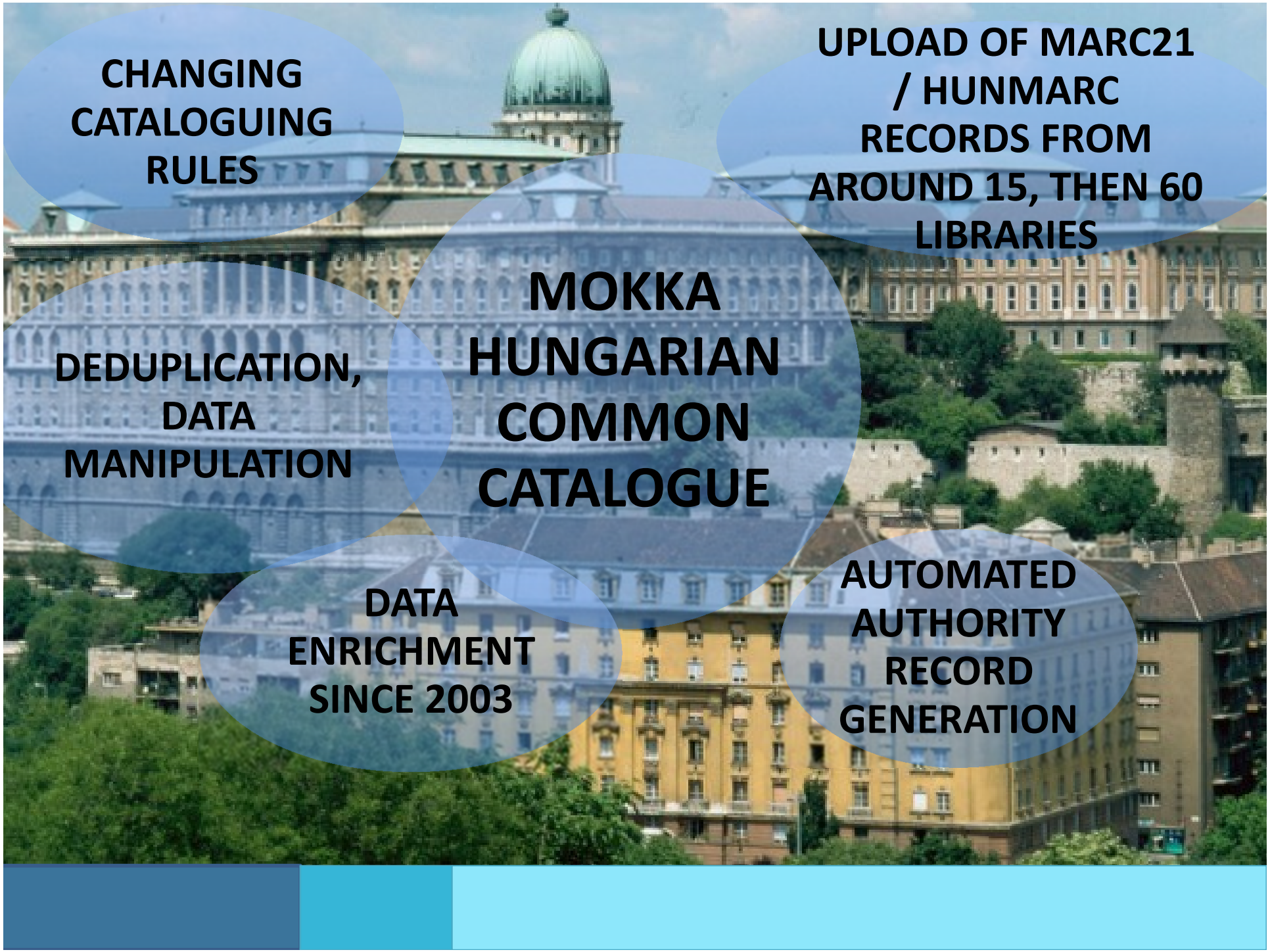
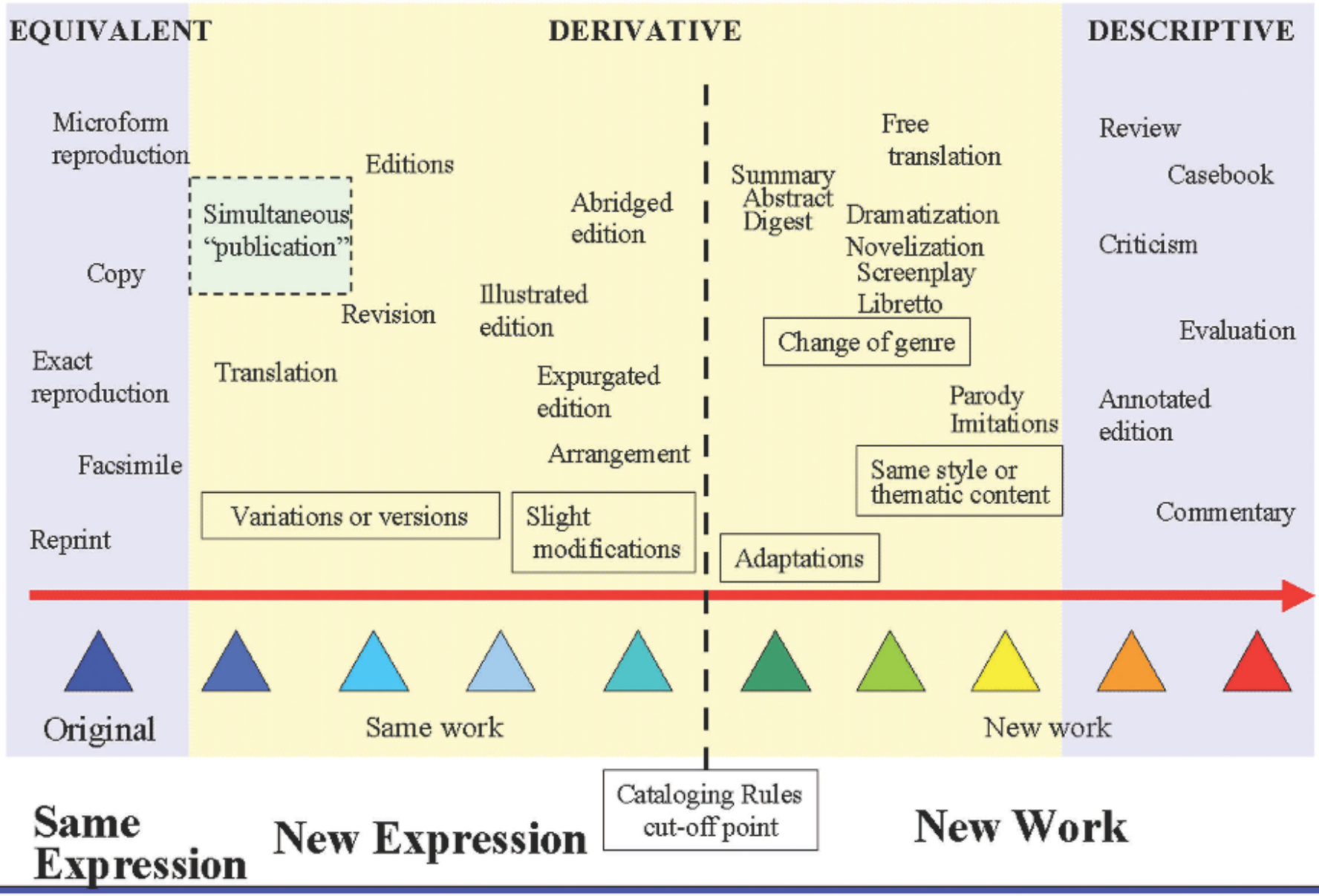

