One Supplier's Approach to BIBFRAME/Linked Data

Tiziana Possemato
Chief Information Officer

ALA Midwinter 2016
BIBFRAME Update Forum Agenda

Current activity and infrastructure

Casalini Libri produces, for publications from Romance language countries, more than 20,000 original bibliographic records in RDA as a member of the Program for Cooperative Cataloguing (PCC)

Bibliographic records are created using the WeCat cataloguing module of the OLISuite ILS (developed by @Cult) in native MARC 21/RDA format

Authority control on names, titles and series headings

Maintenance of the authority databases (NACO and SACO)

The three areas of activities towards the BIBFRAME/Linked Data environment

In order to be ready with the concept of evolution from a web of documents (the traditional web) to a web of data (the semantic web) and with the started transition from MARC formats to Linked Open Data, we are investing in the study and implementation of projects that go in this direction, more recently with emphasis on the model proposed by the Bibliographic Framework Initiative (BIBFRAME).

To this aim, we focus on the following three areas of activities.

The three areas of activities towards the BIBFRAME/Linked Data environment

1. The enrichment of MARC records to simplify BIBFRAME conversion

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

3. The creation of a FRBR/BIBFRAME layer starting from bibliographic and authority records, to help librarians and end users in LOD fruition

1. Enrichment of MARC records to simplify BIBFRAME conversion

Additional MARC tag fulfilment and treatment in order to simplify the conversion into BIBFRAME without losing content: the MARC record is enriched (through manual and automatic processes) with tags and subfields, in particular with the addition of a certain number of local and global identifiers.

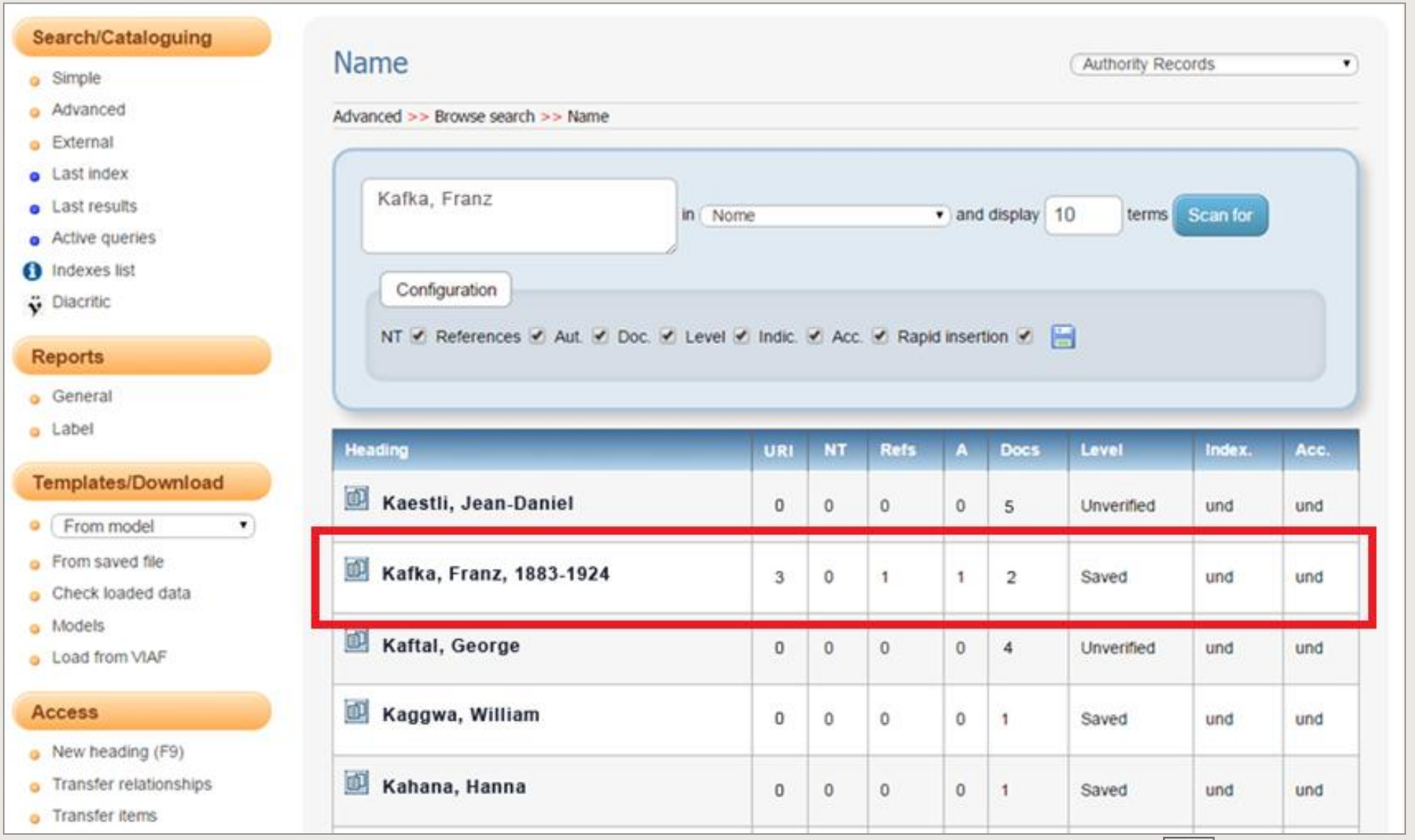
This builds the precondition to allow the conversion of MARC into Linked Open Data by any party.

URI Management System

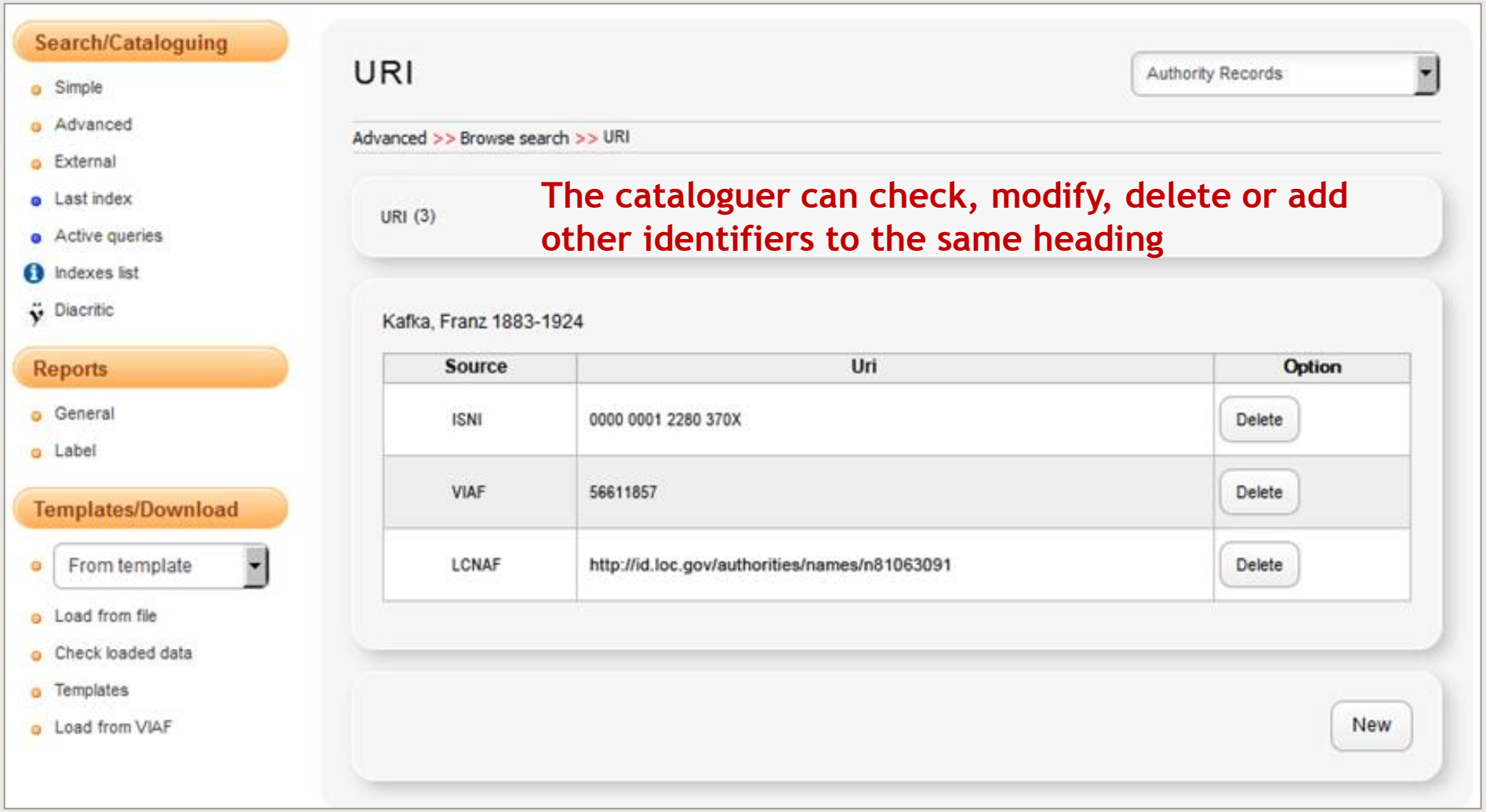
We are improving in the WeCat cataloguing module of OLISuite a «URI Management System», to manage identifiers for each access point or heading.

