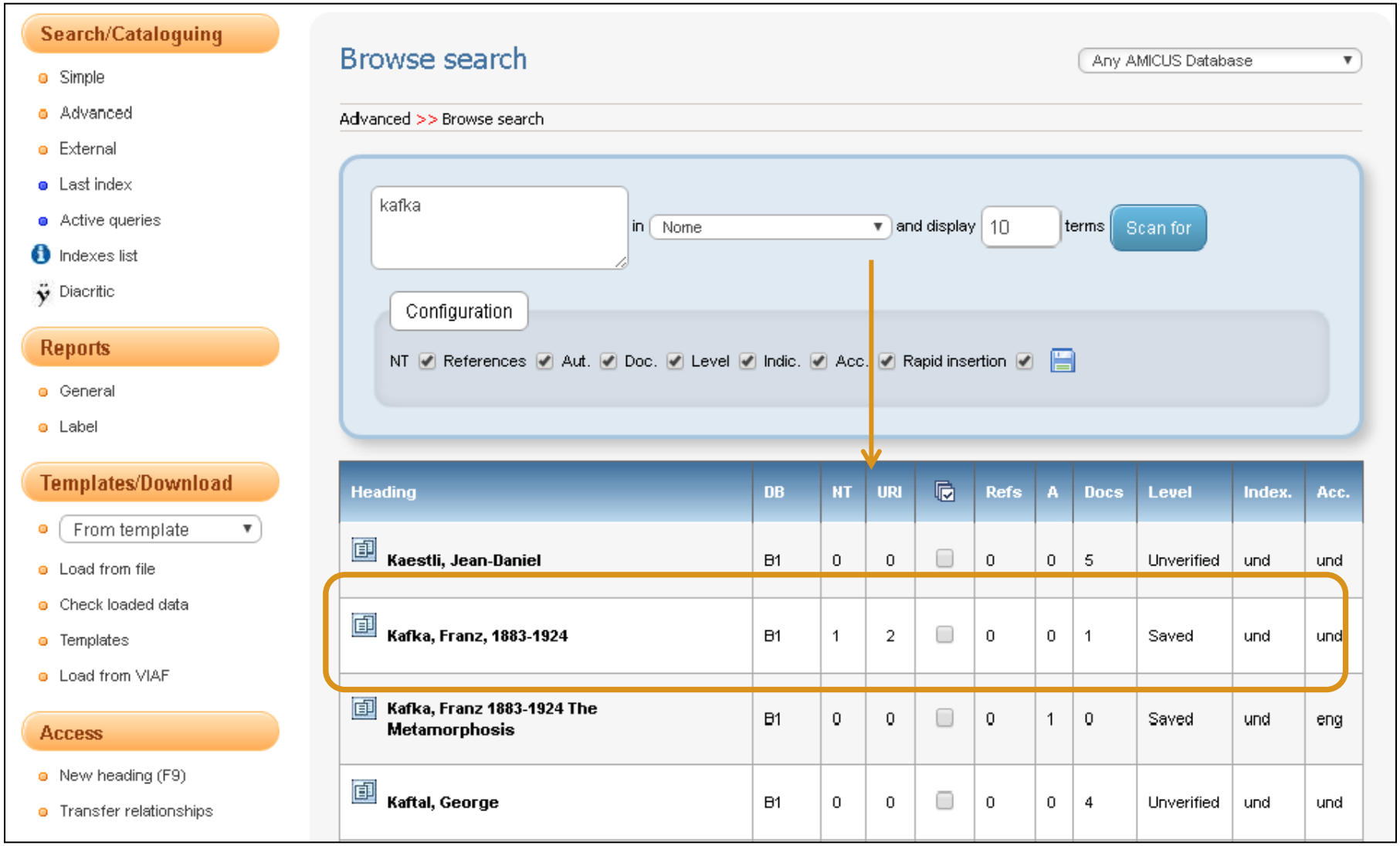
Reconciliation & Enrichment – Manual Procedures

