The Digitized Manuscript as Cultural Heritage:



- Project, Context, Motivation

Prof. Dr. Stefan Gradmann
Humboldt-Universität zu Berlin / School of Library and Information Science
stefan.gradmann@ibi.hu-berlin.de

CC BY SA

Overview

- DM2E Motivation
  - Digitized manuscripts and scholarly context
  - Functionality:
    - Beyond emulation?
    - Scholarly Primitives
    - Contextualisation
- Technology Context
  - RDF, LoD and the EDM
- Digitized Manuscripts to Europeana (DM2E)

CC BY SA

LibreOffice The Document Foundation

Motivation: Scholarly Work

CC BY SA

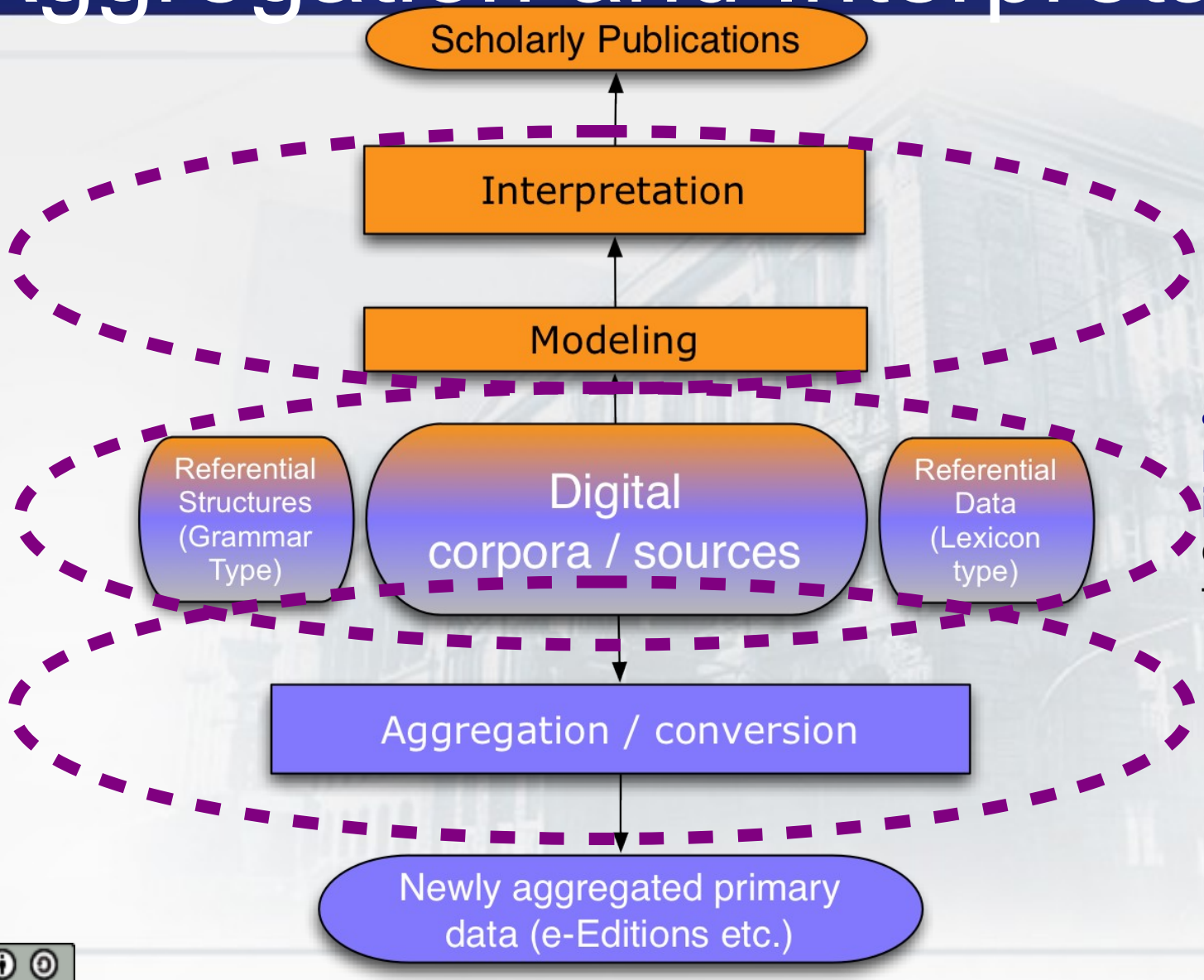
LibreOffice
The Document Foundation

Digitization and Beyond

- DM2E is not about manuscript digitisation ...
- … but about what it takes to turn digitised manuscripts into cultural heritage
- What is 'cultural heritage'?
  - Manuscripts are cultural heritage.
  - Digitised manuscripts are not: they are a means to preserve and/or to represent this cultural heritage!
  - Or might the latter not be true: do digitised manuscripts have a potential of becoming 'cultural heritage' in their own right?
- Building heritage is all about cultural appropriation – and thus about (digital) humanities scholarship

CC BY SA

Humanitites and Corpora: Aggregation and Interpretation

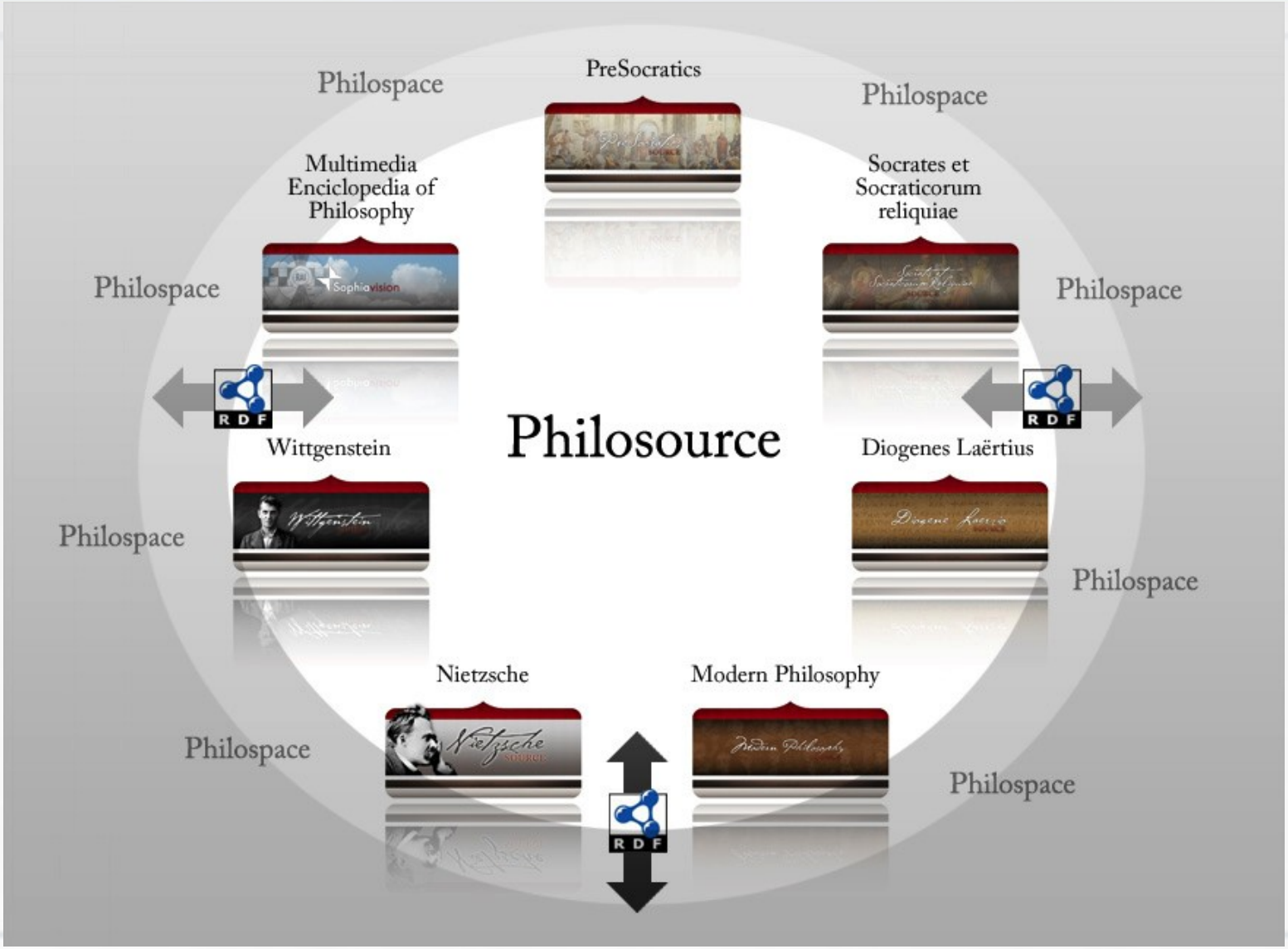


How do **corpora**, **aggregation** and **modeling**/ **interpretation** evolve in the digital domain?