European BIBFRAME workshop 2017
MOKKA UP PROJECT
Conversion of The Common Hungarian Catalogue Into The BIBFRAME / FRBR Structure
MIKLÓS LENDVAY
NATIONAL SZÉCHÉNYI LIBRARY BUDAPEST
26th September 2017 Frankfurt

Országos Széchényi Könyvtár adattérkép

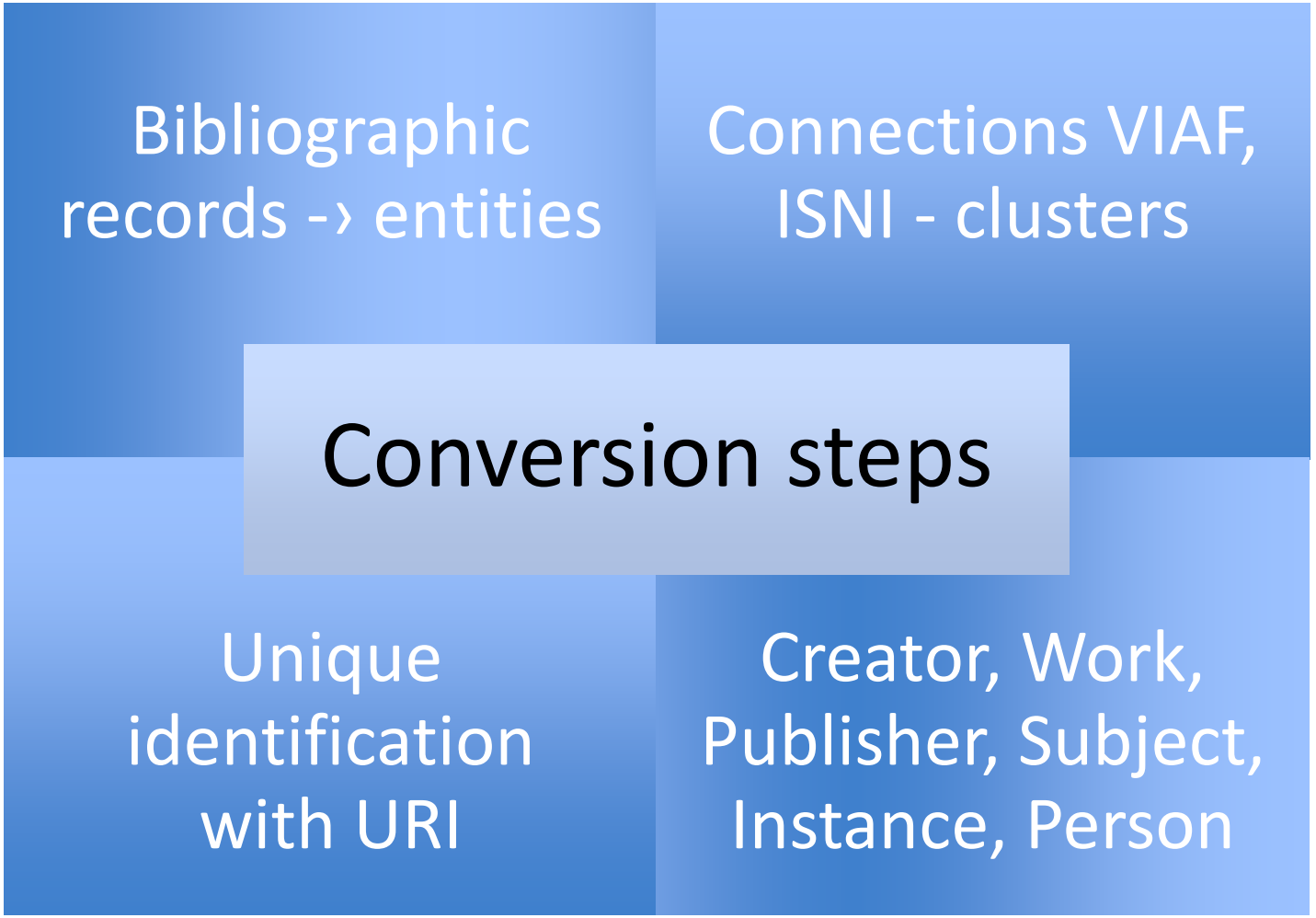
National Széchényi Library Data Map


CHANGING CATALOGUING RULES
UPLOAD OF MARC21 / HUNMARC RECORDS FROM AROUND 15, THEN 60 LIBRARIES
DEDUPLICATION, DATA MANIPULATION
MOKKA HUNGARIAN COMMON CATALOGUE
DATA ENRICHMENT SINCE 2003
AUTOMATED AUTHORITY RECORD GENERATION

