DOing
REusable
MUSical data

http://www.doremus.org

JDEV-2017
Marseille
July 2017

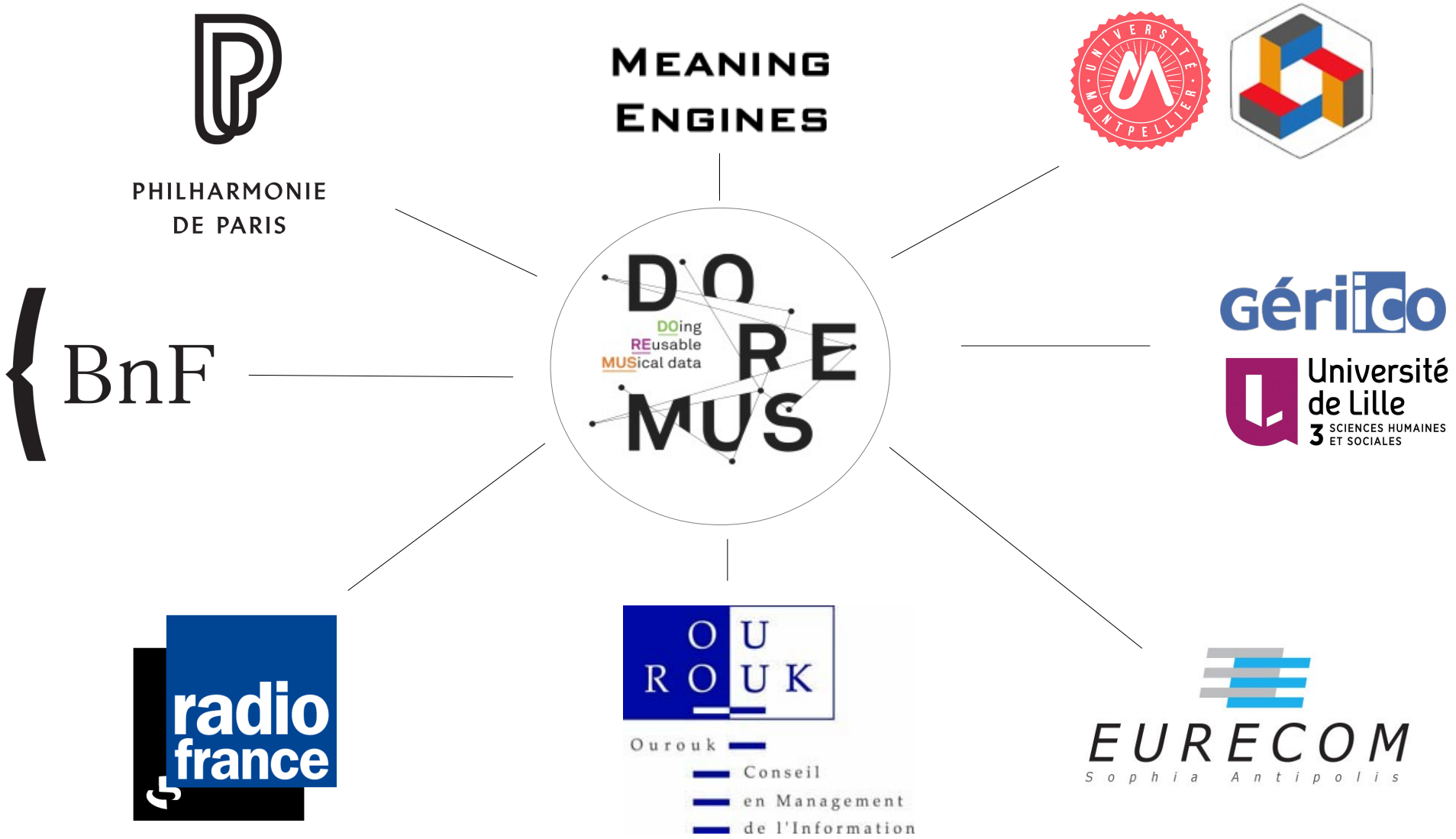
- PROJET FINANCÉ PAR L'ANR
ANR
PROJECT FUNDED BY THE ANR

<ANR-14-CE24-0020>

*Do*ing *Re*usable *Mus*ical Data

Konstantin Todorov
LIRMM / University of Montpellier

Partners


PHILHARMONIE DE PARIS
MEANING ENGINES
UNIVERSITÉ MONTPELLIER
BnF
DOREMUS
DOing REusable MUSical data
Gérico
Université de Lille 3 SCIENCES HUMAINES ET SOCIALES
radio france
OUROUK
Ourouk Conseil en Management de l'Information
EURECOM
Sophia Antipolis

DOREMUS Goals

Provide cultural ***institutions***, ***publishers*** and ***distributors***, ***communities*** of enthusiasts with:

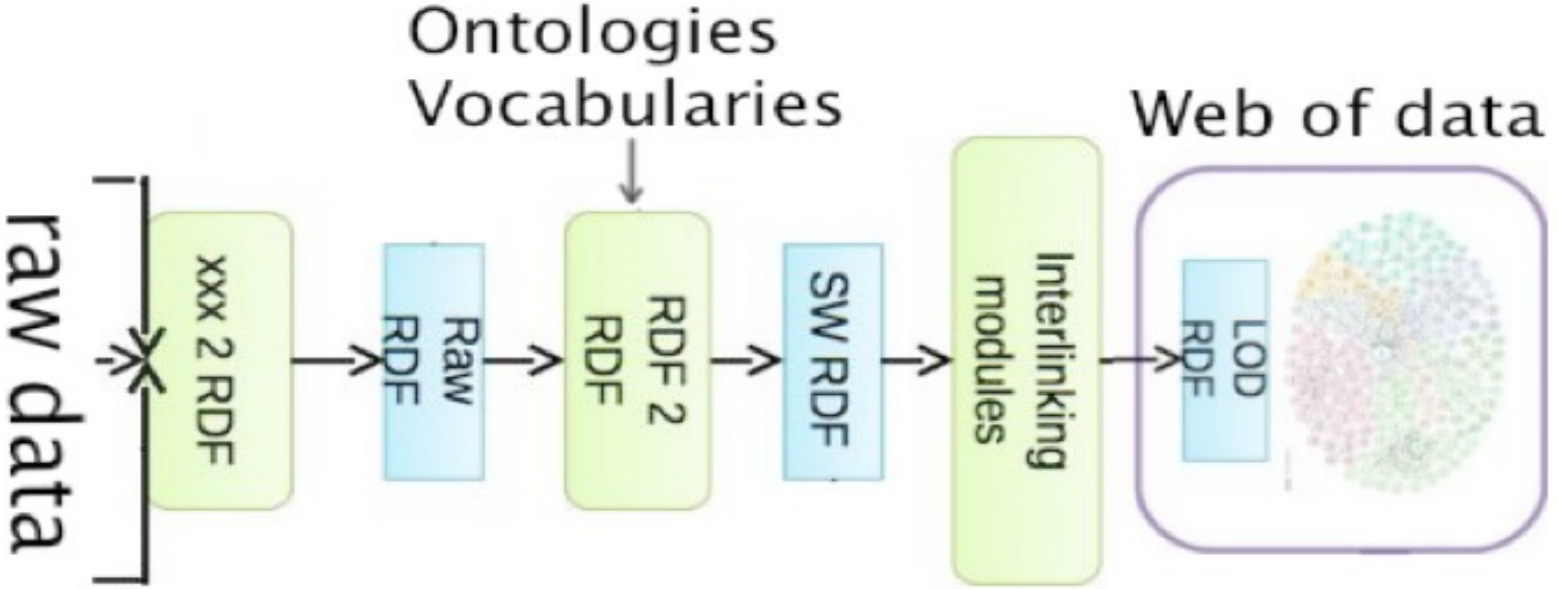
– a common knowledge model – the **DOREMUS ontology**

– shared and **multilingual controlled vocabularies** about music

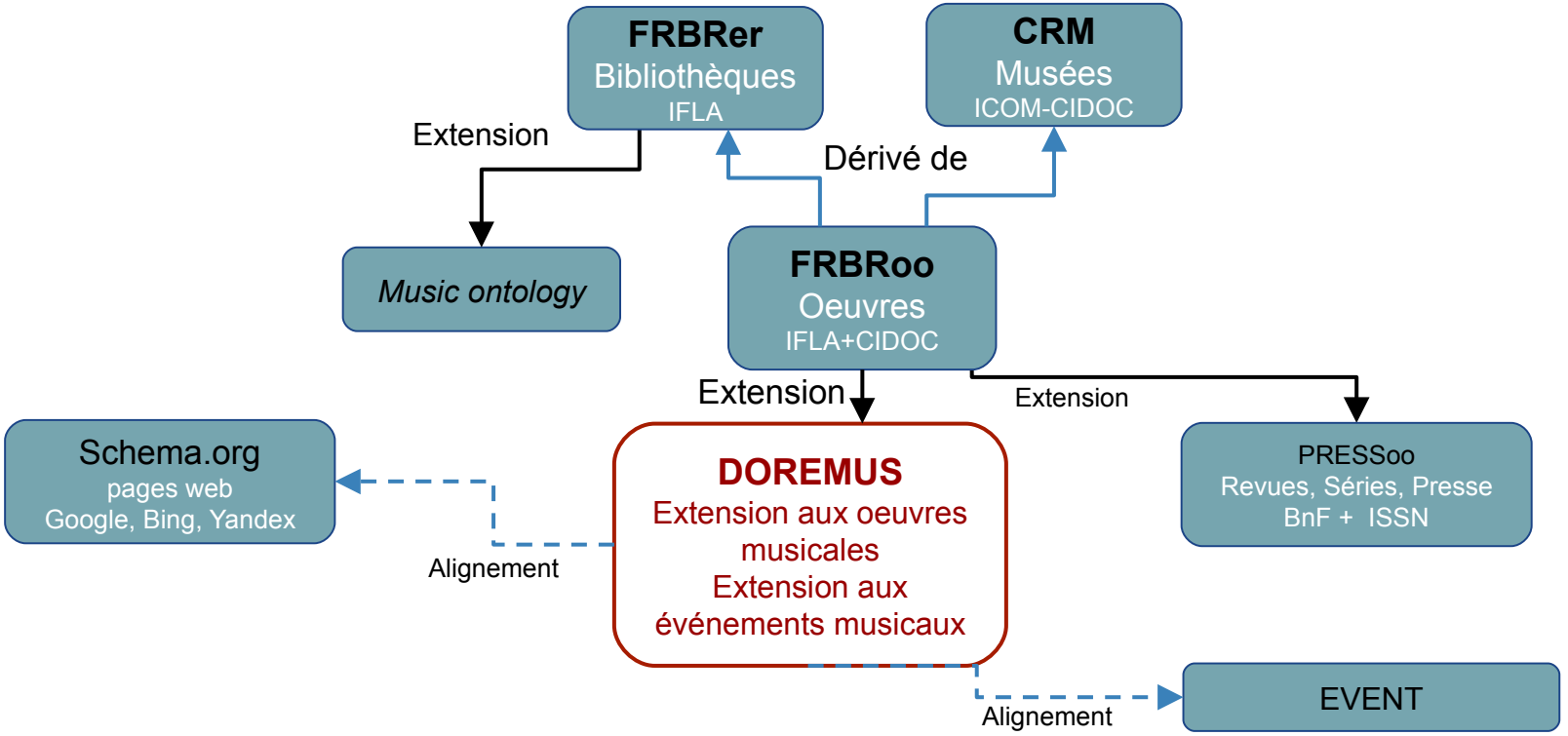
– methods to **RDF-ize, connect, publish**, **share** and **enrich** the catalogs of works and musical events in the web of data.

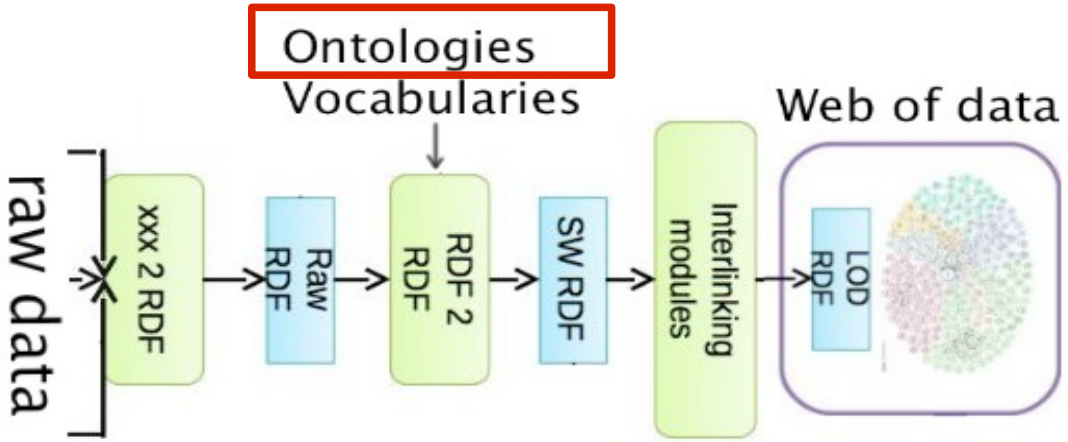
Create services for ***federated search*** and ***recommendation***.