CC BY SA

LibreOffice The Document Foundation

'Discovery' Corpus: Digitised Manuscripts



CC BY SA

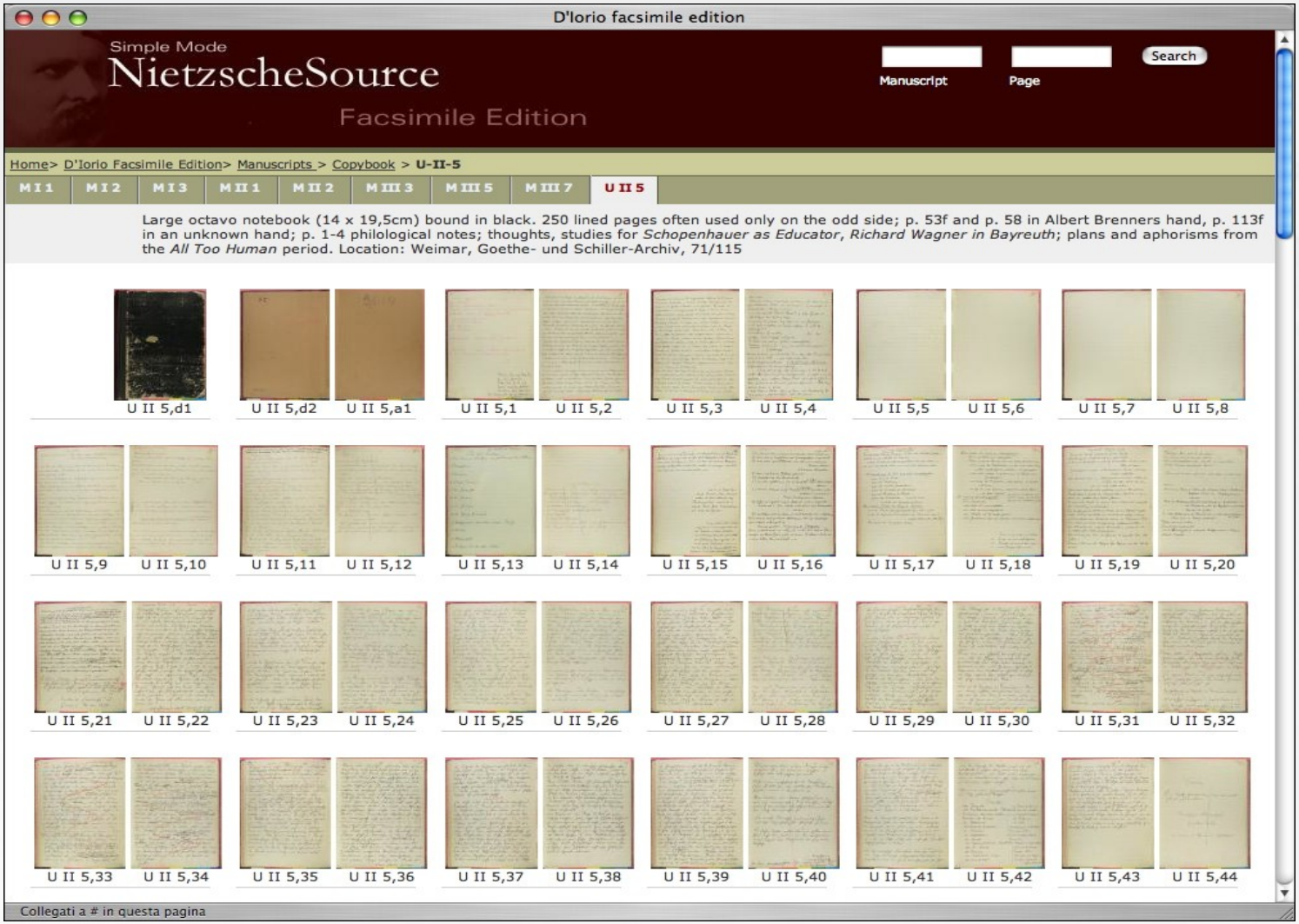
LibreOffice The Document Foundation

Aggregation: Talia

CC BY SA

LibreOffice
The Document Foundation

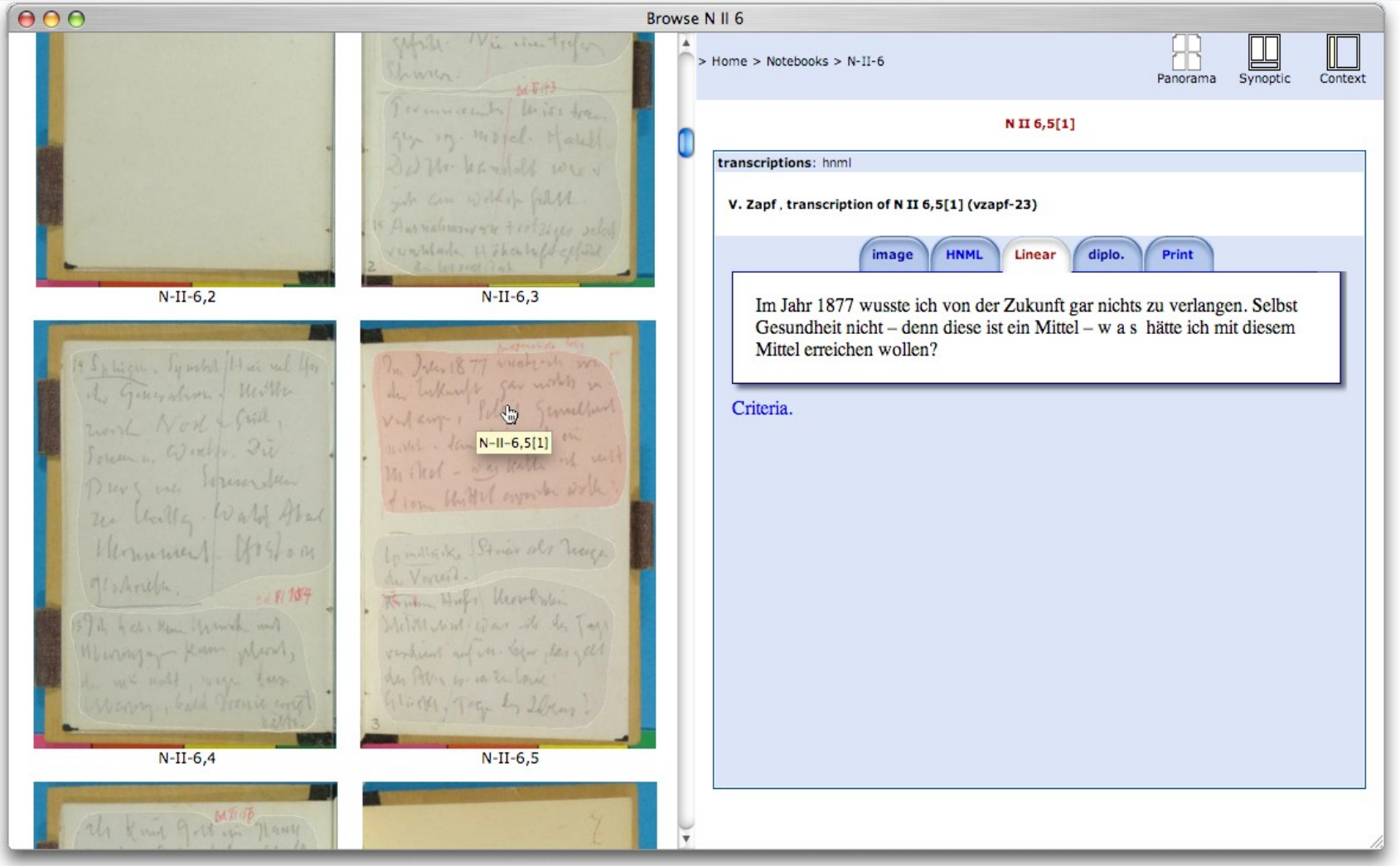
HyperNietzsche: Digitisation, Presentation



CC BY SA

LibreOffice
The Document Foundation

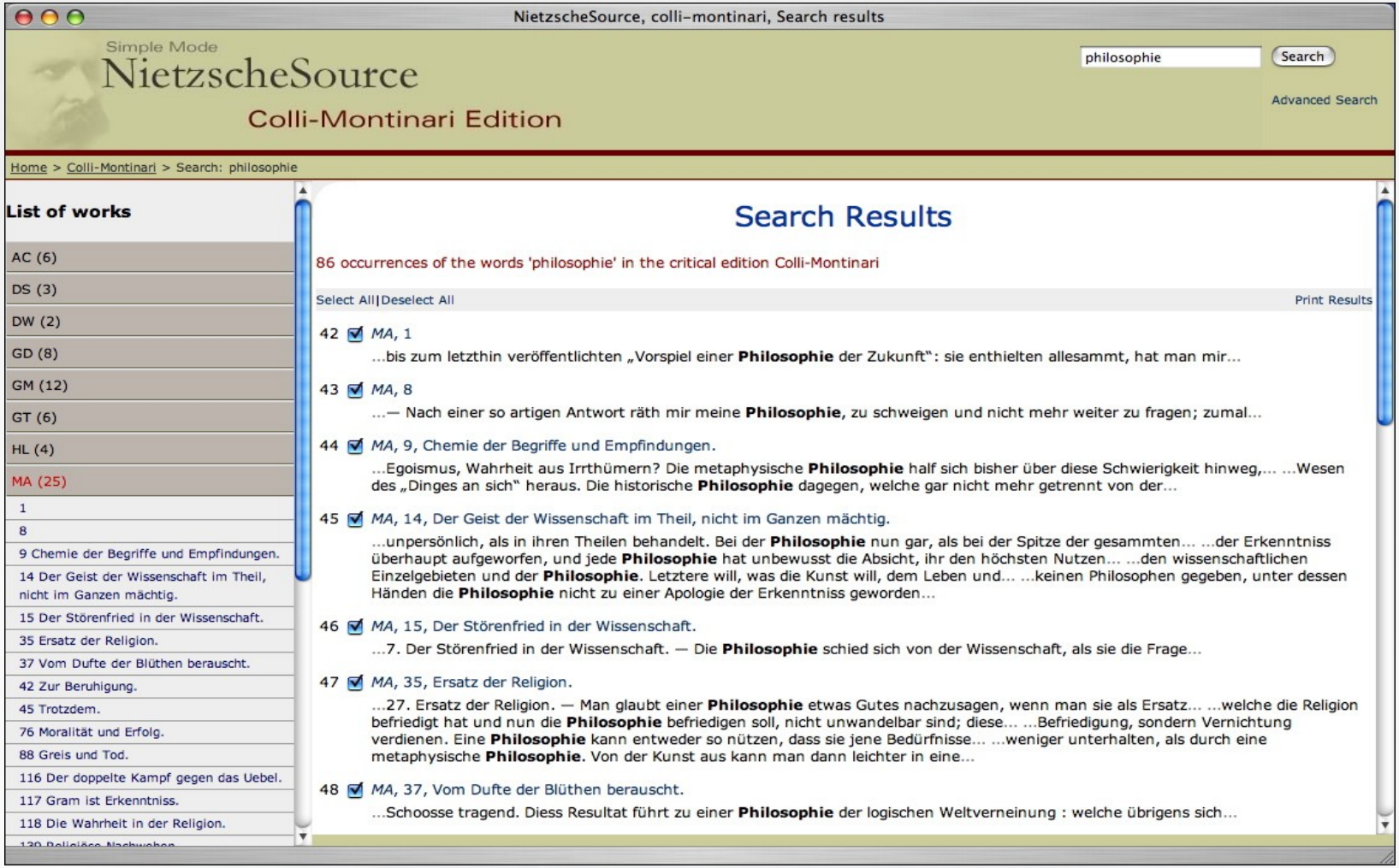
HyperNietzsche: Transcription, Presentation



CC BY SA

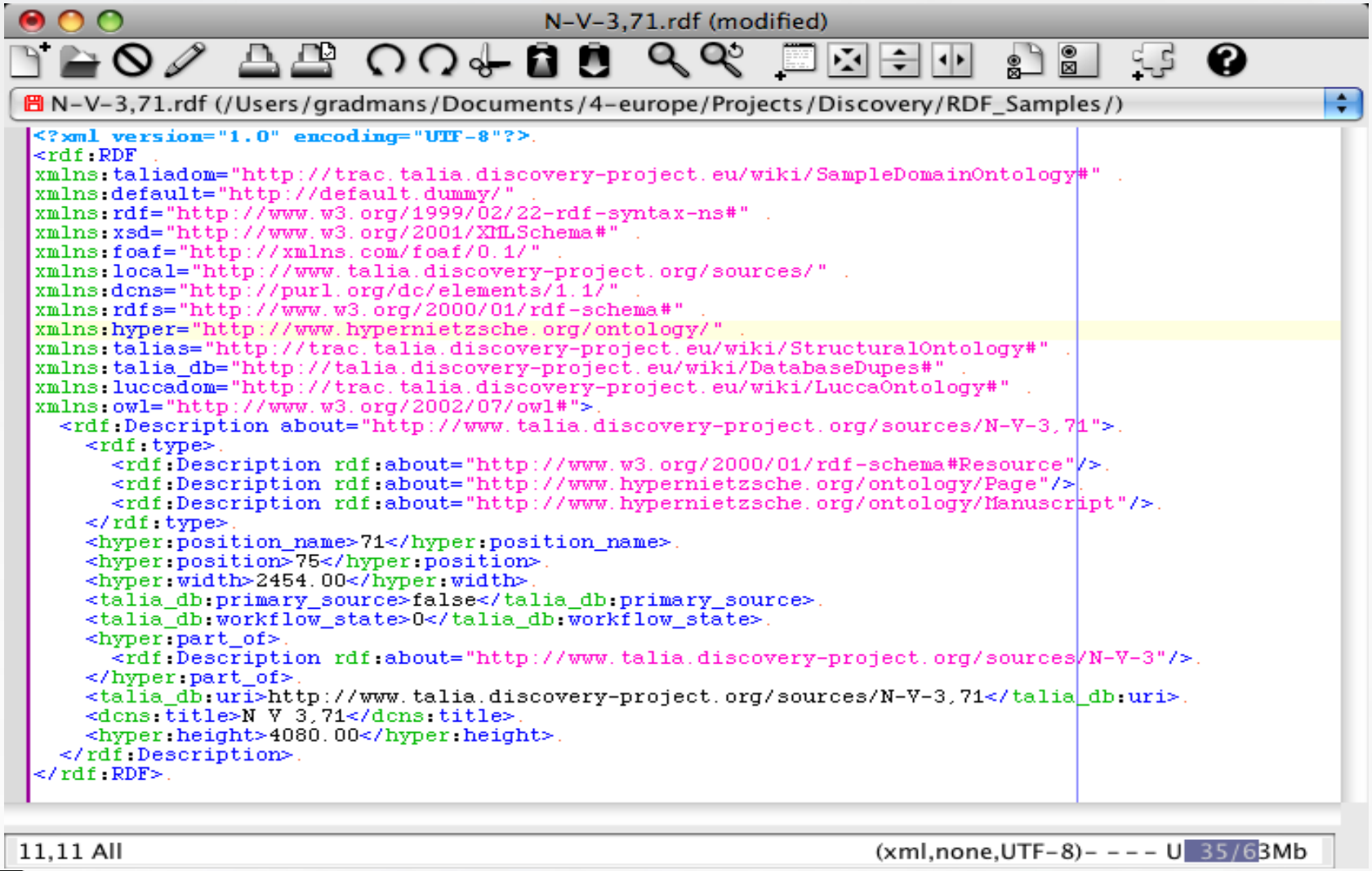
LibreOffice The Document Foundation

HyperNietzsche: Sources, Editions (synoptic)



CC BY SA
LibreOffice The Document Foundation

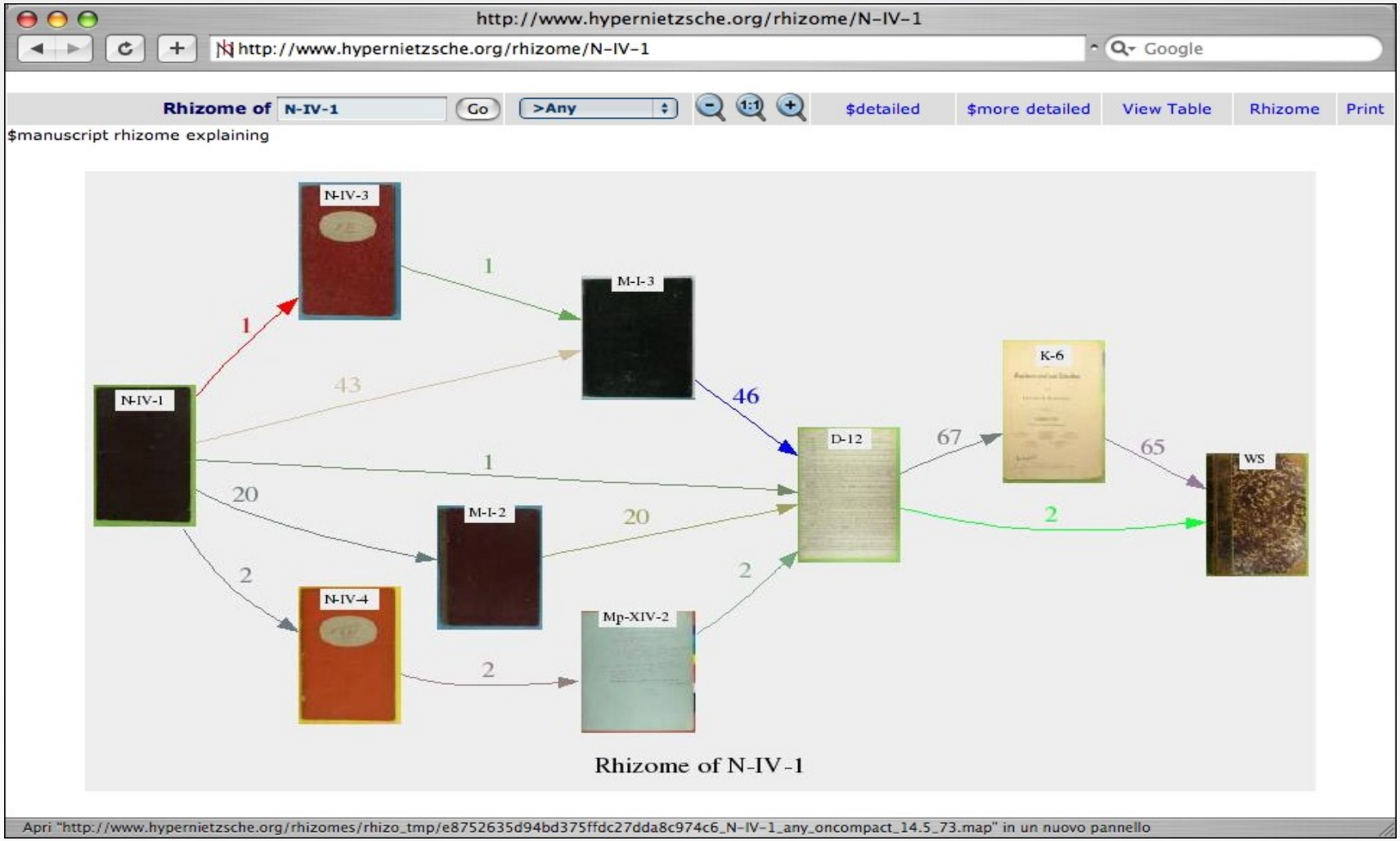
Talia: Refactoring Hyper with Semantic Web Technology




```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
xmlns:taliadom="http://trac.talia.discovery-project.eu/wiki/SampleDomainOntology#"
xmlns:default="http://default.dummy/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:local="http://www.talia.discovery-project.org/sources/"
xmlns:dcns="http://purl.org/dc/elements/1.1/"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:hyper="http://www.hypernietzsche.org/ontology/"
xmlns:talias="http://trac.talia.discovery-project.eu/wiki/StructuralOntology#"
xmlns:talia_db="http://talia.discovery-project.eu/wiki/DatabaseDupes#"
xmlns:luccadom="http://trac.talia.discovery-project.eu/wiki/LuccaOntology#"
xmlns:owl="http://www.w3.org/2002/07/owl#">
  <rdf:Description about="http://www.talia.discovery-project.org/sources/N-V-3,71">
    <rdf:type>
      <rdf:Description rdf:about="http://www.w3.org/2000/01/rdf-schema#Resource"/>
      <rdf:Description rdf:about="http://www.hypernietzsche.org/ontology/Page"/>
      <rdf:Description rdf:about="http://www.hypernietzsche.org/ontology/Manuscript"/>
    </rdf:type>
    <hyper:position_name>71</hyper:position_name>
    <hyper:position>75</hyper:position>
    <hyper:width>2454.00</hyper:width>
    <talia_db:primary_source>false</talia_db:primary_source>
    <talia_db:workflow_state>0</talia_db:workflow_state>
    <hyper:part_of>
      <rdf:Description rdf:about="http://www.talia.discovery-project.org/sources/N-V-3"/>
    </hyper:part_of>
    <talia_db:uri>http://www.talia.discovery-project.org/sources/N-V-3,71</talia_db:uri>
    <dcns:title>N V 3,71</dcns:title>
    <hyper:height>4080.00</hyper:height>
  </rdf:Description>
</rdf:RDF>
```