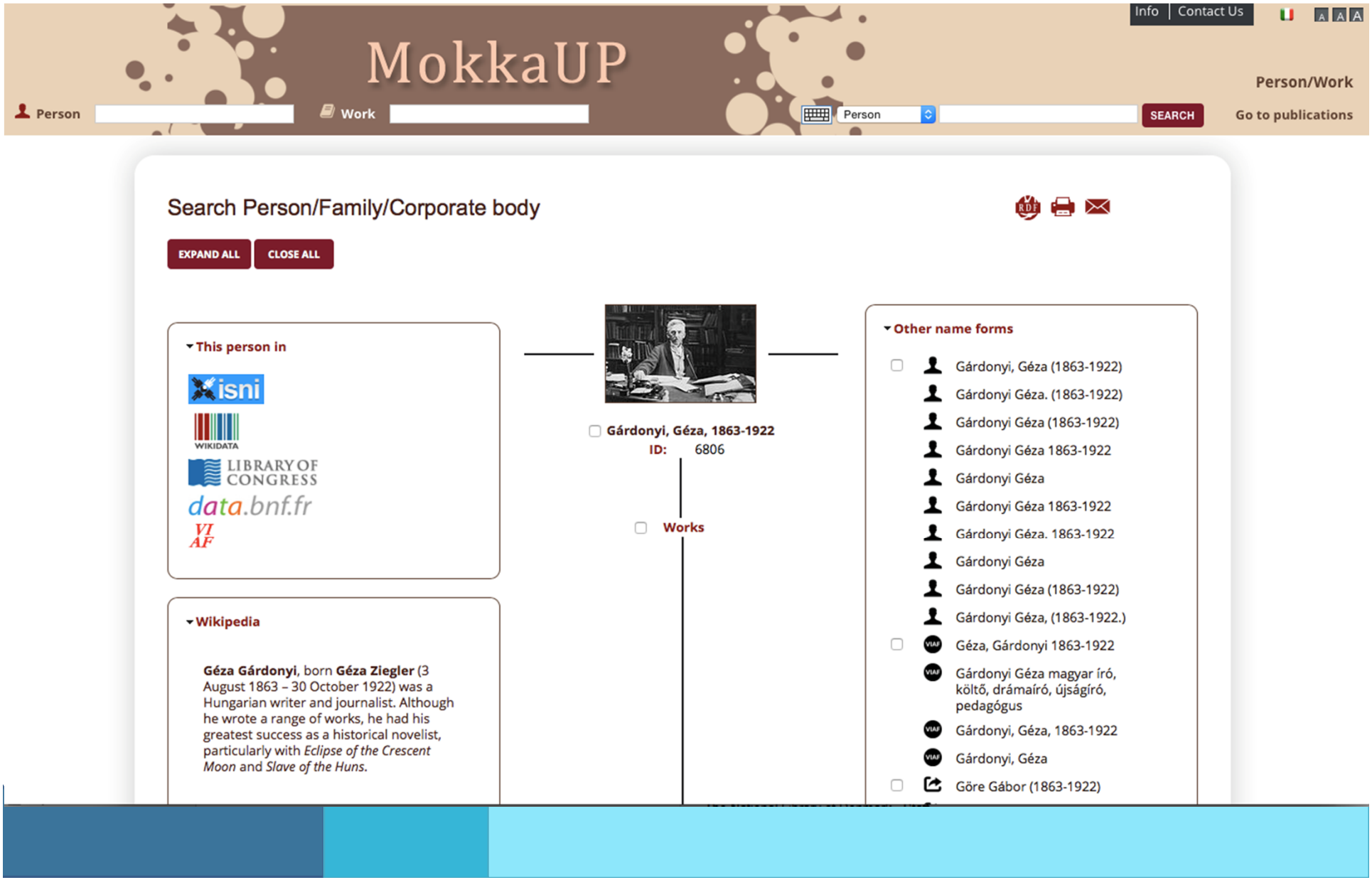
Family of Works



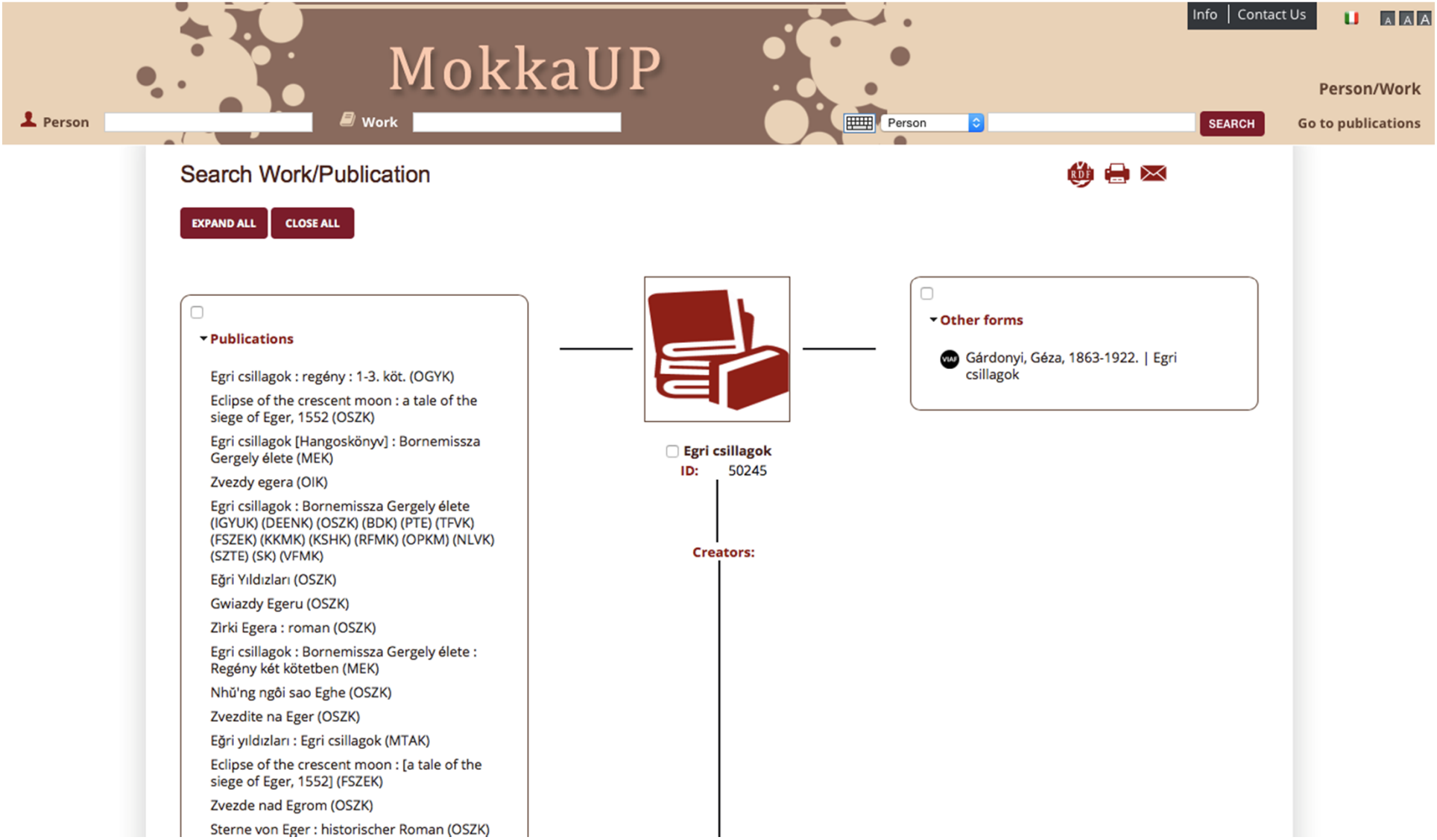
Relationships in the Organization of Knowledge, edited by Carol A. Bean and Rebecca Green, 2001, p. 23, "Bibliographic Relationships" by Barbara B. Tillett, Figure 2, © 2001 Kluwer Academic Publishers Boston, with kind permission of Kluwer Academic Publishers.


