

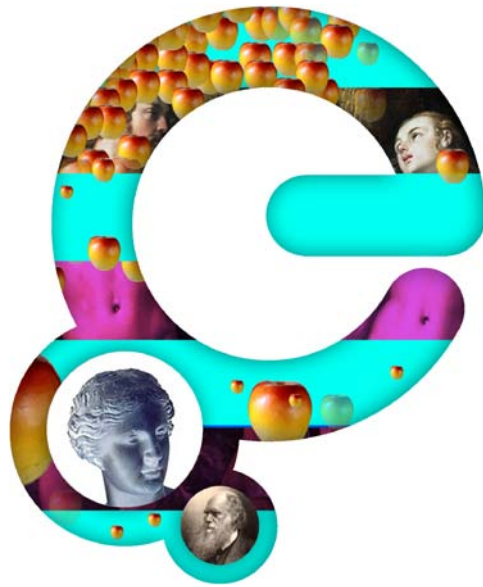
Europeana and semantic alignment of vocabularies

Antoine Isaac

Jacco van Ossenbruggen, Victor de Boer,
Jan Wielemaker, Guus Schreiber

Europeana & Vrije Universiteit Amsterdam

NKOS workshop, Berlin, Sept. 28th 2011



eupepeana
think culture

?

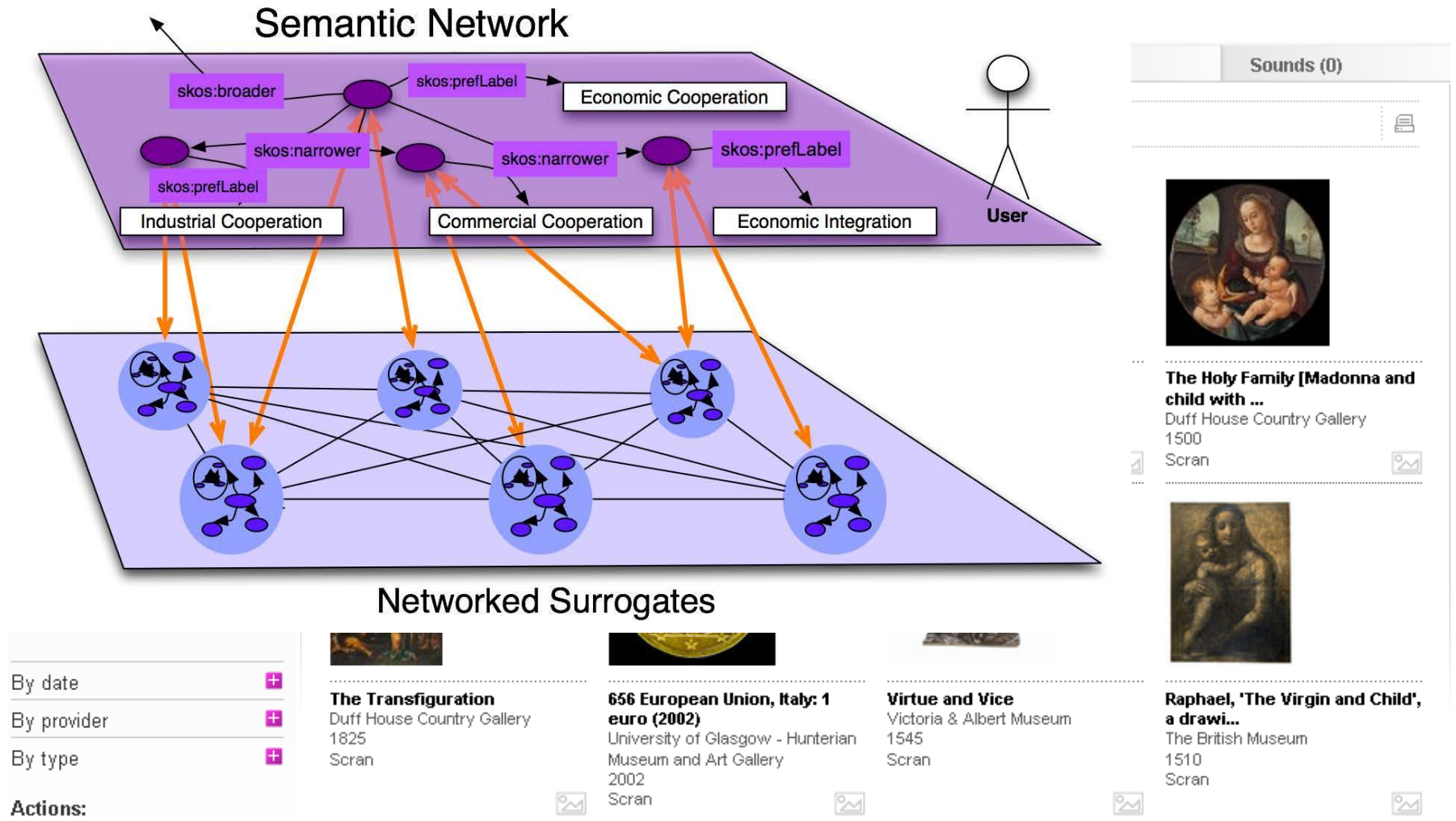


.....
europeana

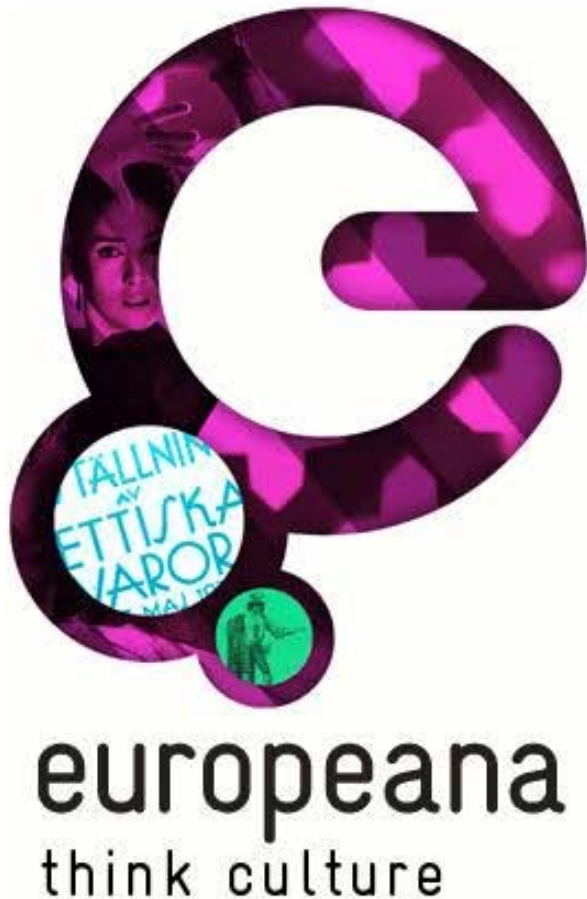
connect

Towards semantics-enabled search

Building a "semantic layer" to help accessing content



Prototype: Europeana Thought Lab



This is a **research prototype** of Europeana's semantic search engine.
Enter a search term, for example: **Egypt**, **Rembrandt**, **window**.

Collections

Thesauri



Rijksmuseum

46,038 artworks



RKD

82,781 artworks



Louvre

11,327 artworks

[home](#) [disclaimer](#) [datacloud](#) [acknowledgments](#)

<http://europeana.eu/portal/thought-lab.html>

keyword: <http://e-culture.multimedien.nl/ns/rijksmuseum/term23826>  [artefact](#) 

▼ **works showing Egypte (14)**



De 57ste Westfriese ...
Bos, Theo



Beklimming van de gr...
Béchar, Henri



Demotische inscripti...
Blanquart-Evrard, Lo...



Portret van 2 Arabis...
Anonymous



Portret Arabisch mei...
Anonymous



Portret van leraar m...
Anonymous

▼ **works showing Egypt (5)**



Een Egyptische weduw...
Alma-Tadema, Lawrenc...



John Howard in een E...
Boer, Otto de



Berglandschap in Egypte
Bauer, Marius



Karnak, Egypte
Bauer, Marius



Landschap met ploege...
Bauer, Marius

▼ **works showing a more specific Egypt (11)**



Wapen van Haarlem



Verovering van Damia...



Verovering van Damia...



