Greening your database of literary corpora

How to avoid reinventing vocabularies, in favor of sustainable, reusable data models

Jean-Baptiste Camps Kelly Christensen

DH 2025
Lisbon, 17th July 2025, *Session SP-31*

École nationale des chartes, PSL

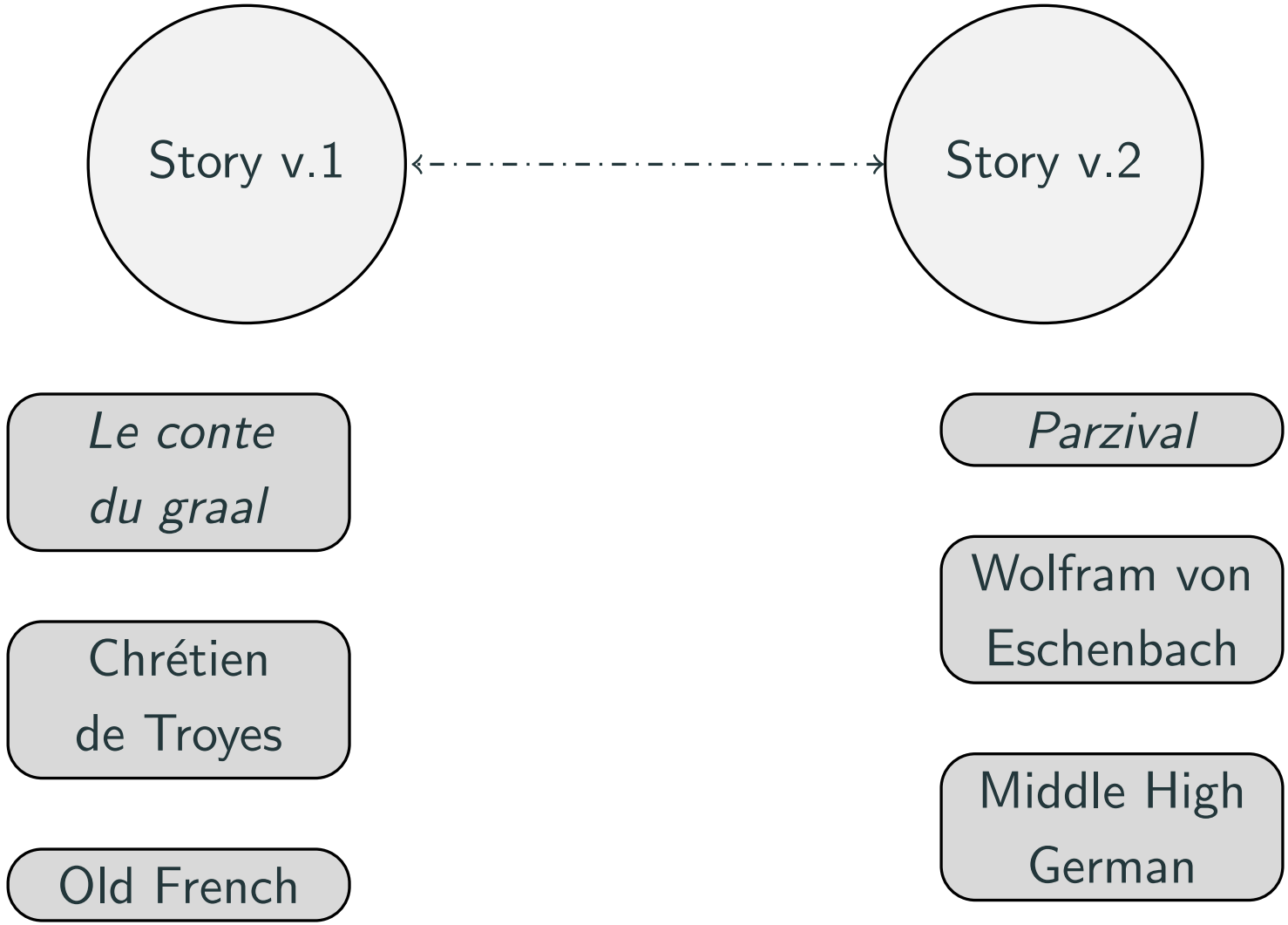
Context: LostMa project

slides: enc.hal.science/hal-05059049

- Project: `LostMa`
  - European Research Council
  - 5 years (all good things come to an end)
- Goal
  - Model the transmission of Medieval stories, texts and manuscripts (and estimate survival rate)
- Problem
  - Group manuscripts (easy to distinguish, libraries do this all the time) by narrative content (more abstract, less standardised)
- Challenge
  - Minimize original ontology. Rely on existing frameworks and digital resources.
    - Findable, Accessible, Interoperable, Reusable (FAIR)