CC BY SA

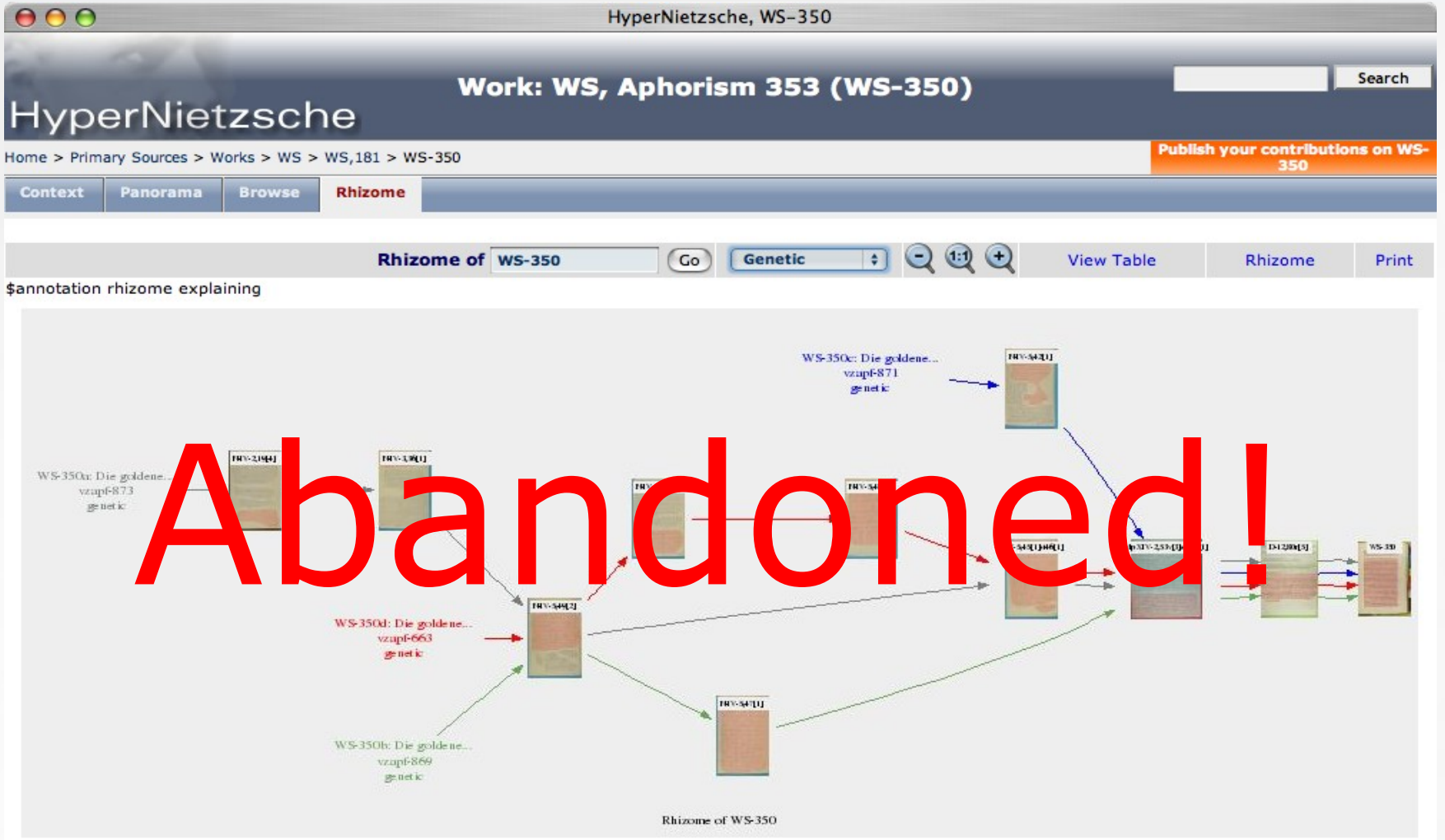
Generating Stemmata based on Inferencing (1)



CC BY SA

LibreOffice The Document Foundation

Generating Stemmata based on Inferencing (2)



Abandoned!

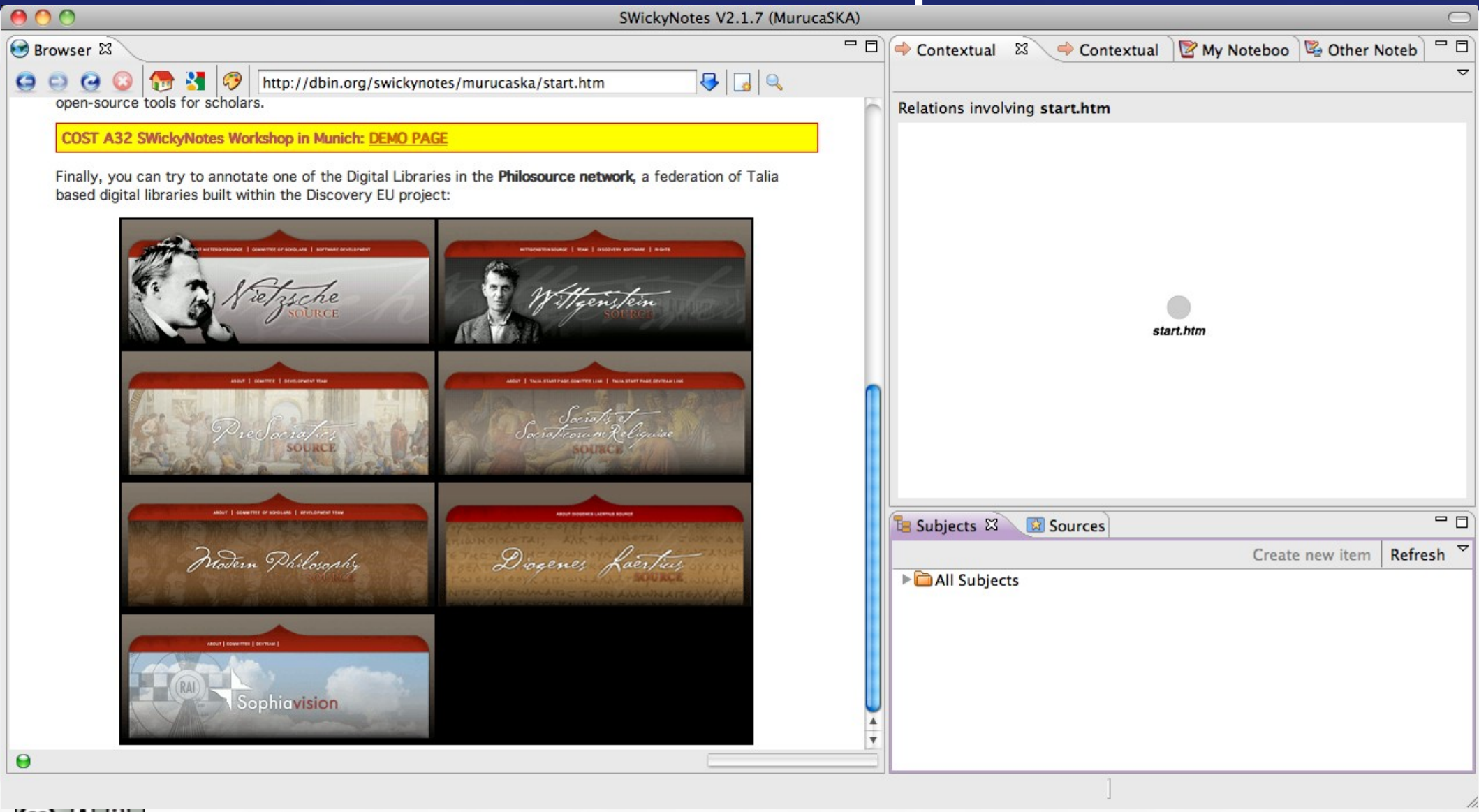
CC BY SA

Interpretation: Muruca / Pundit

CC BY SA

LibreOffice
The Document Foundation

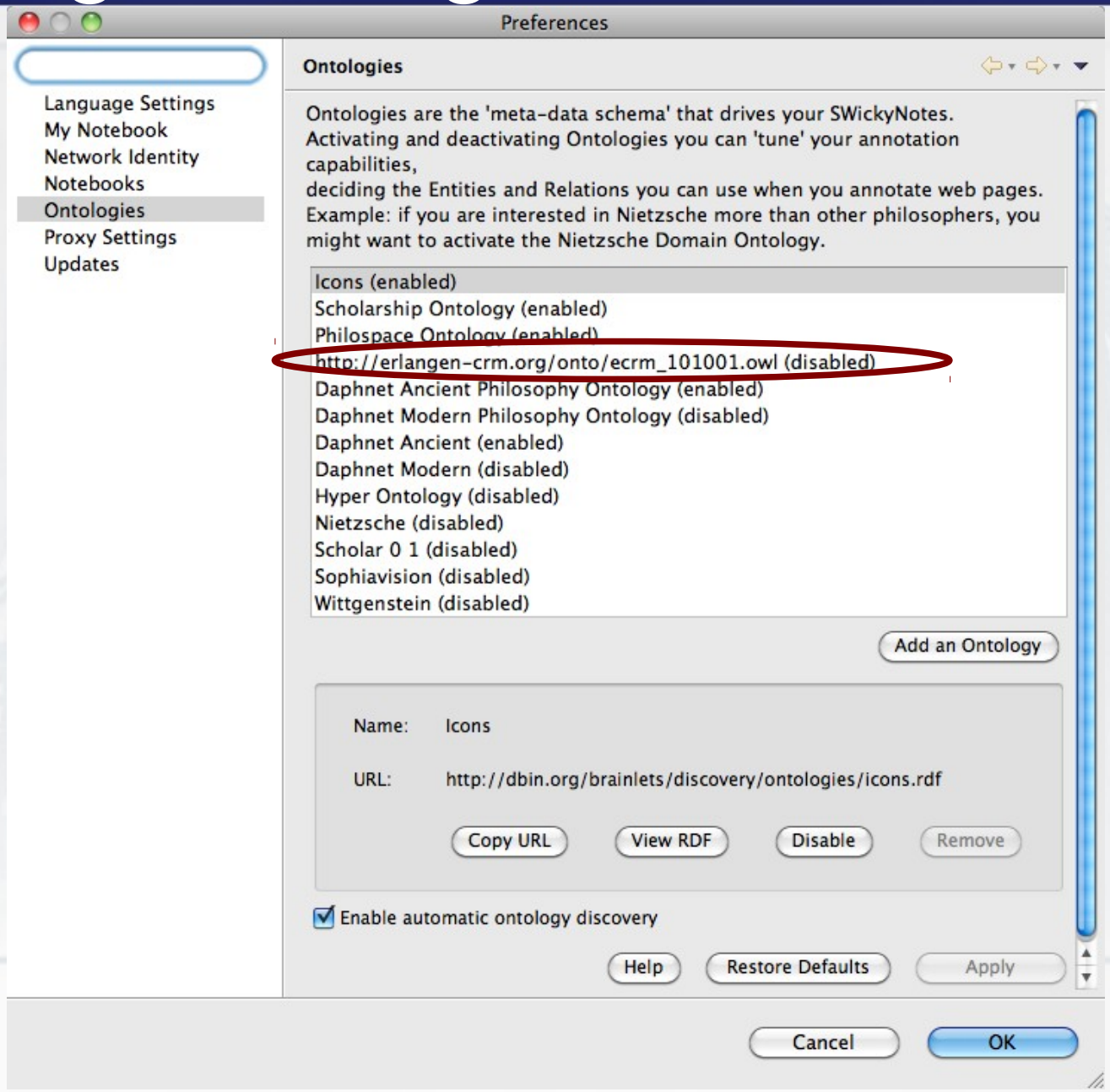
SwickyNotes: Ontology Based Annotation as Linked Open Data