Manual process to produce clusters

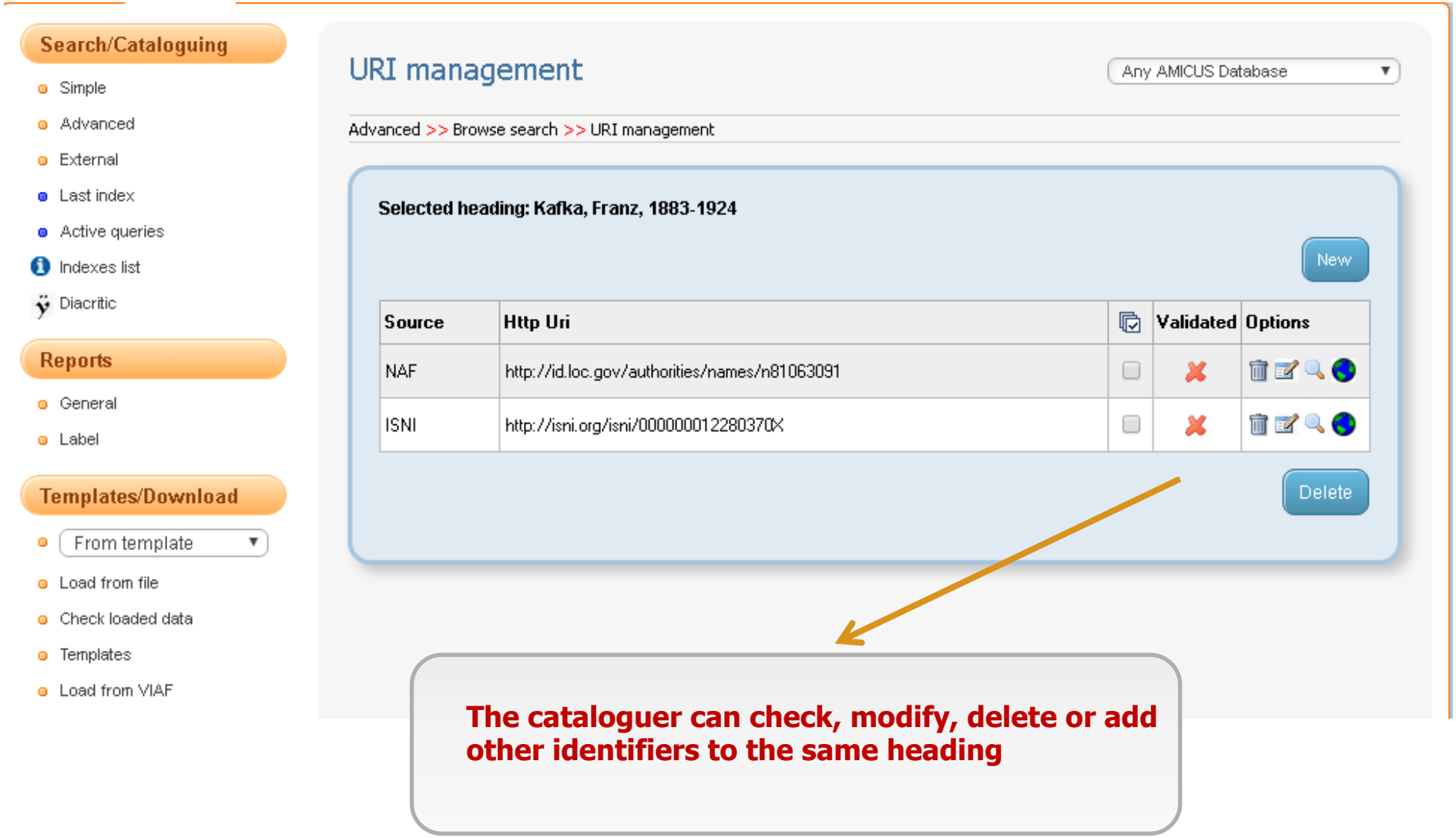
The same resulting entity enrichment, carried out using manual processes in the cataloguing workflow, enables a more precise verification of the results: the WeCat cataloguing module of OLISuite provides a «URI Management System» to manage identifiers for each access point or heading.