Bibliographic records -› entities
Connections VIAF, ISNI - clusters
Conversion steps
Unique identification with URI
Creator, Work, Publisher, Subject, Instance, Person

Hungarian catalogue in BIBFRAME format - @cult
http://test-mokka-up.oseegenius.it/mokka/clusters

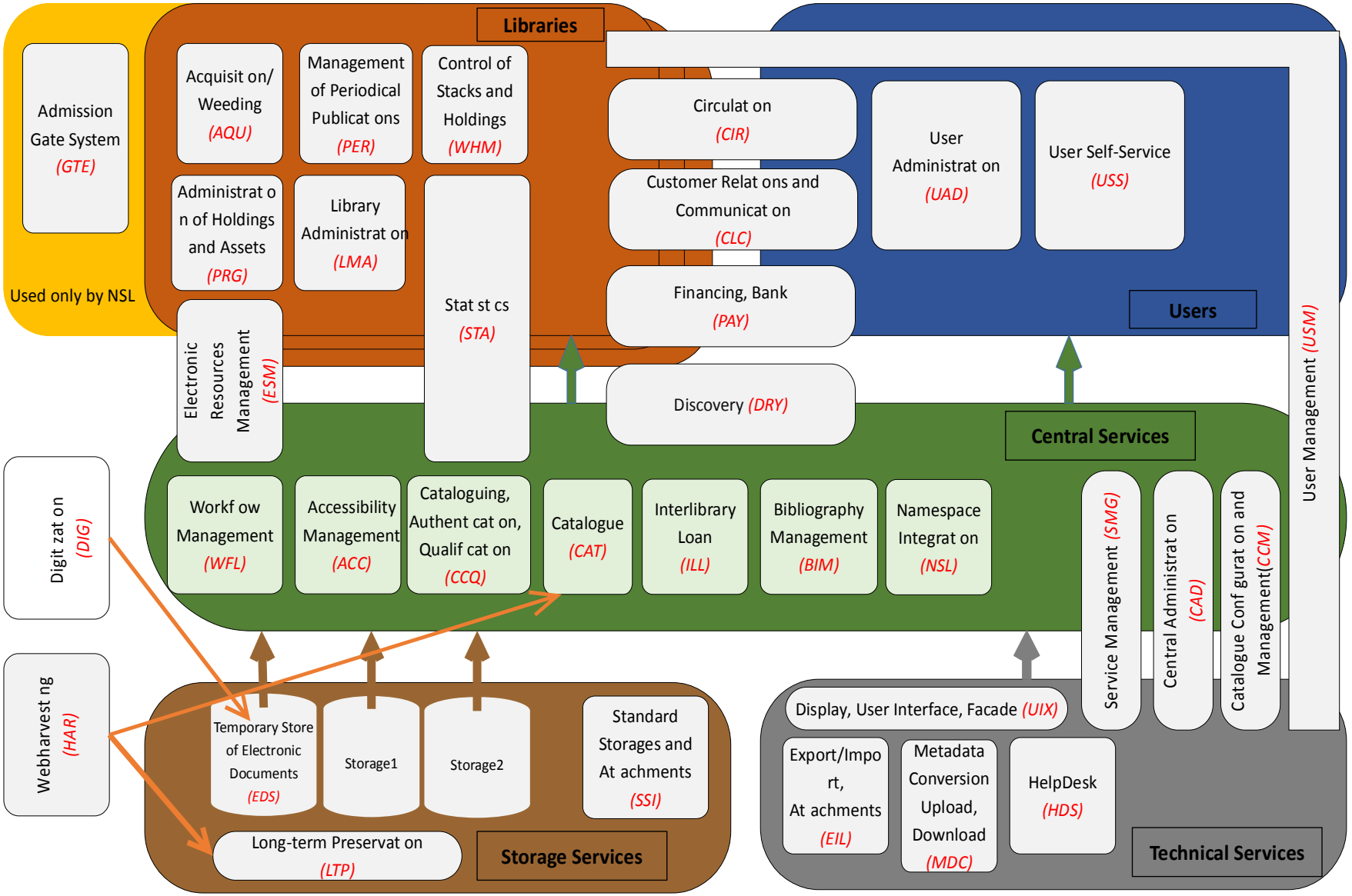


Work – Publication - Instance

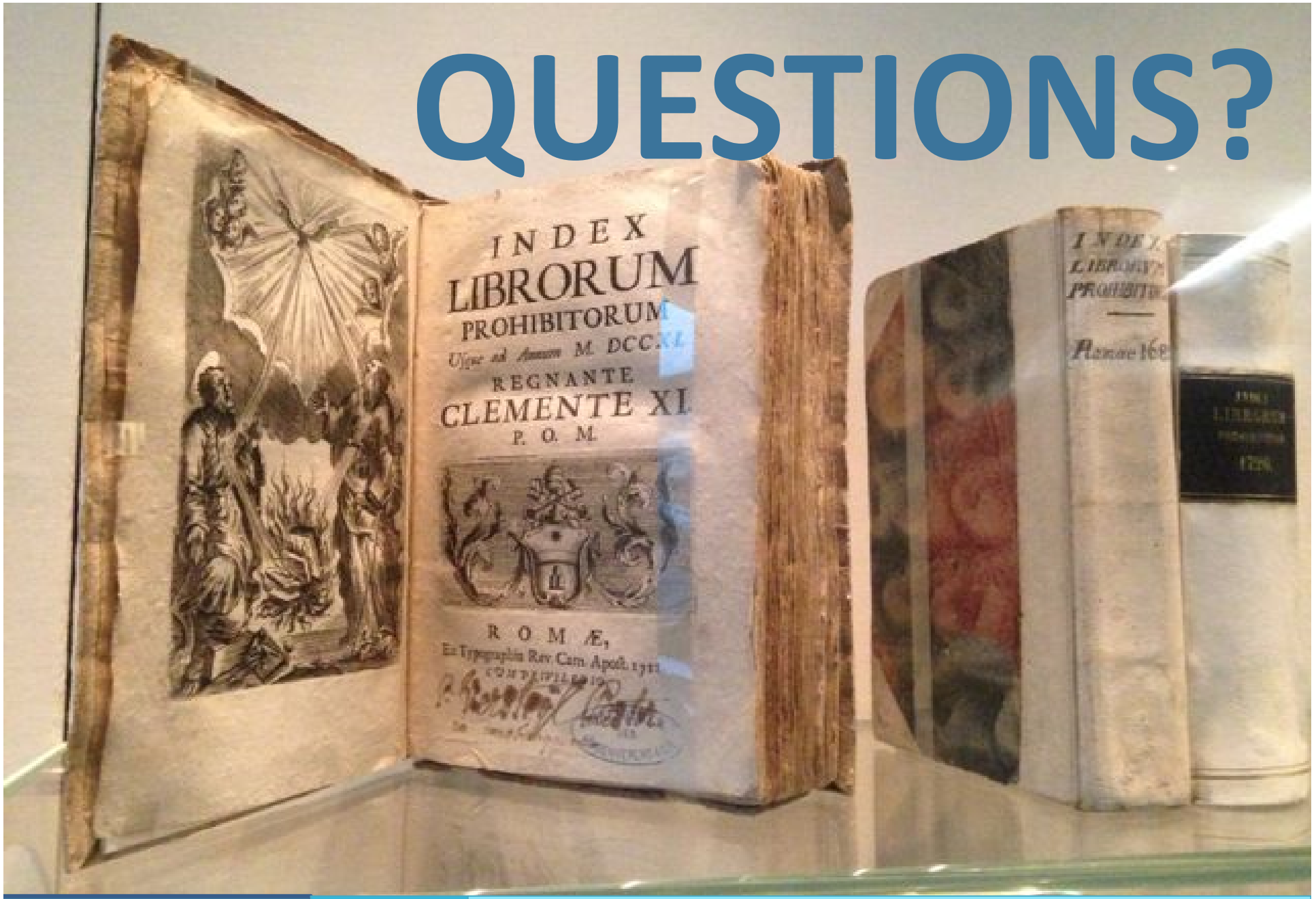


Publication – Libraries, Formats, Publisher

Plan of the National Library Platform



QUESTIONS?