See as an example in the following slides the authorized access point for Franz Kafka and, in the first column (URI), the number of URIs associated to the heading.

URI Management System (WeCat screen)



URI Management System (WeCat screen)



casalinilibri

Access points and URIs

The URIs associated to a heading can be used in varying and useful 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.

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)

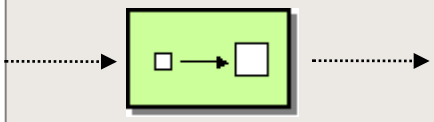
As specific tag in the MARC authority record:

=LDR 00698nz 2200145 4500
=001 000000000617
=005 20160108125155.0
=008 751003s1974\\\\enk\\\\\\\\\\000\1\eng\\
=024 7\$a56611857$2viaf
=024 7\$a000000012280370X$2isni
=040 \\$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)

Access point and URIs
(example 3)

As RDF property in the triples produced in the conversion process:

```
001     000000000617
024 7  $a56611857$2viaf
024 7  $a0000000012280370$2isni
100 1  $aKafka, Franz
```

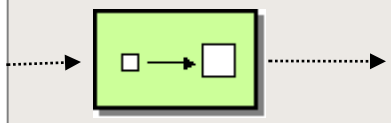


```
<atcult:eb-617>
<rdf:type>
<bf:Identifier>

<atcult:eb-617>
<bf:local>
<atcult:617-kafka-franz>

<atcult:eb-617>
<bf:identifierValue>
“617”

<atcult:eb-617>
<owl:sameAs>
“http://viaf.org/viaf/56611857”

<atcult:eb-617>
<owl:sameAs>
“http://isni-url.oclc.nl/isni/
 000000012280370”
```

Access point and URIs
(example 4)

Another example of identifiers used as RDF property of an entity type *Person*:

```
001     000000000617
024 7  $a56611857$2viaf
024 7  $a0000000012280370$2isni
100 1  $aKafka, Franz
```



```
<atcult:617-kafka-franz>
<rdf:type>
<bf:Person>

<atcult:eb-617>
<rdf:type>
<bf:Identifier>

<atcult:eb-617>
<bf:local>
<atcult:617-kafka-franz>

<atcult:eb-617>
<bf:identifierValue>
“617”

<atcult:617-kafka-franz>
<bf:hasAuthority>
<rdf:resource=
“http://viaf.org/viaf/56611857”>

<atcult:617-kafka-franz>
<bf:hasAuthority>
<rdf:resource=
“http://isni-url.oclc.nl/isni/
 0000000012280370”
```

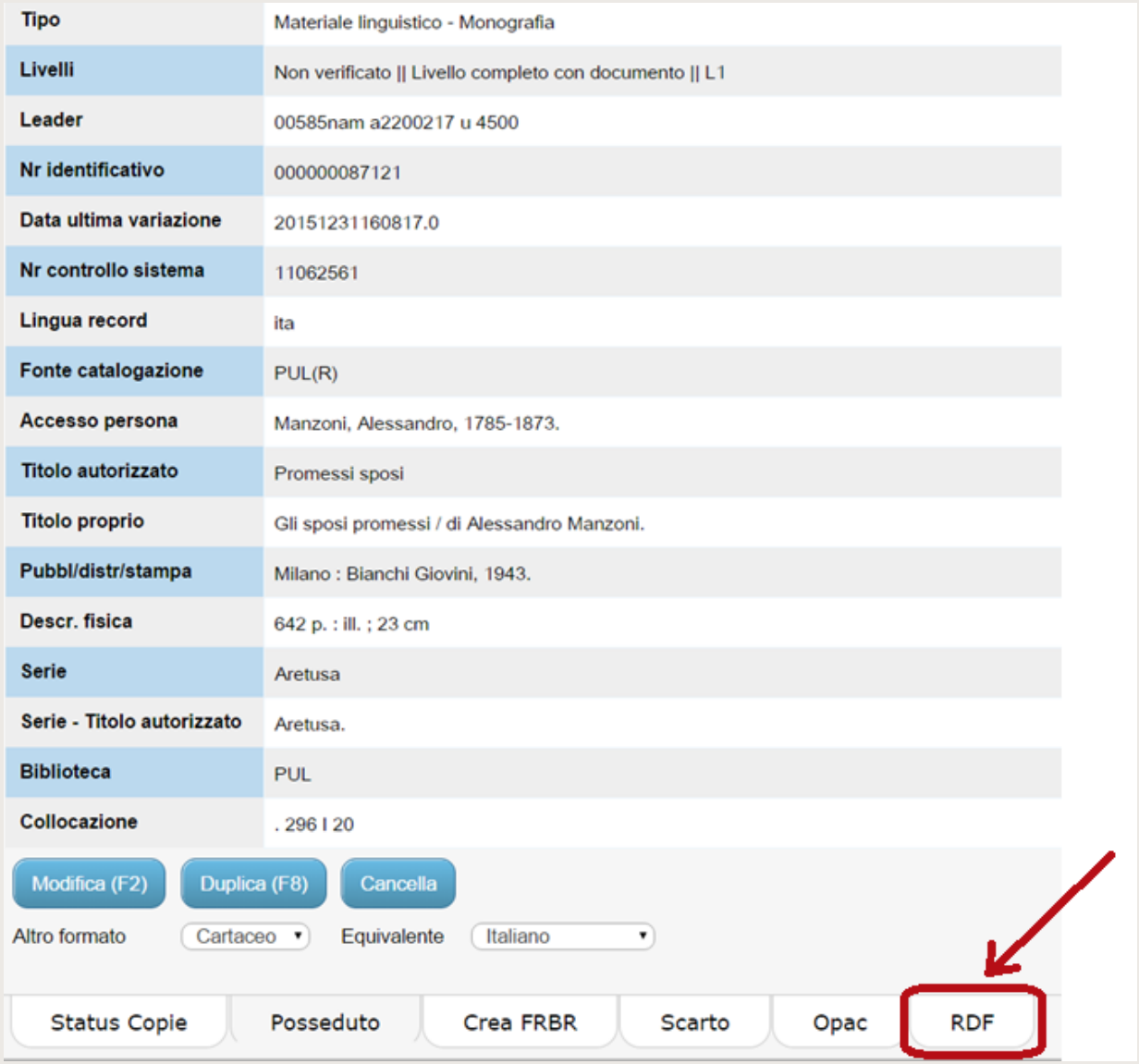
2. Use of a framework to automate the conversion from MARC to RDF

The conversion from different formats to RDF is realized within the WeCat cataloguing module, that embeds micro-agents software, each one mapped on a specific MARC tag/subfield in order to convert and export it as Linked Open Data.