Doing Linked Open Data

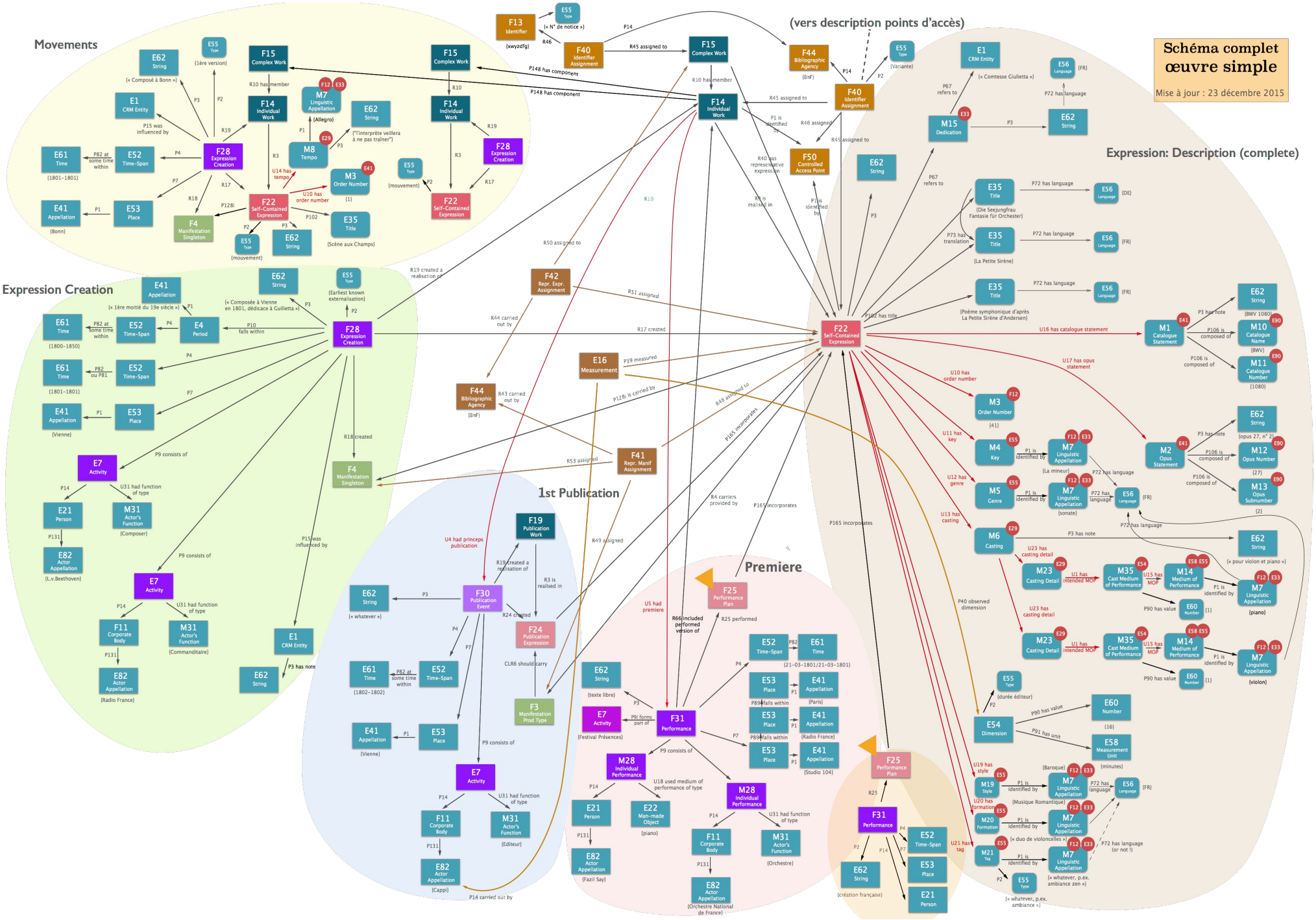


The Environment of Models

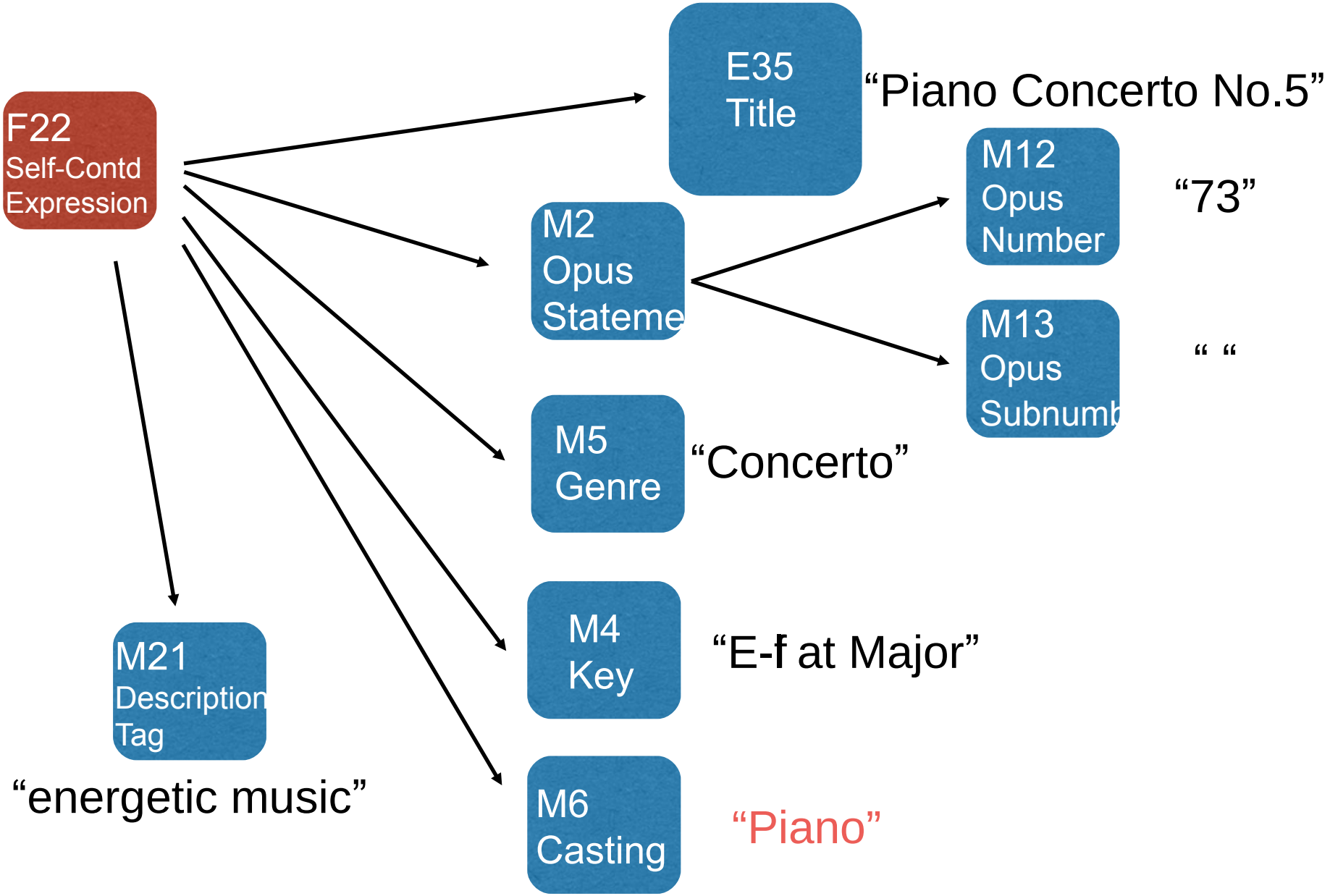




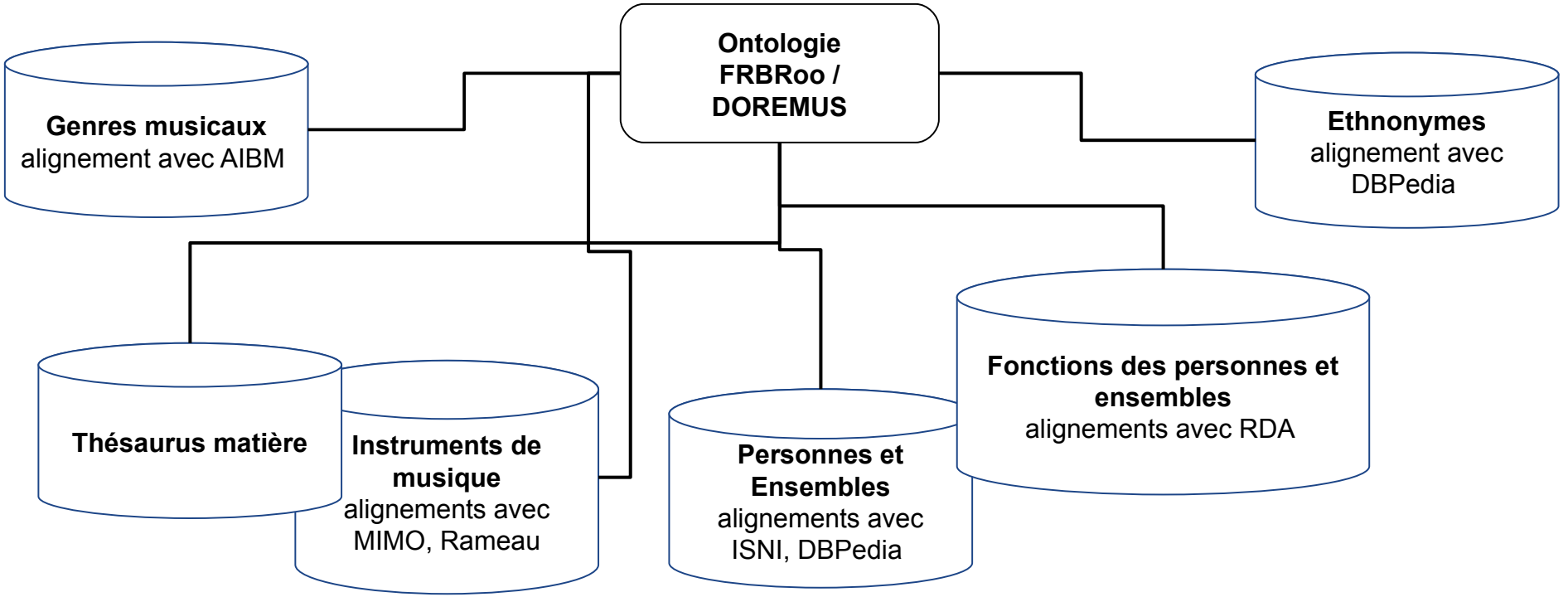
The DOREMUS Model: a Simple Work

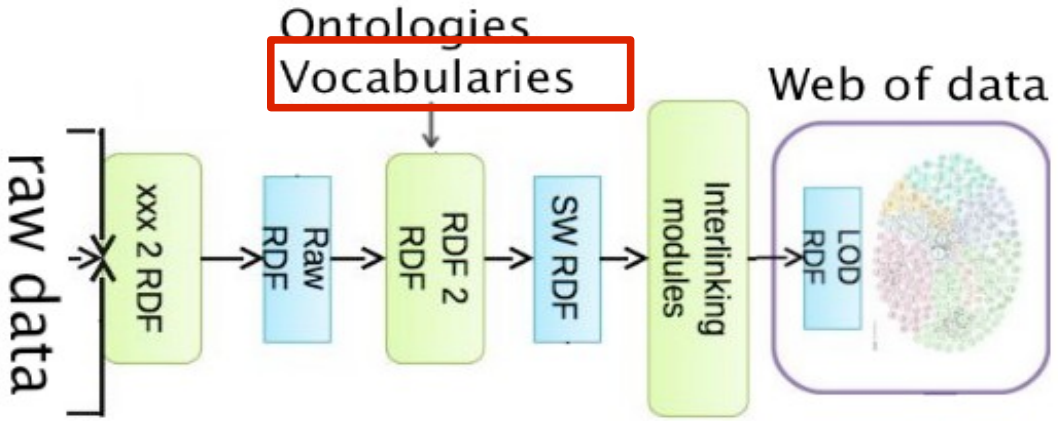


