# POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER


# Mapping from Flat or Hierarchical Metadata Schemas to a Semantic Web Ontology

Justyna Walkowska, Marcin Werla

# Background: the SYNAT Project

- Financed by the National Center for Research and Development (grant No. grant no SP/I/1/77065/10, funding period: 2010 –2013)
- Part of national Strategic Research Program *Interdisciplinary system for interactive scientific and technical information*
- Coordinated by the University of Warsaw – ICM, with 16 leading Polish research and R&D institutions involved in the project.
- PSNC's role in the SYNAT project is to continue the development of the Polish digital libraries infrastructure and enrich this infrastructure with new services aimed to support both users and creators of digital libraries
- PSNC activities are focused on the Humanities domain as the majority of content available in Polish digital libraries is great research material for Humanities

# Our Task Outline

Based on data and metadata sources such as:

- Polish Digital Libraries Federation (80 digital libraries, 1,090,671 publications) http://fbc.pionier.net.pl/owoc/
- NUKAT Union Catalogue (2 million catalogues publications) http://www.nukat.edu.pl/
- Mona system used by National Museum in Warsaw (first package of 15,000 records) http://www.mnw.art.pl/

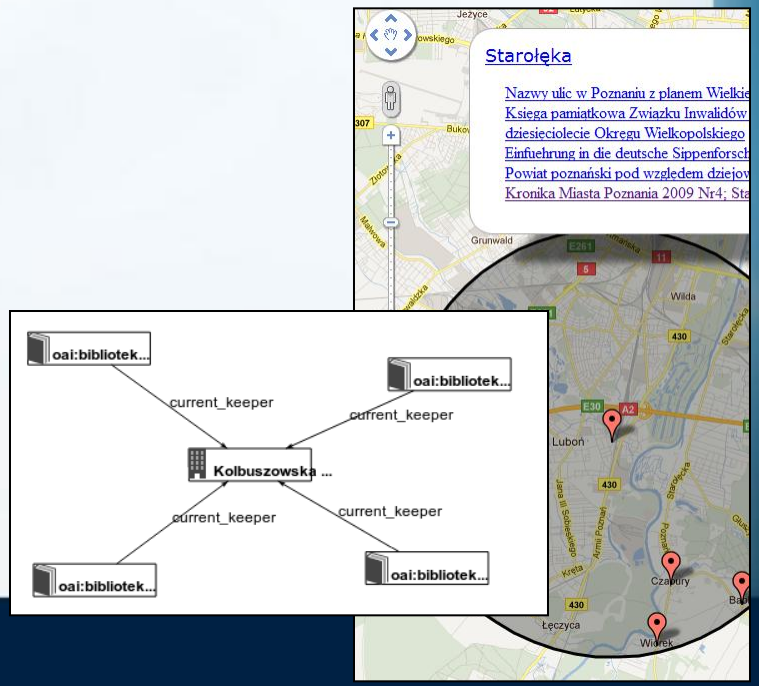
And auxiliary sources such as:

- Virtual International Authority File (http://viaf.org/)
- Geonames geographical database (http://www.geonames.org/)
- National Official Register of Territorial Division of the Country - TERYT (http://www.stat.gov.pl/bip/36_ENG_HTML.htm)

Build an OWL Knowledge Base with a portal over it.

The KB construction steps are:

- metadata ingestion
- cleaning and normalization
- **mapping from flat schemas to OWL ontologies**
- enrichment
- relation detection
- reasoning process & consistency checks
- KB deployment (hundreds of millions of triples)
- KB operation with incremental updates and periodical re-build.