CC BY SA

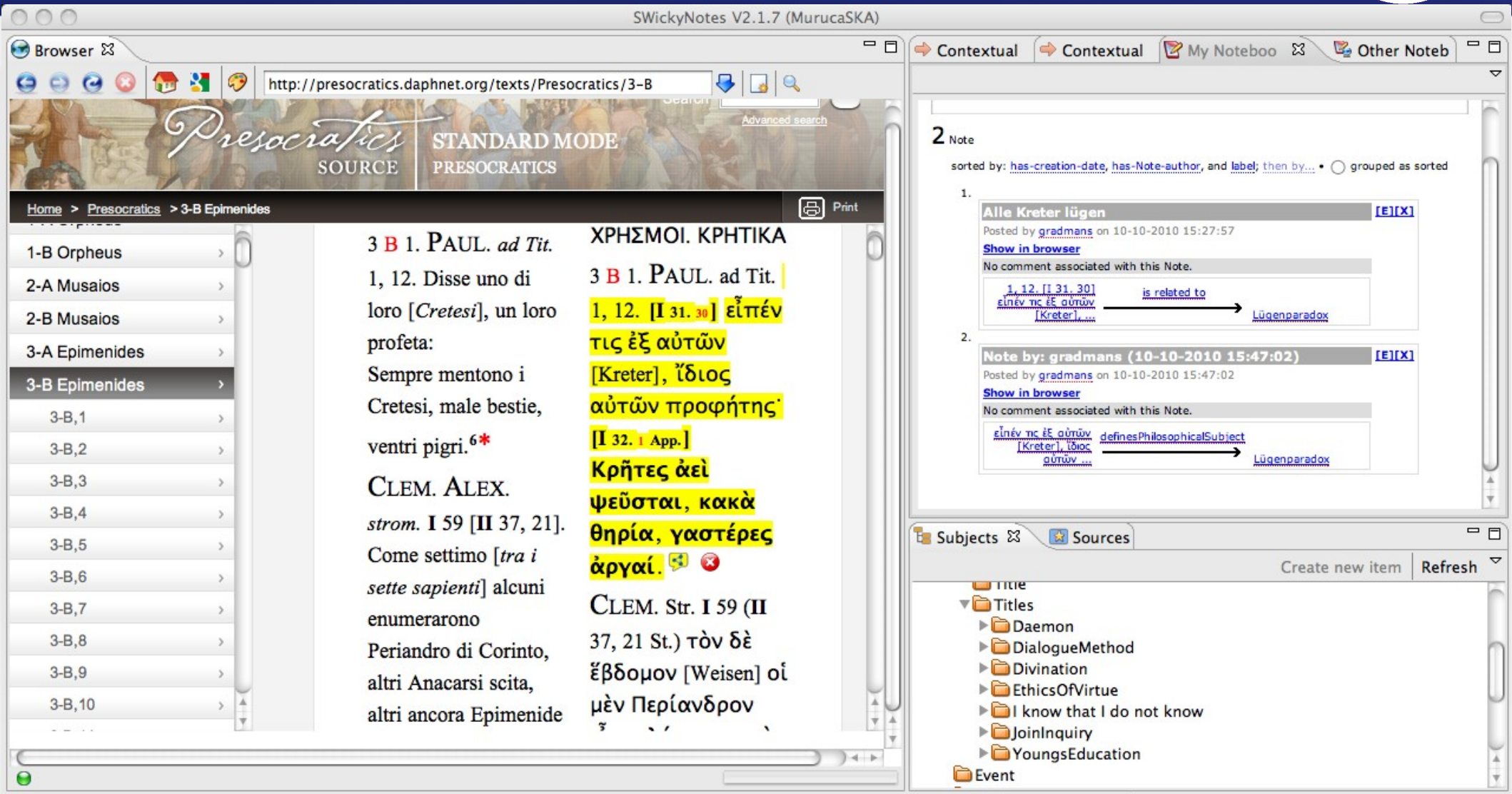
LibreOffice The Document Foundation

SwickyNotes: Selecting Ontologies



CC BY SA
LibreOffice The Document Foundation

“Cretans are always Liars” … annotated



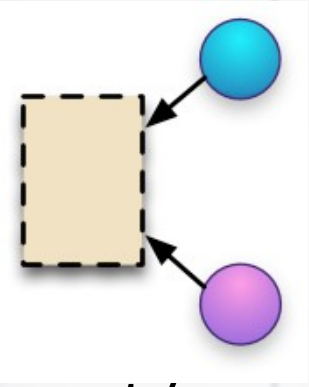
SemLib

- SemLib (http://www.semlibproject.eu/) is a continuation of Discovery in a EU funded project under FP7 working on
  - A Tool to export existing metadata in RDF and publish it as **Linked Data** (Web of Data);
  - A Semantic **Annotation** System (Pundit), to exploit user-generated RDF metadata and publish it as **Linked Data**;
  - A Semantic Recommender System, to use **Linked Data** to improve searching and browsing in the DLs.

CC BY SA
LibreOffice The Document Foundation

Shared Canvas

- Shared Canvas (http://www.shared-canvas.org) is about annotation again – but in a much more sophisticated data model enabling multiple and potentially concurrent layered annotations.



- Demo at http://www.shared-canvas.org/impl/demo1/

- Common traits
  - Use of RDF as underlying technology
  - Emulation of well known functionality on the Web: Annotation

CC BY SA

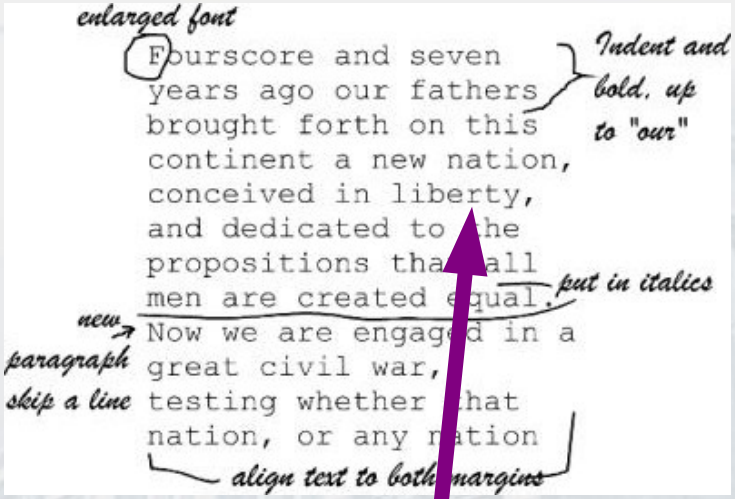
Scholarly Use of Digitisation
… beyond Emulation

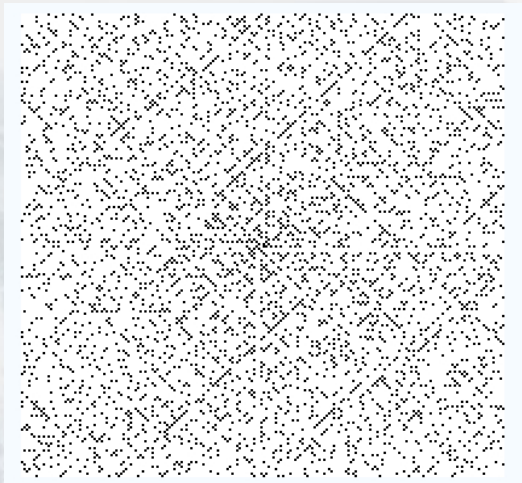
CC BY SA

LibreOffice
The Document Foundation

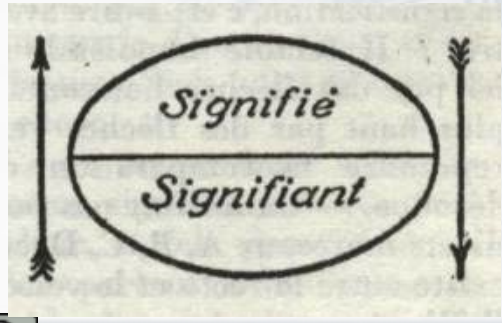
Scholarly Primitives

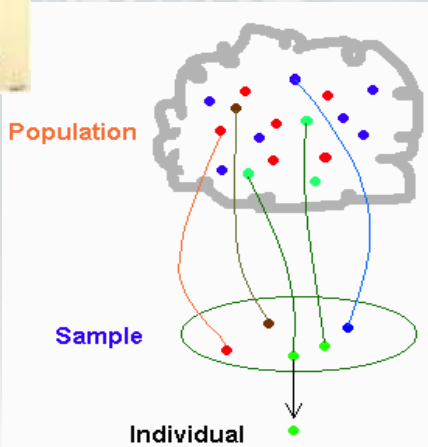
- Discovering
- Annotating
- Comparing
- Referring
- Sampling
- Illustrating
- Representing

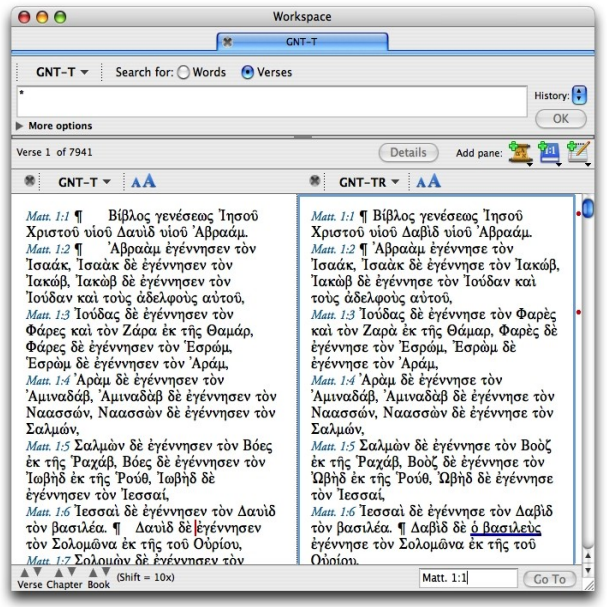








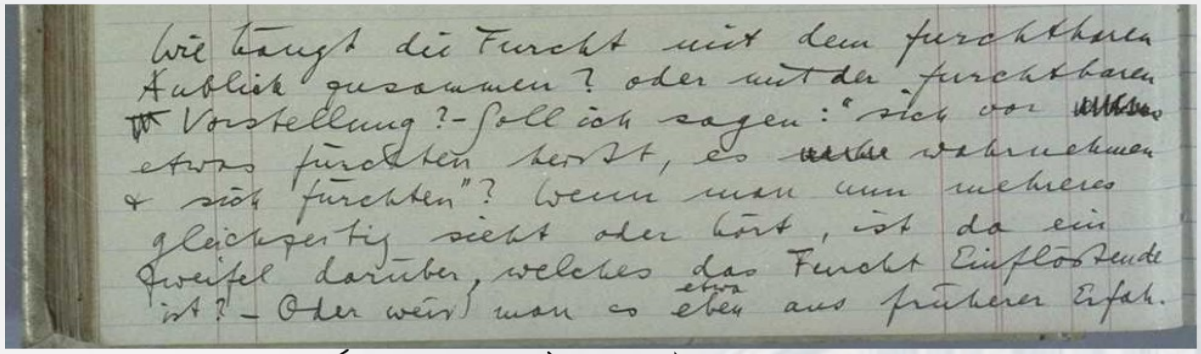




CC BY SA

LibreOffice
The Document Foundation

Massive Contextualisation



CC BY SA

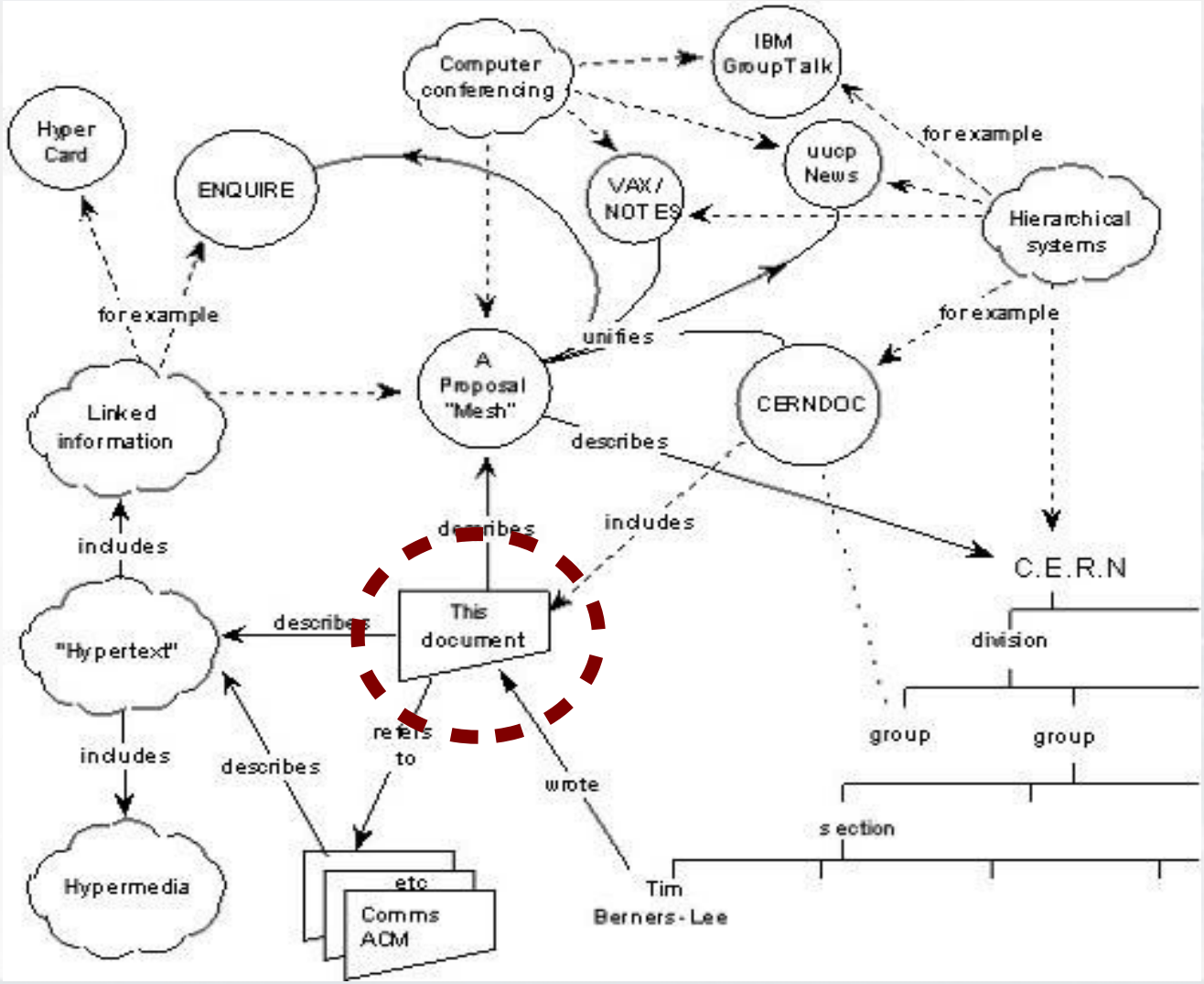
LibreOffice The Document Foundation

More Context:
Resource Description Framework (RDF), Linked open Data (LoD) and the Europeana Data Model (EDM)

CC BY SA

LibreOffice
The Document Foundation

The Web of Documents



Information Management: A Proposal (TBL, 1989)