The DOREMUS Model: Describing a Simple Work

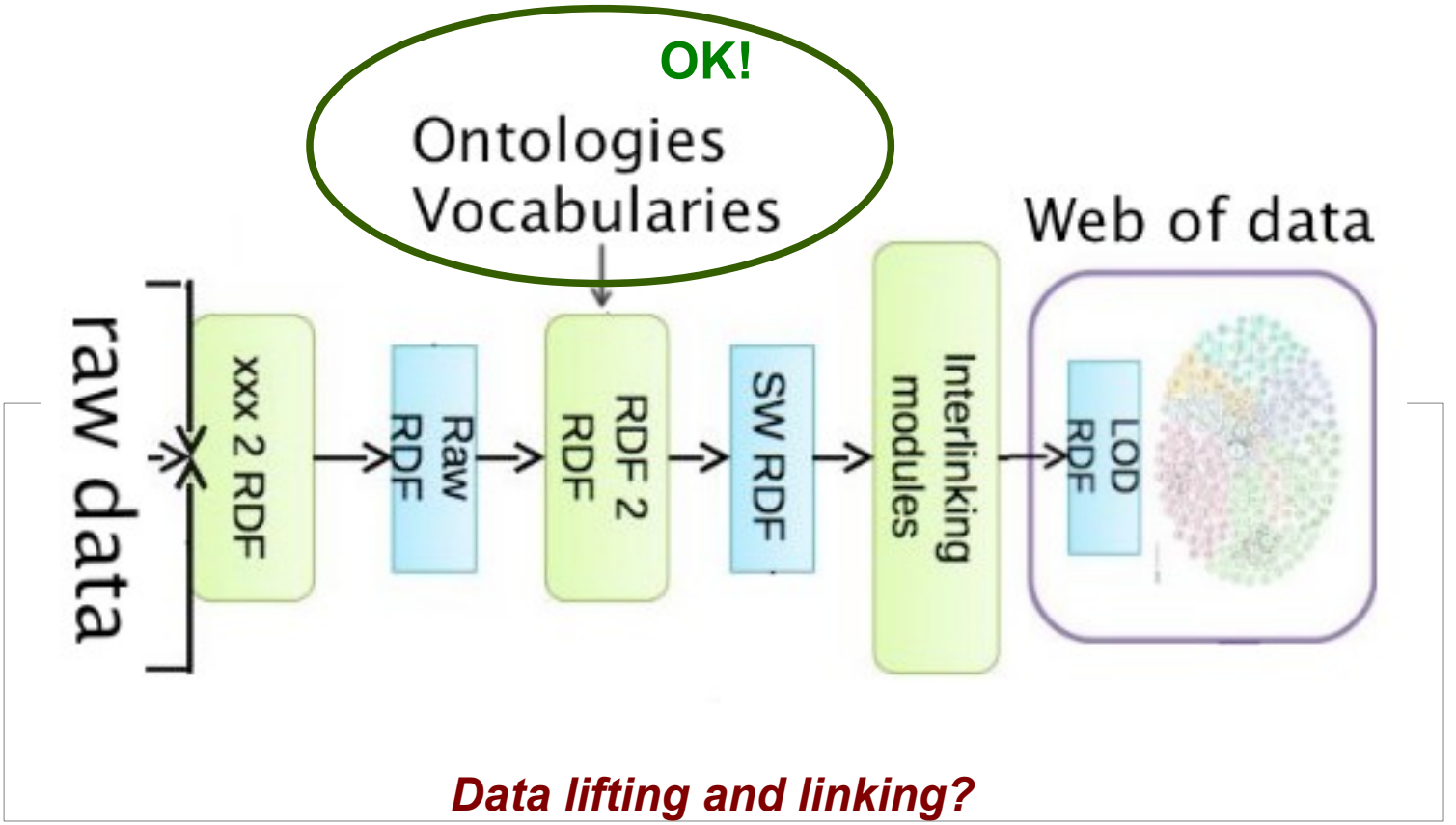


Controlled SKOS-ified Mutilingual Vocabularies





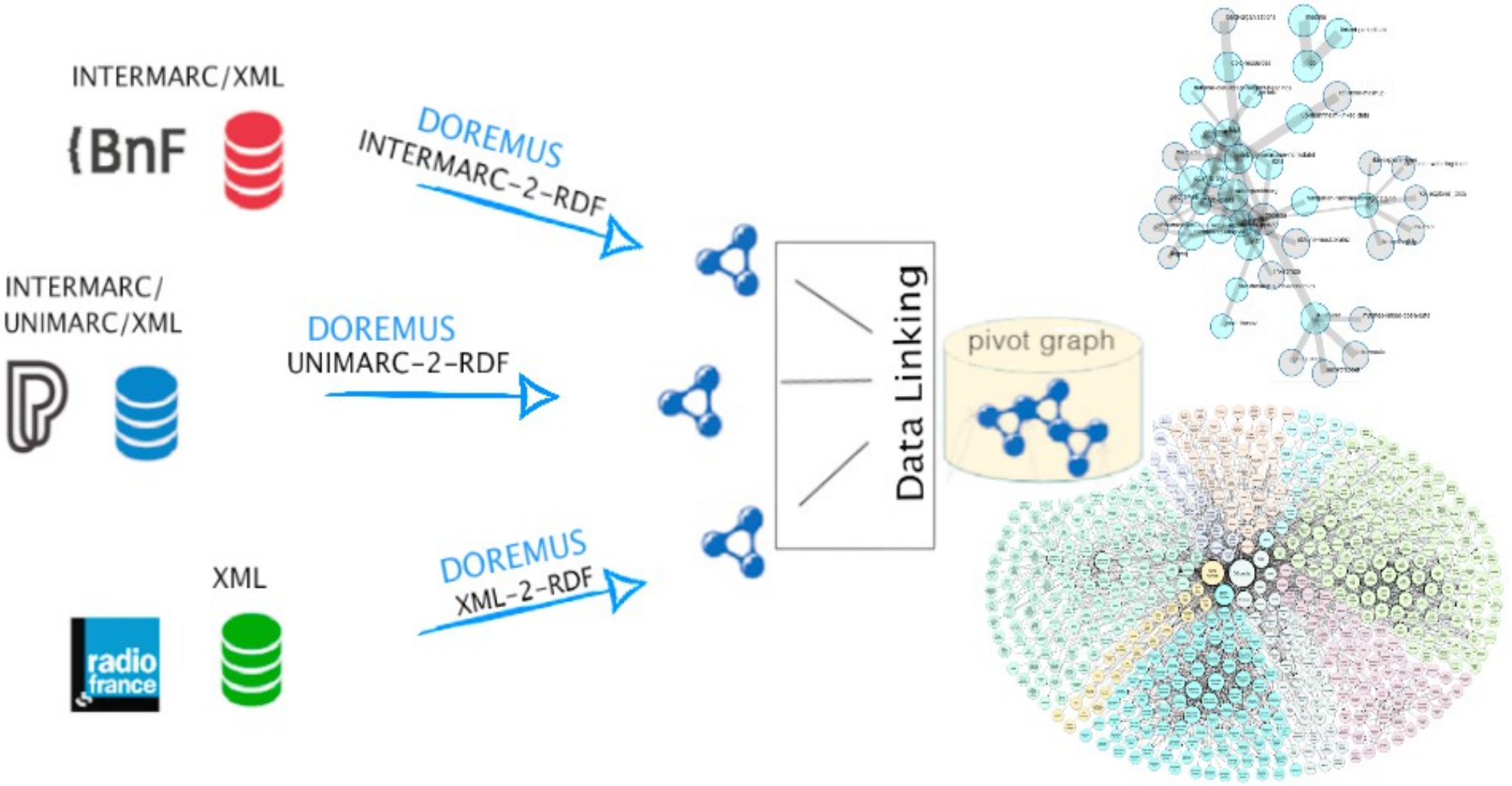
Data Life Cycle



We have our model and our controlled vocabulaires,
let's lift our data.