Problem: Modelling narrative traditions

What data model makes versions of a story findable (groups them) but also reuses the main bibliographic ontologies?



Challenge: Reusability and (project) resource management

FAIR Guiding Principles for scientific data management and stewardship (2016)

F	A	I	R
Findable	Accessible	Interoperable	Reusable

- Greener digital data storage and distribution[1]
- Minimal single-use documentation
- Efficient integration in other projects

[1] Marion Ficher et al. **"Assessing the carbon footprint of the data transmission on a backbone network".** In: *2021 24th Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*. Paris, France, Mar. 2021.

Methodology: Models for F.A.I.R. literary projects

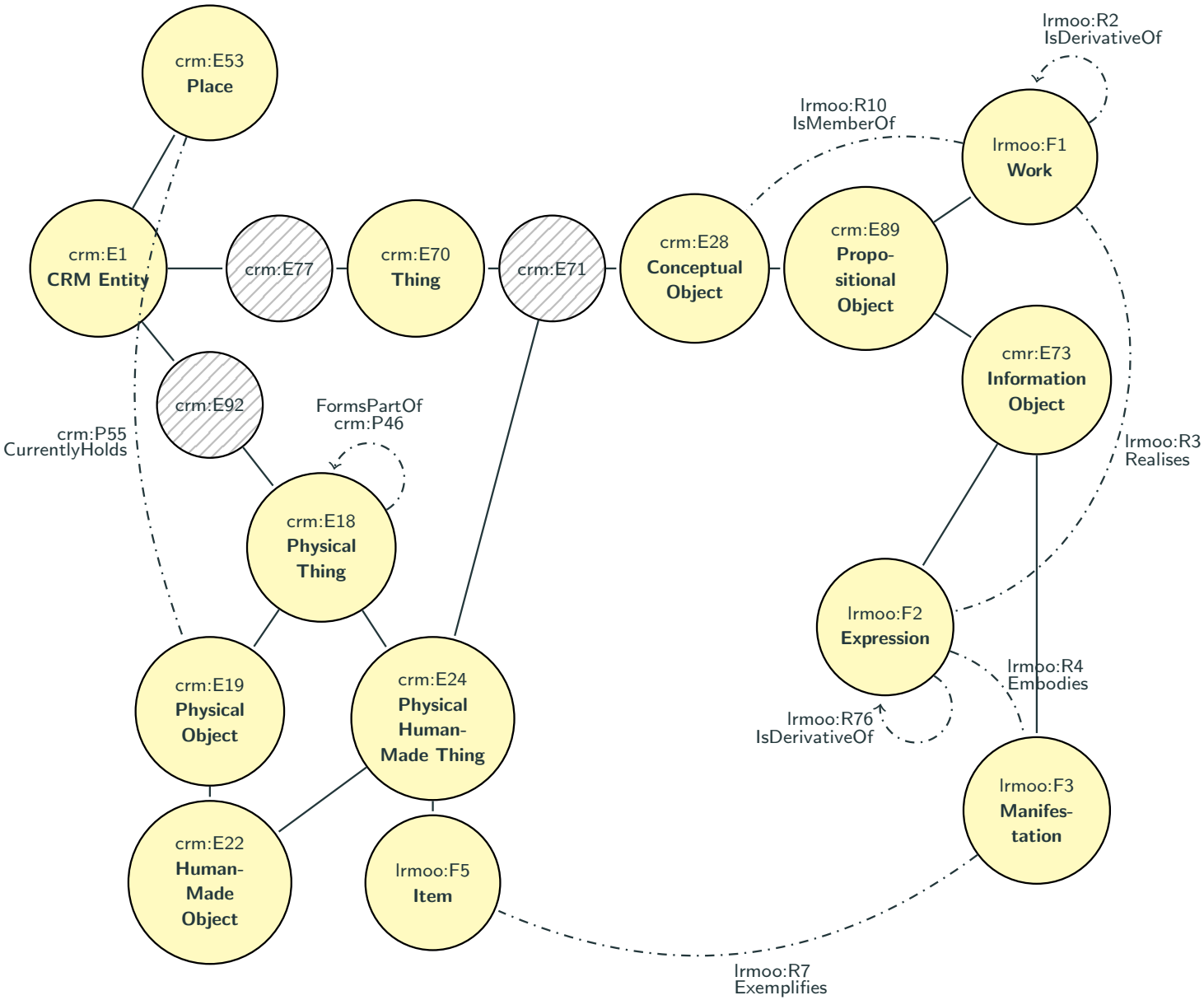
The big groups:

- Comité International pour la DOCumentation (CIDOC)
- International Federation of Library Associations (IFLA)
- Dublin Core (DC)

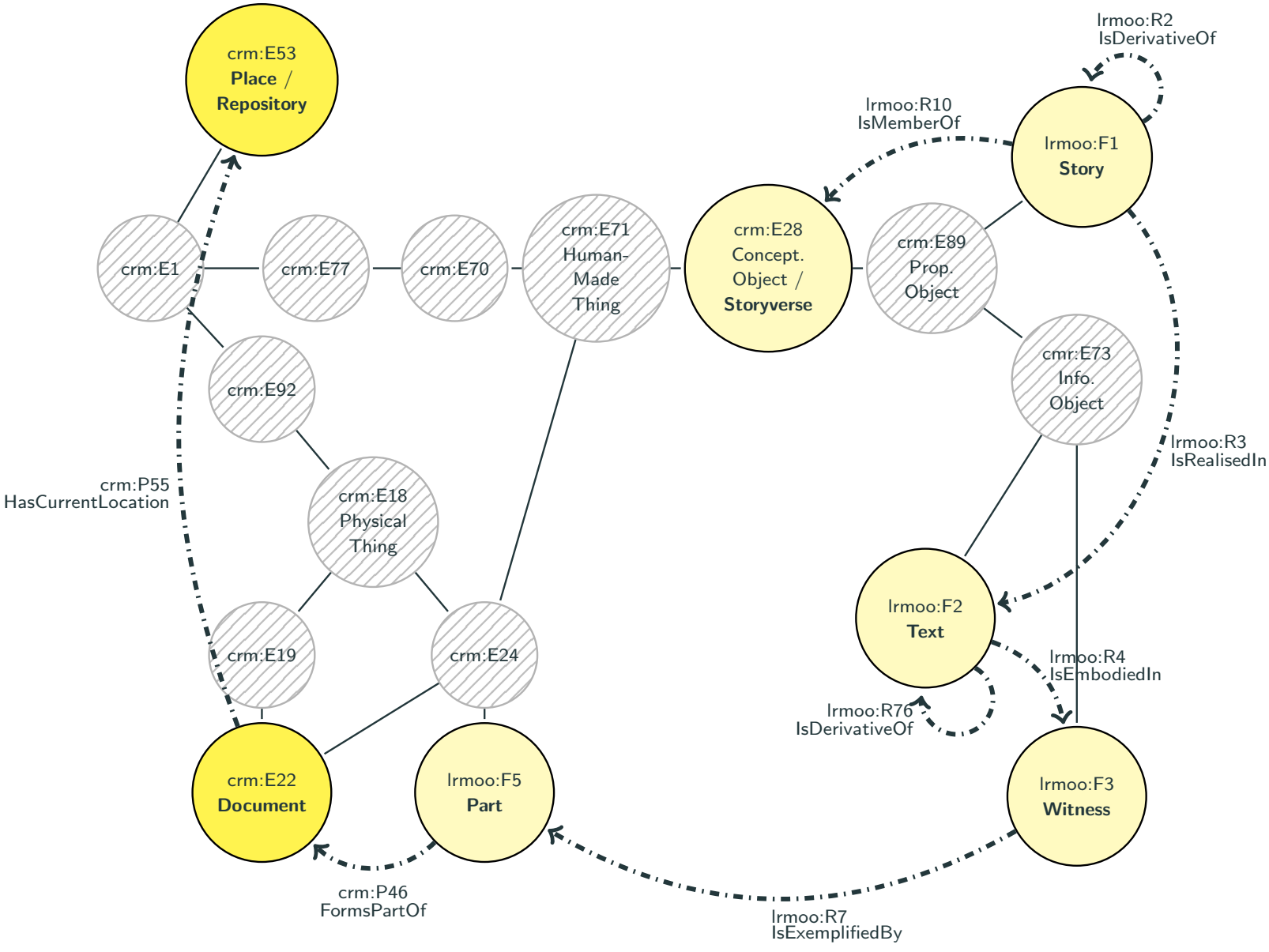
The main ontologies:

Org.	Year	Ver.	Model
IFLA	1998	1.0	Functional Requirements for Bibliographic Records (FRBR)
CIDOC	2006	1.0	Conceptual Reference Model (CRM)
IFLA	2016	2.4	Functional Requirements for Bibliographic Records Object-Oriented ($FRBR_{OO}$)
IFLA	2017	1.0	Library Reference Model (LRM)
CIDOC	2022	7.1.12	CRM
IFLA	2024	1.0	Library Reference Model Object-Oriented (LRM_{OO})

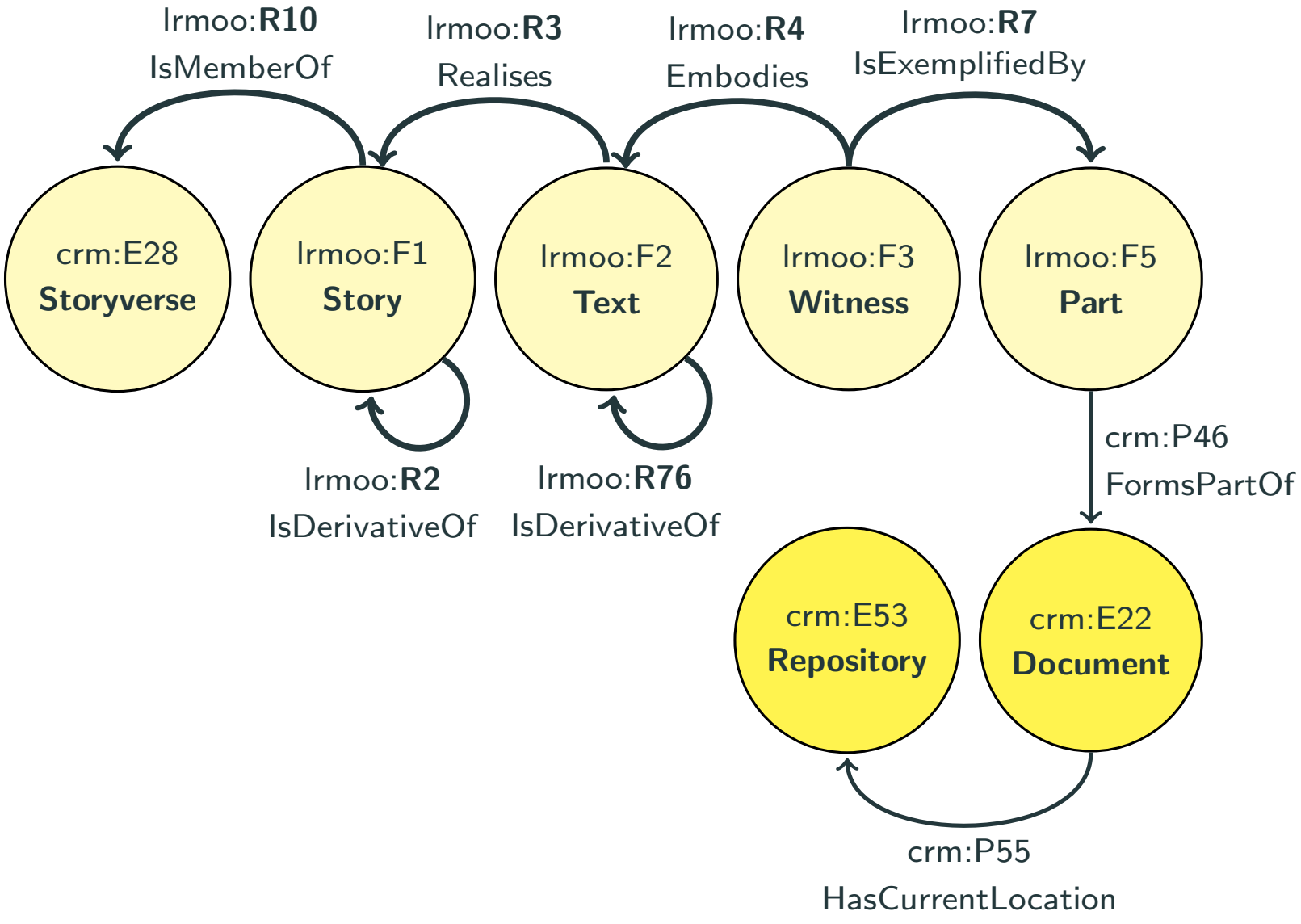
Library Reference Model Object-Oriented (LRM_{OO}), built atop the CRM