The same conversion process can be activated independently from an ILS, using data in different formats (MARC, xml, Lido, etc.).

Automatic conversion in RDF is realized through the **ALIADA** framework, applying the BIBFRAME vocabulary.

OliSuite/WeCat: from MARC 21 to RDF



Tipo	Materiale linguistico - Monografia
Livelli	Non verificato || Livello completo con documento || L1
Leader	00585nam a2200217 u 4500
Nr identificativo	000000087121
Data ultima variazione	20151231160817.0
Nr controllo sistema	11062561
Lingua record	ita
Fonte catalogazione	PUL(R)
Accesso persona	Manzoni, Alessandro, 1785-1873.
Titolo autorizzato	Promessi sposi
Titolo proprio	Gli sposi promessi / di Alessandro Manzoni.
Pubbl/distr/stampa	Milano : Bianchi Giovini, 1943.
Descr. fisica	642 p. : ill. ; 23 cm
Serie	Aretusa
Serie - Titolo autorizzato	Aretusa.
Biblioteca	PUL
Collocazione	. 296 I 20

Example of the conversion process activated in the ILS: at the end of the cataloguing workflow, the user can click on the RDF button to convert/publish the record as RDF triples, using the ALIADA framework

15

ALIADA: the RDF conversion & publication framework

The framework used to convert and publish data in RDF is ALIADA:
Automatic publication under LInked DAta Paradigm of library Data.

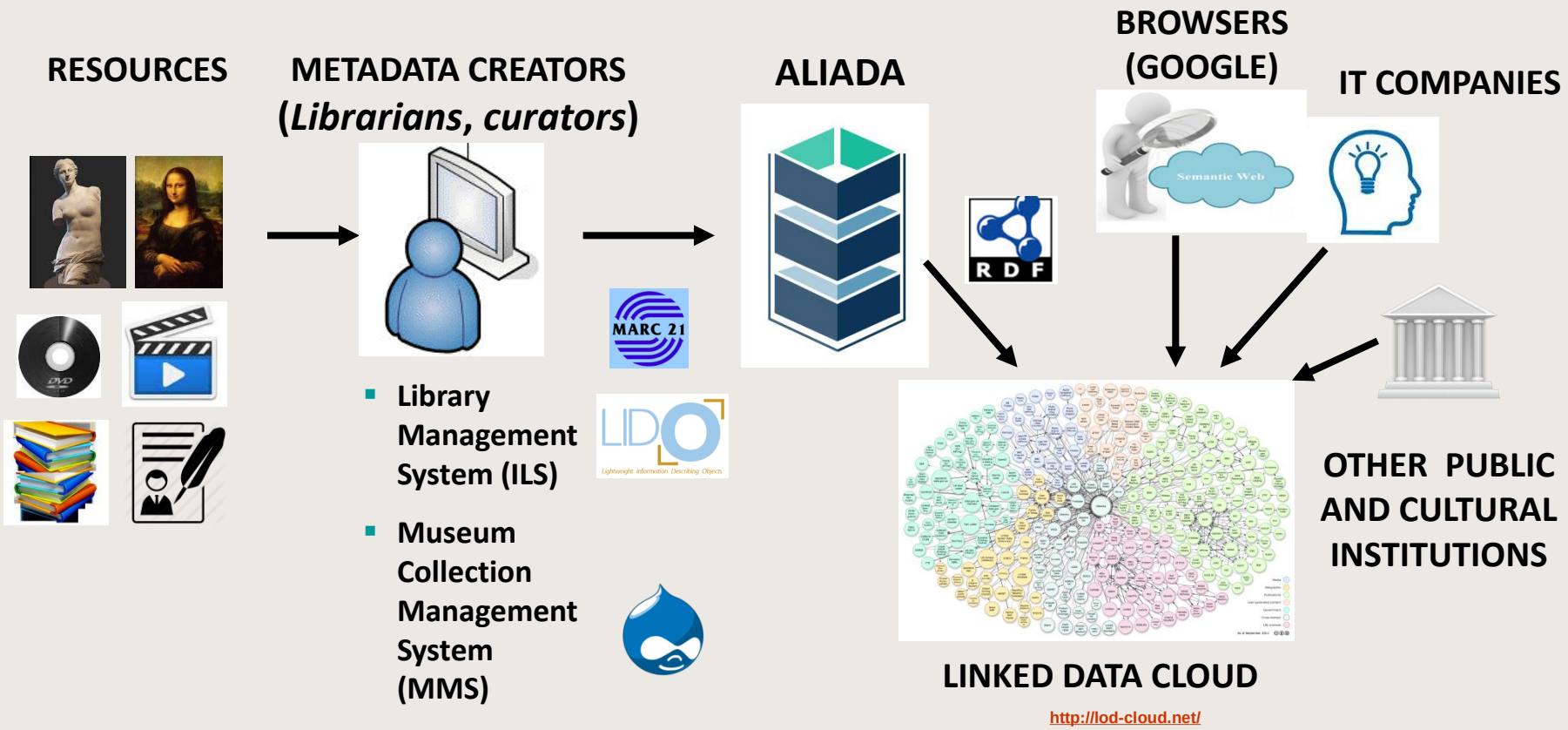
The project is co-financed by the **European Union's Research and Innovation funding programme for 2007-2013 (FP7).**

- 5 partners from 3 different countries (**Italy**, **Spain**, **Hungary**)
- 2 IT companies: **@CULT**, **SCANBIT**
- 2 museums: **ARTIUM** (Spain), **Museum of Fine Arts Budapest** (Hungary)
- 1 research institute: **TECNALIA** (Spain)