DOREMUS Data Lifting: Roadmap

1. Input Data
2. Conversion to Doremus RDF
3. Data Linking
4. Connect to the Web of Data



DOREMUS Data Lifting: Roadmap

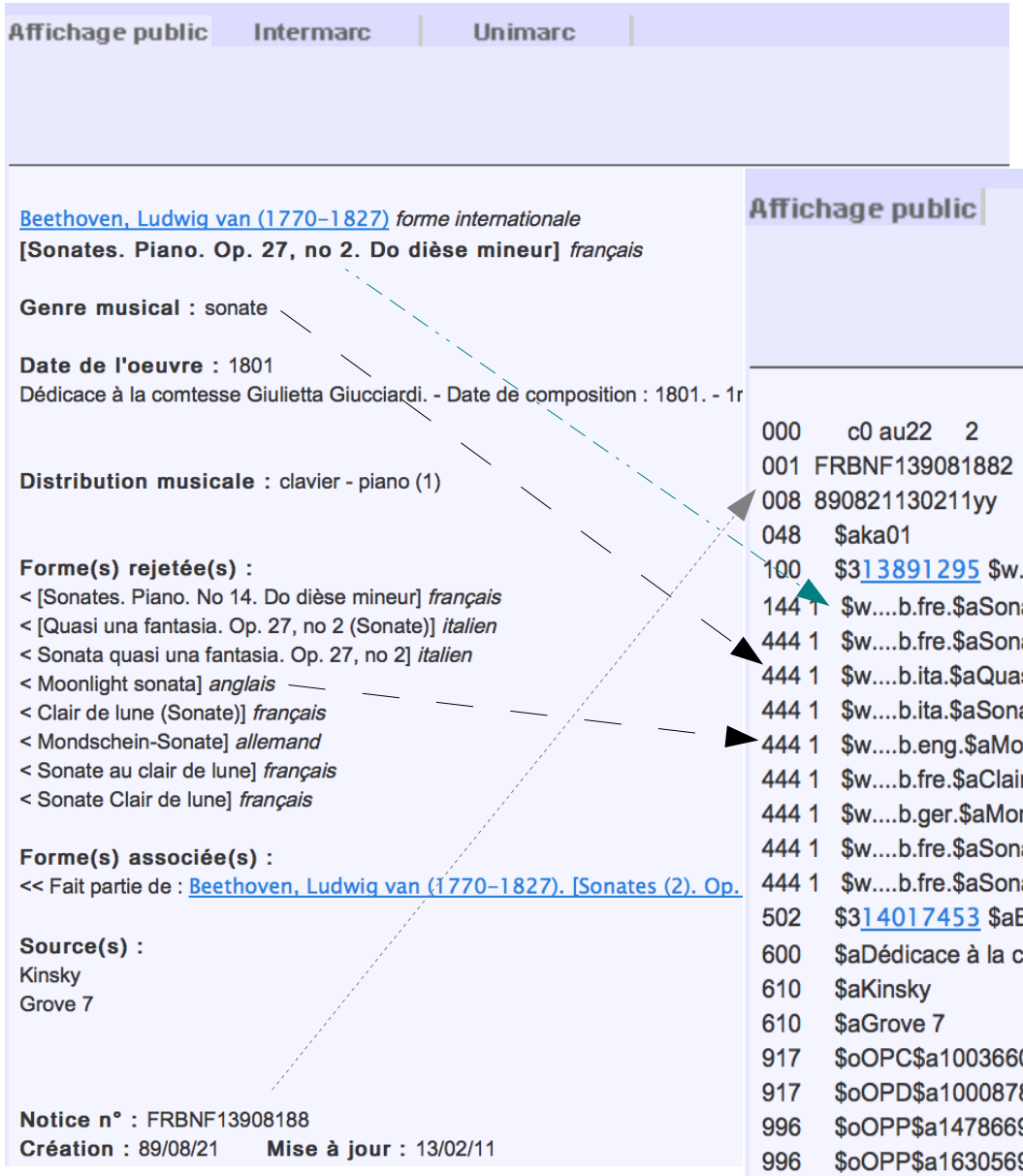
1. Input Data

INTERMARC/XML

UNIMARC/XML

XML

Input Data



PUBLIC VIEW

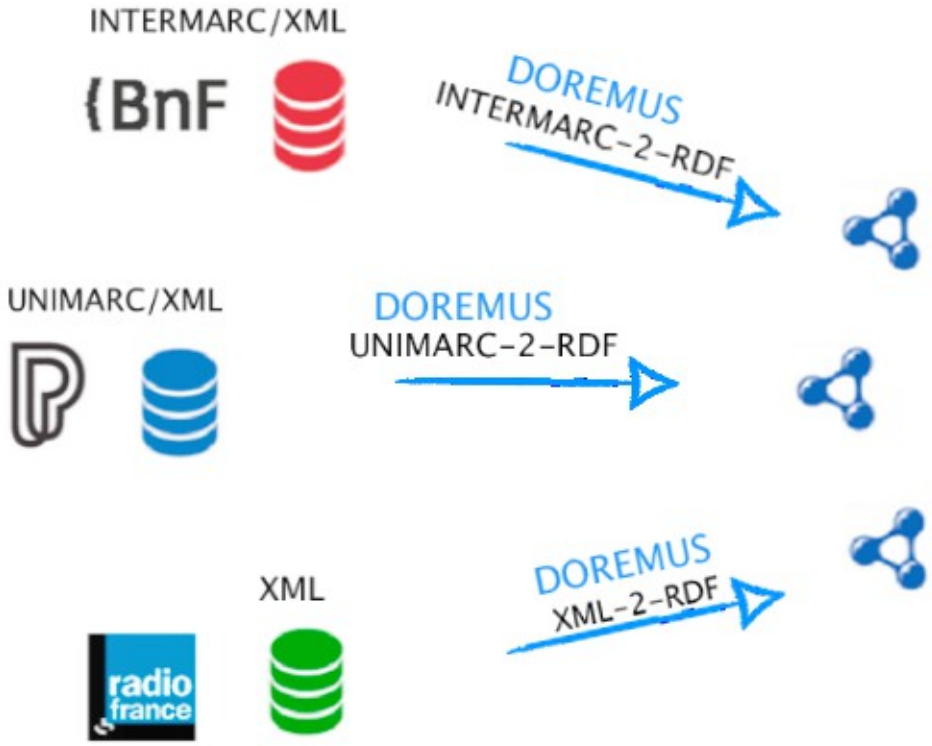
INTERMARC

Affichage public | Intermarc | Unimarc

```
000     c0 au22    2
001  FRBNF139081882
008  890821130211yy   sn      1801                    010
048    $aka01
100    $313891295 $w.0..b.....$aBeethoven$mLudwig van$d1770-1827
144 1  $w....b.fre.$aSonates$bPiano$pOp. 27, no 2$tDo dièse mineur
444 1  $w....b.fre.$aSonates$bPiano$nNo 14$tDo dièse mineur
444 1  $w....b.ita.$aQuasi una fantasia$pOp. 27, no 2$eSonate
444 1  $w....b.ita.$aSonata quasi una fantasia$pOp. 27, no 2
444 1  $w....b.eng.$aMoonlight sonata
444 1  $w....b.fre.$aClair de lune$eSonate
444 1  $w....b.ger.$aMondschein-Sonate
444 1  $w....b.fre.$aSonate au clair de lune
444 1  $w....b.fre.$aSonate Clair de lune
502    $314017453 $aBeethoven$mLudwig van$d1770-1827$t[Sonates (2). Op. 27]
600    $aDédicace à la comtesse Giulietta Giucciardi$aDate de composition : 1801$a1re éd. : Vienne : Cappi, 1802
610    $aKinsky
610    $aGrove 7
917    $oOPC$a100366020
917    $oOPD$a100087890$bATUM
996    $oOPP$a14786691$d20060411
996    $oOPP$a16305693$d20130211
```

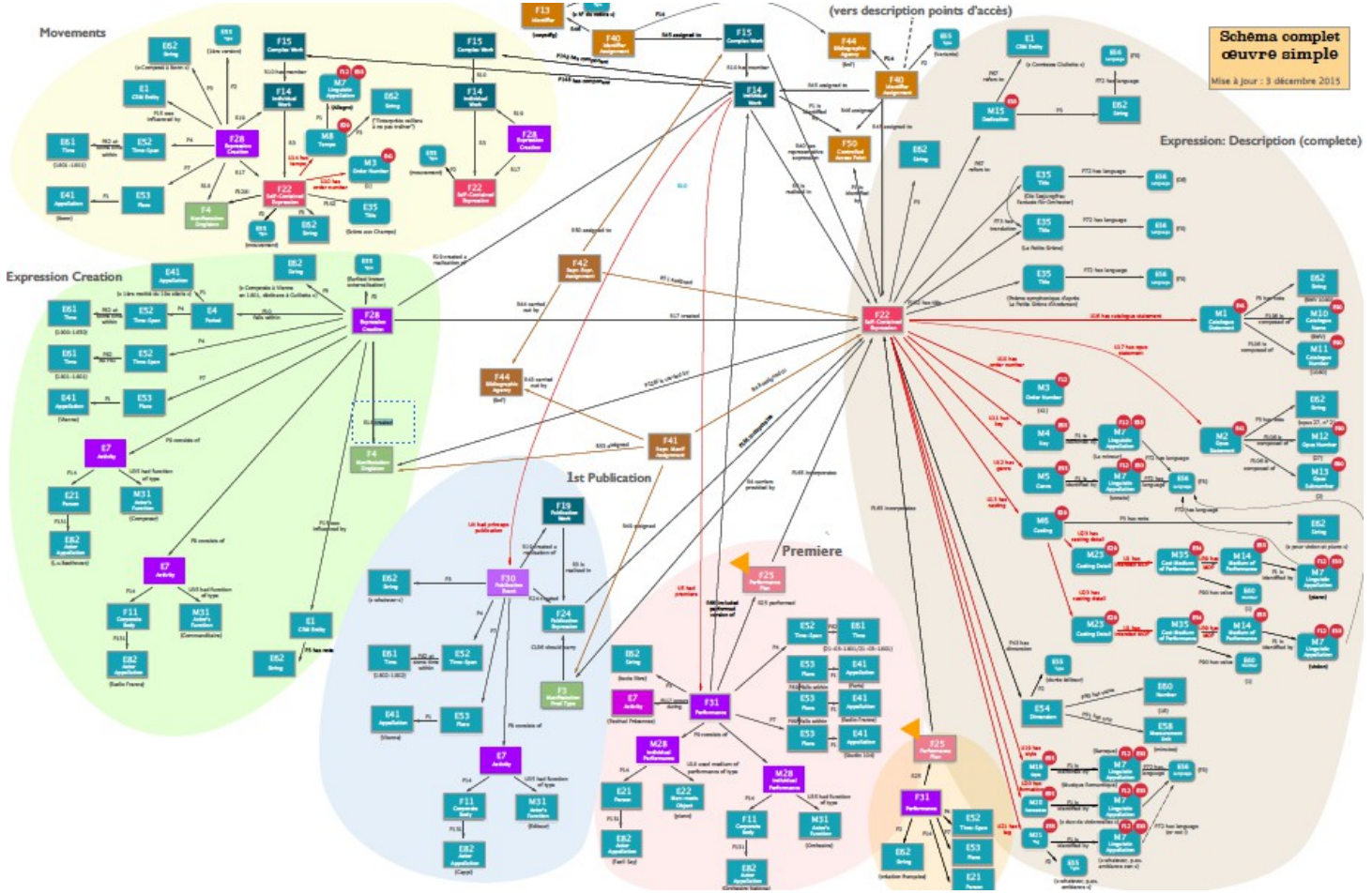
DOREMUS Data Lifting: Roadmap

2. Conversion to Doremus RDF



Data Conversion to RDF

Remember the Doremus model?...



Let's do DOREMUS RDF!

Data Conversion to RDF

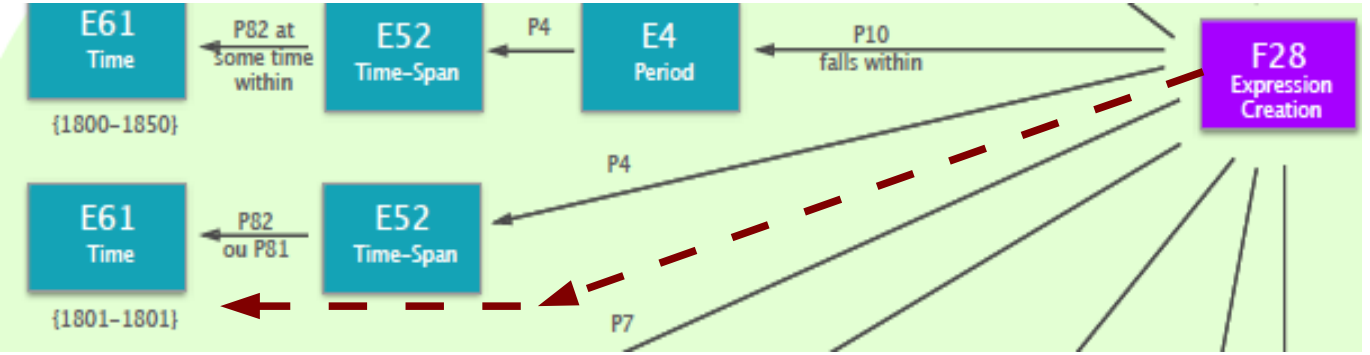
Expert-defined mapping rules

- Where to look for information and how to interpret it
- Implementing the DOREMUS model
- Reflect the practices of each institution: a mapping table *per* institution

Data Conversion to RDF

Expert-defined mapping rules

An example



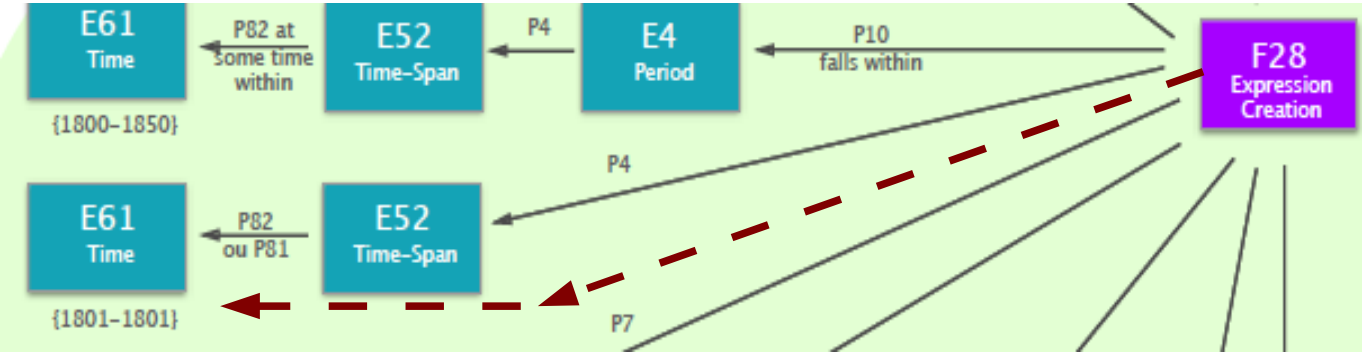
Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

What to look for?

Data Conversion to RDF

Expert-defined mapping rules

An example



Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

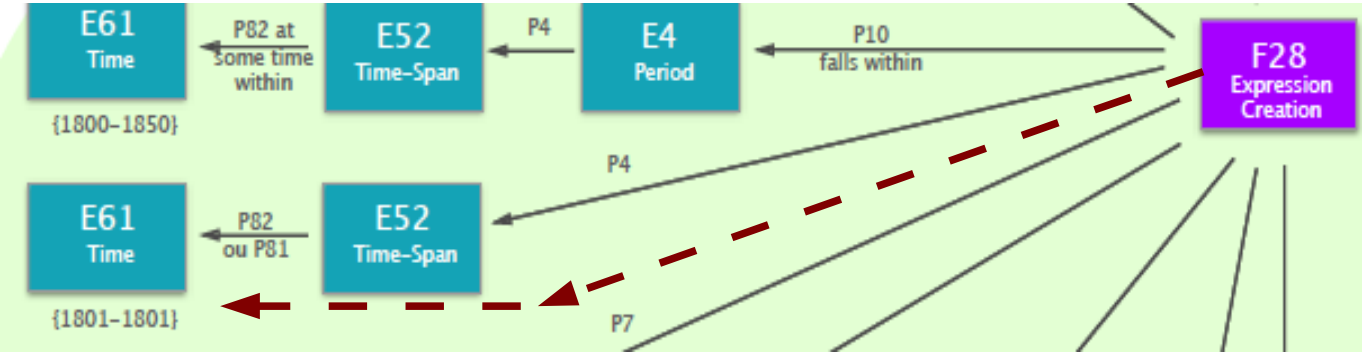
What to look for?

Where to look?

Data Conversion to RDF

Expert-defined mapping rules

An example



Model

Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive

What to look for?

MARC file

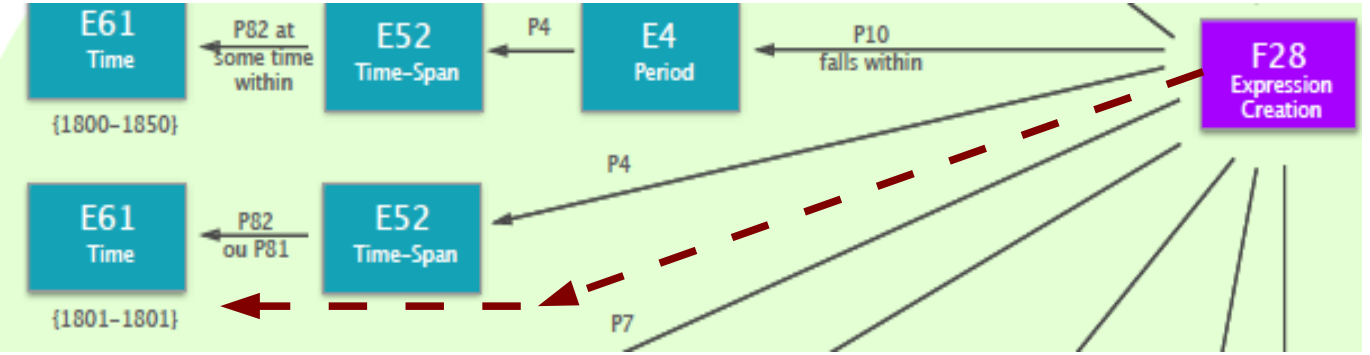
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

Where to look?

Data Conversion to RDF

Expert-defined mapping rules

An example



Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

What to look for?

Where to look?