# Mapping Sub-Task Outline

Based on data in the following schemas:

- Dublin Core (http://dublincore.org/documents/dces/)
- PLMET (http://dl.psnc.pl/community/display/FBCMETGUIDE)
- MARC 21 (MAchine-Readable Cataloging,

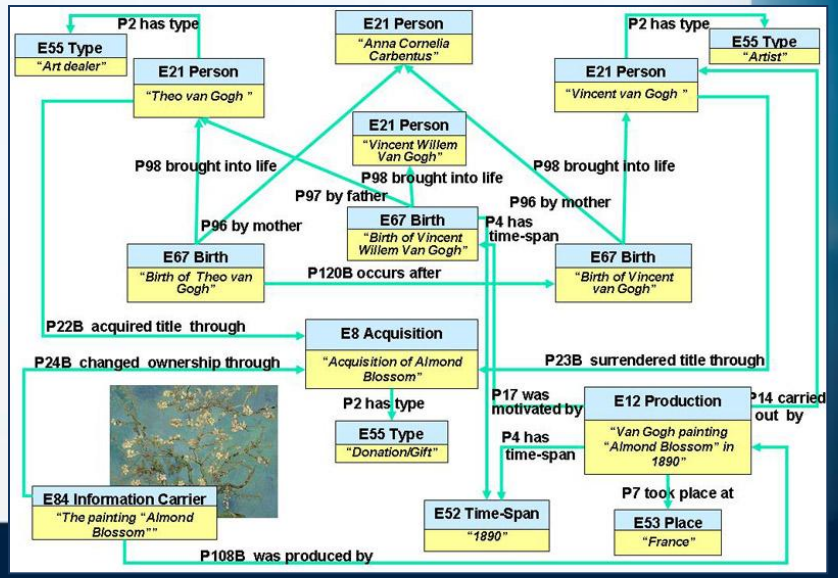
e.g. http://www.loc.gov/marc/bibliographic/)

- Mona System format (http://www.jws.com.pl/mona/mona_lista.html)

Generate RDF triples consistent with:

- CIDOC CRM for museum objects (http://www.cidoc-crm.org/)
- FRBRoo for library objects (http://www.cidoc-crm.org/frbr_inro.html)

Another possible target ontology is the Europeana Data Model (EDM) (http://www.europeana.eu/schemas/edm/) now being introduced by Europeana.