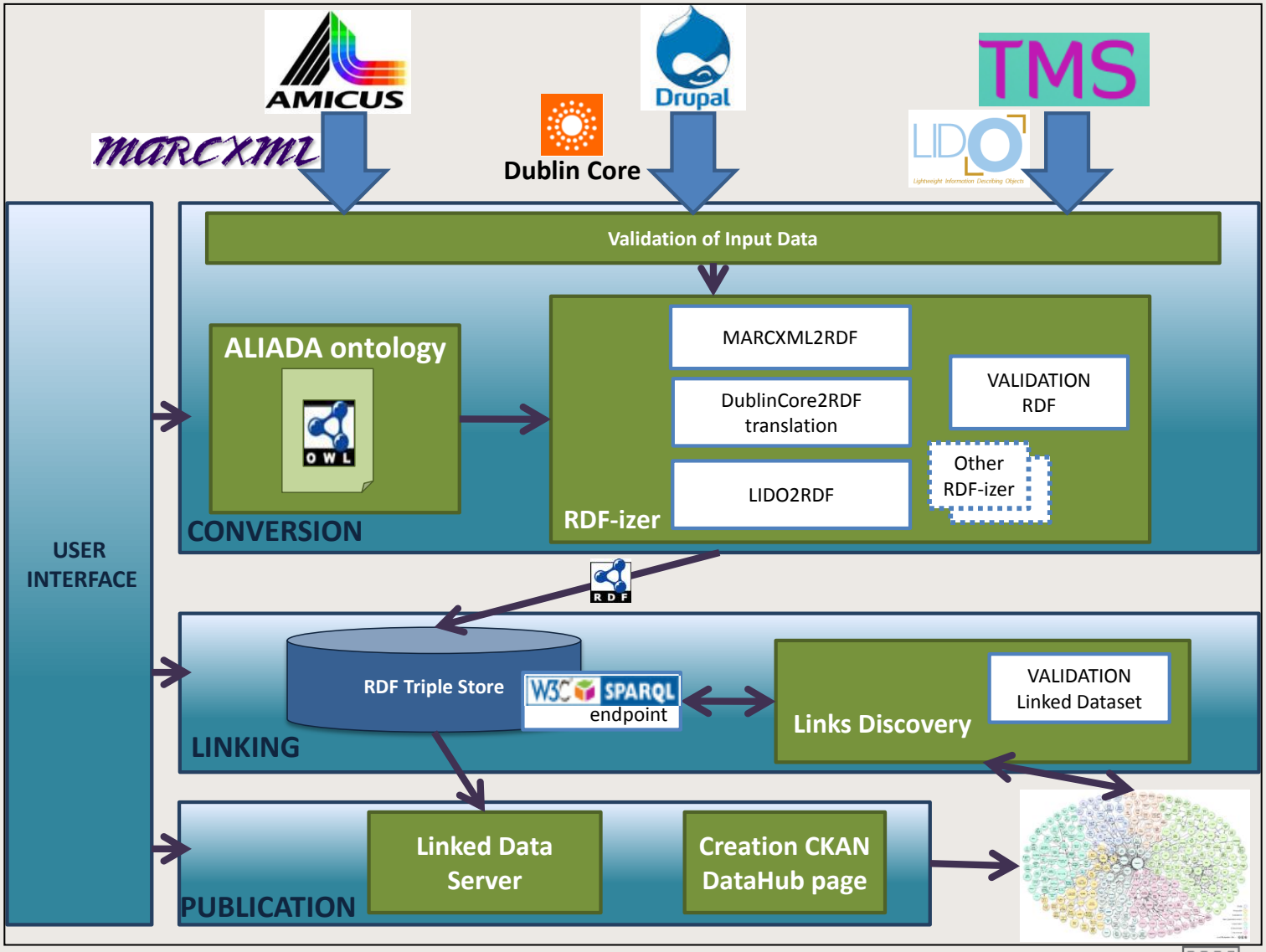
Project duration: 24 months (from November 2013 to October 2015)

Results available as open-source at www.aliada-project.eu

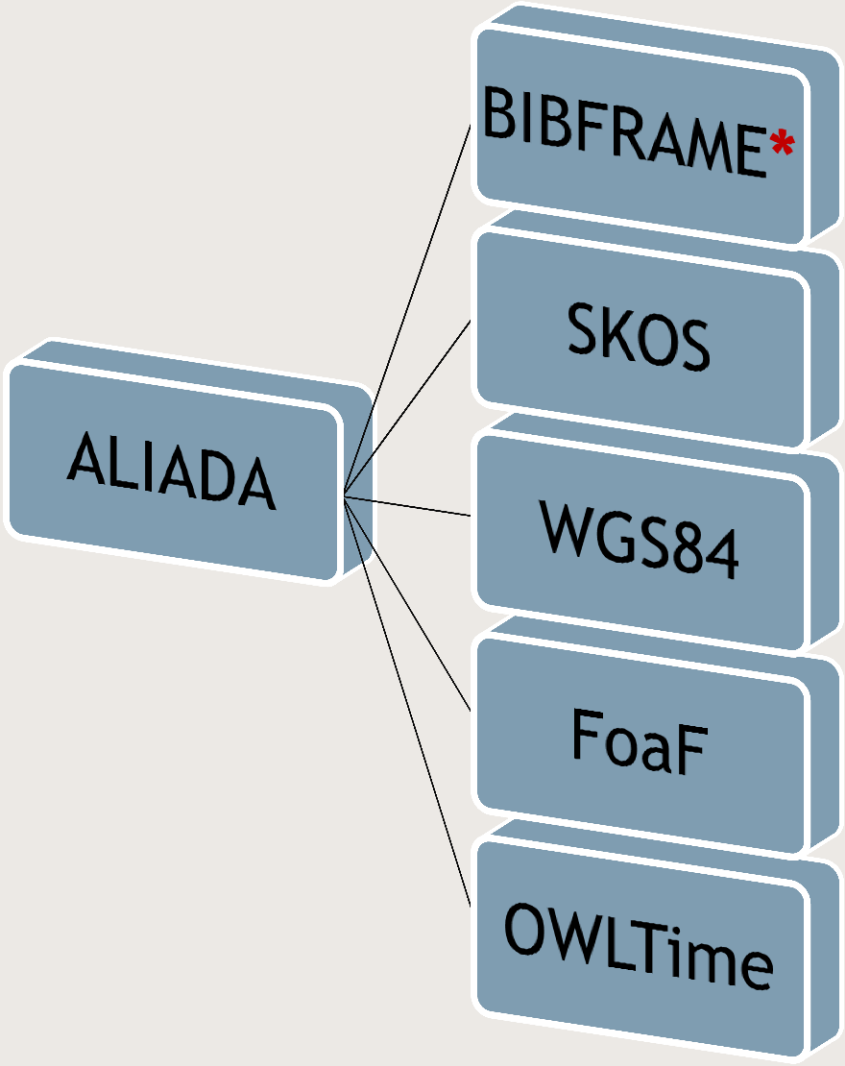
The conversion process from any format to RDF



ALIADA conversion & publishing layers



Ontologies used in the framework



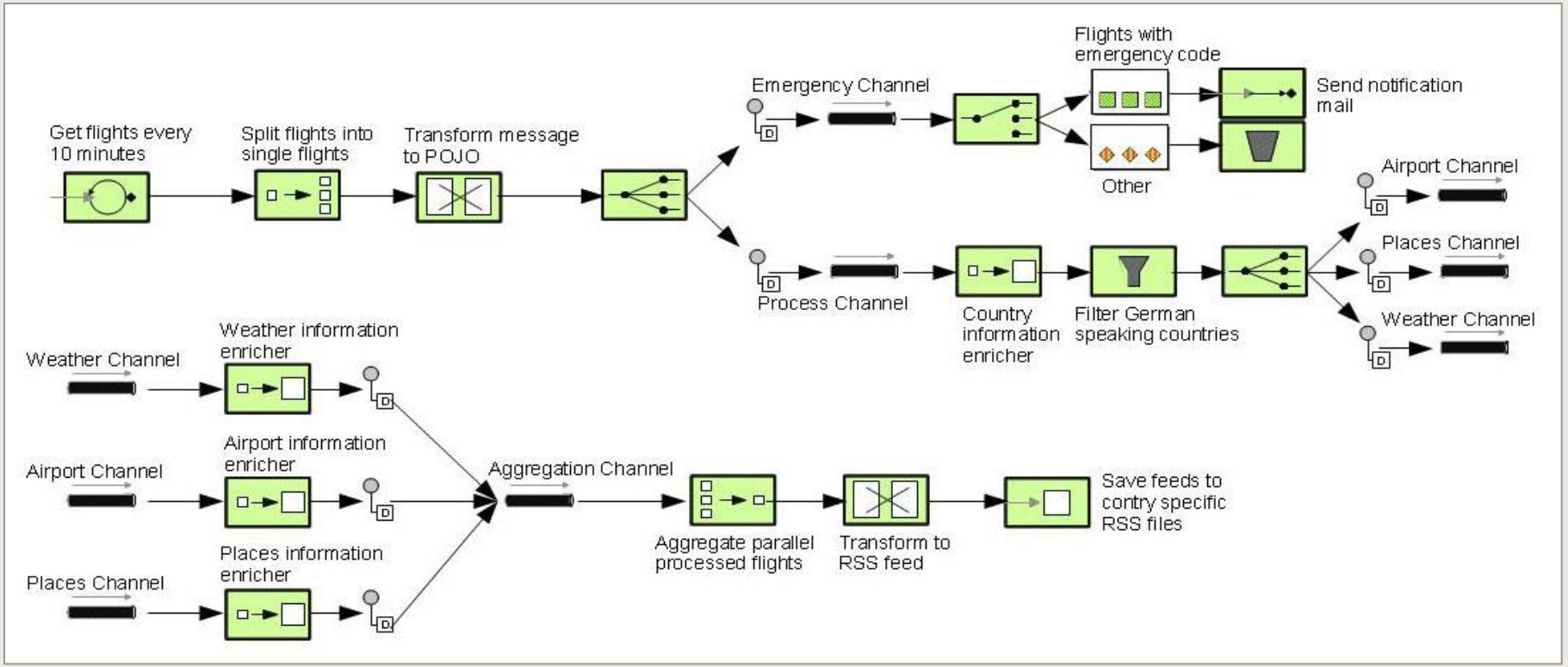
Additional ontologies used in the ALIADA framework:

- FRBRoo (part of the first release)
- DCMI Metadata Terms
- RDF Schema
- RDA elements

**BIBFRAME added in the current release in progress*

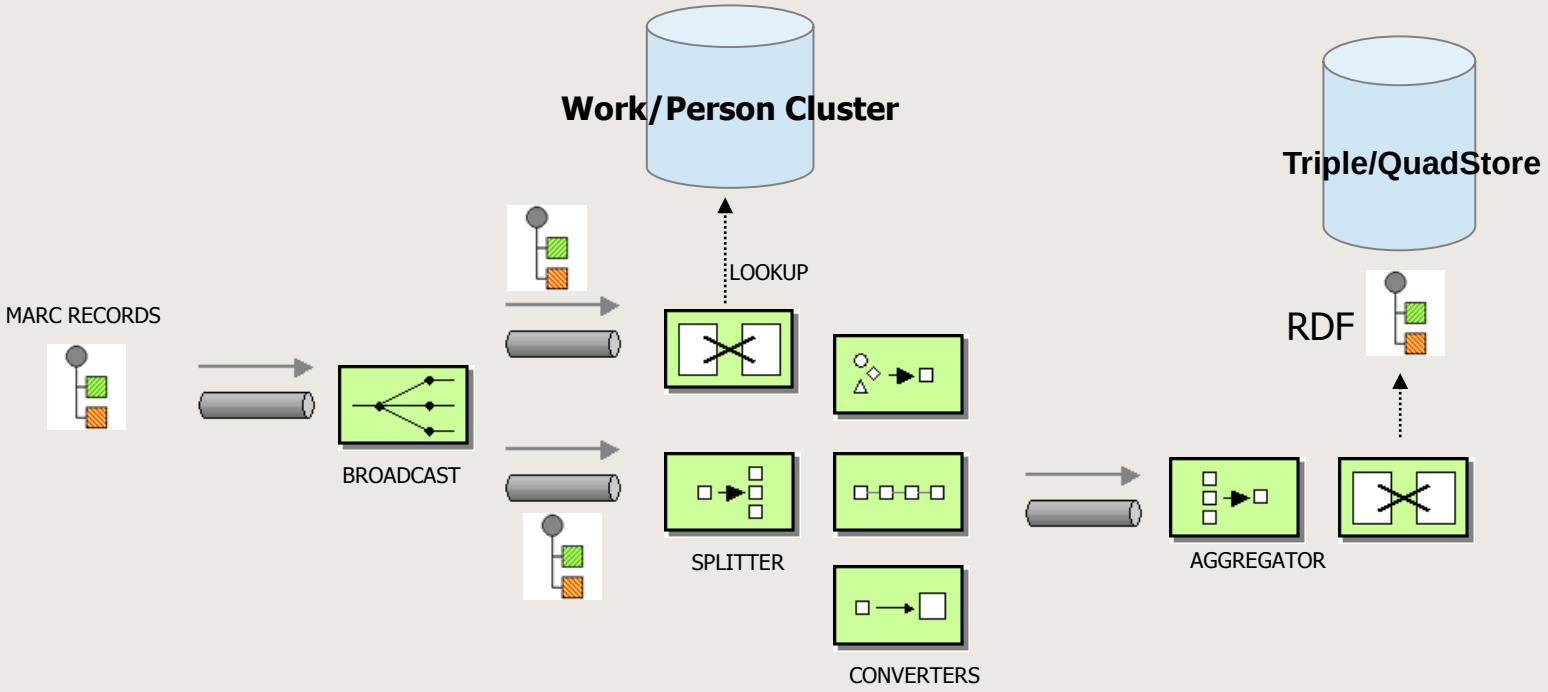
The asynchronous pipeline

ALIADA 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.



It's just an asynchronous pipeline!

The high-level workflow in ALIADA is as follows: before proceeding with the conversion of a record, the pipeline looks up the Work/Person cluster to gather information about a given entity, in order to disambiguate and uniquely identify things in the out-coming dataset.

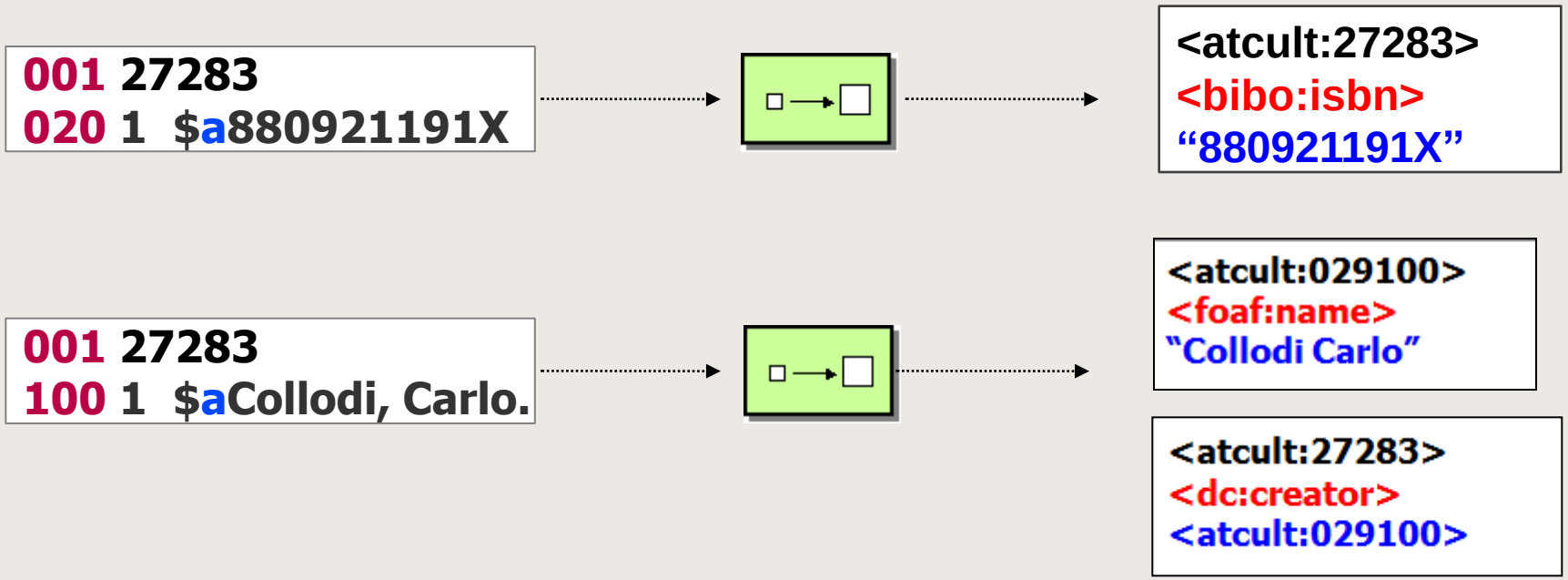


A set of MARC records go through the pipeline, which splits, processes and converts them.