Data Conversion to RDF

DOREMUS resource URI naming convention

The DOREMUS convention combines the *best practices* (see the DataLift project [6]) with the *DOREMUS model*

the Doremus convention

http://data.doremus.org/human-readable-name /UUID

*the **class** from the DOREMUS model*

Example:

http://data.doremus.org/expression/b90b3b97-2526-4152-95bb-273

Data Conversion to RDF

The DOREMUS property naming convention

Properties in the DOREMUS ontology:
three namespaces

- CIDOC-CRM **cidoc-crm**: <http://www.cidoc-crm.org/cidoc-crm/>
- FRBRoo **frbroo**: <http://erlangen-crm.org/efrbroo/>
- DOREMUS **mus**: <http://data.doremus.org/ontology/>

Constructing a property URI: concatenate the namespace URI and the property identifier (code + name in the model)

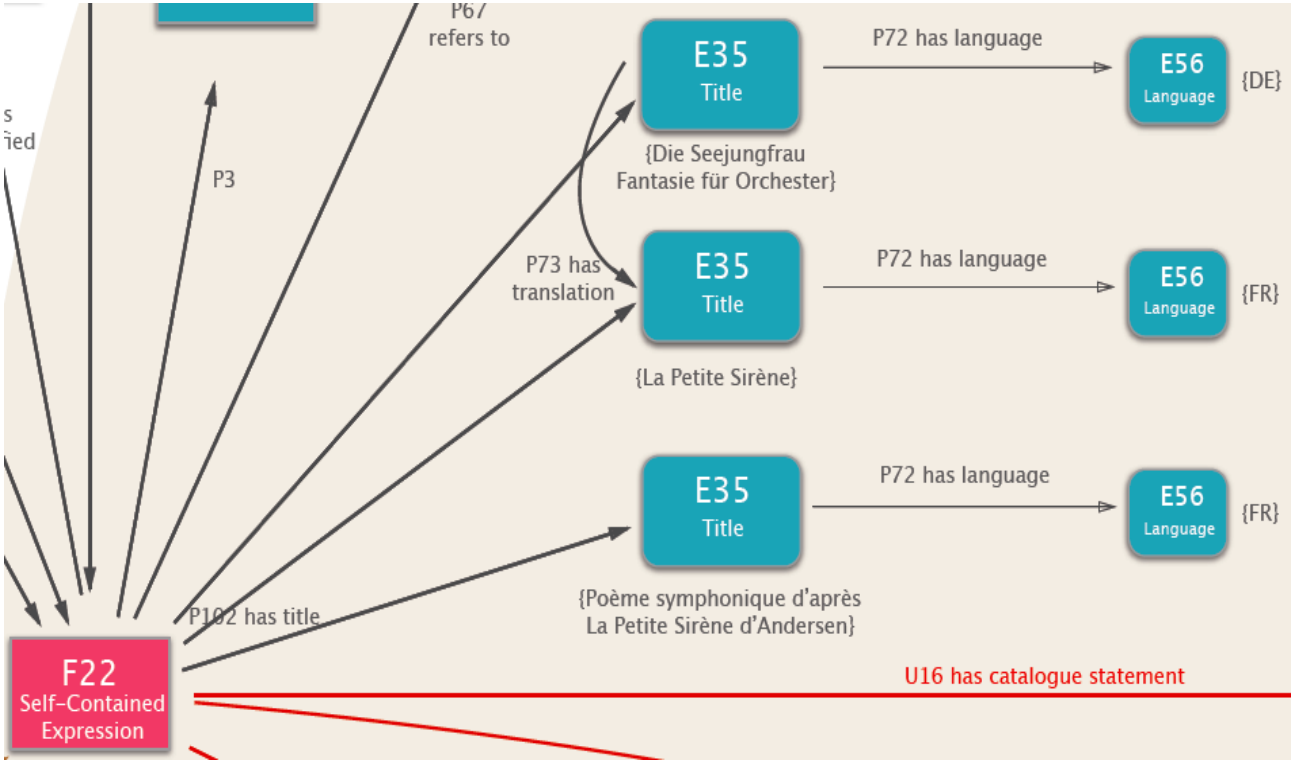
→ *see next slide.*

Data Conversion to RDF

The DOREMUS property naming convention

Properties are identified by their codes followed by their names.

CIDOC-CRM properties:



P102_has_title

P72_has_language

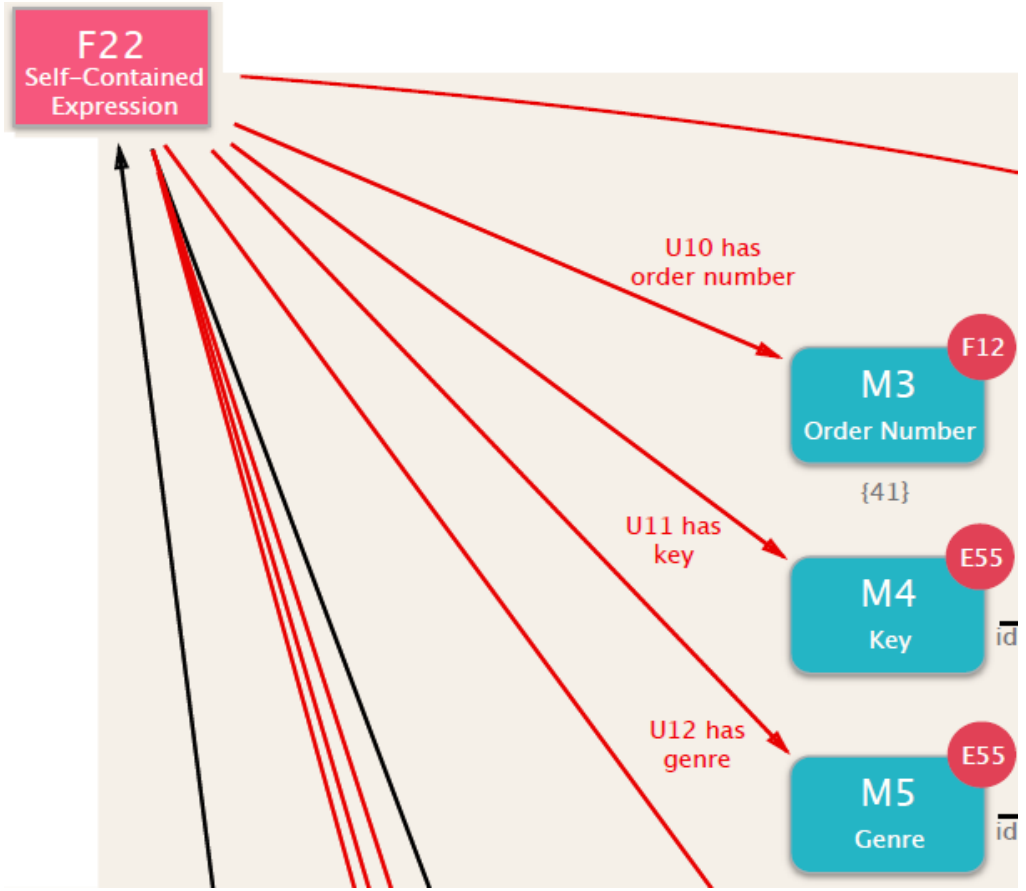
P73_has_translation

The CIDOC-CRM ns: @prefix **cidoc-crm**: <http://www.cidoc-crm.org/cidoc-crm/>

Data Conversion to RDF

The DOREMUS property naming convention

DOREMUS properties:



U11_has_key

U12_has_genre

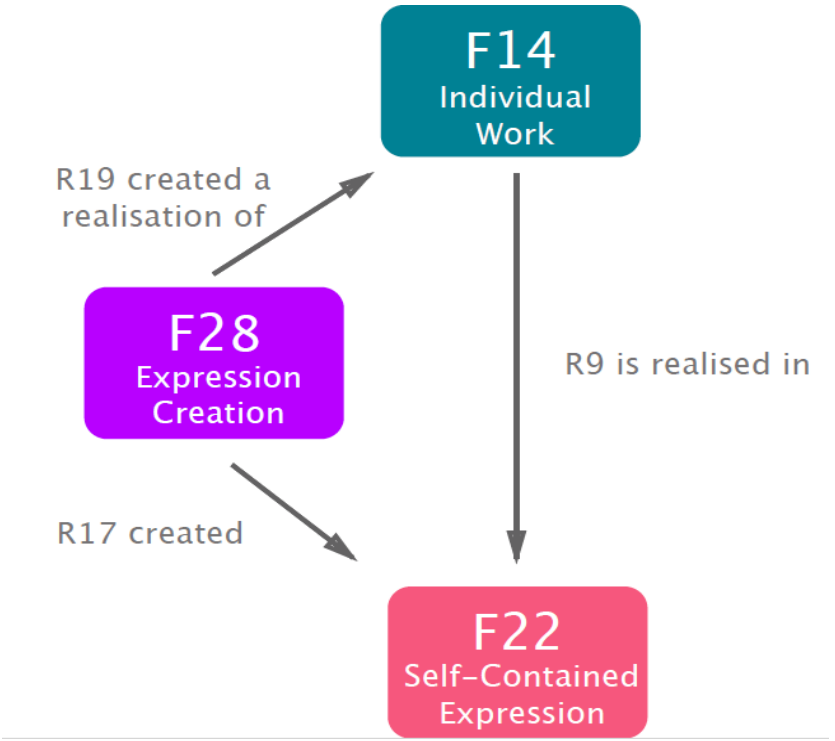
P10_has_order_number

The DOREMUS namespace: @prefix **mus**: <http://data.doremus.org/ontology/>

Data Conversion to RDF

The DOREMUS property naming convention

FRBRoo properties:



R17_created

R9_is_realized_in

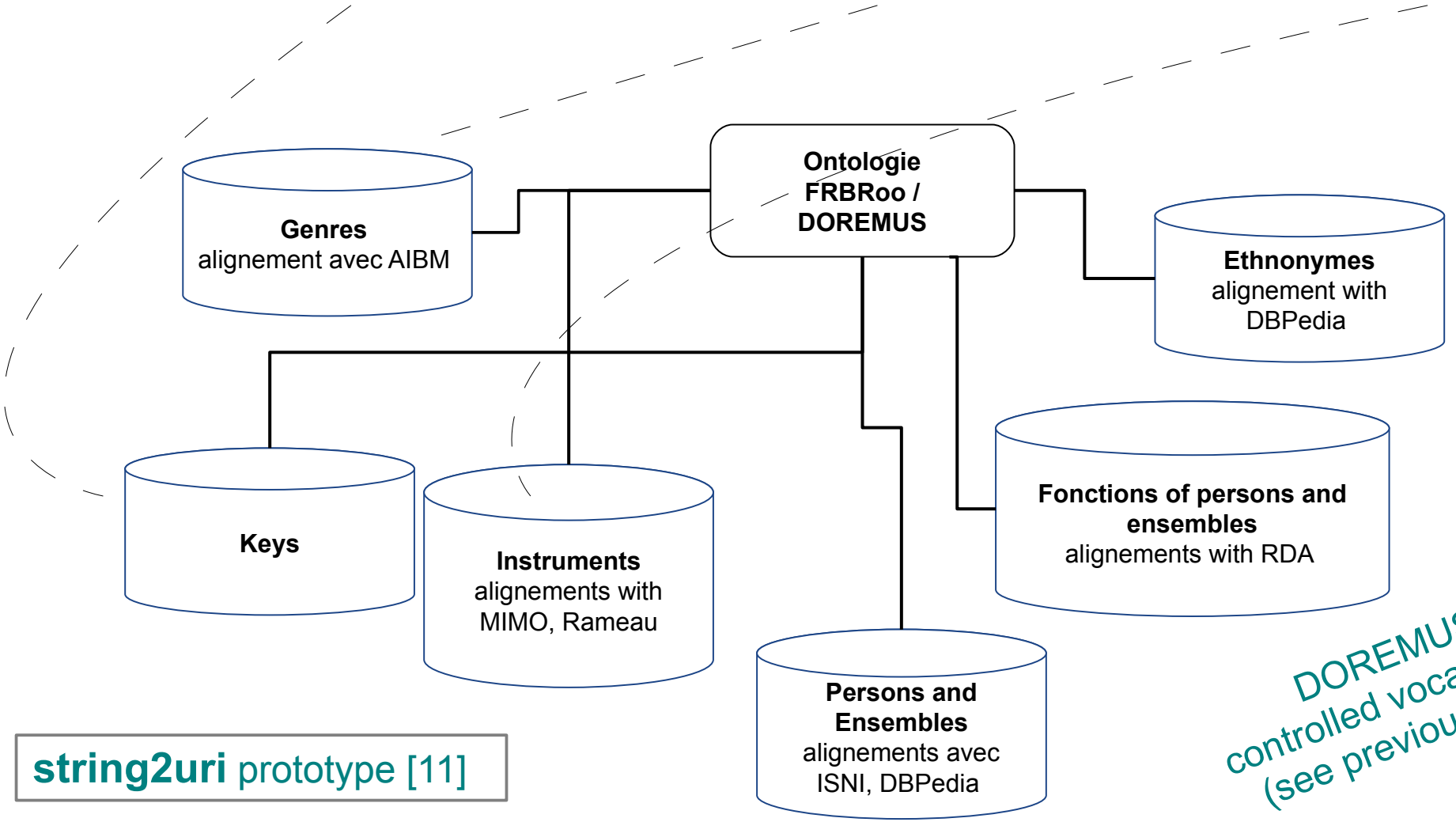
R19_created_a_realisation_of

The FRBRoo namespace: @prefix **frbroo**: <http://erlangen-crm.org/efrbroo/>

Data Conversion to RDF

Data type properties / Object properties

U11_has_key “C-sharp” U12_has_genre “symphony” U13_has_casting “piano”

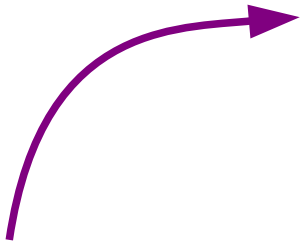


string2uri prototype [11]

DOREMUS controlled vocabularies (see previous talks).