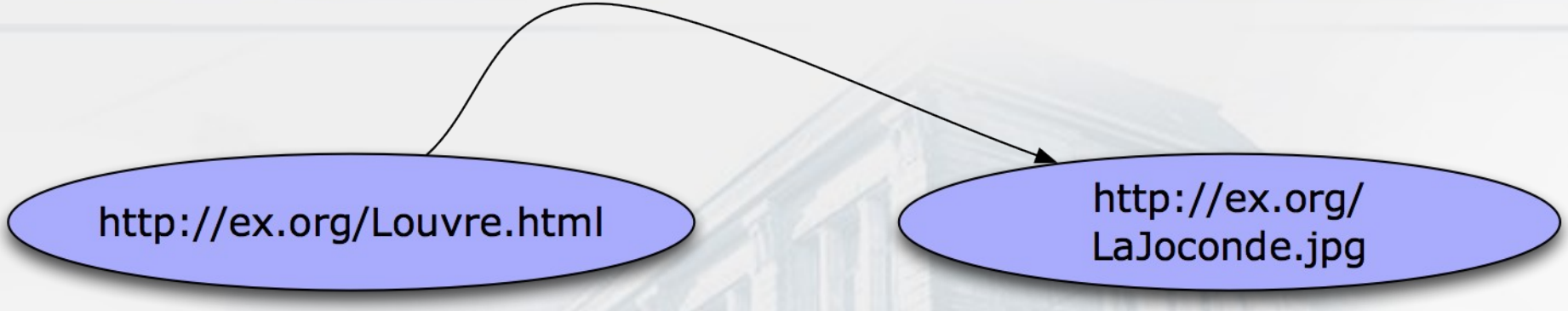
The Linked Data Web is a double Extension of this model

- in **syntax**
- in **scope**

CC BY SA

LibreOffice The Document Foundation

Resources and Links in the Document Web




```
<a href="http://ex.org/LaJoconde.jpg">b</a>
```

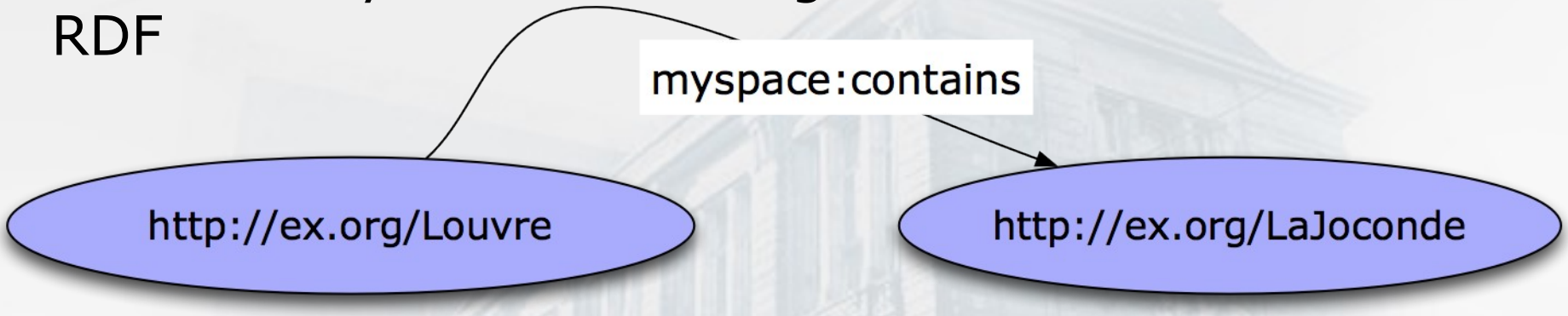
- We have HTTP URIs to identify resources and links between them – but we are missing a few things!
- What kinds of resources are 'Louvre.html' and 'LaJoconde.jpg'?
  - A machine cannot tell.
  - Humans can: we recognize implied context!
- How exactly do they relate to each other?
  - A machine cannot tell.
  - Humans can: again we recognize implied context!

CC BY SA

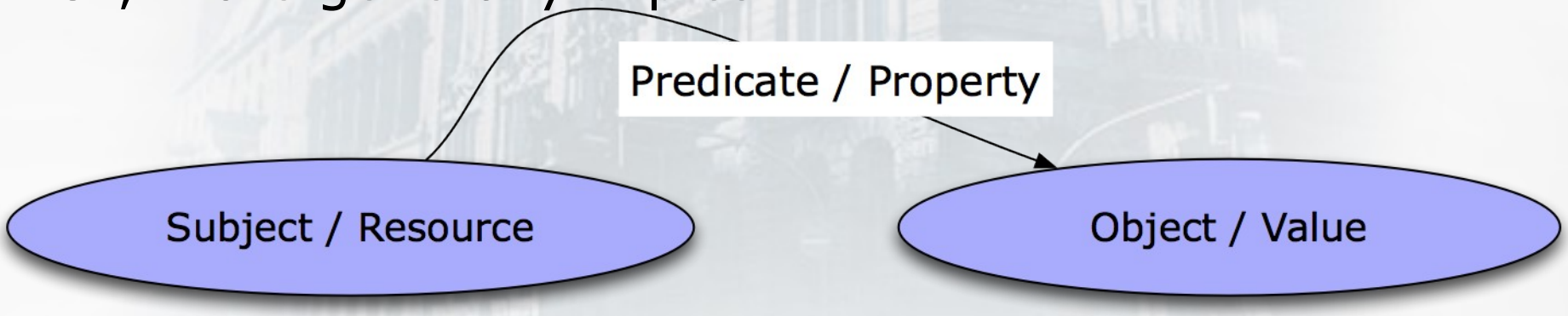
LibreOffice
The Document Foundation

Syntactically Extending the Document Web: RDF (1)

- We add a syntax for making statements on resources: RDF



- Or, more generally triples ...



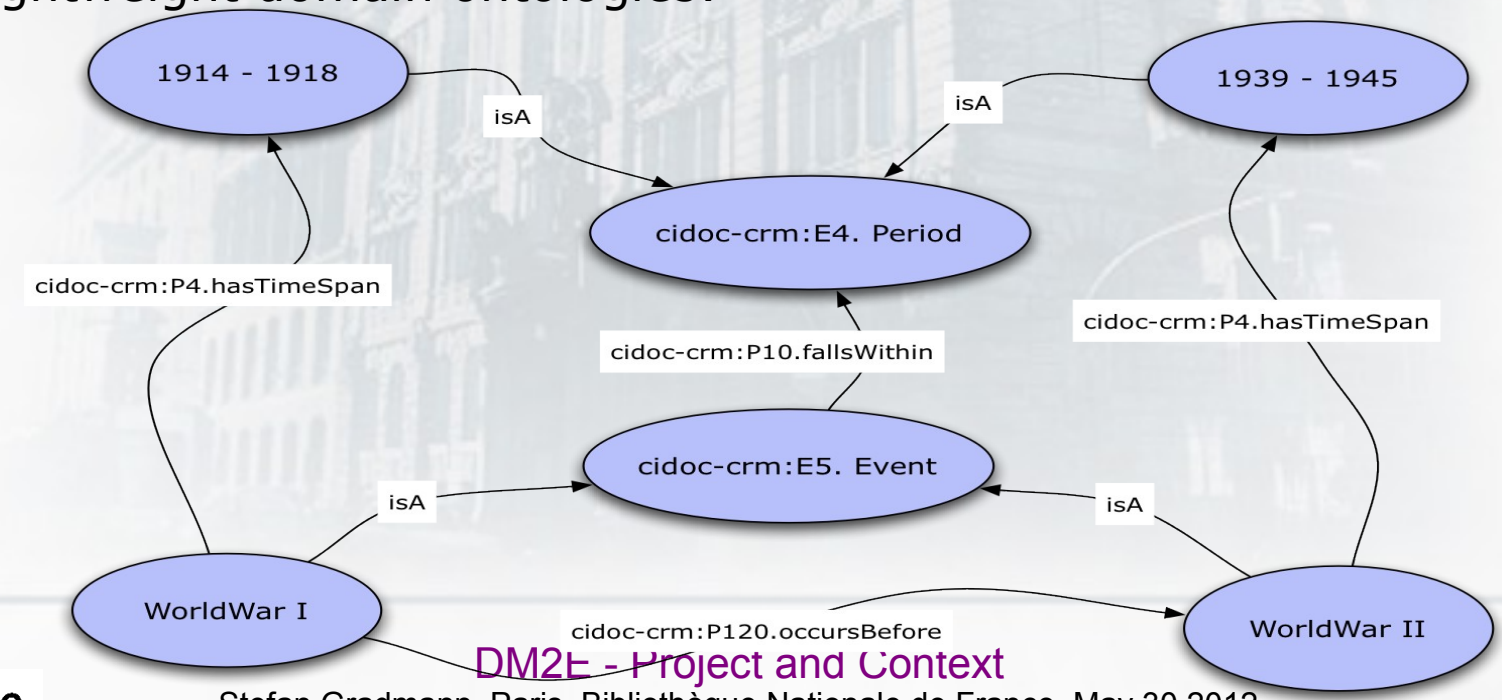
- … where S and P are web resources (identified using URIs) and O is either a web resource or a literal

CC BY SA

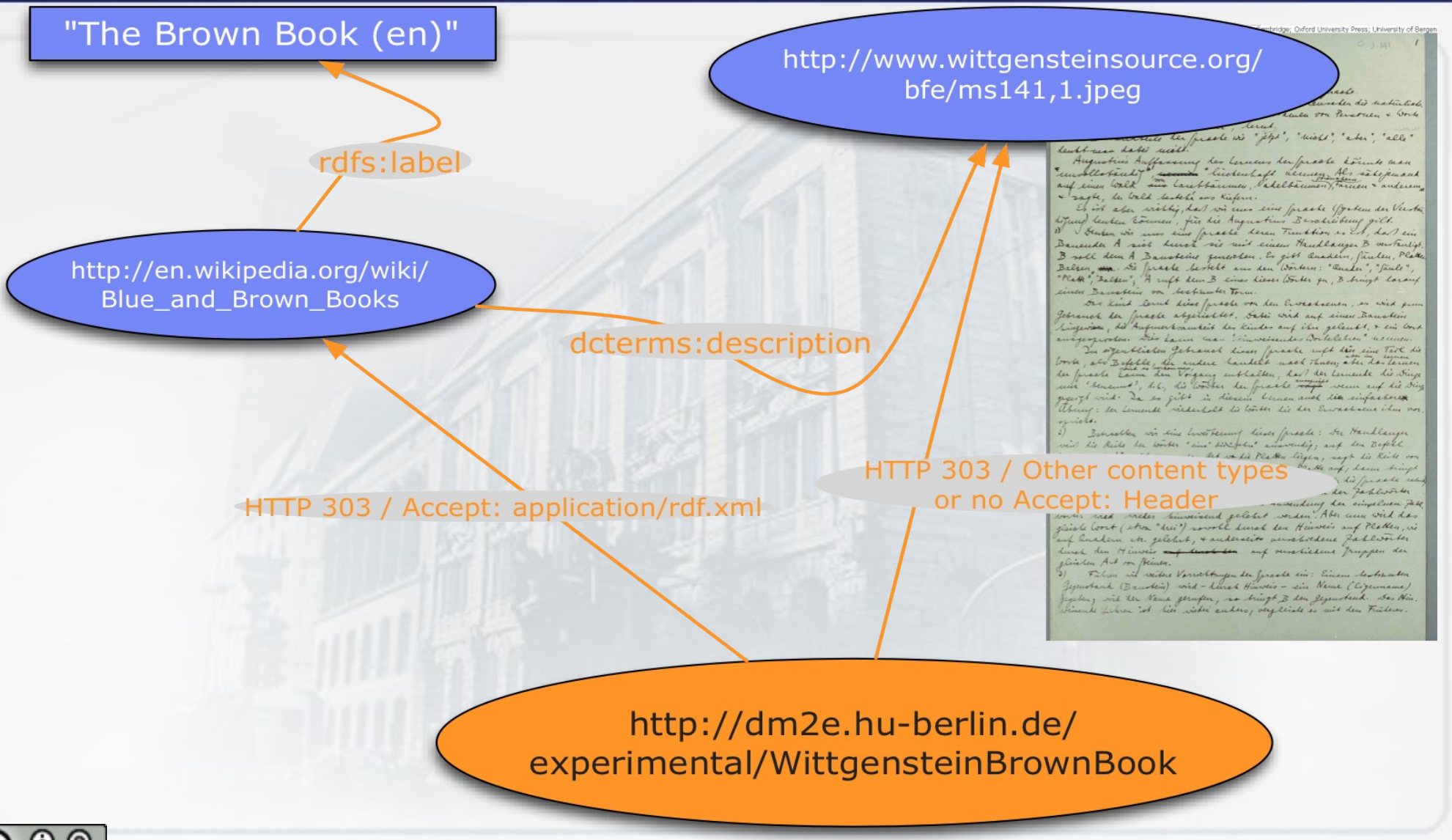
LibreOffice The Document Foundation

Syntactically Extending the Document Web: RDF (2)

- We add a schema language (RDFS) with elements such as
  - classes (chair' as instance of chairs),
  - hierarchies of classes and properties (chairs are a subclass of furniture, 'teaches' is a sub-property of 'communicates')
  - inheritance (communication based on language → teaching also is)
  - support for basic inferencing.
- And thus are able to establish structures in triple aggregations resulting in lightweight domain ontologies:



… and the Way we extend the Web in scope to make it a 'Web of Things'



CC BY SA

LibreOffice The Document Foundation

And we get … Linked Data

Linked Data essentials

1. Use URIs — Standard **Identifiers**
2. Use HTTP URIs — Standard **Pointers**
3. Serve useful information using SPARQL, RDF standards — Standards for **Queries** and **Statements**
4. Mention URIs of related objects — **Link** to Context