The availability of API and web services allows the use of external sources (such as NAF, ISNI and VIAF) and the association of the heading with the URIs that identify it in each of the projects.

URI Management System (OLISuite/WeCat screen)



URI Management System (OLISuite/WeCat screen)



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, or removed from it as well as modifications to resources

• the date of the update

• the particular operation performed

• the status of an URI (for instance valid or invalid)

• the URI Aliases

Access Points and URIs

Access points and URIs

The URIs associated with a heading can be used in various advantageous ways.

In the data export/conversion process we can choose how many URIs to make available for each heading, how to associate them to the heading, how to show them in relation to data use and formats.

This export considers different customers profiles (so that each one can choose which sources to use and how to register URI).

Access point and URIs (example 1)

As $0 associated to access point in the MARC bibliographic record:

=LDR 00560nam a2200181 4500
=001 000000127573
=003 CaOOAMICUS
=005 20160108094931.0
=008 160107s\\\\\\\\it\\\\\\\\\\\\000\u\ita\r
=040 \\$aAtCult$bita
=100 1\$aKafka, Franz,$d1883-1924$0(isni) 0000 0001 2280 370X.
=245 03$aLa metamorfosi /$cFranz Kafka.
=260 \\$aMilano :$bLa spiga,$c2002.
=300 \\$a61 p.; $c18 cm
=336 \\$atext$2rdacontent
=337 \\$aunmediated$2rdamedia
=338 \\$avolume$2rdacarrier
=997 \\$aPS

Access point and URIs (example 2)

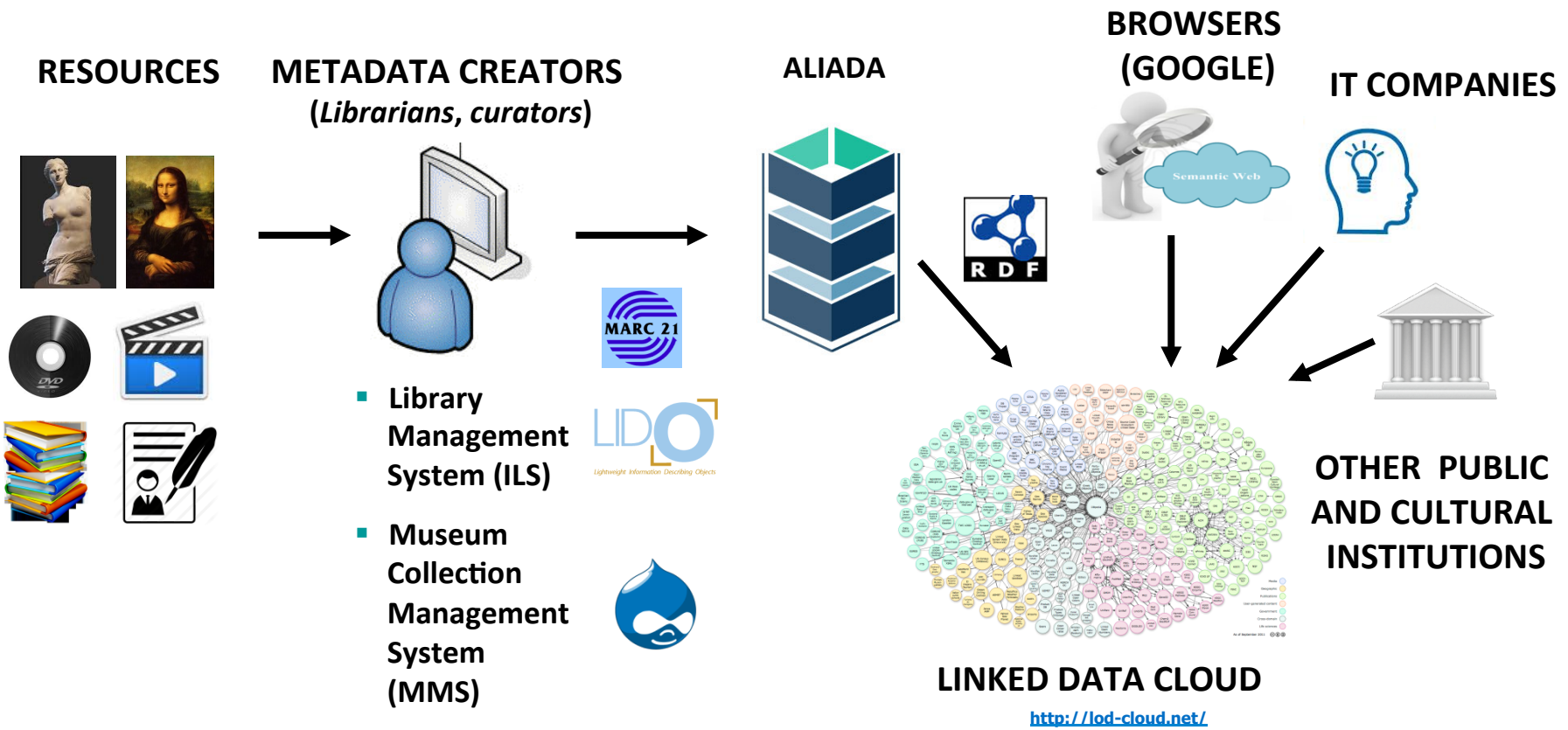
As specific tag in the MARC authority record:

=LDR 00698nz 2200145 4500
=001 000000000617
=005 20160108125155.0
=008 751003s1974\\\\enk\\\\\\\\\\000\1\eng\\
=024 7\$a56611857$2viaf
=024 7\$a0000000122803 70X$2isni
=040 \\$aPS$bita
=100 1\$aKafka, Franz$d1883-1924
=400 1\$aKafka, F.$q(Franz)$d1883-1924
=670 \\$aWikipedia, Oct. 25, 2012$bFranz Kafka; born 3 July 1883 in Prague; died 3 June 1924 Kierling near Vienna; an influential German-language writer of novels and short stories, regarded by critics as one of the most influential authors of the 20th century. Kafka was a Modernist and heavily influenced other genres, including existentialism)

Conversion in RDF/BIBFRAME

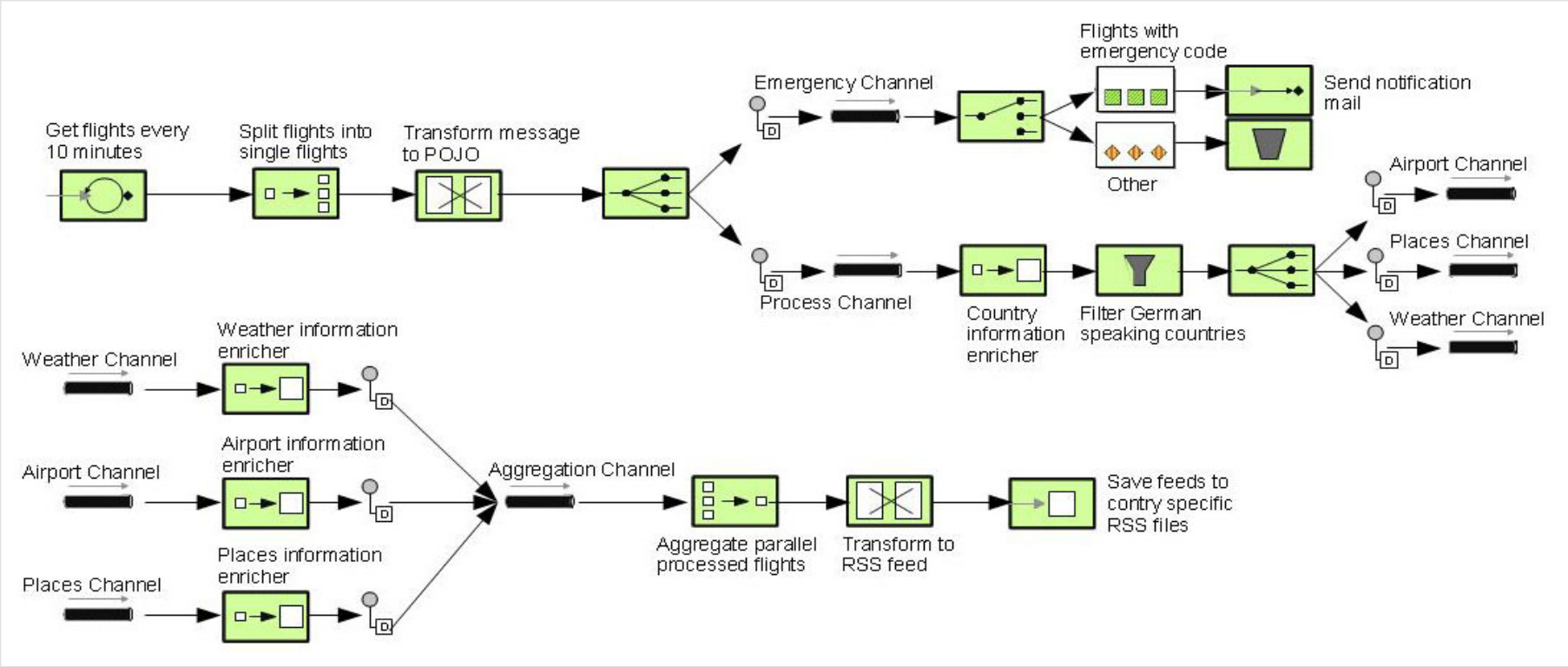
Lodify: the evolution of Aliada for BIBFRAME conversion