Data Conversion to RDF

Example: a BNF TUM



```
001 FRBNF139081882
008 890821130211yy
048   $aka01
100   $313891295$w.0..
144 1 $w....b.fre.$aSc
444 1 $w....b.fre.$aSc
444 1 $w....b.ita.$aQu
444 1 $w....b.ita.$aSc
444 1 $w....b.eng.$aMc
444 1 $w....b.fre.$aCl
444 1 $w....b.ger.$aMc
444 1 $w....b.fre.$aSc
444 1 $w....b.fre.$aSc
502   $314017453$aBeet
600   $aDédicace à la comtesse Giulietta Giucciardi$aDate de composition : 1801$a1re éd. : Vienne : Cappi, 1802
610   $aKinsky
610   $aGrove 7
917   $oOPC$a100366020
917   $oOPD$a100087890$bATUM
996   $oOPP$a14786691$d20060411
996   $oOPP$a16305693$d20130211
400   $w....b.....$aBeethoven$mLudwig von$d1770-1827
```

```
<http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>
  a    efrbroo:F22_Self-Contained_Expression ;
  mus:U70_has_title    "Clair de lune"@fr, " Sonate Clair de lune"@fr, "Quasi una fantasia"@it,
  "Mondschein-Sonate"@de, "Sonates"@fr, "Quasi una fantasia"@it, "Sonata quasi una fantasia"@it,
  "Moonlight sonata"@en ;
  mus:U10_has_order_number   "14"^^xsd:int ;
  mus:U11_has_key   <http://data.doremus.org/vocabulary/key/cxm> ;
  mus:U12_has_genre   <http://data.doremus.org/vocabulary/iaml/genre/sn> ;
  mus:U13_has_casting   <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea/casting/1> ;
  mus:U17_has_opus_statement   <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea/opus/27-2> ;
  dcterms:identifier   "13908188" ;
<http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892>
  a    efrbroo:F28_Expression_Creation ;
  efrbroo:R17_created   <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>;
  efrbroo:R19_created_a_realisation_of   <http://data.doremus.org/work/30256b51-d277-3688-ad62-560ae982ff2f> ;
  ecrm:P9_consists_of   <http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892/activity/1> ;
  ecrm:P4_has_time-span   <http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892/time> ;
```

Data Conversion to RDF

You can find the current version of the converter MRAC-2-RDF here.

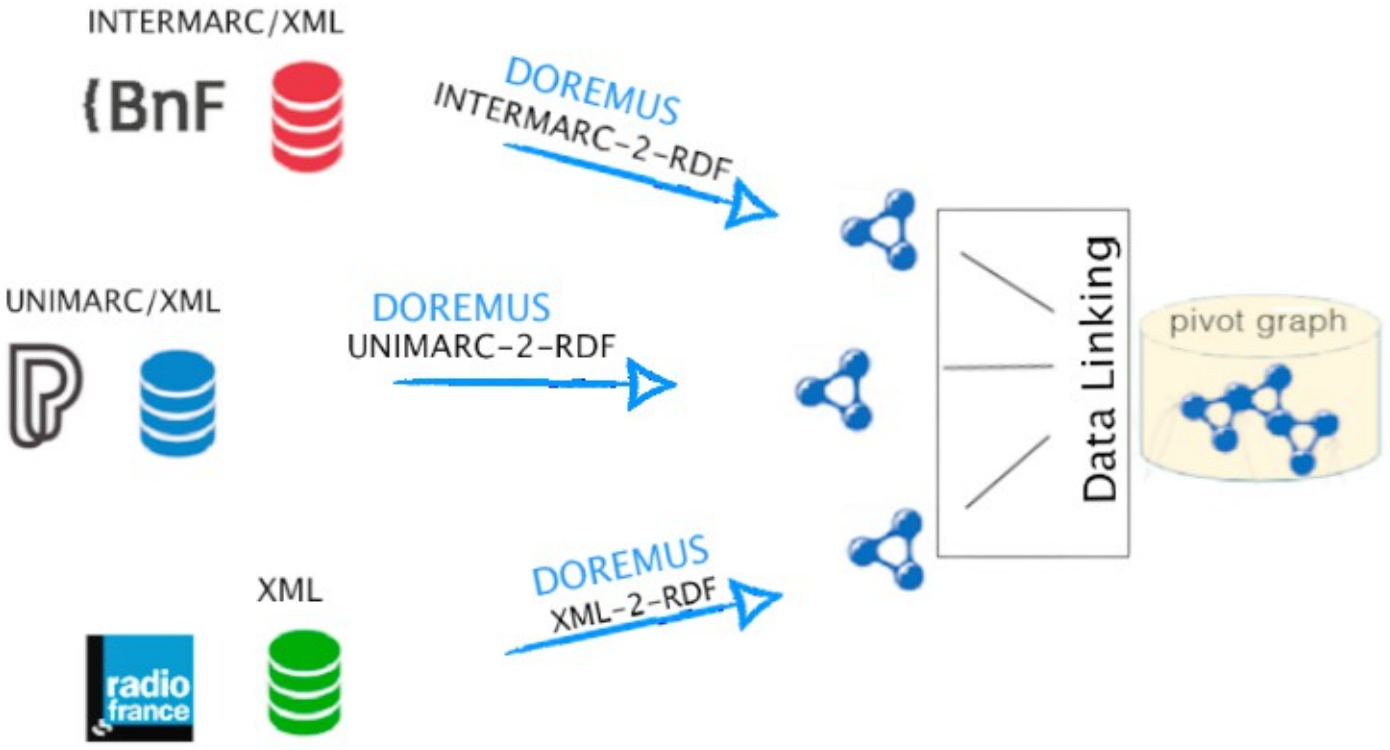
https://github.com/DOREMUS-ANR

```
001 FRBNF139081882
008 890821130211yy    sn
048   $aka01
100   $313891295$w.0..b.....$aF
144 1 $w....b.fre.$aSonates$bPi
444 1 $w....b.fre.$aSonates$bPi
444 1 $w....b.ita.$aQuasi una f
444 1 $w....b.ita.$aSonata quas
444 1 $w....b.eng.$aMoonlight s
444 1 $w....b.fre.$aClair de lu
444 1 $w....b.ger.$aMondschein-
444 1 $w....b.fre.$aSonate au c
444 1 $w....b.fre.$aSonate Clai
502   $314017453$aBeethoven$mLu
600   $aDédicace à la comtesse
610   $aKinsky
610   $aGrove 7
917   $oOPC$a100366020
917   $oOPD$a100087890$bATUM
996   $oOPP$a14786691$d20060411
996   $oOPP$a16305693$d20130211
400   $w....b.....$aBeethoven$mLudwig von$d1770-1827
```

```
<http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>
  a    efrbroo:F22_Self-Contained_Expression ;
  mus:U70_has_title    "Clair de lune"@fr, " Sonate Clair de lune"@fr, "Quasi una fantasia"@it,
  "Mondschein-Sonate"@de, "Sonates"@fr, "Quasi una fantasia"@it, "Sonata quasi una fantasia"@it,
  "Moonlight sonata"@en ;
  mus:U10_has_order_number  "14"^^xsd:int ;
  mus:U11_has_key  <http://data.doremus.org/vocabulary/key/cxm> ;
  mus:U12_has_genre  <http://data.doremus.org/vocabulary/iaml/genre/sn> ;
  mus:U13_has_casting  <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea/casting/1> ;
  mus:U17_has_opus_statement  <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea/opus/27-2> ;
  dcterms:identifier  "13908188" ;
<http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892>
  a    efrbroo:F28_Expression_Creation ;
  efrbroo:R17_created  <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>;
  efrbroo:R19_created_a_realisation_of  <http://data.doremus.org/work/30256b51-d277-3688-ad62-560ae982ff2f> ;
  ecrm:P9_consists_of  <http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892/activity/1> ;
  ecrm:P4_has_time-span  <http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892/time> ;
```

DOREMUS Data Lifting Roadmap

3. Data Linking

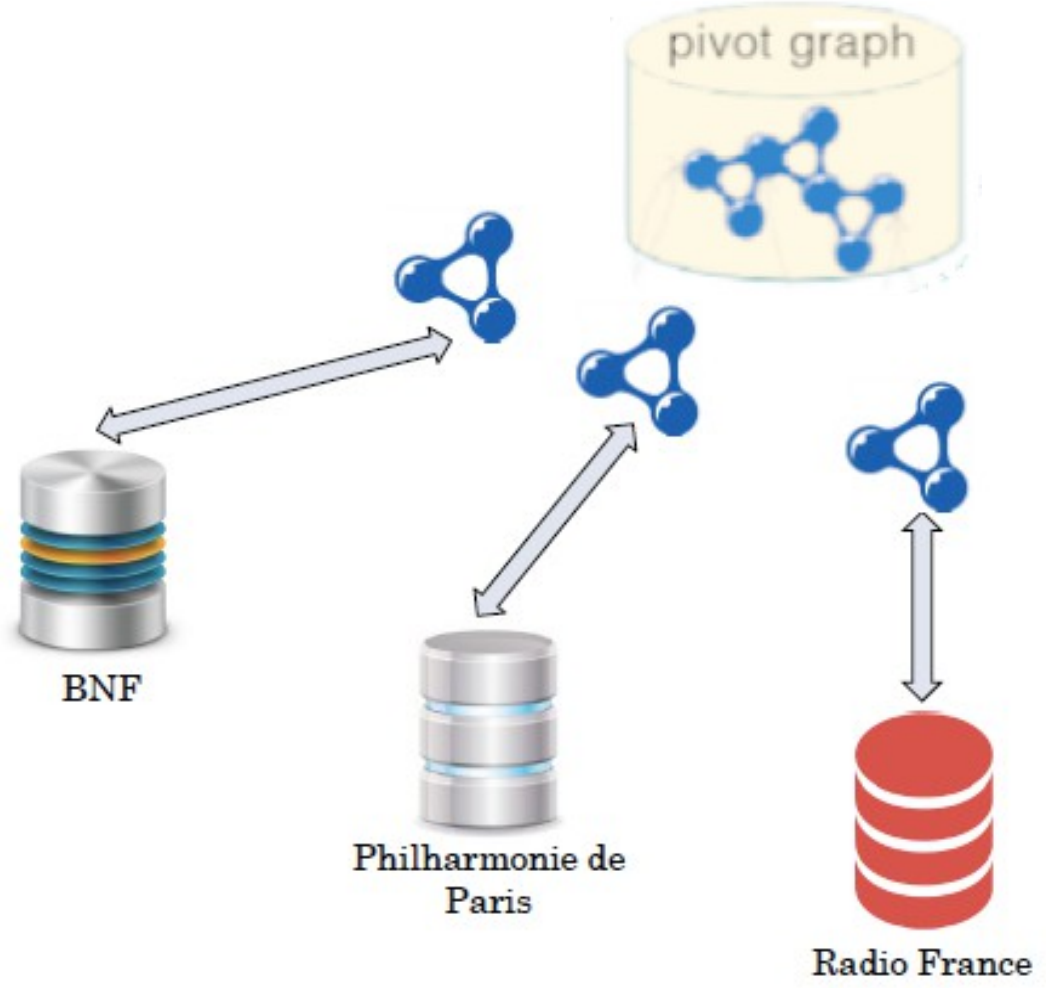


Data Linking

DOREMUS: What are we linking?

An RDF graph of music works per institution.

+ A pivot graph containing all works.



A work exists potentially in each of the 3 RDF datasets identified by different URIs.