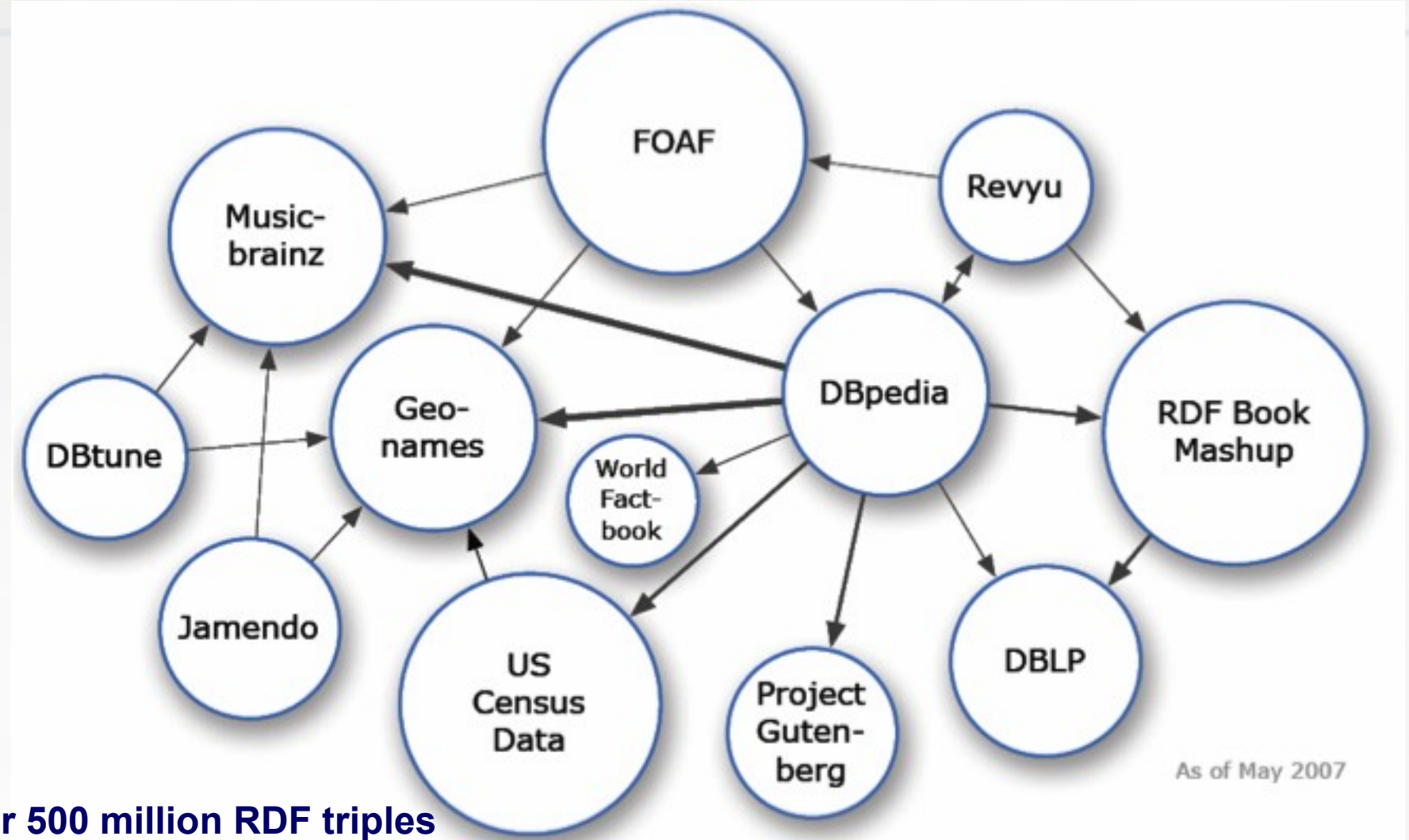
Copyright © 2008 W3C (MIT, ERCIM, Keio)

http://www.w3.org/2008/Talks/0617-lod-tbl/#(4)

CC BY SA

LibreOffice The Document Foundation

A few Bubbles: 5/2007

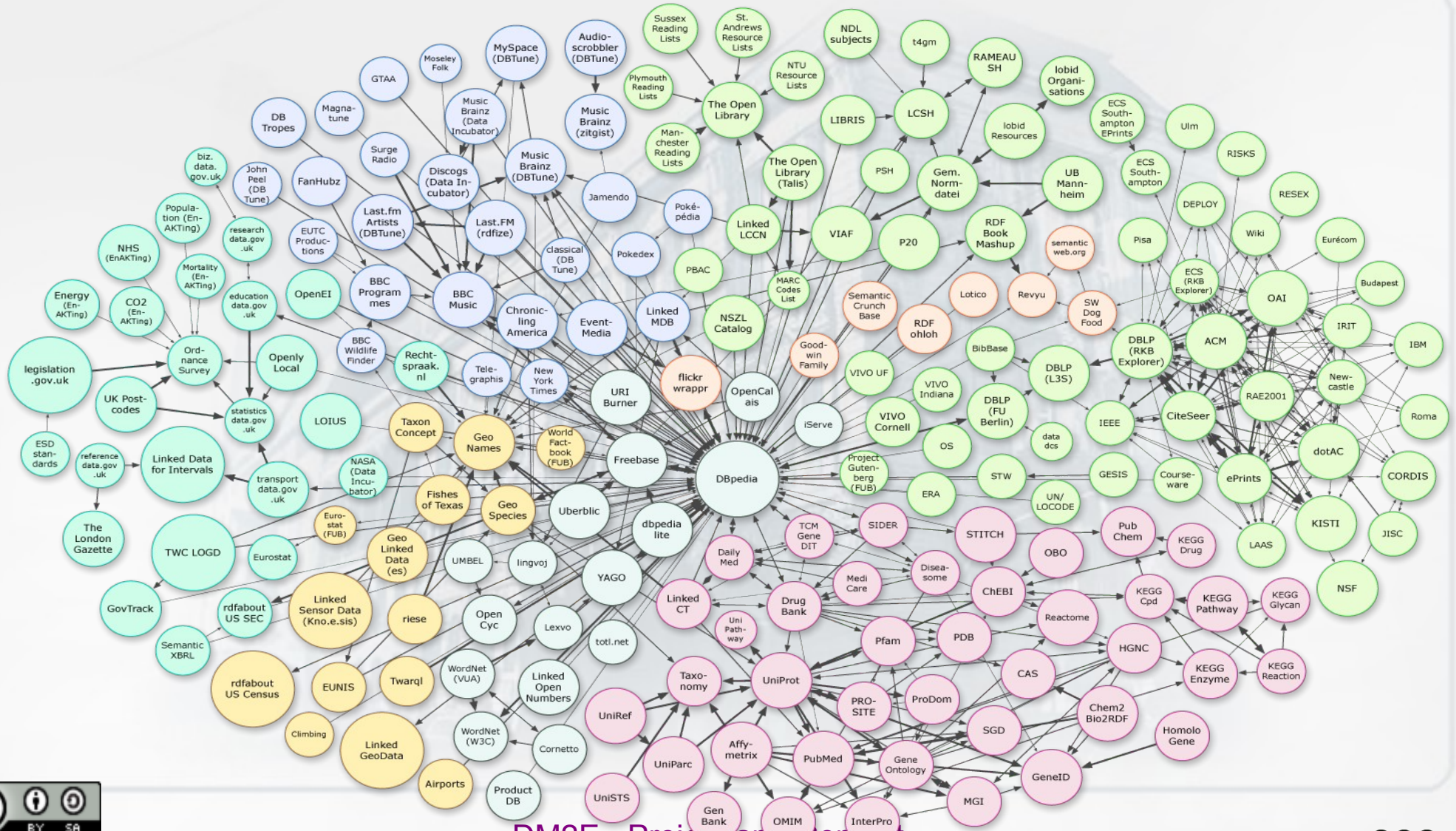


- **Over 500 million RDF triples**
- **Around 120,000 RDF links between data sources**

© Richard Cyganiak

A Whole Lot of Bubbles: 09/2010

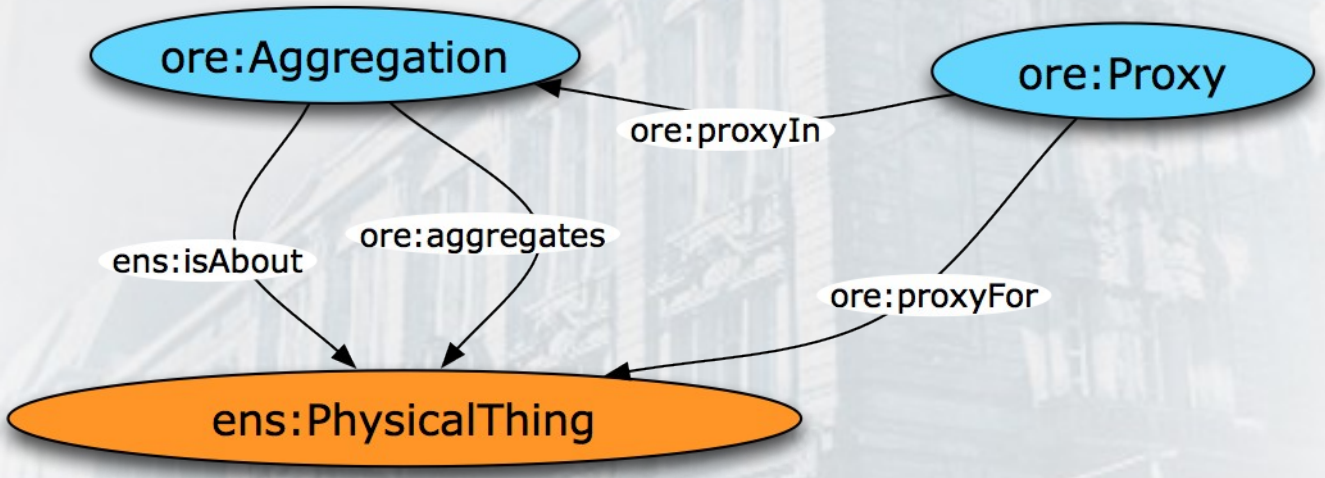
http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html



CC BY SA

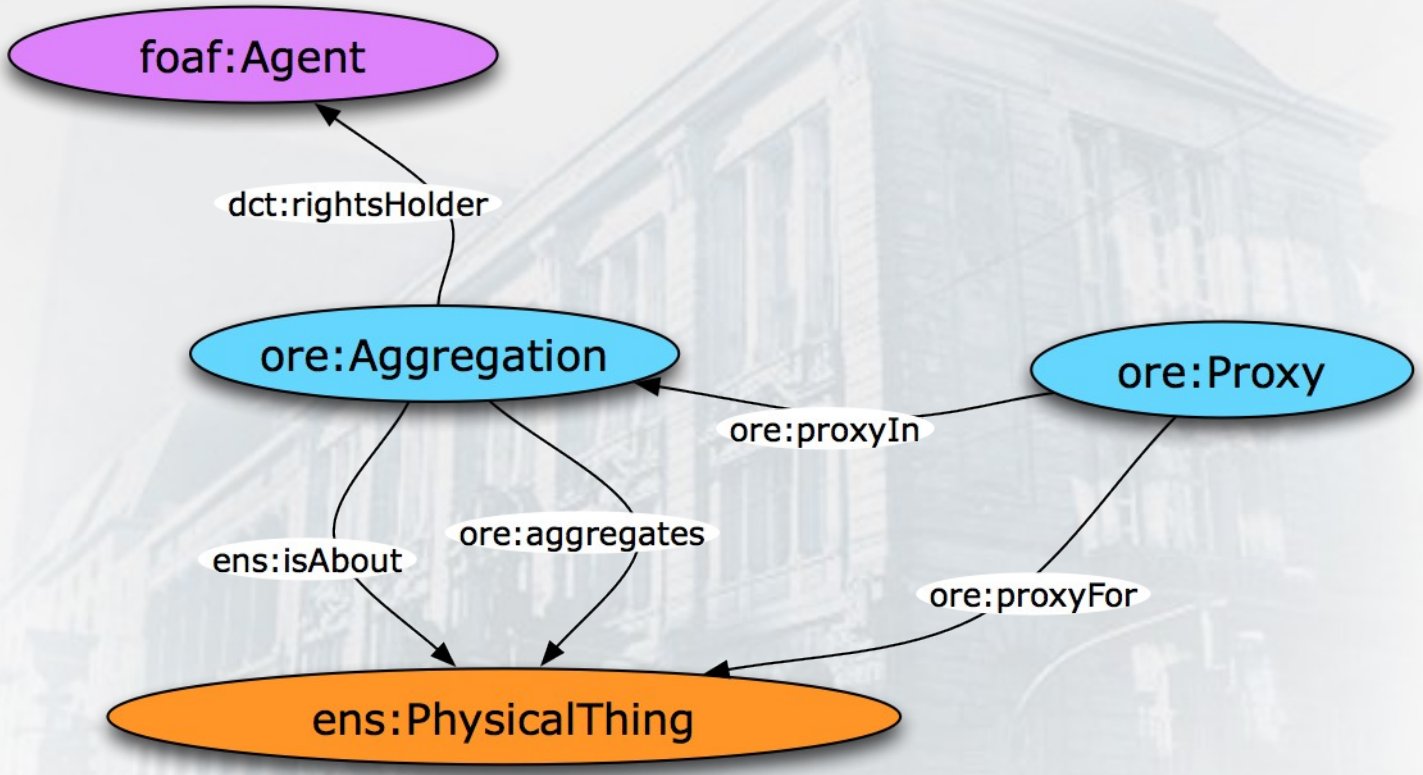
LibreOffice The Document Foundation

An EDM Aggregation ...