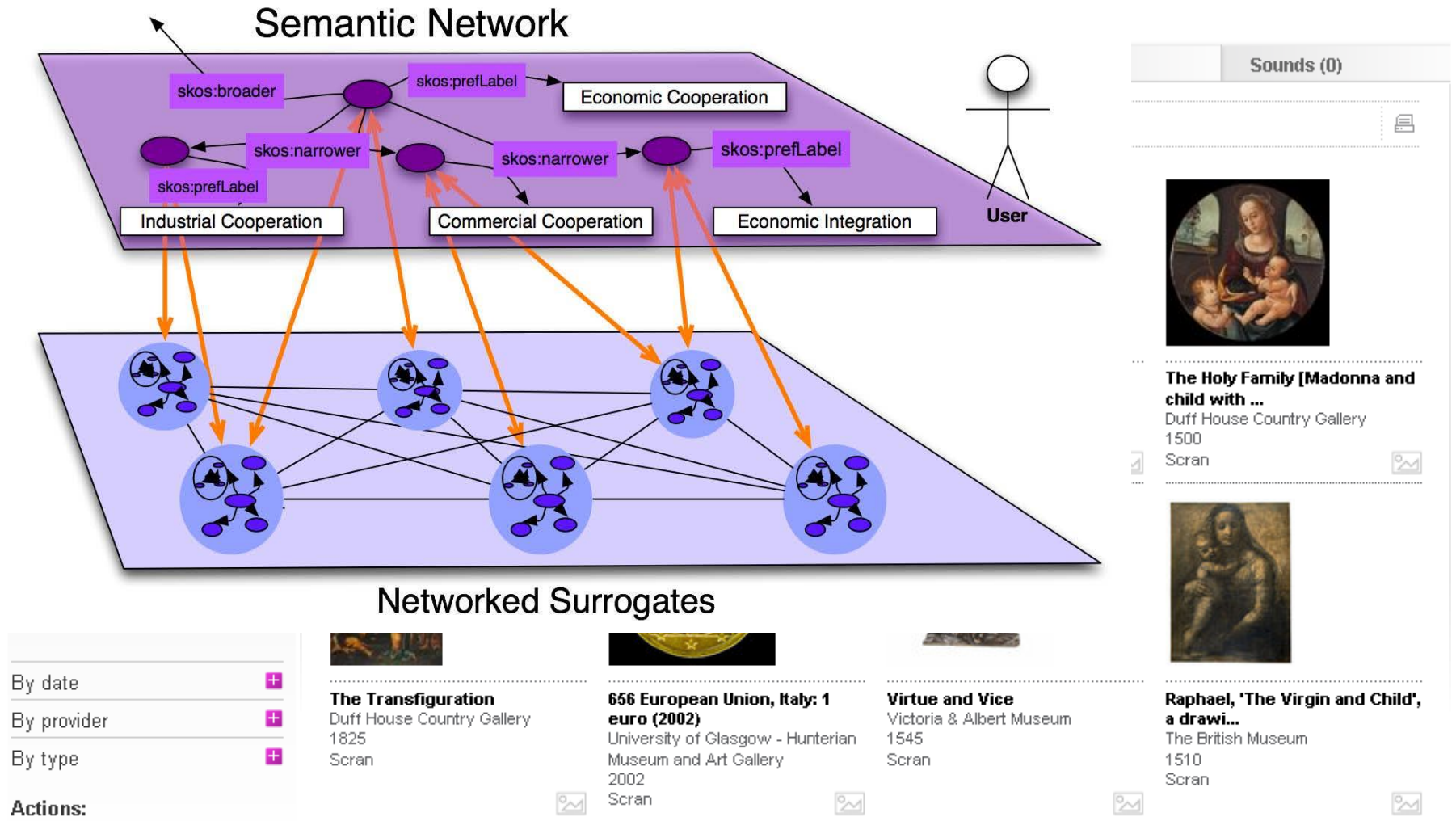
Beleg van Damiate. 121...

Towards semantics-enabled search

- Enhance access to Europeana content by semantics
Query expansion, clustering of results
- Exploiting various types of relations
"located in", "lived in", "is more specific concept"...
- Semantics are already there
Thesauri, classifications...
- No need to convert KOSs to formal ontologies for now
Too much overhead, too difficult to exploit

Towards semantics-enabled search

Building a "semantic layer" to help accessing content



Interoperability challenges?