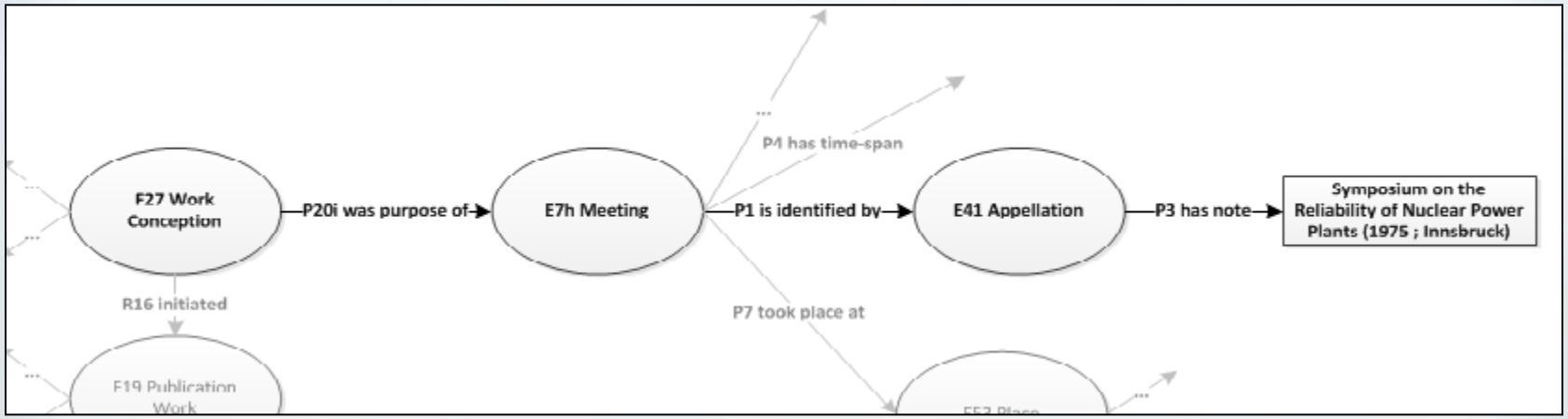

http://www.cidoc-crm.org/crm_core/core_examples/van_gogh.htm

# Our Mapping Requirements in a Nutshell

To go from this:

```
<datafield tag="111" ind1="2" ind2=" ">
    <subfield code="a">Symposium on the Reliability of Nuclear Power Plants</subfield>
    <subfield code="d">(1975 ;</subfield>
    <subfield code="c">Innsbruck).</subfield>
</datafield>
```

To this:



# Mapping Requirements List

**General metadata mapper requirements:**

- Element 1 to Element 2 mapping
- Attribute-value-based mapping
- XML structure-based mapping rules
- XML Structure in the Output
- Patterns
- Substring Mapping
- Concatenation
- Value Maps

**Semantic Web mapper requirements:**

- Ontology Paths
- Record Level Identifiers
- Element Level Identifiers
- Iterable Identifiers
- Static Paths
- Classes vs. Instances on Path
- Repeatable URI's
- Unique vs. Shared URI's
- Existential Conditions

# General Requirements: Structure

- Element 1 to Element 2 mapping

```
<a>AAA</a>
<b>BBB</b>
```

```
<c>AAA</c>
<d>BBB</d>
```

- Attribute-value-based mapping
- XML structure-based mapping rules

```
<datafield tag="050" ind1="0" ind2="0">
    <subfield code="a">DS113.3</subfield>
    <subfield code="b">.K56 2001</subfield>
</datafield>
<datafield tag="100" ind1="1" ind2=" ">
    <subfield code="a">Kimmerling, Baruch</subfield>
    <subfield code="d">(1939- ).</subfield>
</datafield>
```

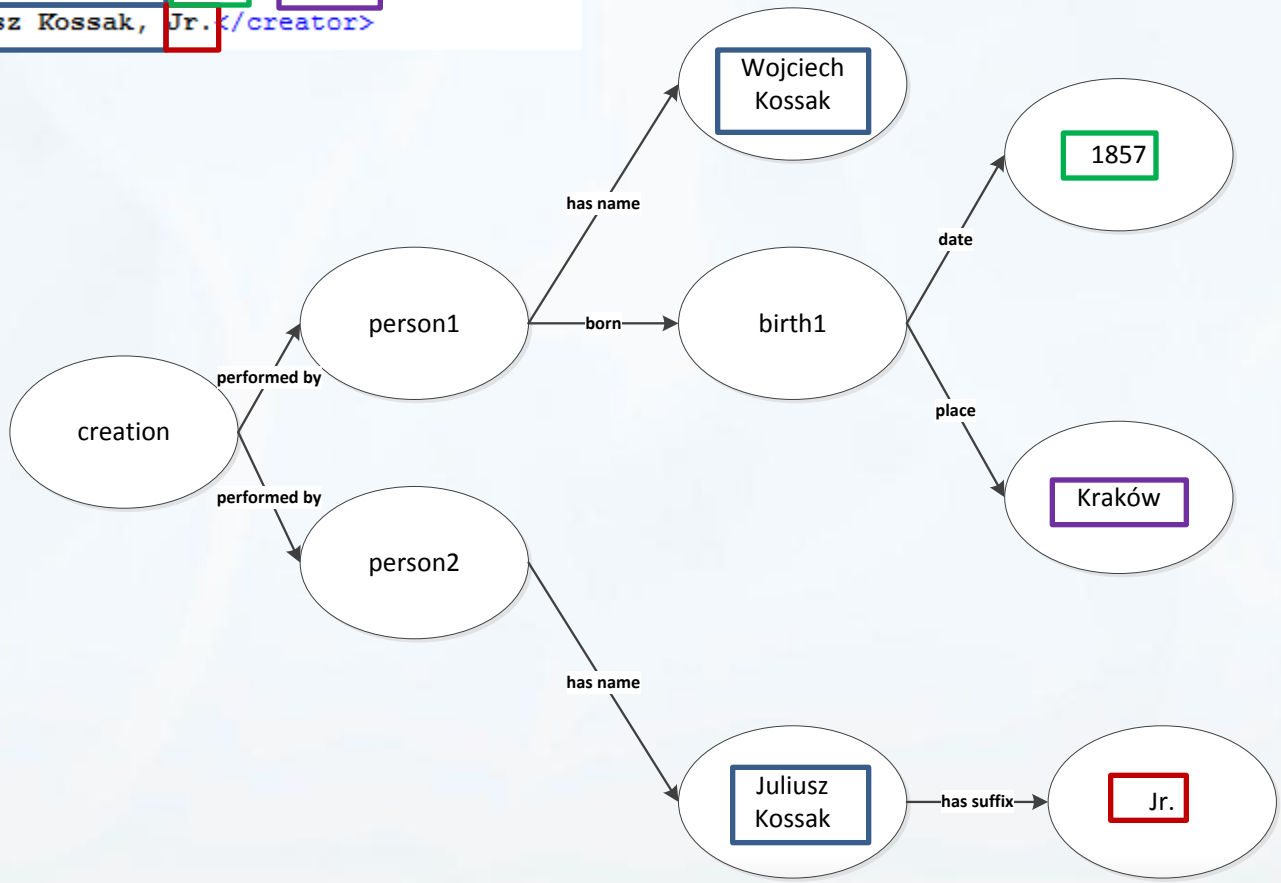
- XML Structure in the Output

```
<a>AAA</a>
<b>BBB</b>
```

```
<c>AAA</c>
<d>
  <e>BBB</e>
</d>
```