Among the reasons for this decision:

– the descriptions of a given work across institutions are not uniform (see following slides)

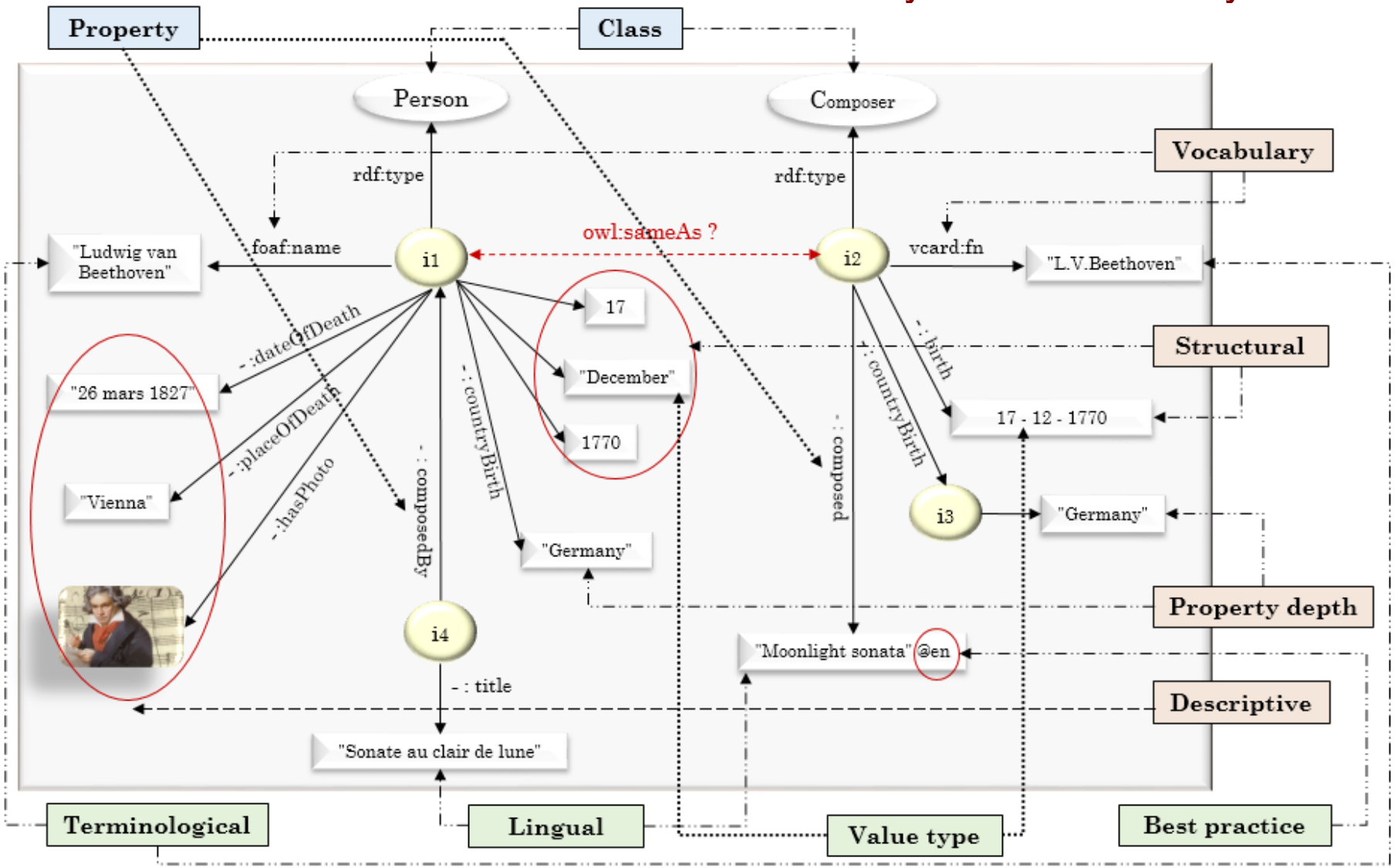
– not always a 1:1 correspondance

– independence of representation

So, we need to link these datasets!

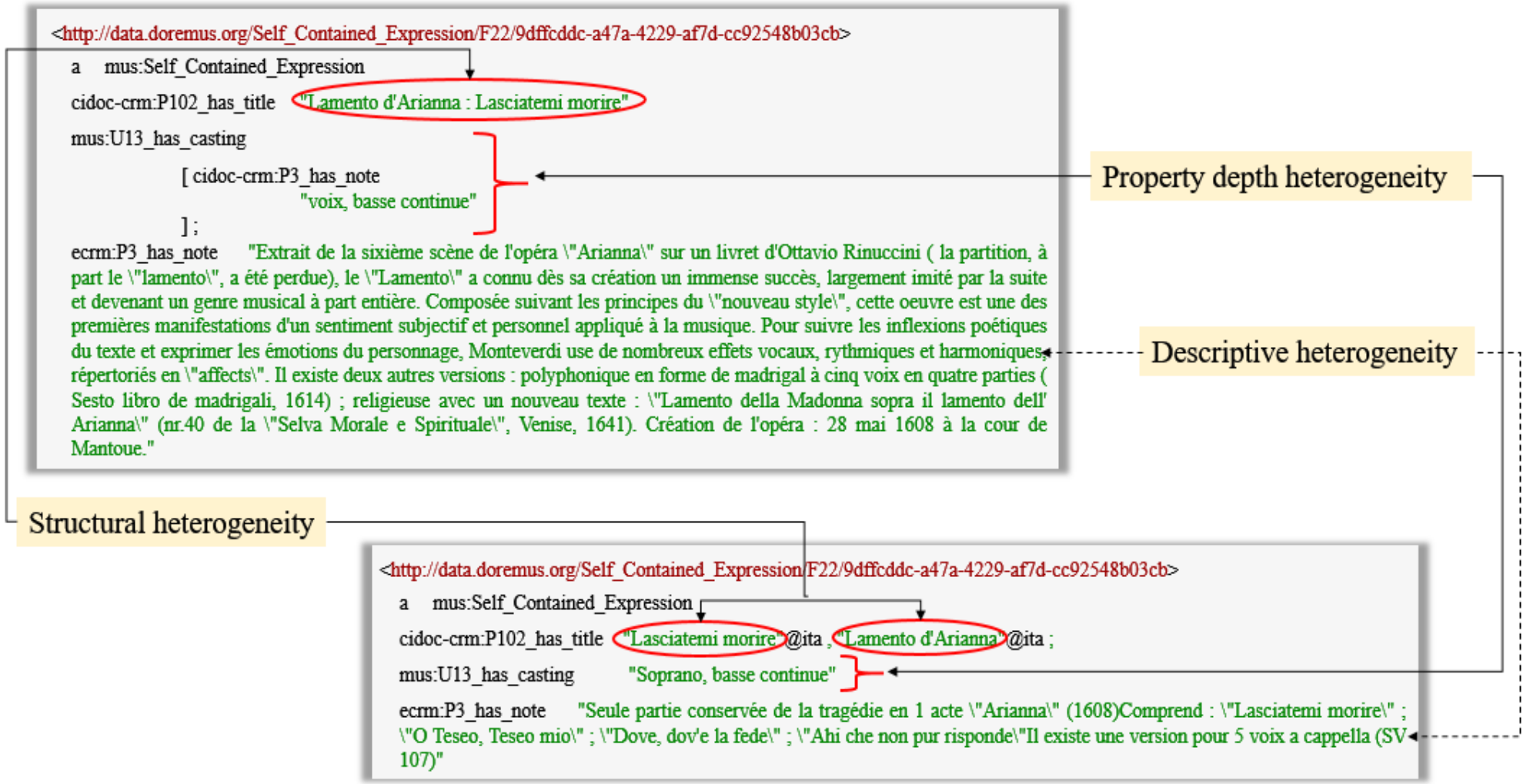
Data Linking

Why is it not that easy...



Datasets can be highly heterogeneous!

Data Linking



Data Linking

using state-of-the-art tools

Lessons learned

– SILK is the only off-the-shelf tool that returns results without any data re-writing

– Heterogeneities in titles appear to be very problematic

– Missing values of key-properties (e.g., catalog number) are problematic

– Multilingual information is hard to handle correctly

A novel dedicated tool *Legato*

– Each work → a bag of litteral values

– Indexing techniuqes combined with key detection and ranking

– Effective data cleaning and link filtering

Data Linking

Aligning with ***Legato***

	DS_SM		
	F-Measure	Precision	Recall
LEGATO	0.98	1.0	0.96
AML	0.70	0.68	0.72
SILK	0.32	0.7	0.21

	DS_HT		
	F-Measure	Precision	Recall
LEGATO	0.9	0.96	0.85
AML	0.69	0.63	0.77
SILK	0.45	0.98	0.29

The DOREMUS Playground

For those of you who would like to try all that out, check the DOREMUS Playground.
https://github.com/DOREMUS-ANR/doremus-playground

You will find sample data, benchmarks and configuraiton files used for SILK.

If you'd like to try linking with our dedicated tool ***Legato***, all is there: https://github.com/DOREMUS-ANR/legato

Coming up:

DOREMUS Instance Matching track at OAEI 2017
jointly with ISWC2017, Vienna, Austria.

From the user perspective

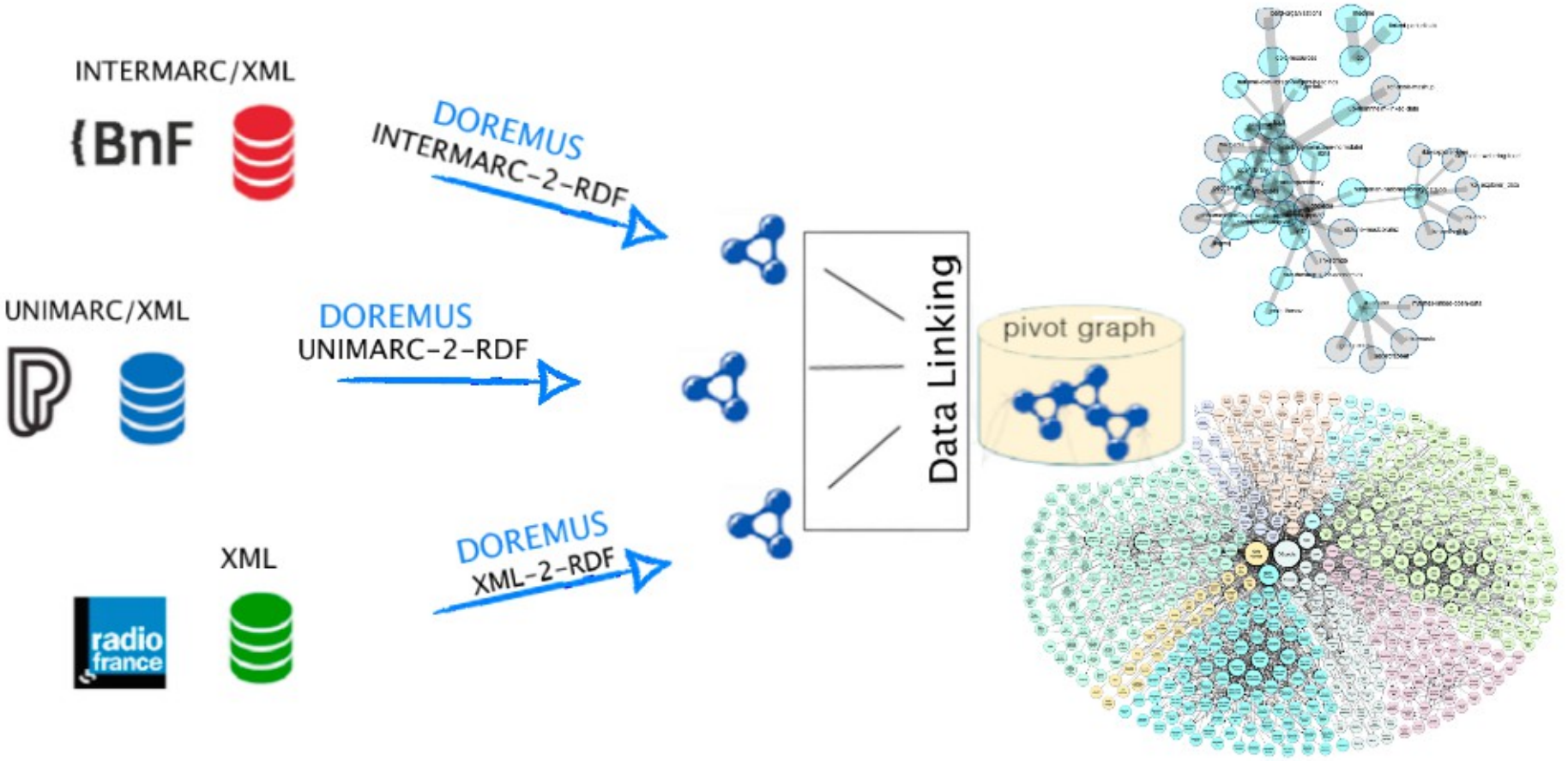
- Example queries
  - https://github.com/DOREMUS-ANR/knowledge-base/tree/master/query-examples
- Query & explore: http://data.doremus.org
  - SPARQL endpoint
  - Exploring the connected catalogs with Overture
  - Sample search results with Overture:

- https://github.com/DOREMUS-ANR/overture/blob/master/EXAMPLES.md

Thank you for listening.

DOREMUS Data Lifting: Roadmap

4. Connecting to the web of data



References and Links

[1] Ferrara, A., Nikolov, A., & Scharffe, F. (2013). Data linking for the semantic web. *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications*, *169*.