www.oszk.hu , lendvay.miklos@oszk.hu

Detailed Structural And Technical Information Provided By @cult

Reconciliation and Enrichment

Automatic procedures in MOKKA – UP Project

Data Reconciliation

Data reconciliation and enrichment is obtained by:

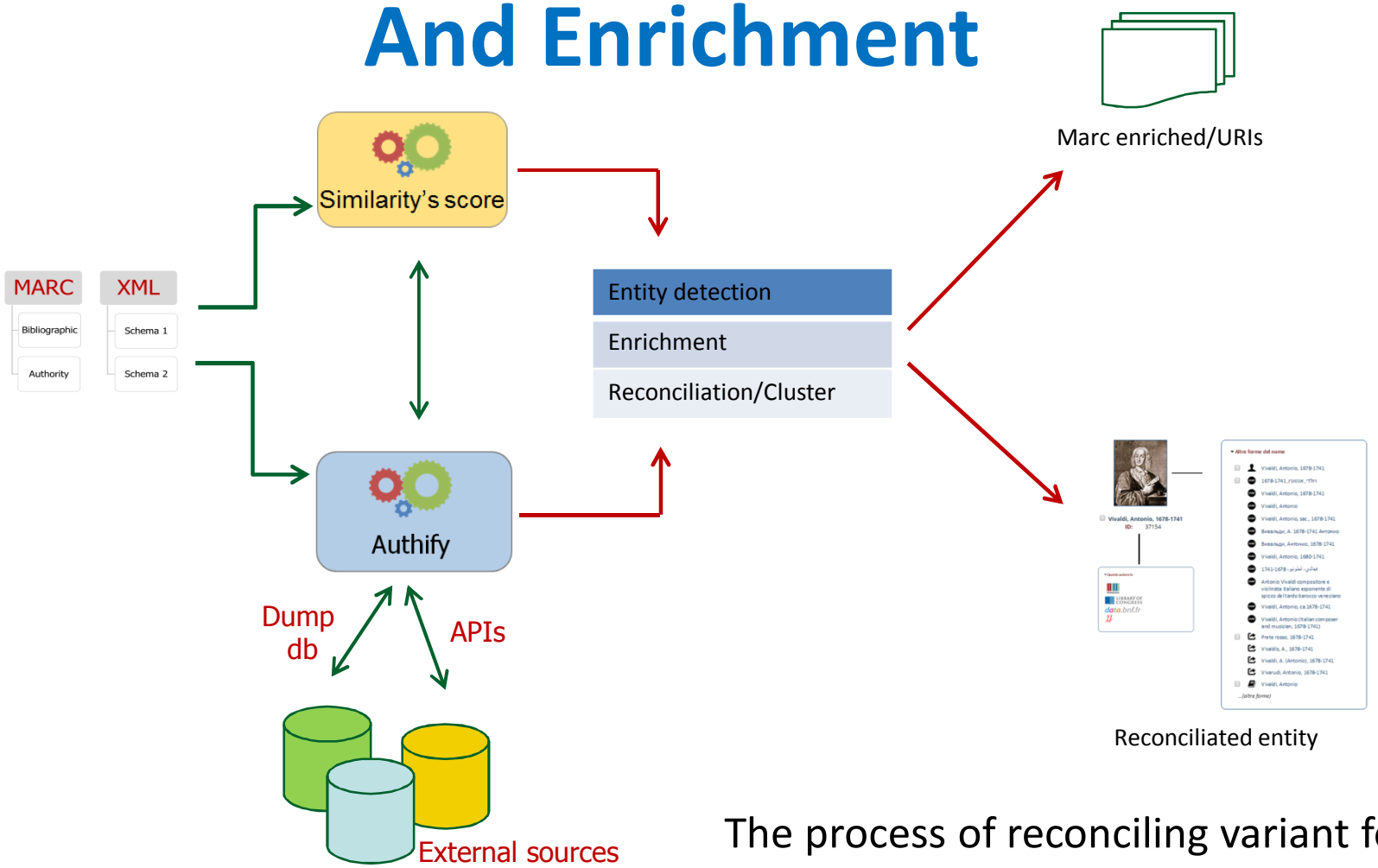
- automated processes
- manual processes

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

- Authify is a RESTFul module, that offers several search and detection services. The project started at the very beginning for overcoming some limitation of the public VIAF Web API.

- VIAF, being a public project, doesn't allow a massive invocation of its API: for those use cases where such requirement is needed, the project provides a download of the whole dataset.

- That was mainly the reason why we started implementing Authify: index and store the VIAF clusters dataset and provide, on top of that, powerful full-text and bibliographic search services.

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

Authify – Cluster Search Services

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

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

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

Authify – Cluster Search Services

- The system has been built with extensibility in mind, so the mentioned chain is fully configurable; for instance, here's 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.
- **FullText 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 some entities (e.g. birth and death dates).
- As last chance, 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 short form. Same as the previous point, this could lead to a response with minor 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 (whatever is the specific dialect) the system analyses all (configured) tags that contain a name and, for each of them, tries to figure out (using the statements of responsibility of the input record) what is the corresponding role within the work represented by the given record.

- 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 desumed 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 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 – Authity / 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