# General Requirements: Patterns

```
<creator>Wojciech Kossak (1875, Kraków) </creator>
<creator>Juliusz Kossak, Jr.</creator>
```



# General Requirements: Substrings & Concatenation

Depending on the source and target schema granularity, it may be necessary to either SPLIT or CONCATENATE the contents of a source metadata element.

```
<datafield tag="100" ind1="1" ind2=" ">
    <subfield code="a">Camus, Albert</subfield>
    <subfield code="d">(1913-1960).</subfield>
</datafield>
```

```
<dc:creator>Camus, Albert (1913-1960)</dc:creator>
```

# General Requirements: Value Maps

Depending on the source and target schema granularity, it may be necessary to either split or concatenate the contents of a source metadata element.

```
<datafield tag="048" ind1=" " ind2=" ">
    <subfield code="a">vn</subfield>
    <subfield code="a">ka01</subfield>
    <subfield code="a">pz</subfield>
</datafield>
```

What do you want in your target data?

- *original code*
- *full instrument name*
- *URI from a LOD vocabulary*
- *target schema code*

```
py - Percussion--Ethnic
pz - Percussion--Other
sa - Strings, bowed--Violin
sb - Strings, bowed--Viola
sc - Strings, bowed--Violoncello
sd - Strings, bowed--Double bass
se - Strings, bowed--Viol
sf - Strings, bowed--Viola d'amore
sg - Strings, bowed--Viola da gamba
sn - Strings, bowed--Unspecified
su - Strings, bowed--Unknown
sy - Strings, bowed--Ethnic
sz - Strings, bowed--Other
ta - Strings, plucked--Harp
tb - Strings, plucked--Guitar
tc - Strings, plucked--Lute
td - Strings, plucked--Mandolin
tn - Strings, plucked--Unspecified
tu - Strings, plucked--Unknown
ty - Strings, plucked--Ethnic
tz - Strings, plucked--Other
va - Voices--Soprano
vb - Voices--Mezzo Soprano
vc - Voices--Alto
vd - Voices--Tenor
ve - Voices--Baritone
vf - Voices--Bass
vg - Voices--Counter tenor
vh - Voices--High voice
vi - Voices--Medium voice
vj - Voices--Low voice
vn - Voices--Unspecified
vu - Voices--Unknown
```