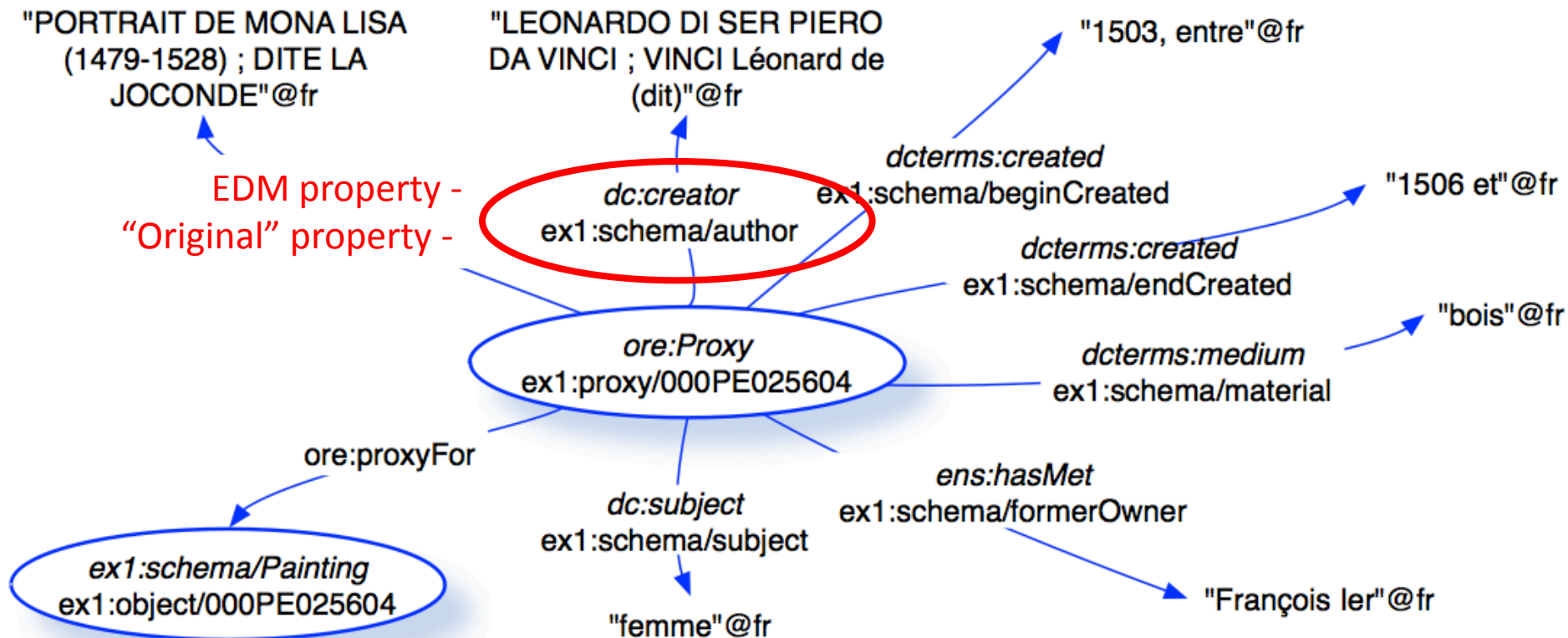
- Starting from heterogeneous data sources
- Conversion of data in RDF
- Semantic alignments
 - Between metadata element schemes
`vra:depicts rdfs:subPropertyOf dc:subject`
 - Between concepts in value vocabularies
`iconclass:bird skos:closeMatch ddc:bird`