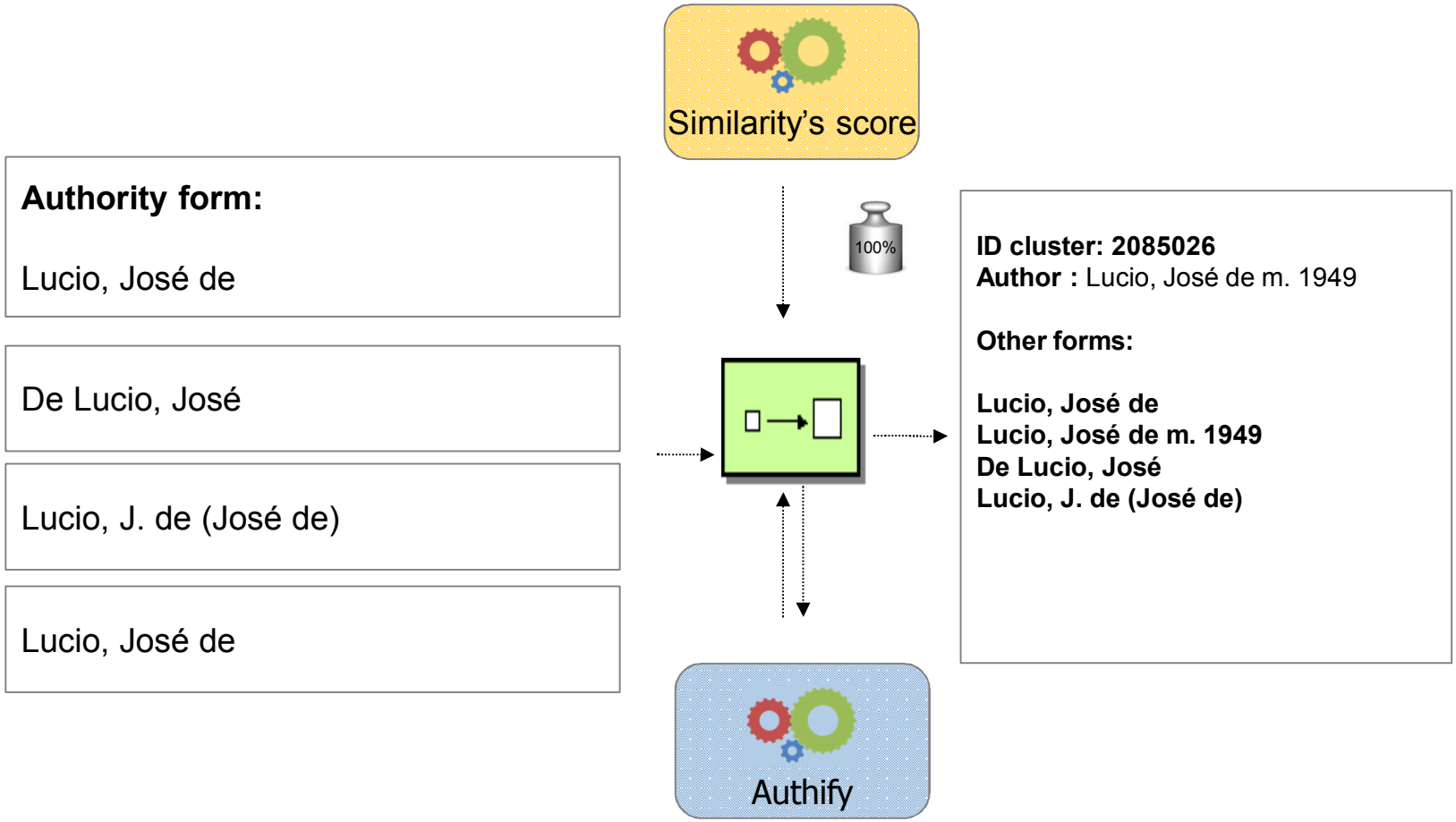
- =LDR 01145nam a2200241 i 4500
- =001 LE01988135
- =005 20020503105244.0
- =008 010702s1999\\\\it\\\\\\\\\\\\000\0\lat\\
- =020 \\$a882092868X
- =040 \\$aDip.to Beni Arti e Storia$bita
- =082 0\$a264.024
- **=245** 00$aBreviarium Romanum :$beditio princeps, 1568 **/$cedizione anastatica, introduzione 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