[2] Achichi, M., Bailly, R., Cecconi, C., Destandau, M., Todorov, K., & Troncy, R. (2015). DOREMUS: Doing Reusable Musical Data. *ISWC2015 P&D track.*

[3] Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk-A Link Discovery Framework for the Web of Data. *LDOW, 538*.

[4] Soru, T., Marx, E., & Ngonga Ngomo, A. C. (2015, May). ROCKER: A refinement operator for key discovery. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1025-1033). International World Wide Web Conferences Steering Committee.

[5] Symeonidou, D., Armant, V., Pernelle, N., & Saïs, F. (2014). SAKey: Scalable almost key discovery in rdf data. In *The Semantic Web–ISWC 2014* (pp. 33-49). Springer International Publishing.

[6] The DataLift project: http://datalift.org

[7] The DOREMUS github repository and playground: https://github.com/DOREMUS-ANR, https://github.com/DOREMUS-ANR/doremus-playground

[8] UNIMARC (authority records): http://www.ifla.org/publications/ifla-series-on-bibliographic-control-38

[9] UNIMARC (bibliographical records): http://www.ifla.org/publications/ifla-series-on-bibliographic-control-36

[10] INTERMARC: http://www.ifla.org/node/4858

[11] String2URI prototype: https://github.com/ThibWeb/stringtouri

[12] Instance matching track DOREMUS at OAEI: http://oaei.ontologymatching.org

[13] Destandau, M., Troncy, R., Todorov, K., Cecconi, C., Voisin, M., Canno; I., Leresche, F. (2016). Linked Data Approach for Structuring and Interlinking Musical Catalogs: How Three Major French Cultural Institutions Finally Came to an Agreement. *IFLA's satellite event : Data in Libraries: the big picture.*

1. Input Data

	BnF	**PP** Médiathèque	**PP** Concerts	**Radio France** Disco-thèque	**Radio France** Docu-mentation musicale	**Radio France** Docu-mentation sonore	**Target entity** ↓
Format	XML/ INTER MARC	XML/ UNIMARC	XML	XML	XML	XML	
Uniform Music Titles (TUM) & work entries	135 940	6 846			62 550		**Work**
Scores	89 184	30 319			9 154		**Expression**
Books		21 035					
CD/DVD/ Vinyls		8 602		340 609			**Performance**
Concerts		2 447	2 717		7 700	1 800	

1. Input Data

Introducing the MARC family

MARC:

Machine **R**eadable **C**ataloging

a bibliographical data exchange format

```
001 FRBNF139081882
008 890821130211yy    sn        1801
048   $aka01
100   $313891295$w.0..b.....$aBeethoven$mLudwig van$d1770-1827
144 1 $w....b.fre.$aSonates$bPiano$pOp. 27, no 2$tDo dièse mineur
444 1 $w....b.fre.$aSonates$bPiano$nNo 14$tDo dièse mineur
```

A **MARC** file is

- a succession of fields of different lengths, each carrying a label (a 3 digit number)

- each field is a succession of sub-fields (also of variable lengths)

- a sub-field is delimited by the “$” symbol

- sub-fields can repeat in order to “host” data of the same kind

Different variants of MARC...

- USMARC in the United States, CANMARC in Canada, UKMARC in the UK,...

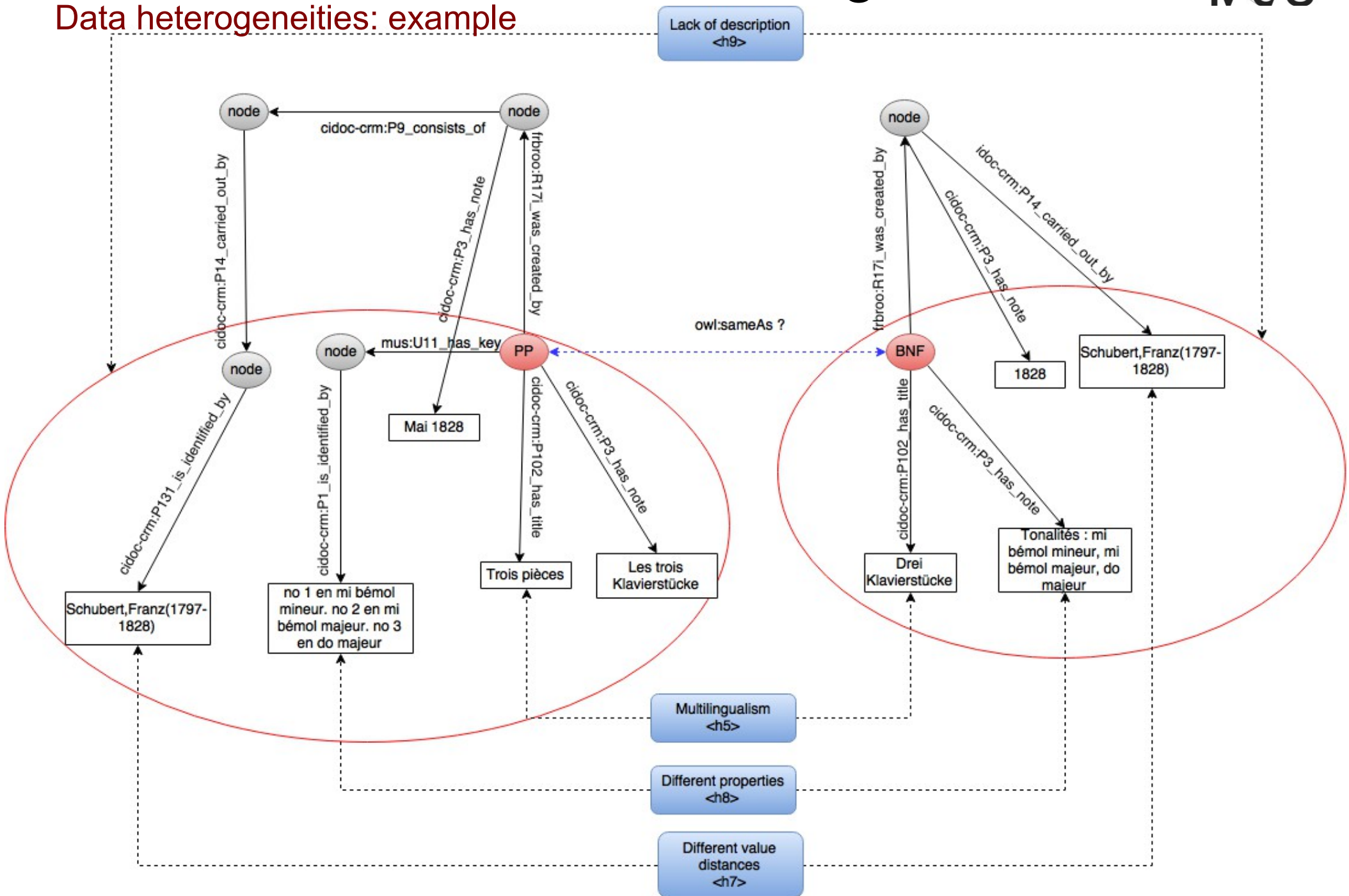
- MARC21 unifies USMARC, AUSMARC, UKMARC, CANMARC.

- INTERMARC is used by the BNF and other libraries in Paris and Lyon in France.

- UNIMARC was initially designed as a unique format for exchange between the different MARCs, it became the official french MARC format.

3. Data Linking

Data heterogeneities: example



Data Linking

...Anyone?

The 4th principle of the web of data:
when publishing data, provide links to other, already published data!

Link datasets on the web!

Data Linking

Links

A link-statement is a triple (as any other) that

links an instance from one dataset (*the subject*)

to an instance of another dataset (*the object*)

via a *link-predicate* coming from established vocabularies, such as owl:sameAs (meaning that the 2 instances are equivalent), but also skos:closeMatch, rdf:seeAlso, or other.

Example:

`http://yago-knowledge.org/resource/Ludwig_van_Beethoven`,
owl:sameAs, `http://dbpedia.org/resource/Ludwig_van_Beethoven`

Data Linking

Some basics:
The data linking processing chain

(1) preprocessing → (2) instance matching → (3) post-processing

– reduce the search-space, identify a set of pairs of linking candidates, identify key properties

– make instances comparable: models of representation, handling multilingualisme

See [4],[5].

– discover a link between two resources, give it a type and a confidence degree

See [1].

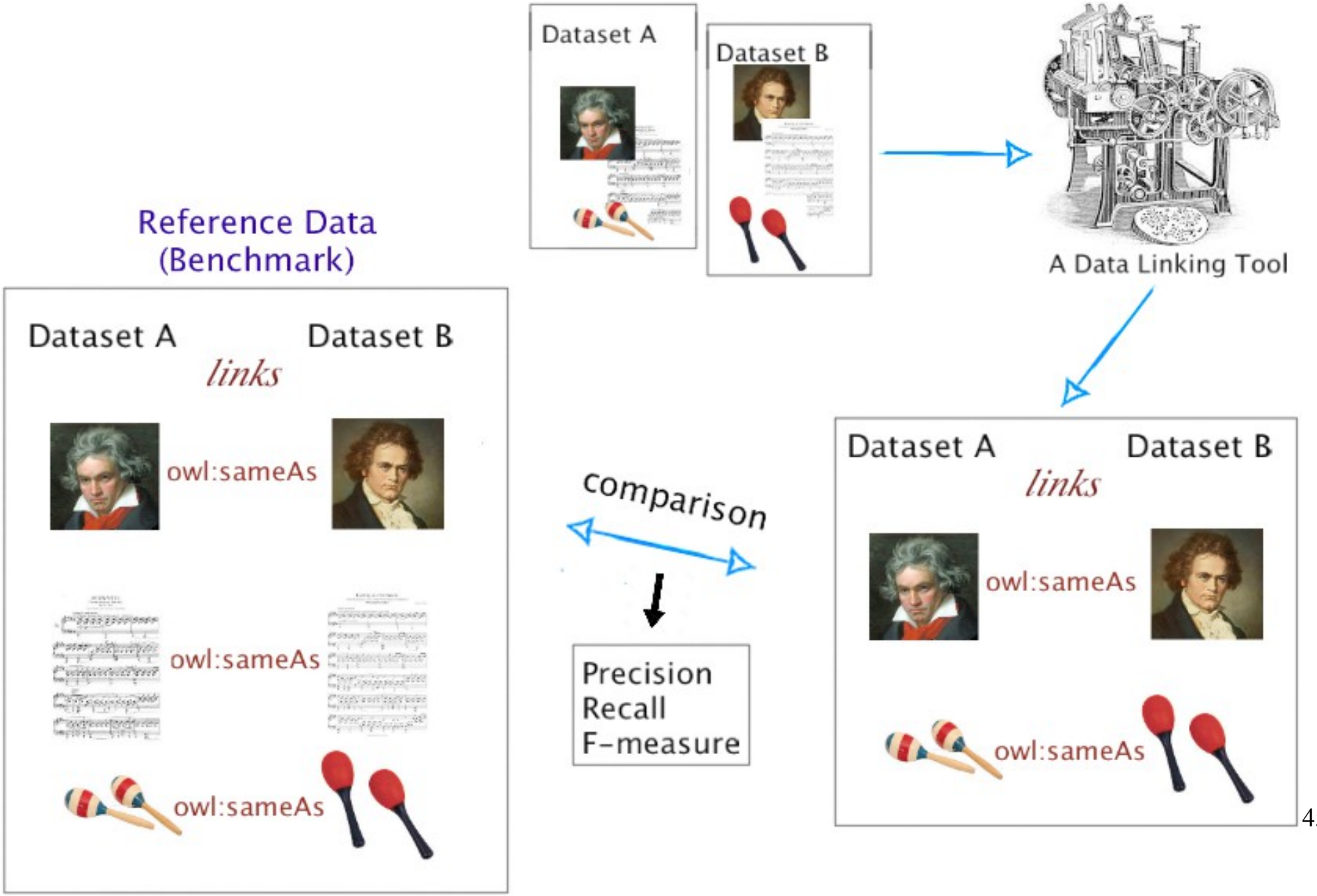
– filter out erroneous matches
– infer new ones

A plethora of tools:
LIMES, SILK, RiMOM, RDF-AI,... [1]

From a user perspective, the tool configuration is 90% of the task.

Data Linking

A common approach to develop and evaluate linking tools



Data Linking

The DOREMUS benchmark data

What are the **heterogeneities** manifested by music bibliographical data?

1. We asked experts to identify the most current problems that may appear
→ **Dataset DS-HT**

H1. Letters or numbers in the property values (particularly titles)
H2. Differences in spelling (terminological heterogeneity)
H3. Missing catalog numbers and/or opus numbers
H4. Multilingual titles
H5. Letters with diacritical signs
H6. Different value distances
H7. Different properties describing tonalities or instruments
H8. Missing properties (lack of description)
H9. Missing titles

2. We also asked experts to give us examples of very similar yet distinct works in order to test the capacity of tools to disambiguate between instances
→ **Dataset DS-SM.**