ALIADA Conversion templates

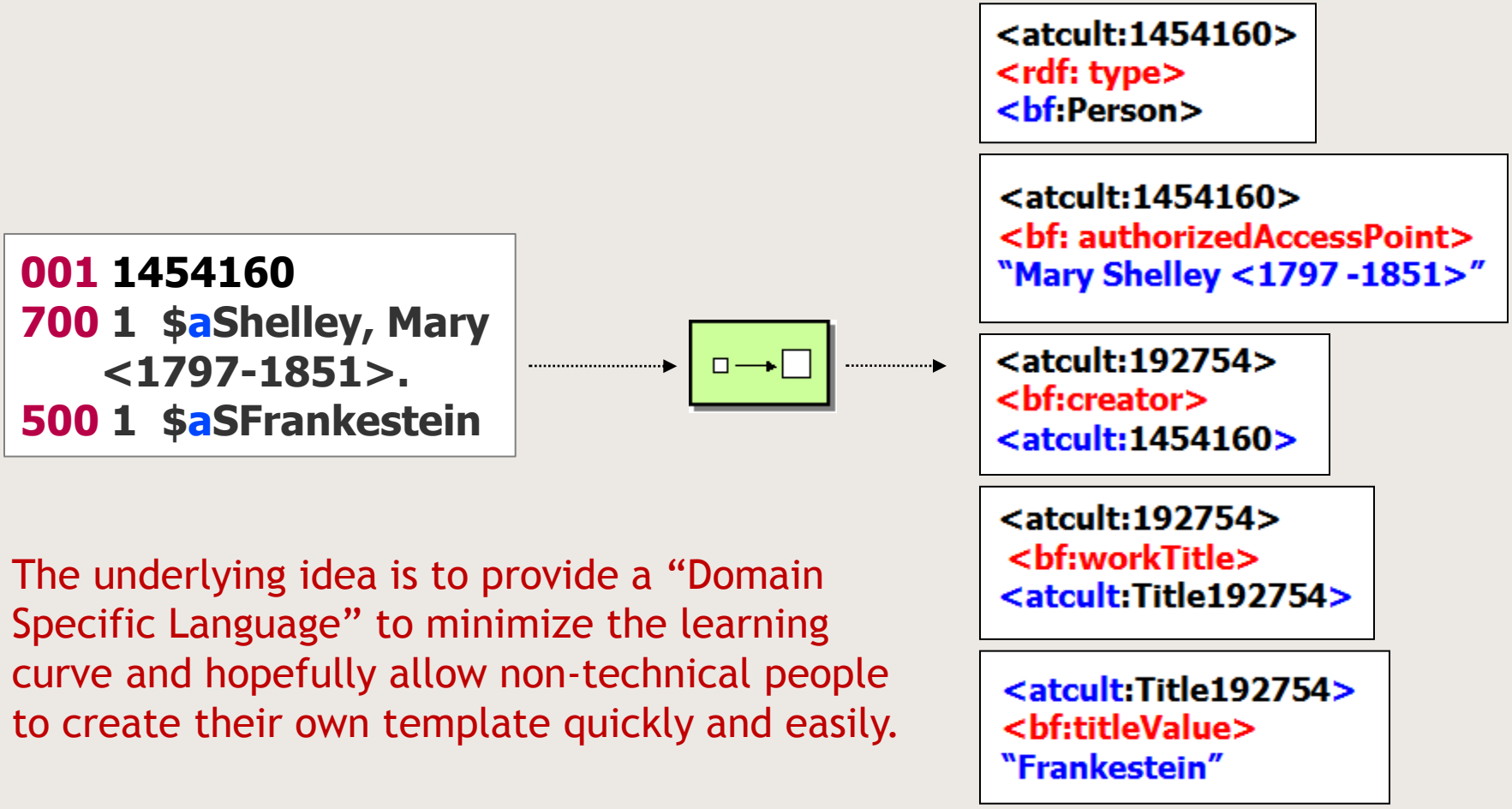
ALIADA converts each incoming record by means of *Conversion templates*.
Each template associates:

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



ALIADA Conversion templates

Another example of the conversion process from UniMARC to BIBFRAME



The underlying idea is to provide a "Domain Specific Language" to minimize the learning curve and hopefully allow non-technical people to create their own template quickly and easily.

ALIADA Conversion rules

Technically, a conversion template is a file containing conversion rules, expressed in a high-level programming language.

For instance, the rule:

```
#set ($s = #uri('Work' 1643)
$s $is_a #bf("Work) .
```

produces the following:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix bf: <http://bibframe.org/vocab/>

<http://rdf.atcult.it/Work/1643> <rdf:type> <bf:Work> .
```

The conversion rules can be centralized and then reused, in order to gain speed for the implementation of new rules, e.g. adding more mappings with different ontologies.

3. The creation of a FRBR/BIBFRAME layer from bibliographic and authority records

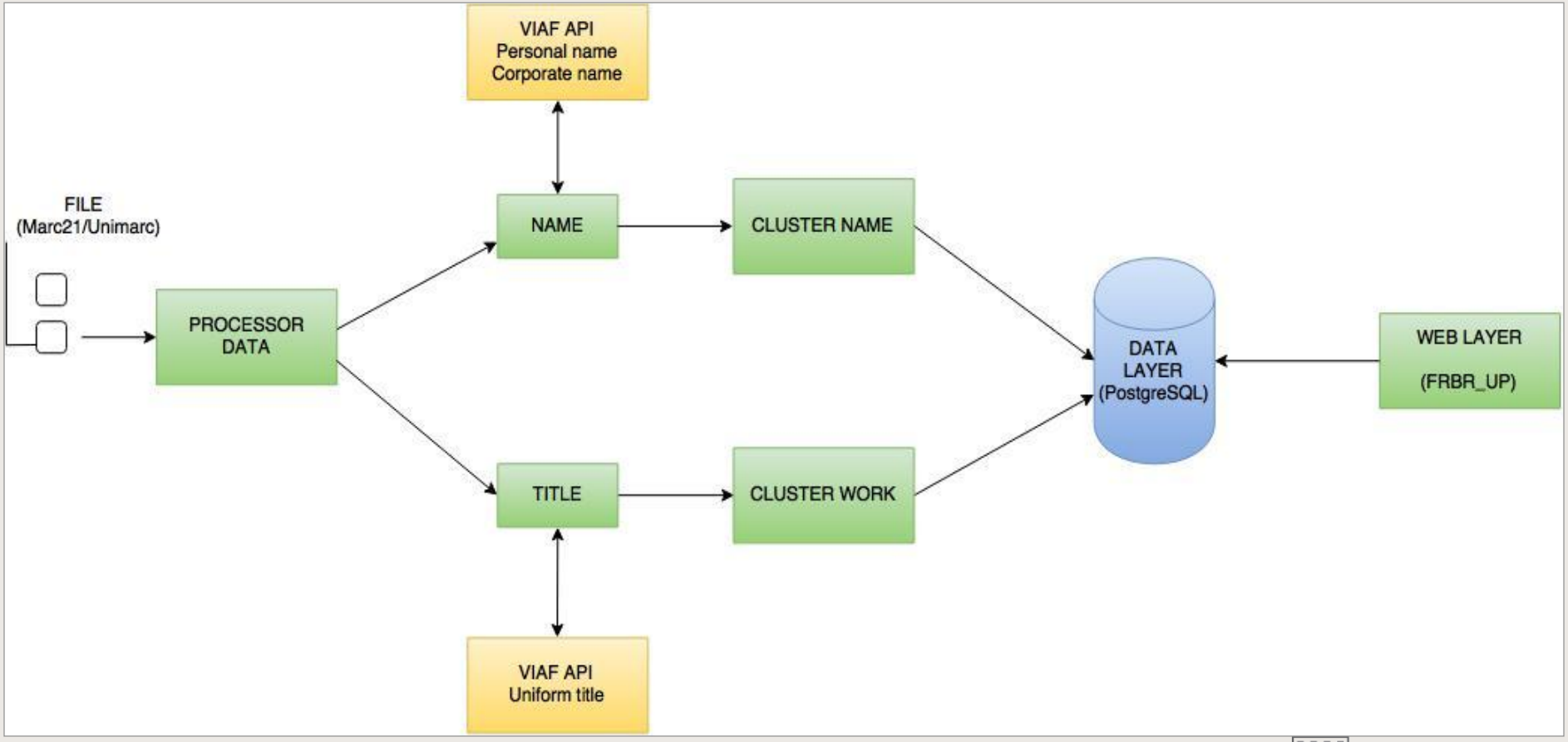
The existent catalogues are description, above all, of manifestations/instances. We tried to give an answer to the requirement to re-design the data model with a system that derives data from existent records to produce a new Person/Work layer. The process creates for each Person entity a 'cluster' of possible variant forms, and does the same for associated Works.

- Person cluster: creation of a unique *name access point* for Person names.
- Work cluster: each Person is associated to his Work.

Each Work cluster is linked to Instance titles.

The loading process and creation of clusters

The loading processor and creation of Person/Work clusters: an important step of the process retrieves data from external authority files, such as VIAF, using the specific APIs.



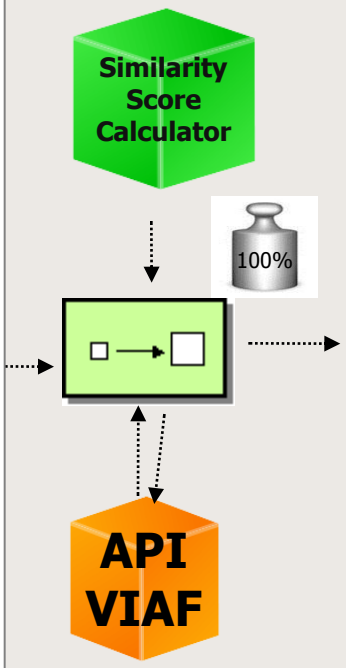
Cluster makers - Person (example 1)