Europeana Data Model



Mapping metadata element schemes

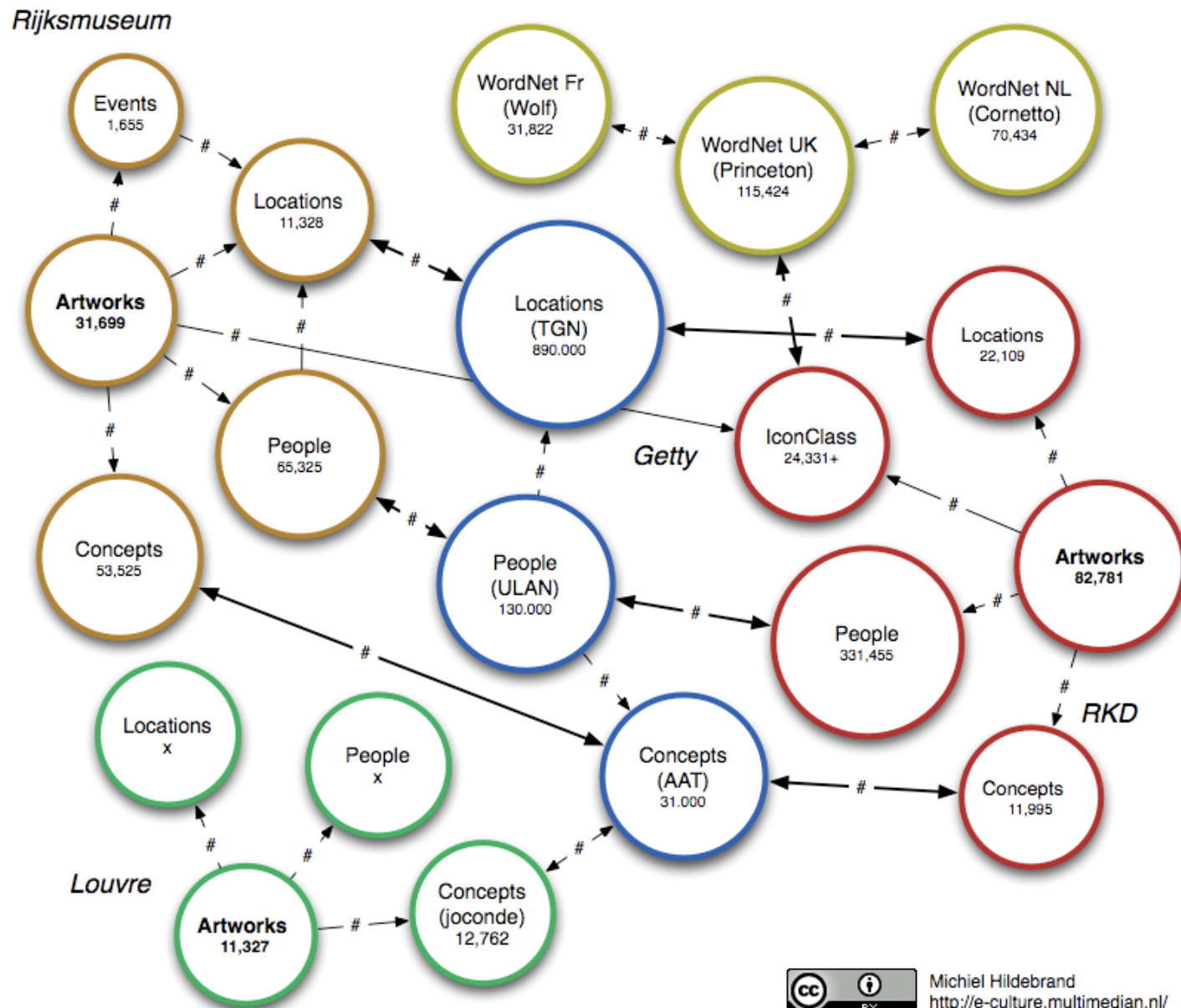
(to EDM)



Mapping metadata element schemes

- Providers will do it themselves!
Yes, manually
- Conversion is automatic, but the specification of the conversion is manual
mapping file using OWL,
SPARQL CONSTRUCT query
whatever XSLT/XMLRDF script...

Building a network of semantic references



A brief story of EuropeanaConnect's efforts on the “semantic layer”...



Where to start from?