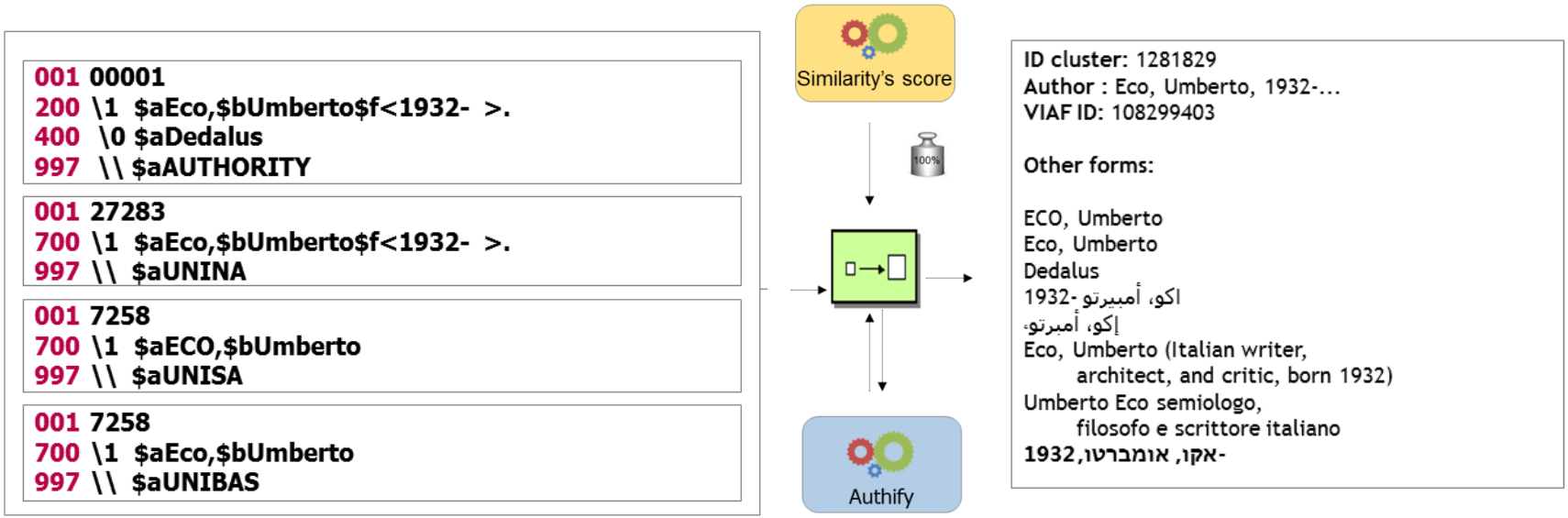
Response (3)


```
{
 "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



Name cluster process



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

Manual Process To Produce Clusters

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 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.

Conversion into RDF / BIBFRAME

Lodify: the evolution of Aliada for BIBFRAME conversion

The conversion process from any format to RDF

RESOURCES

METADATA CREATORS
(*Librarians, curators*)

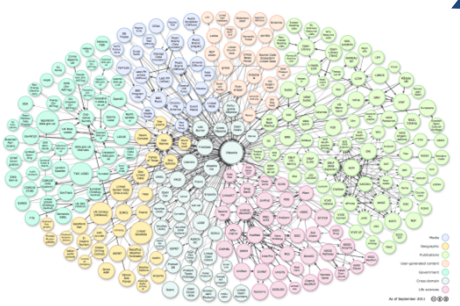
- Library Management System (ILS)
- Museum Collection Management System (MMS)
- Content Management System (CMS)

ALIADA

BROWSERS (GOOGLE)

IT COMPANIES

OTHER PUBLIC AND CULTURAL INSTITUTIONS

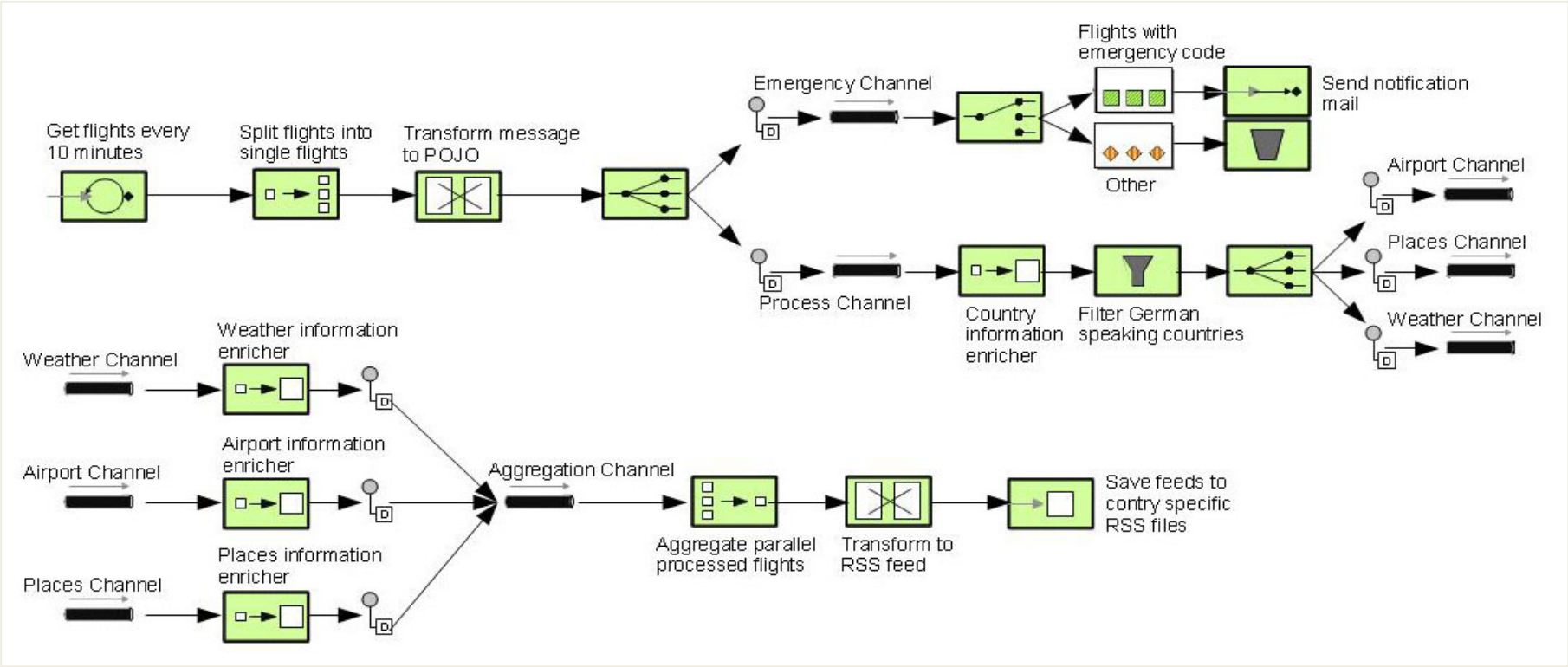


LINKED DATA CLOUD

http://lod-cloud.net/

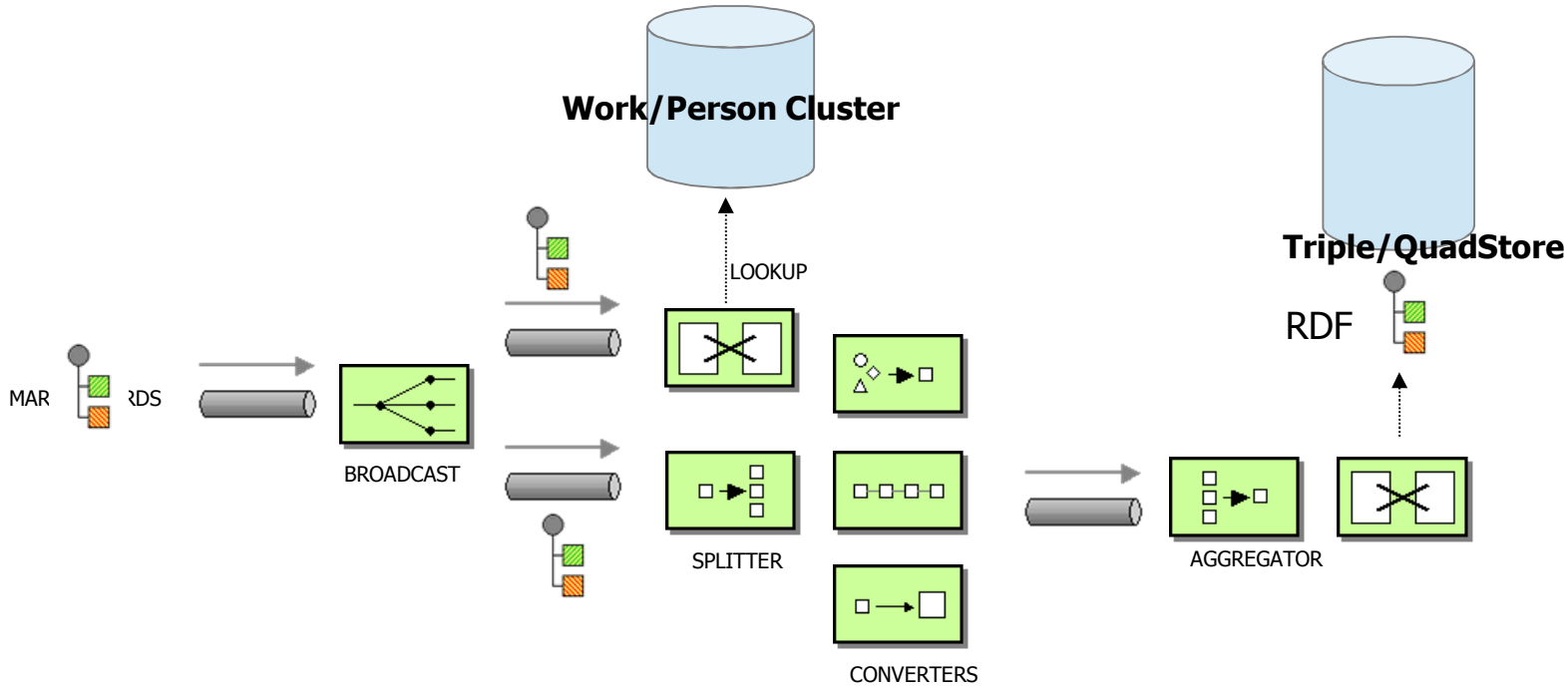
Lodify – The Asynchronous Pipeline

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