The conversion process from any format to RDF



Lodify - The Asynchronous pipeline of the tool

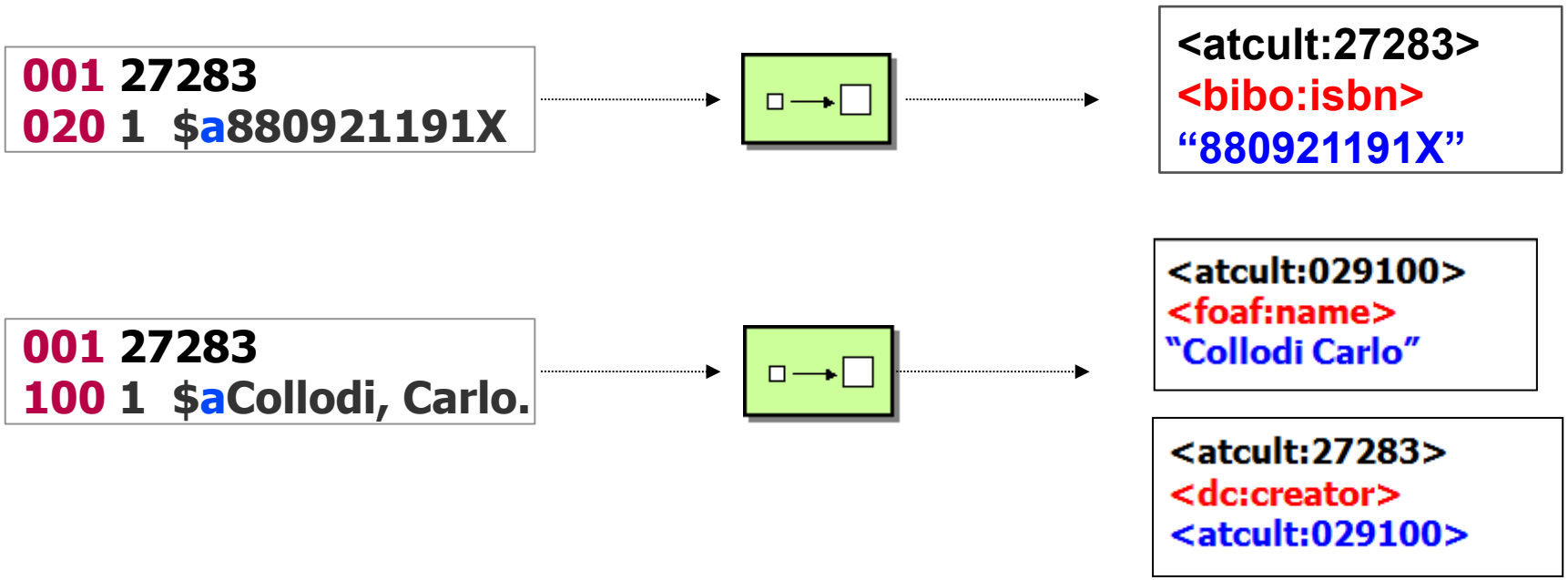
Lodify building block, realized through Apache Camel. The process is split into atomic pieces (processors), each of these responsible for a small part of the overall task. Each processor can act as a splitter or aggregator and can achieve content manipulation of the incoming message.