# Semantic Web Requirements: Ontology Paths

Source:

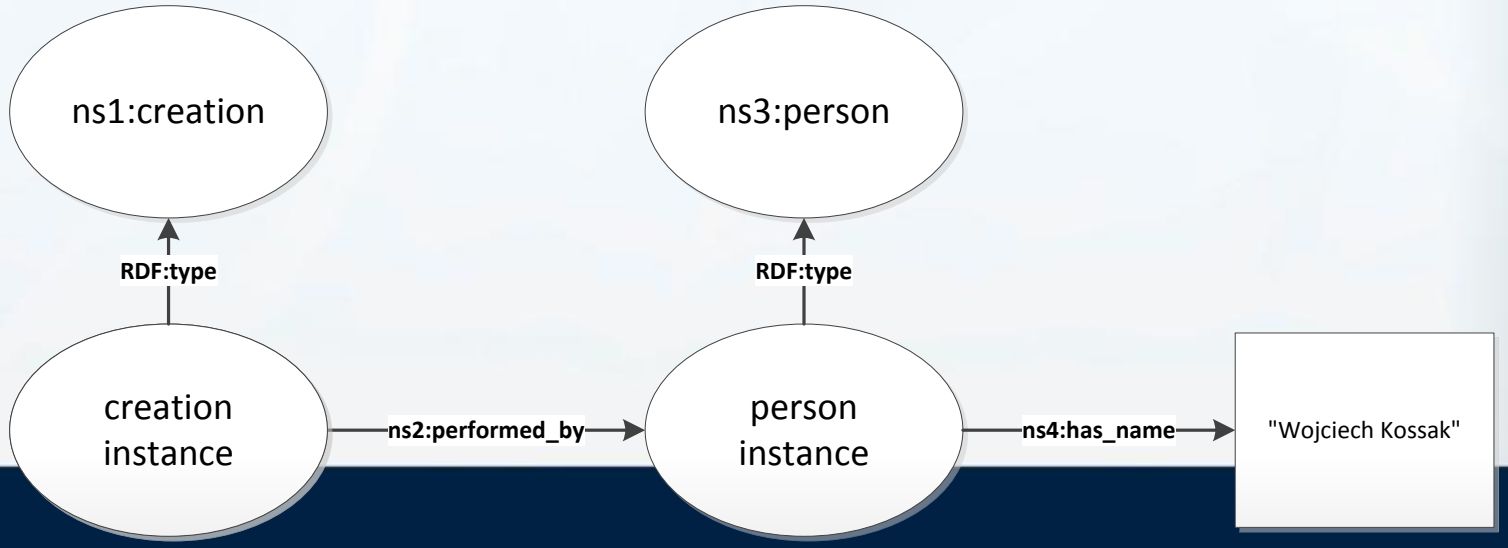
```
<creator>Wojciech Kossak</creator>
```

Rule (path):

creator

ontNS:creation -- ontNS:performed_by -- ont NS:person -- ontNS:has_name

Target:



# Semantic Web Requirements: Identifiers

```
<creator>Wojciech Kossak (1875, Kraków)</creator>
<creator>Juliusz Kossak (1824, Kraków)</creator>
```

Record level

- Both creators participate in the same creation event of the same resource (e.g. painting)
- ontNS:creation [REC_ID:creation] -- ontNS:performed_by -- ont NS:person -- ontNS:has_name

Element level

- The date and place within one creator element correspond to the same birth event of the same person.
- But the person id is only valid within the creator element; the next occurence of the element within the record resets the identifier.