```
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:

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-

This chart and the following example show the mechanism for associating names from different records in a single Person cluster

Cluster makers - Person (example 2)

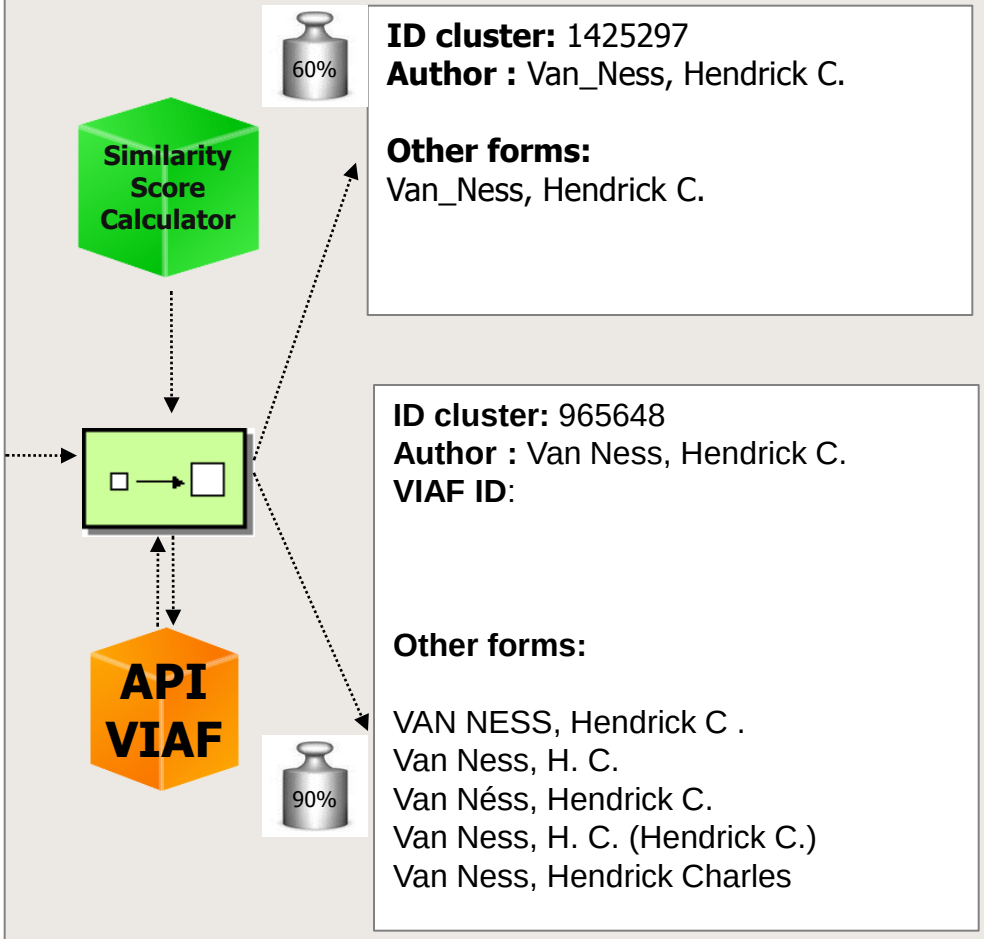
001 00002
200 \1 $aVan Ness,$bHendrick C.
997 \\ $aAUTHORITY

001 8379
701 \1 $aVan_Ness,$bHendrick C.
997 \\ $aUNIBAS

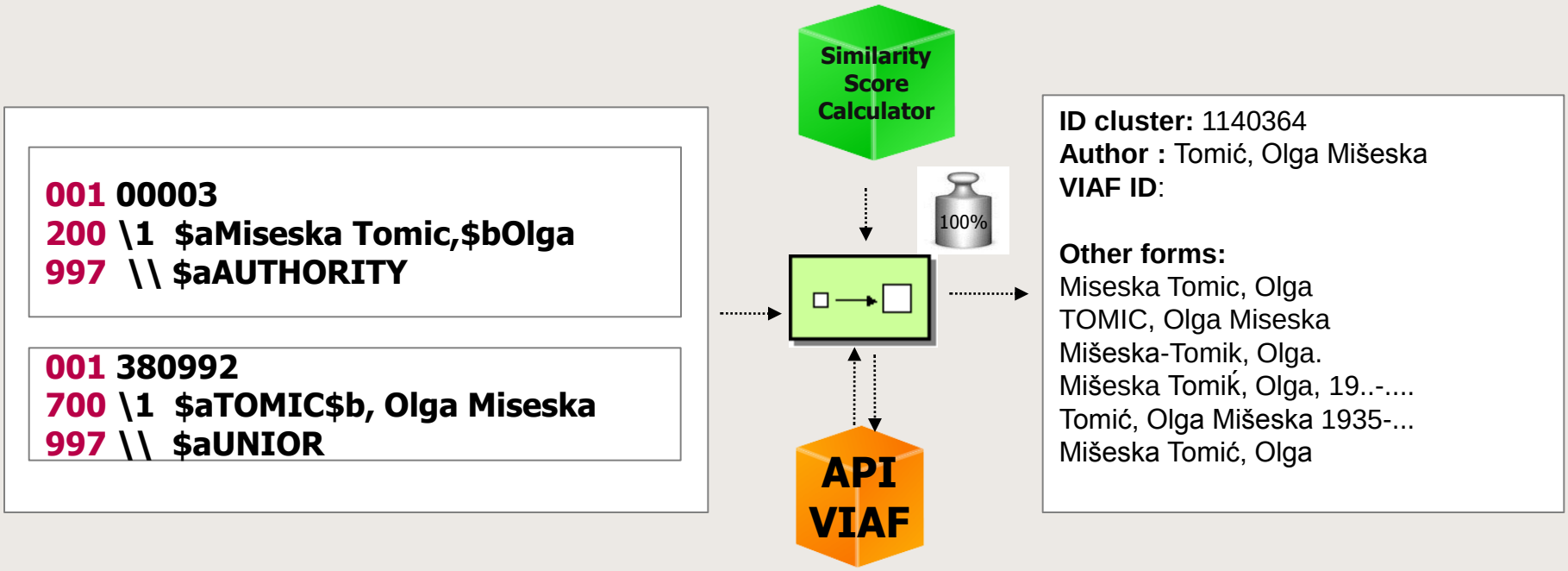
001 173506
701 \1 $aVan Ness,$bHendrick C.
997 \\ $aUNINA