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.



Blazegraph: the actual triple-store used in the project

”Blazegraph is an ultra-scalable, high-performance graph database with support for the Blueprints and RDF/SPARQL APIs. It supports up to 50 Billion edges on a single machine”*

* https://www.blazegraph.com/

Trust and Provenance

Guarantee of authority and quality in the new Linked Data environment

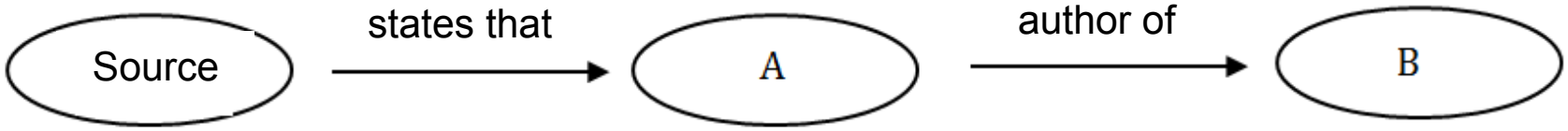
Need to guarantee the accuracy of this information.

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 and quality in the new Linked Data 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.

Link 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 the 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

Reconciliation of the same instances present in different catalogues (attention, it's in the test db):
http://dev-vde.atcult.it/sharevde/search?t_cluster_id=7961;Bufera%20e%20altro&v=ll&dls=true&l=en

Triple Store Query Examples in Blazegraph

Blazegraph: query examples

http://share-vde.org:9999/blazegraph/#query

1. Select all entities (type Person) that cointain the term 'Federico' :

```
select *
where {
  ?s rdfs:label ?o .
  ?s          rdf:type      <http://bibframe.org/vocab/Person> .
     FILTER regex(?o, "federico", "i")
  }
```

2. Select all tripes that have as Subject the work ID xxxxxx (example, the Work ID 95194):

```
select *
where {
  <http://rdf.share-vde.org/Work/95194> ?p ?o
  }
```

3. Select all variant labels for Person with ID xxxx (example, 546261)

```
select *
where {
  <http://rdf.share-vde.org/Agent/546261> rdfs:label ?o
  }
```

Blazegraph: query examples

4. Retrieve the ID of Person wich VIAF ID is equal to xxxx (example http://viaf.org/viaf/42027007/

```
construct
where
 {
   ?s owl:sameAs <http://viaf.org/viaf/42027007/>
     }
```

5. Retrieve all labels and predicate owl:sameAs for Person «xxx» (example Pirandello)

```
construct {
      <http://rdf.share-vde.org/Agent/138504> owl:sameAs ?o .
       <http://rdf.share-vde.org/Agent/138504> rdfs:label ?p .
  }
where
 {
   <http://rdf.share-vde.org/Agent/138504> owl:sameAs ?o .
   <http://rdf.share-vde.org/Agent/138504> rdfs:label ?p .

     }
```

Please note: taking into account the suggestion coming from Libraries during the Phase 1 of the project, we have discarded the use of the predicate owl:sameAs in favor of http://www.loc.gov/mads/rdf/v1#isIdentifiedByAuthority

PCC Operations Committee Meeting (OpCo)
May 4-5, 2017
Library of Congress
Washington, DC

PCC
Program for Cooperative Cataloging
BIBCO / CONSER / NACO / SACO

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

Michele Casalini
Managing Director, Casalini Libri
michele@casalini.it

Tiziana Possemato
Chief Information Officer, Casalini Libri
Director, @Cult
tiziana.possemato@casalini.it
tiziana.possemato@atcult.it

casalinilibri

@Cult