It's just an asynchronous pipeline!

•The high-level workflow in Lodify 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.

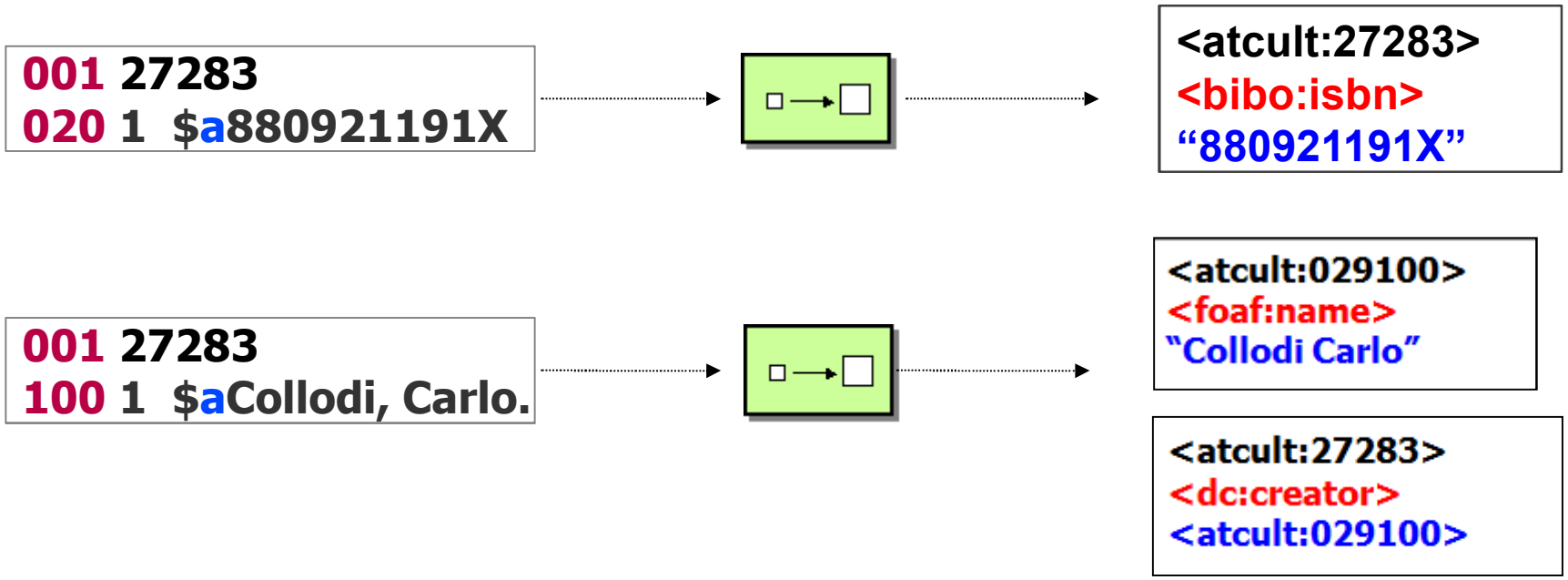


A set of MARC records go through the pipeline, which splits, processes and converts them.

Lodify - Conversion templates

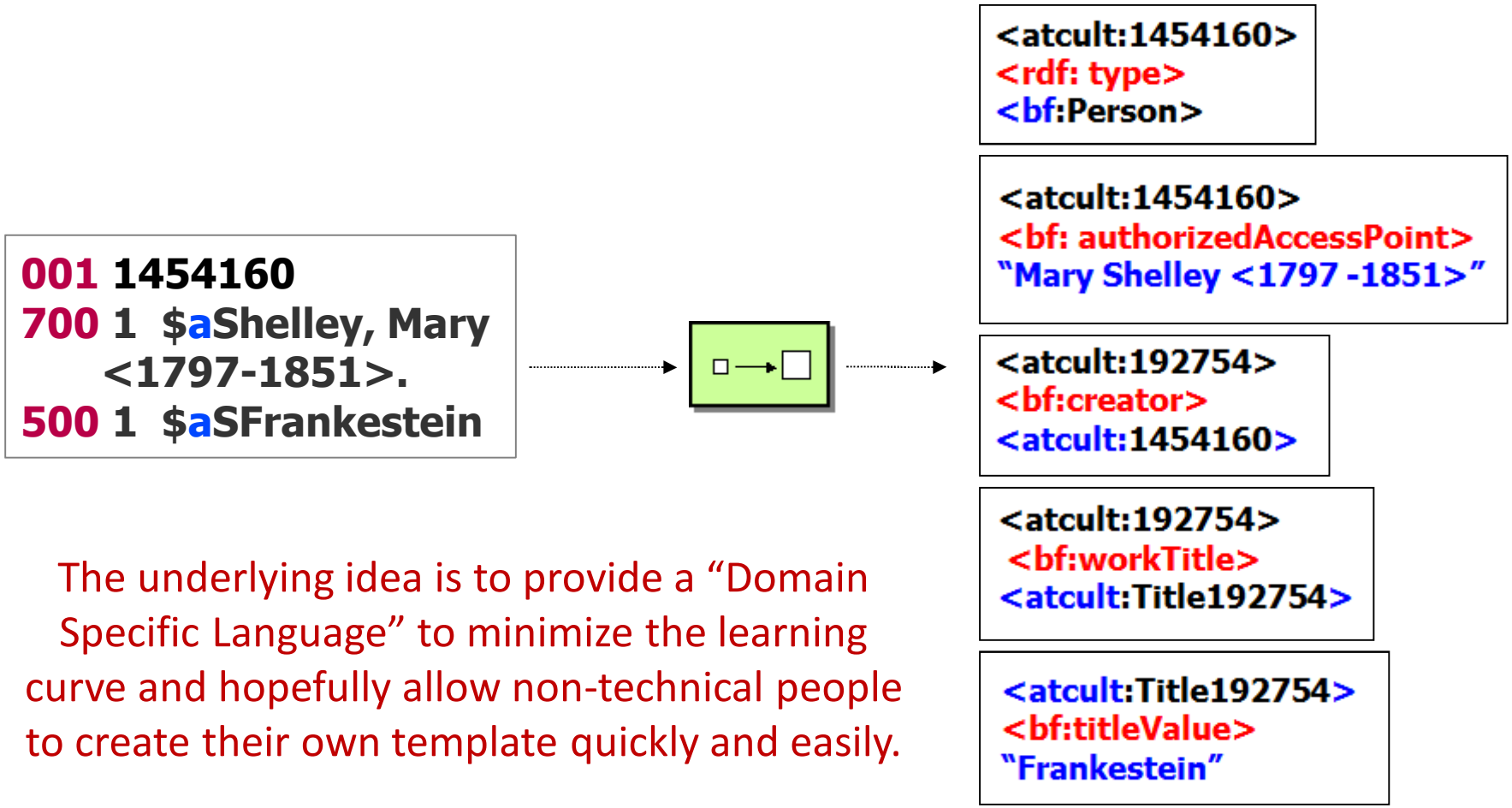
Lodify converts each incoming record by means of Conversion templates.
Each template associates:

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



Lodify - 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.

Lodify – 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.

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/

Further Information About The MOKKA UP Project

General information about scope and goal:
Miklós Lendvay **lendvay.miklos@oszk.hu**

Information about the data-models and connections:
Tiziana Possemato **tiziana.possemato@atcult.it**

Technical realisation:
Andrea Gazzarini **andrea.gazzarini@atcult.it**

www.oszk.hu www.atcult.it