001 1317
701 \1 $aVan Néss$b, Hendrick C.
997 \\ $aUNISANNIO

001 56522
701 \1 $aVAN NESS,$bHendrick C
997 \\ $aUNISA



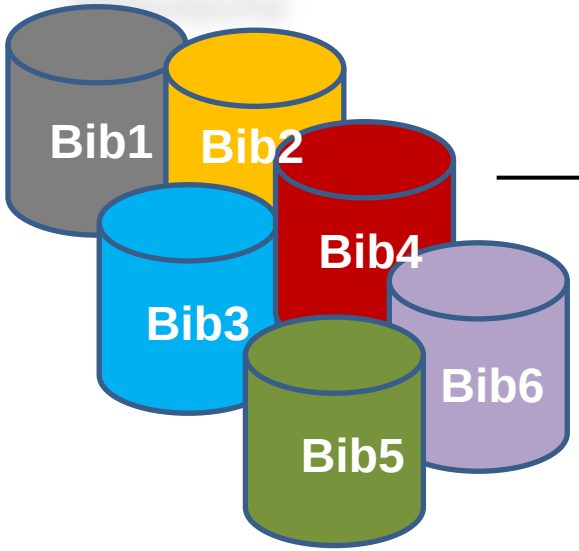
Cluster makers – Person (example 3)



Reading the MARC record we obtain:

- the association of names through a weighted algorithm for comparison
- the identification of already existing clusters or creation of new clusters
- the aggregation of different forms of names through VIAF APIs

Biblioteche



- Individuazione
- Raccolta
- Selezione
- Elaborazione

VIAF
Virtual International Authority File

Authority file locali

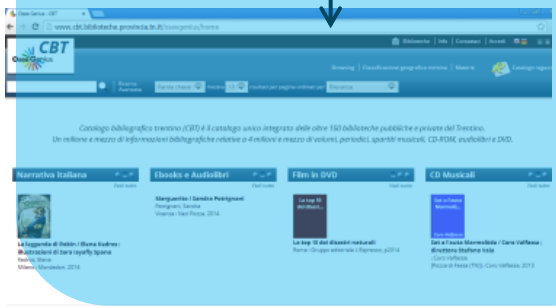
ALIADA

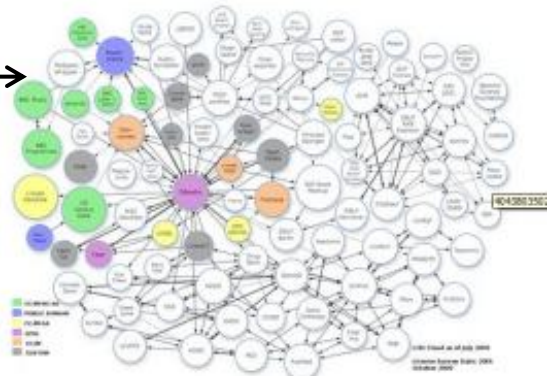
(Elaborazione dati in RDF)

Search Engine

BIBFRAME-UP

RDF Store





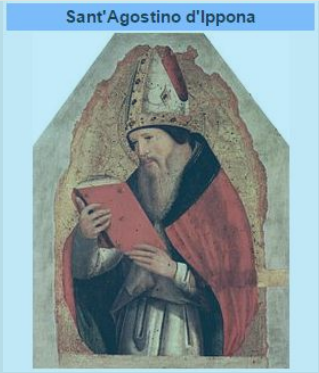
LOD Cloud

Linked Open Services Platform

The BIBFRAME-UP: a three layer architecture

Person/Works

WIKIPEDIA
L'enciclopedia libera



VIAF
Virtual International Authority File

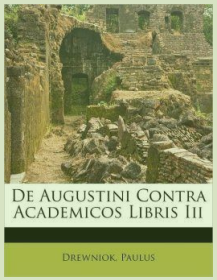
LIBRARY OF CONGRESS AUTHORITIES

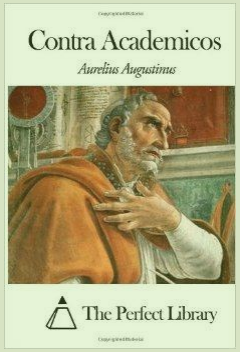
Contra academicos

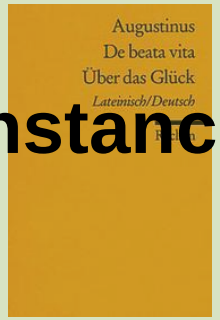
De Beatâ Vitâ

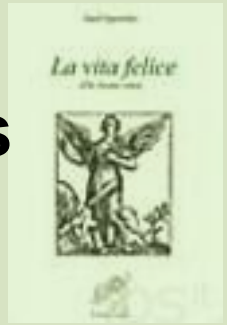
De civitate Dei contra paganos

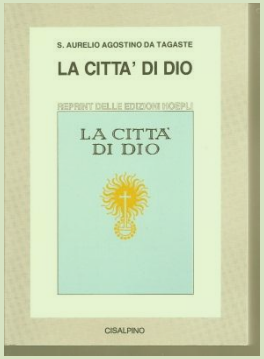
Instances

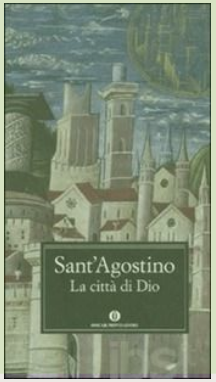












Item

UNIVERSITÀ DEGLI STUDI DI SALERNO

Università degli Studi del Sannio

Biblioteca Centrale di Ateneo

OpacUnina.it
CATALOGO DI ATENEO

"L'Orientale"
Università degli studi di Napoli

Università degli Studi di Napoli "Parthenope"
Catalogo generale

casalinilibri

The BIBFRAME-UP Portal: synthesis

1st layer - Person/Work: the set of data related to Person and Work, in RDF, saved in a SPARQL endpoint and made available by specific search and presentation functions.

2nd layer - Instance: bibliographic data indexed in SOLR search engine, that is able to produce new different data aggregations in facets (such as publication date, language, publisher, edition, etc.). This layer provides users with a wide range of search and navigation functions.

3rd layer - Item: holdings data, related to copy information, coming from the local OPAC or local system of each specific library.

Conclusions

Where we are now?

The action plan of the three exposed areas of activities is in progress.

We are in a phase of analysis and development where opportunities to share experiences, doubts and input from the community about expected priorites can be crucial.

Any cooperation and manifesation of interest by institutions and people, to share and disseminate activities and results, is very welcome.

Thank you

Tiziana Possemato
Chief Information Officer

tiziana.possemato@casalini.it
www.casalini.it

ALA Midwinter 2016
BIBFRAME Update Forum Agenda