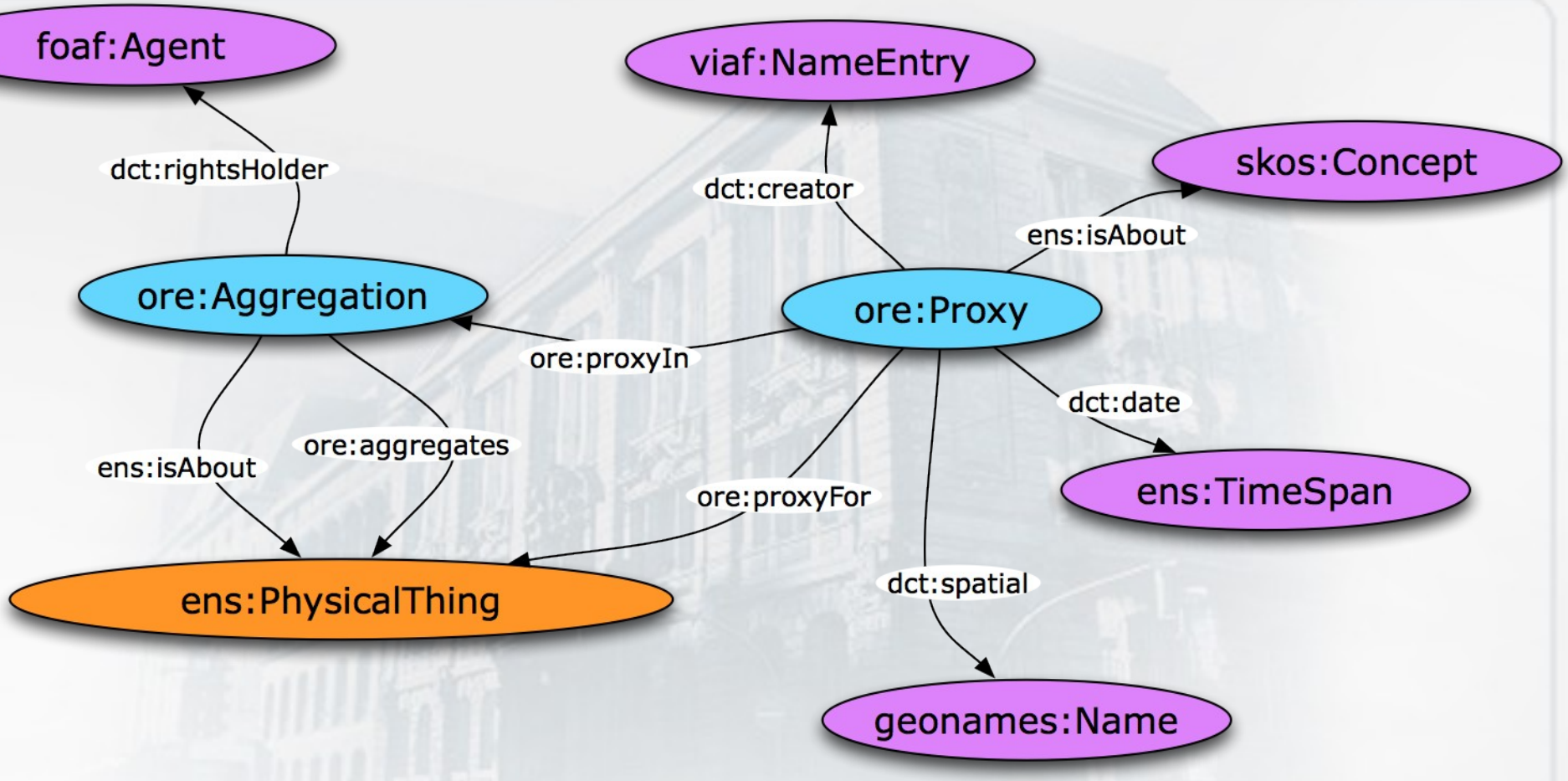
CC BY SA
LibreOffice The Document Foundation

… some context



CC BY SA

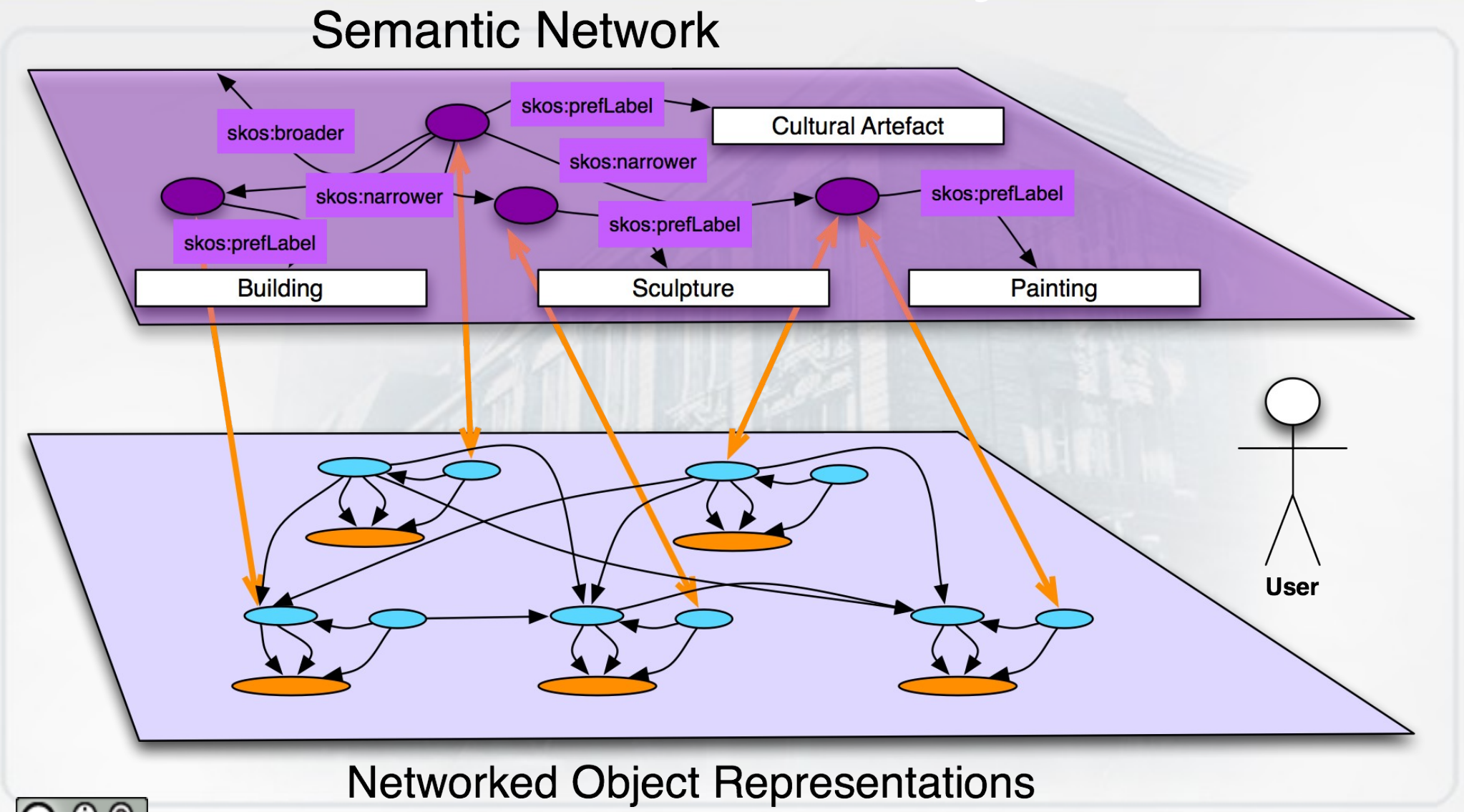
… more context



CC BY SA

LibreOffice The Document Foundation

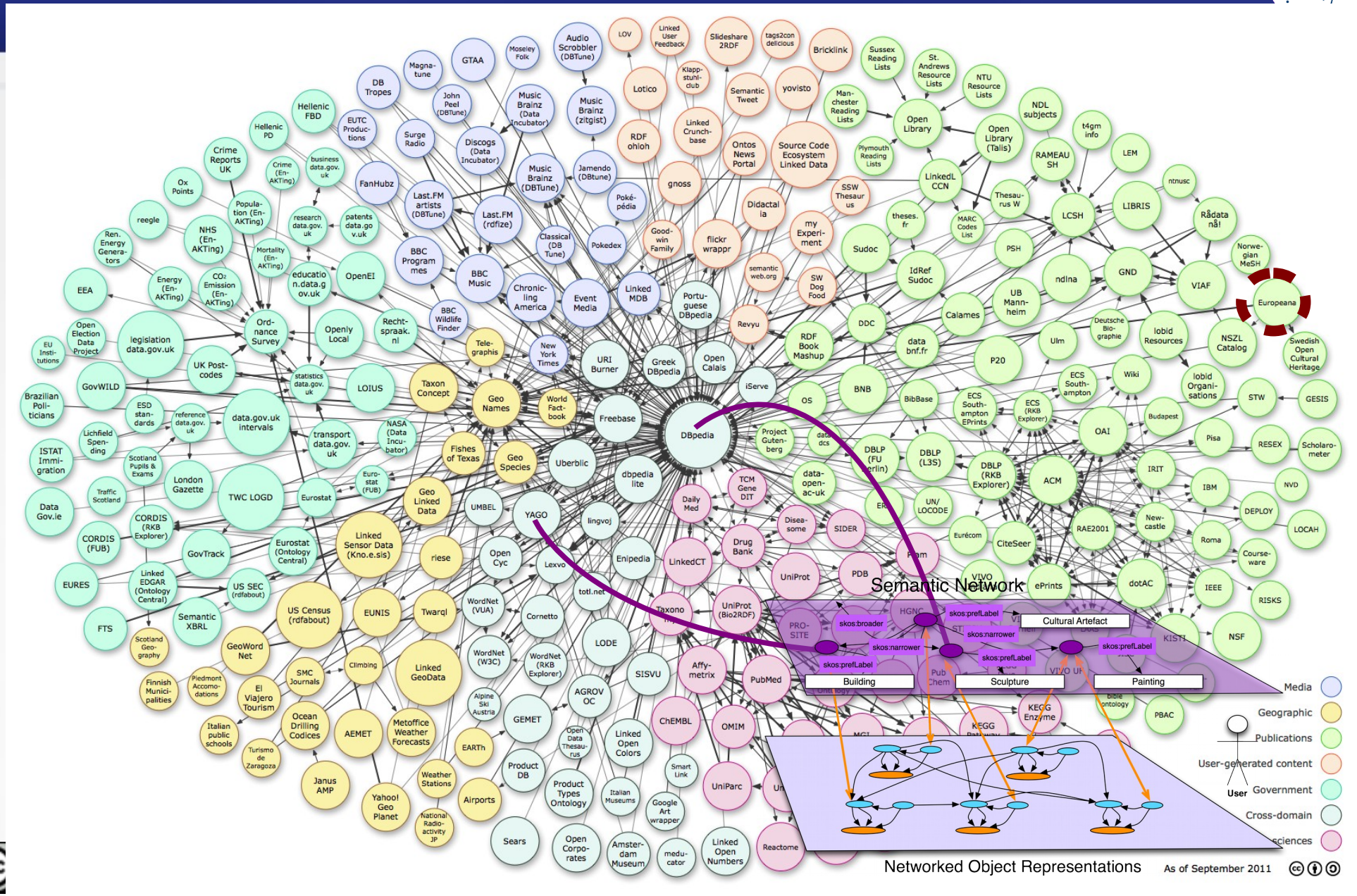
… and the Big Picture: The Semantic Data Layer



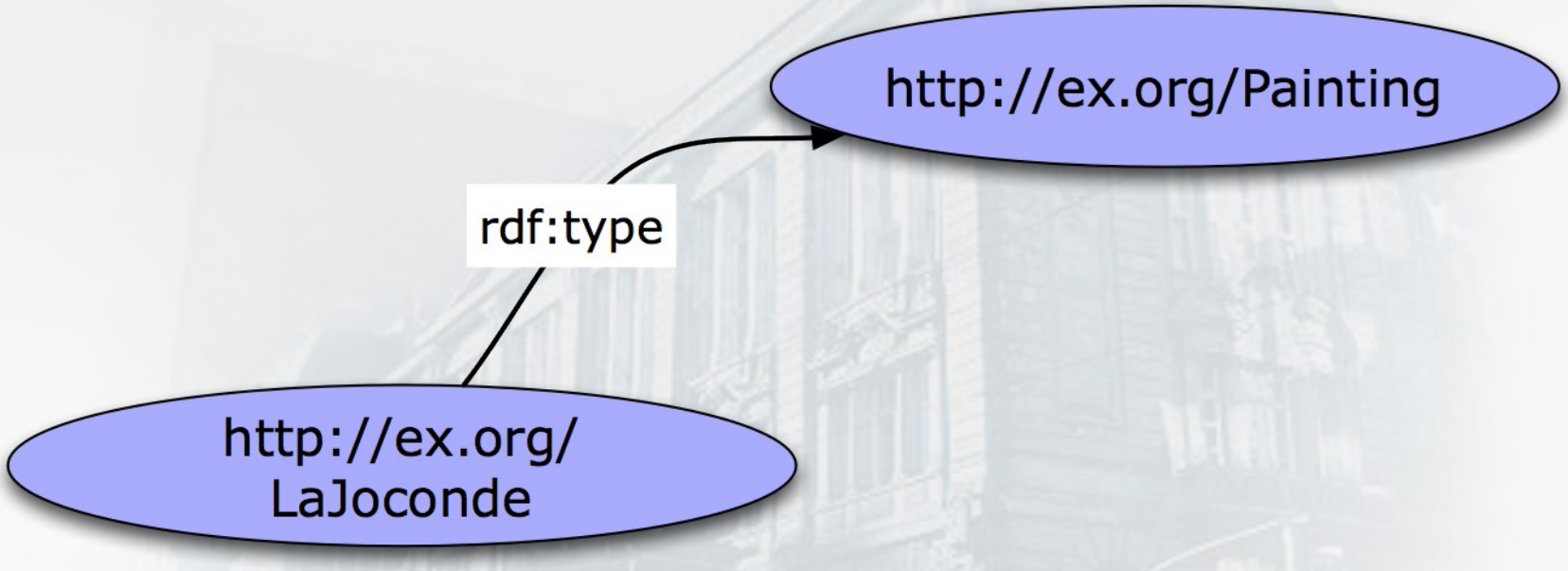
CC BY SA

LibreOffice The Document Foundation

EDM and Linked Open Data



… and new opportunities: Triple Sets and Inferencing (1)



CC BY SA

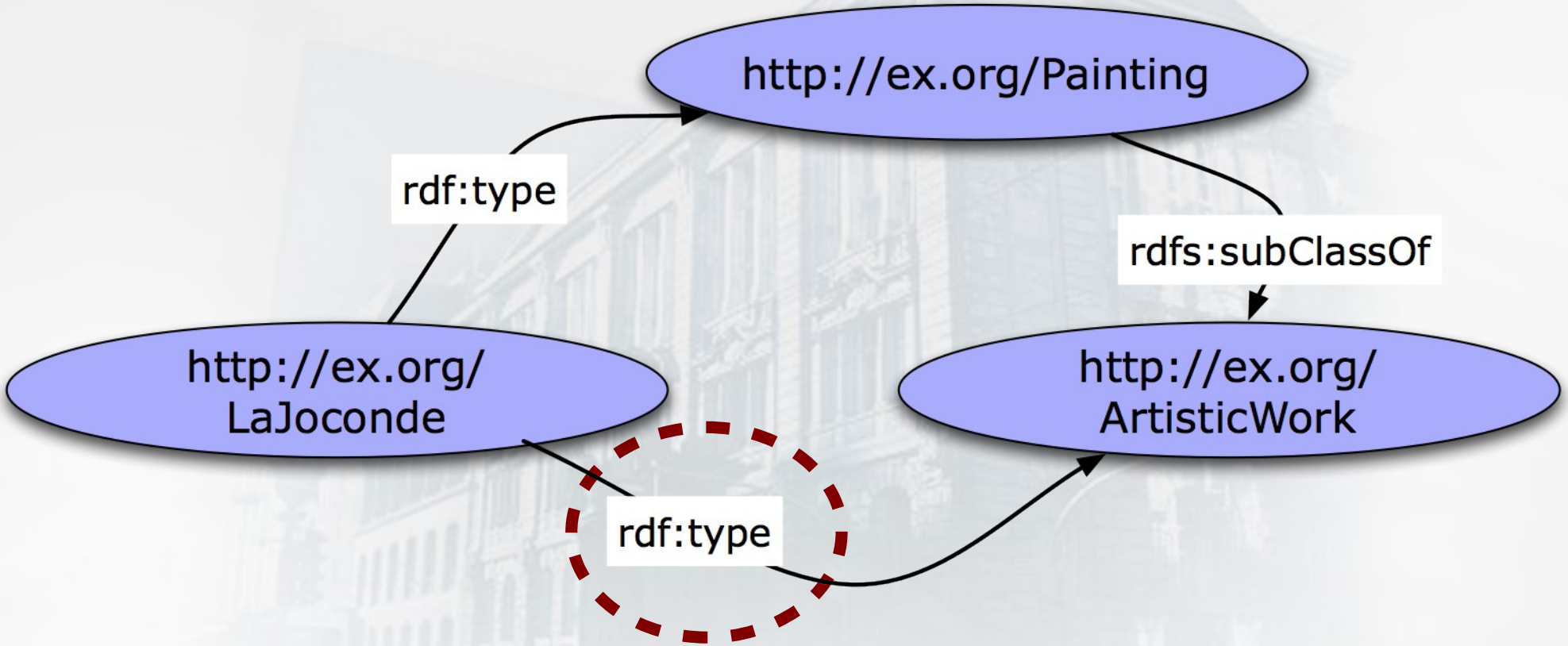
LibreOffice
The Document Foundation

Triple Sets and Inferencing (2)



CC BY SA
LibreOffice
The Document Foundation

Triple Sets and Inferencing (3)



→ Potential of novel digital heuristics!