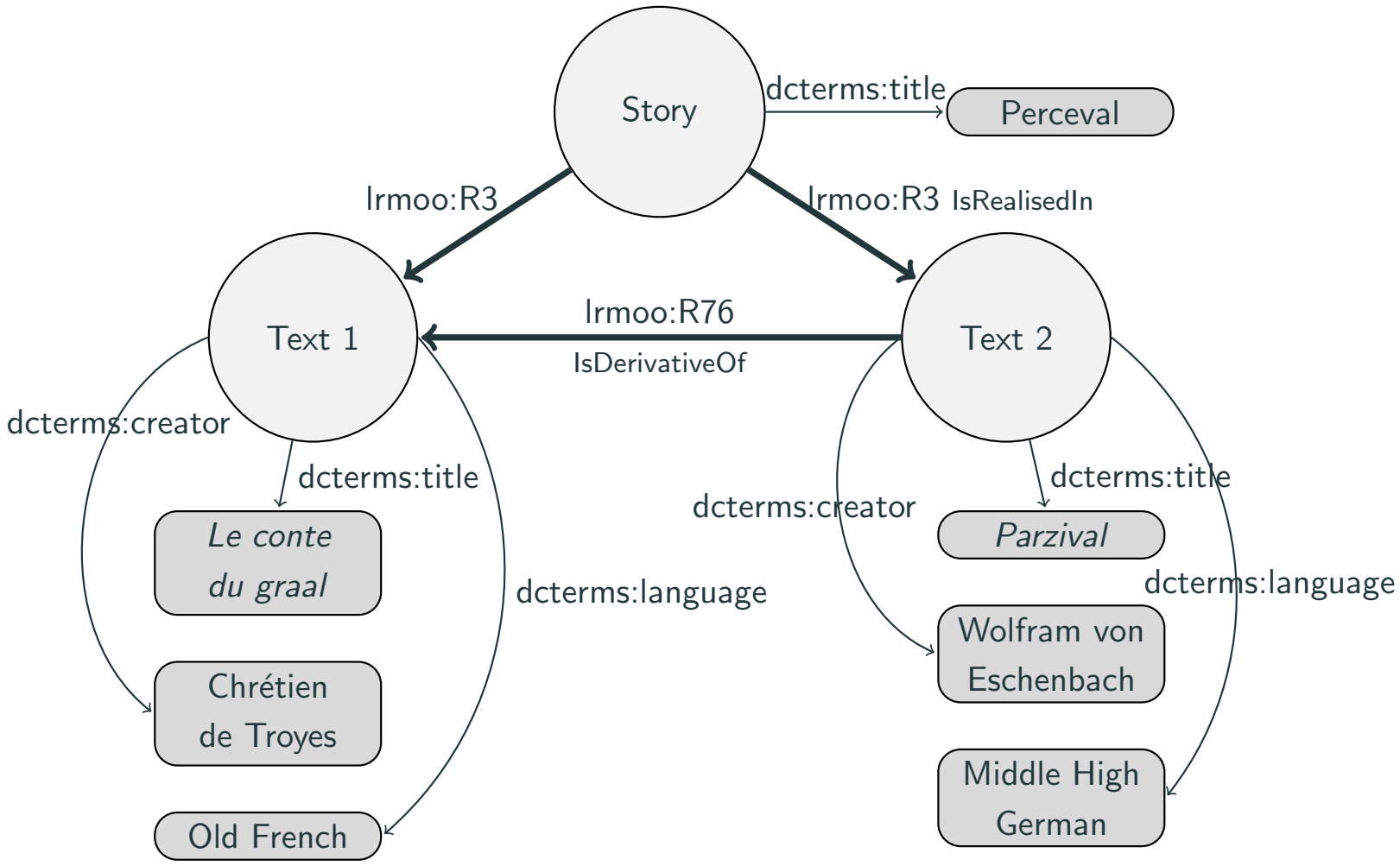
LostMa Ontology, built atop the LRM$_{oo}$