Iterable record level

- Those are necessary when in some metadata formats the Nth occurence of an element corresponds to the sam entity as the Nth occurence of another element.

```
<elem1>
    <subel>Name and surname 1</subel>
    <subel>Name and surname 2</subel>
    <subel>Name and surname 3</subel>
</elem1>
...
<elem2>
    <subel2>date of birth 1</subel2>
    <subel2>date of birth 2</subel2>
    <subel2>date of birth 3</subel2>
</elem2>
```

# Semantic Web Requirements: URI's

- Repeatable URI's

The mapper has to be able to generate predictable, consistent URI's for the objects it creates.

Also:

- Classes vs. Instances on Path

```
ontNS:class -- ontNS:property -- ontNS:class -- ontNS:property
```

Normally, classes are found on such paths, but in some cases a particular instance (e.g. representing a type from an external vocabulary) may be included. A related concept is a *static path* which does not end with a literal from the source element contents, but is added as-is after occurence of an element.

- Unique vs. Shared URI's

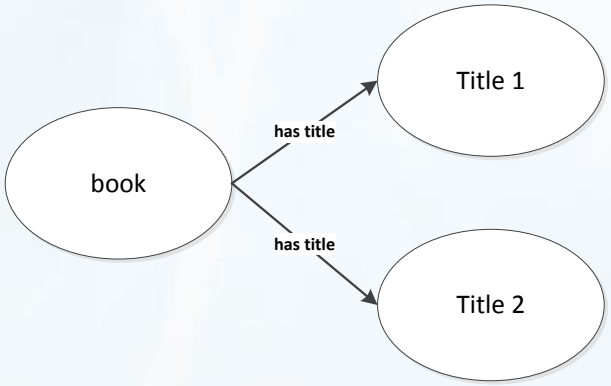
If I keep finding "Larnaka" in the *plmet:placeOfPublishing* element, do I want the generated URI to be the same of different at each occurence?