CC BY SA

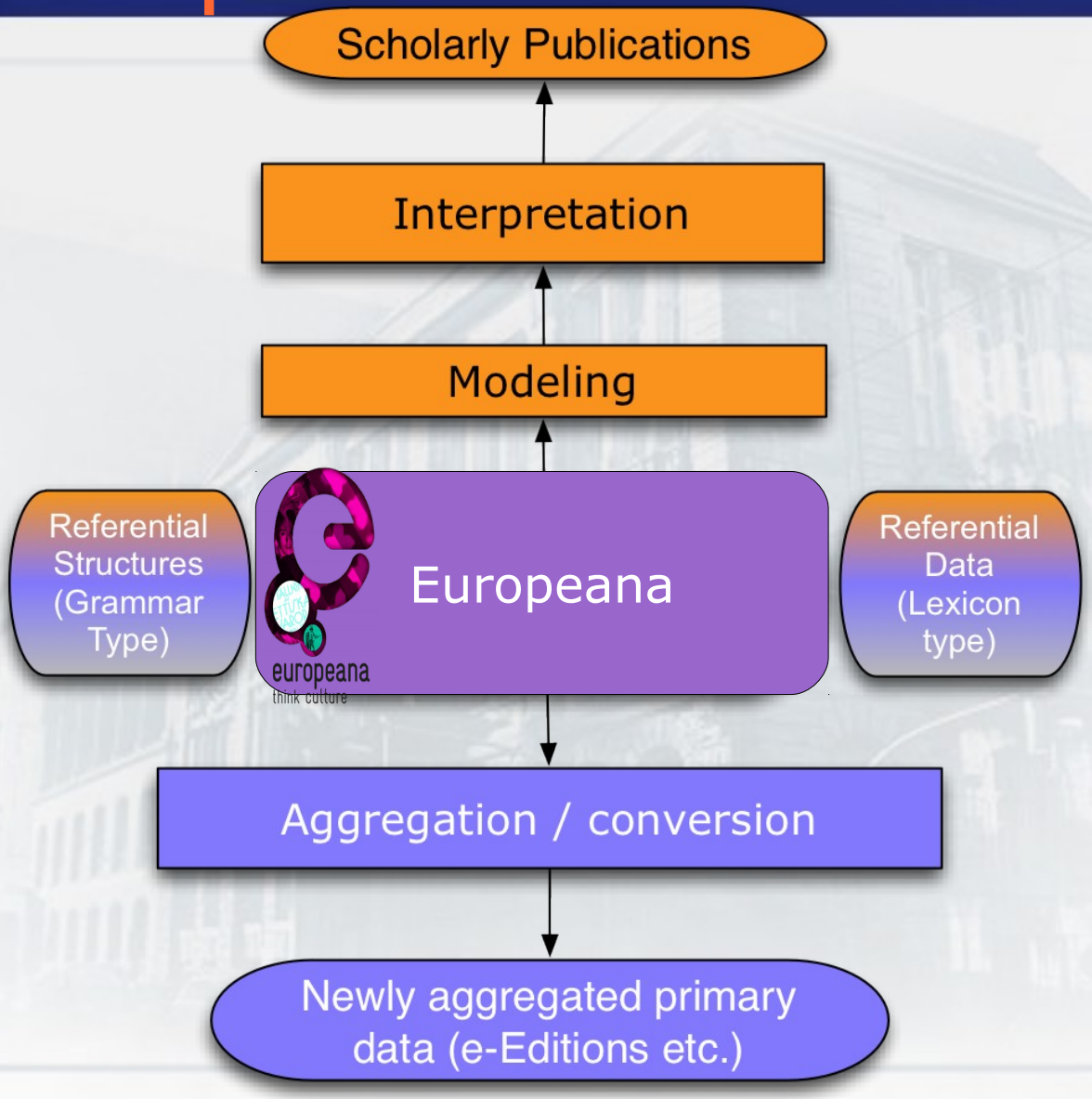
EDM + Muruca
+ Digital Humanities =



CC BY SA

LibreOffice
The Document Foundation

Digital Humanities
... and Europeana?



CC BY SA

LibreOffice
The Document Foundation

Digitised Manuscripts to Europeana (DM2E): What?

- WP1: Provide substantial amounts of **digital content** to Europeana with a focus on **digitised manuscripts**
- WP2: Integrate existing **technical building blocks**
  - from Europeana development
  - as well as from generic LoD oriented development
  - into a **generic production chain** for migrating data from various sources to the EDM as well
  - as for the **contextualisation** of the object representations.
- WP3: Explore usage scenarios of such content in a specialised platform for **humanities research** generating digital heuristics and making data as well as heuristics available to specialised visualisation or reasoning environments.

CC BY SA

Digitised Manuscripts to Europeana (DM2E): Who (1)?

- **Content Providers**
  - European Association for Jewish Culture (Judaica)
  - Max-Planck-Institut für Wissenschaftsgeschichte (ECHO)
  - Österreichische Nationalbibliothek (Google)
  - Staatsbibliothek zu Berlin (Kalliope)
  - University of Bergen (Wittgenstein)
  - CNRS ITEM (Nietzsche)
  - National Library of Israel (Judaica)
  - Berlin Brandenburgische Akademie (German Text Archive)
  - Humboldt-Universität zu Berlin (Polytechnisches Journal)
- **Technology Providers**
  - ExLibris (Aleph, MARC sources management)
  - Freie Universität Berlin (dBPedia, LoD, LoD2, D2R, SILK)
  - Max-Planck-Institut für Wissenschaftsgeschichte (ECHO)
  - Net7 S.r.l. (Muruca/Pundit)
  - National Technical University of Athens (MINT)

CC BY SA

LibreOffice The Document Foundation

Digitised Manuscripts to Europeana (DM2E): Who (2)?

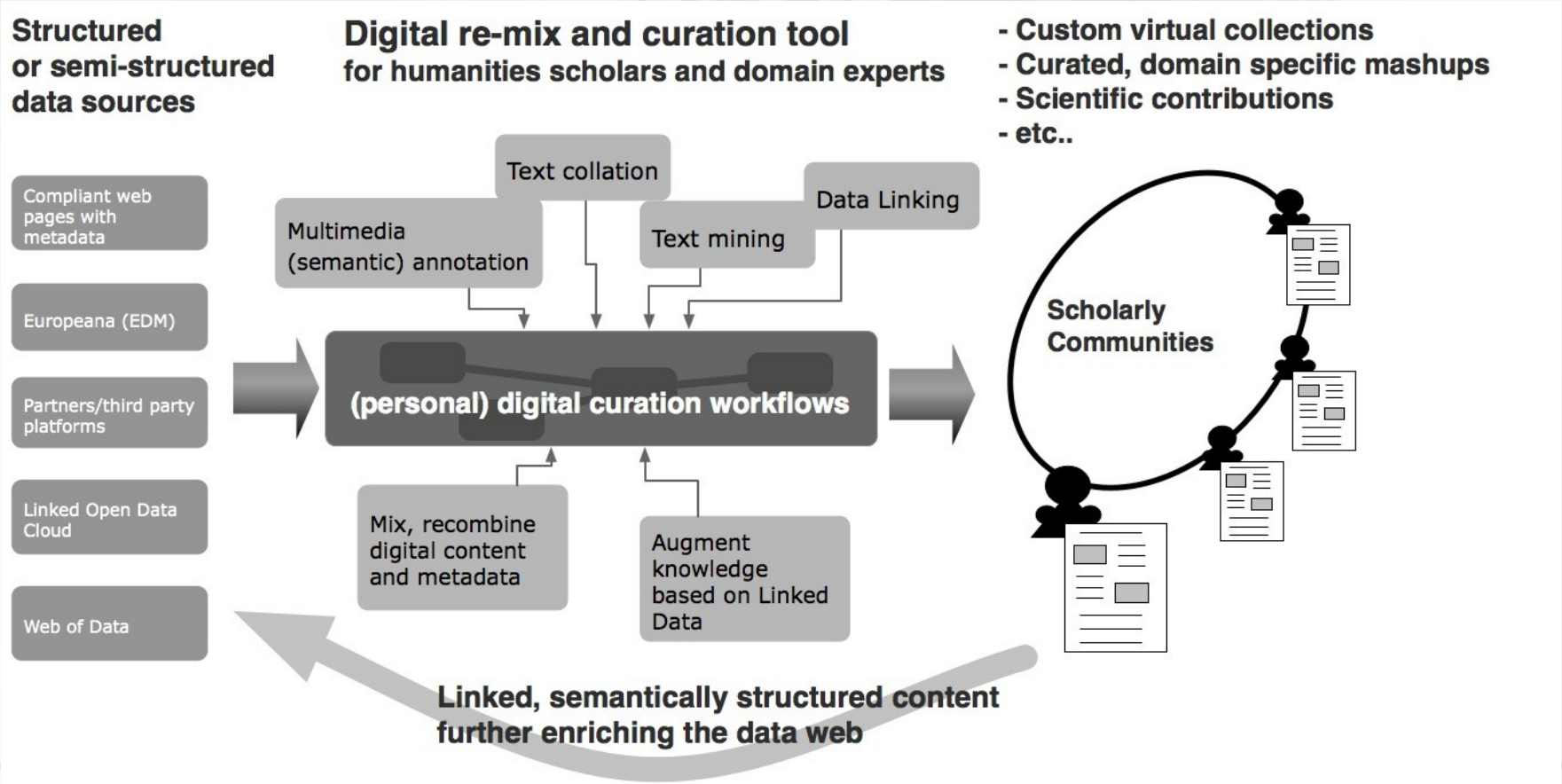
- **Digital Humanities Community**
  - Dr. Tobias Blanke (King's College, London)
  - Sally Chambers (The European Library / DARIAH-D)
  - Prof. Dr. Stefan Gradmann (HUB)
  - Prof. Dr. Gerhard Lauer (Göttingen Center for the Digital Humanities)
  - Dr. Alois Pichler (UIB)
  - Dr. Jürgen Renn (MPIWG)
  - Dr. Laurent Romary (HUB)
  - Prof. Dr. Susan Schreibman (Trinity College Dublin)
  - Dr. Claire Warwick (University College, London)
- **Community Building**
  - Open Knowledge Foundation (OKFN)
- **Coordination, Management & Information Science**
  - Humboldt-Universität zu Berlin (HUB)

CC BY SA

LibreOffice The Document Foundation

WP3: Digital Humanities Requirements and Related Engineering - Context

Goal: lower the barriers for **digital content curation** by providing an integrated, flexible, semantic based environment targeted to **digital humanities scholars**



CC BY SA

LibreOffice The Document Foundation

Digital Humanities Functionality (WP3)

- How to enable **digital scholarship** building on combined EDM RDF metadata and digital surrogates ...
- … building on an ontological, granular representation of John Unsworth's **scholarly primitives** (or their successors according to Blanke/Hedges 2011) …
- ... resulting in a increasingly complex **social semantic scholarly graph** containing RDF statements such as
  - $Version_A$ – isSuccessorOf – $Version_B$
  - $Statement_1$ – contradicts - $Statement_2$
  - $Scribe_Y$ – copiedFrom – $Scribe_Z$
- → generate **scholarly heuristics** from inferences on this graph ...

CC BY SA

LibreOffice The Document Foundation

Conclusion

- Digitised manuscripts are **cultural heritage objects** in their own right!
- They enable doing things scholars weren't able to do before – if we can leave behind emulation!
- Needs of **scholarly users** should be the driving force
  - → We need to better understand what scholars are actually doing
  - → Are scholars interested in 'infrastructure'?
- **RDF** and its future extensions are key technology: they enable us to leave the "walled gardens"
- Try to avoid some terms librarians are so fond of such as '**Catalogue**' or '**Collections**': they are deeply rooted in the walled garden paradigm!

CC BY SA

Selected Reading

- Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, Herbert van de Sompel (2010): The Europeana Data Model. IFLA 2010 (Gothenburg). Session on „Libraries and the Semantic Web". http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf
- Stefan Gradmann (2010): Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. Europeana White Paper 1. http://www.scribd.com/doc/32110457/Europeana-White-Paper-1
- John Unsworth (2000): Scholarly Primitives. What methods do humanities researchers have in common, and how might our tools reflect this? In the seminar on Humanities Computing, King's College, London. http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html
- Tobias Blanke, Mark Hedges (2011): Scholarly primitives. Building institutional infrastructure for humanities e-Science, Future Generation Computer Systems, In Press (Science) http://www.sciencedirect.com/science/article/pii/S0167739X11001178#br000085

CC BY SA

LibreOffice The Document Foundation

Have a look at http://dm2e.eu from time to time!

Thank you!

CC BY SA

LibreOffice
The Document Foundation