LostMa Ontology, simplified



Solution: Perceval example (revisited)



Solution: Multi-use documentation (text)

Namespaces: `lrmoo` (IFLA), `dcterms` (DCMI), `dc` (DCMI)

```
<lrmoo:F2_Expression id="48572">
 <dcterms:title>Perceval</dcterms:title>
 <dcterms:language>fro</dcterms:language>
 <dcterms:created>1196-01-01/1210-12-31</dcterms:created>
 <dc:date>Vers 1200</dc:date>
 <lrmoo:R4_is_embodied_in>
  <lrmoo:F3_Manifestation id="54725">
   <!-- Witness metadata -->
 </lrmoo:R4_is_embodied_in>
</lrmoo:F2_Expression>
```

Solution: Multi-use documentation (witness)

Namespaces: `lrmoo` (IFLA), `dcterms` (DCMI), `crm` (CIDOC), `bibo` (DCMI)

```
<lrmoo:F3_Manifestation id="54725">
 <dcterms:created>1301-01-01/1400-12-31</dcterms:created>
 <lrmoo:R7_is_exemplified_by>
  <lrmoo:F5_Item id="53068">
   <bibo:locator>
    <bibo:pages>
     <bibo:pageStart>96r</bibo:pageStart>
     <bibo:pageEnd>126v</bibo:pageEnd>
    </bibo:pages>
   </bibo:locator>
   <crm:P46i_forms_part_of>
    <crm:E22_Human-Made_Object id="46077"/>
     <!-- Document metadata -->
   </crm:P46i_forms_part_of>
  </lrmoo:F5_Item>
 </lrmoo:R7_is_exemplified_by>
</lrmoo:F3_Manifestation>
```

Conclusion

slides: enc.hal.science/hal-05059049

In progress

- Off-load documentation of our data model to existent digital resources by linking our ontology to others' documentation
  - Side goal: save on resources for maintaining online documentation (esp. after project ends)
- Map more properties to our core (LRM_{OO}, CRM) entities

Future work

- Accommodate more uncertainty in values, i.e. dates:
  - earliest possible date
  - estimated earliest date
  - estimated latest date
  - latest possible date
- Accommodate source citation on properties