# Semantic Requirements: Existential Conditions

```
<dc:title>Title 1<dc:title>
<dc:title>Title 2<dc:title>
```

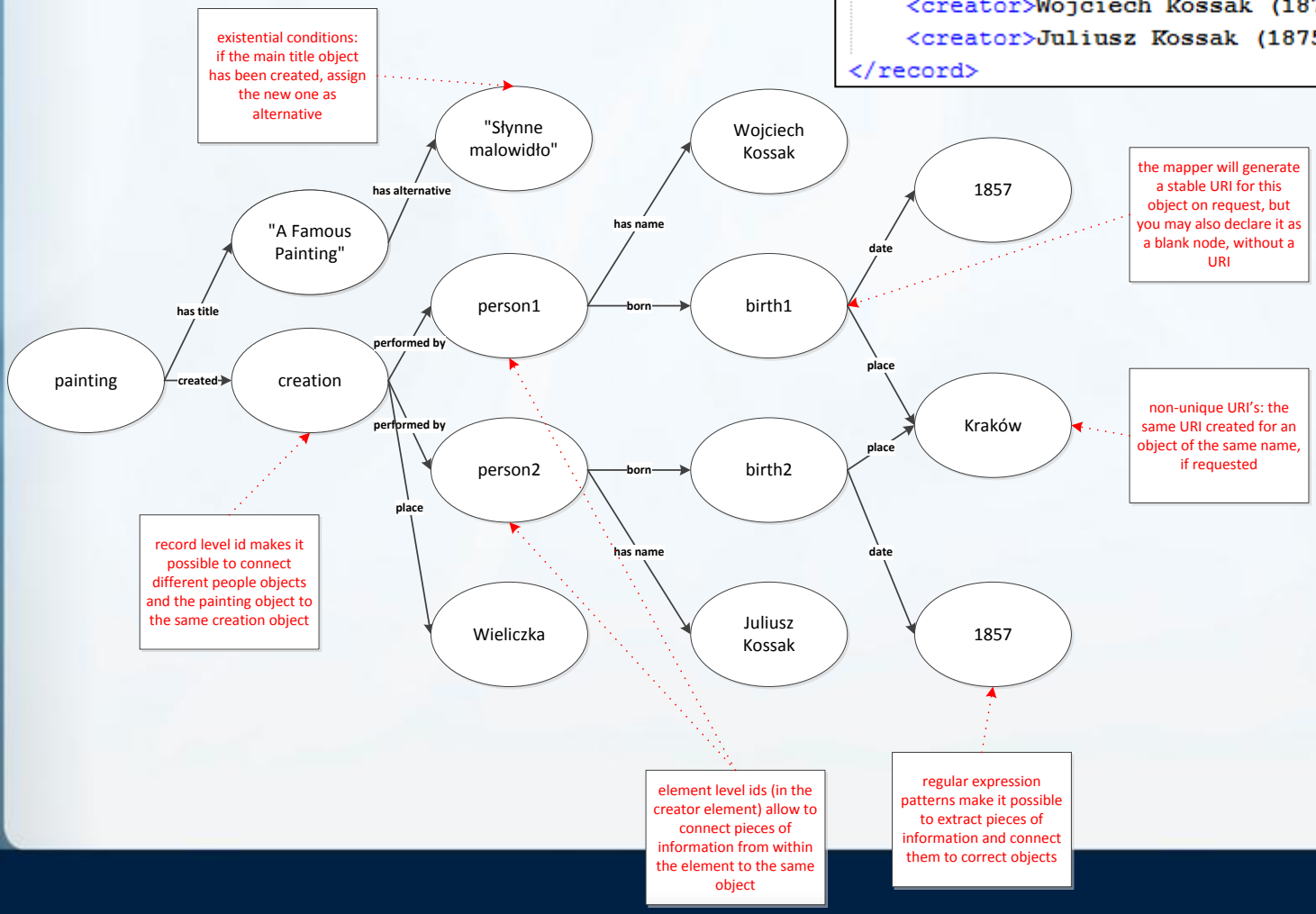
Depending on the format, you may want one of the following.

In the 2nd case, the mapping rule defines different behaviour in an object with a given id already exists.




# Requirements: Example

```
<record>
    <title>A Famous Painting<title>
    <title>Słynne malowidło</title>
    <placeOfCreation>Wieliczka</placeOfCreation>
    <creator>Wojciech Kossak (1875, Kraków)</creator>
    <creator>Juliusz Kossak (1875, Kraków)</creator>
</record>
```



# jMet2Ont Final Remarks

- No programming knowledge required
- Regular expressions may come in handy
- No GUI at this point (in the works)
- Available (with excessive user documentation) as a tool or as a programming library here:

  http://fbc.pionier.net.pl/pro/jmet2ont/
- jMet2Ont is a mapper, not an enricher and not a metadata digestion tool
- Good performance: mapping of full DLF (1 million records) and NUKAT (2 million records) data took 3 hours on a standard desktop PC and created 200 million explicit triples
- Rule example:

```
<!--dimensions of a museum object-->
<elementMapping elementName="GAB">
    <patternToPaths>
        <pattern prior="1">(?s)(.*)</pattern>
        <path no="1" type="literal">
            <pathElement recordLevelId="mainDescribedObject">cidoc:E22_Man-Made_Object</pathElement>
            <pathElement>cidoc:P43_has_dimension</pathElement>
            <pathElement elementLevelId="dimension">cidoc:E54_Dimension</pathElement>
            <pathElement>cidoc:P90_has_value</pathElement>
        </path>
    </patternToPaths>
</elementMapping>
```

# POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

## THANK YOU FOR YOUR ATTENTION

Poznań Supercomputing and Networking Center
affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences,
ul. Noskowskiego 12/14, 61-704 Poznań, POLAND,
Office: phone center: (+48 61) 858-20-00,
fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.man.poznan.pl