- What is already available as RDF data (Linked or not)

SKOS/Datasets

[< SKOS](#)

Contents [\[hide\]](#)

[I Find & Share SKOS Data](#)

[1.1 STW Thesaurus for Economics](#)

[1.2 Thesaurus for the Social Sciences](#)

[1.3 Country Codes](#)

[1.4 Drug Administration Forms](#)

[1.5 Vocabularies for GeoSciML Geoscience information interchange](#)

[1.6 DDC Dewey Decimal Classification](#)

[1.7 Libris' vocabularies](#)

[1.8 National Széchényi Library's vocabularies](#)

[1.9 Library of Congress' vocabularies](#)

[1.10 German national Library' subject headings](#)

[1.11 French National Library's subject headings](#)

[1.12 VIAF person authorities](#)

[1.13 Wikipedia categories](#)

[1.14 New York Times subjects](#)

[1.15 IVOA astronomy vocabularies](#)

[1.16 NASA taxonomy](#)

[1.17 GEMET General Multilingual Environmental Thesaurus](#)

Navigation

- [Main Page](#)
- [Recent changes](#)
- [Tools](#)
- [Books](#)
- [Validators](#)

Other w3c resources

- [Activity news](#)
- [Publications](#)
- [Logos, buttons](#)
- [Activity home page](#)

w3c rss feeds

- [Activity newsfeed](#)
- [W3C blogs](#)
- [Use cases, case studies](#)



Search - the Data Hub

[Search](#)

67 datasets found

cablegate

US Embassy cables from Wikileaks

[torrent](#) [image/svg+xml](#) [application/rdf+xml](#) [api/sparql](#)

Not Openly Licensed

NHS Jargon RDF

Extracted data from the 'NHS Jargon Buster' site (<http://www.ic.nhs.uk/jargon-buster>). Contains definitions for commonly used acronyms in NHS (British National Health Service) documents.

[example/rdf+xml](#)[OPEN](#) [DATA](#)

Yleinen suomalainen asiasanasto - YSA

The General Finnish Thesaurus (YSA) contains general, commonly used terms. The thesaurus is widely used in Finnish libraries and other organisations.

[application/rdf+xml](#) [application/rdf+xml](#)[OPEN](#) [DATA](#)

ESD Standards: Controlled lists for the UK public sector

Standard controlled vocabularies used within the UK public sector.

[example/rdf+xml](#) [example/rdf+xml](#)[OPEN](#) [DATA](#)

Federal Reserve Economic Data RDF

This is an RDF snapshot of the Federal Reserve Bank of St. Louis's FRED2 database, which contains a variety of economic data.

[api/sparql](#) [meta/void](#) [example/rdf+xml](#)[OPEN](#) [DATA](#)

Tags

- [lod](#) (66)
- [format-skos](#) (66)
- [format-rdf](#) (48)
- [published-by-producer](#) (43)
- [format-dc](#) (35)

Resource Format

- [example/rdf+xml](#) (61)
- [api/sparql](#) (41)
- [application/rdf+xml](#) (26)
- [meta/void](#) (22)
- [meta/rdf-schema](#) (12)

Groups

- [LOD Cloud](#) (58)
- [Library Linked Data](#) (26)
- [Bibliographic Data](#) (11)
- [Linking Open Data](#) (5)
- [Civil Society](#) (4)

Other access

<http://thedatahub.org/dataset?q=format-skos>

Where to start from?

- What is already available as RDF data (Linked or not)
- What can be made available for Europeana by its partners
Survey in 2009 by HUB

	A	B	C	D
1	ID from vocs	Institution	Voc	Links Voc - Metadata
2	20	DISMARC	3 sample vocs in SKOS	yes
3	8	Fondazione Federico Zeri - Università di Bologna	txt	no
4	9	Slovak National Library	authority file in MARC	no links
5	10	Vlaamse Kunstcollectie (Flemish Art Collection)	AAT, RKD artist	
6	12	Musées de la Ville de Lausanne	xls	no
7	16	CulturaItalia, The Italian Culture Portal	http://www.culturaItalia.it/pico/thesaurus/4.1/thesaurus_4.1.0.skos.xml	yes
8	27	NATIONAL LIBRARY OF ESTONIA	MARC21	no
9	29	National Library of Poland	MARC	said application specific
10	31	Österreichische Mediathek	txt	yes
11	32	SCRAN	XML	yes
12	24	Istituto e Museo di Storia della Scienza	end of Sept - RDF	http
13	34	Slovak National Gallery	end of Sept - RDF	URN
14	14	Bibliothèque nationale de France	INTERMARC - in Sept	

Criteria for selection

- Languages covered

French	German	English	Portuguese	Hungarian
Italian	Spanish	Polish	Swedish	Dutch
Core languages			Secondary languages	

- Institutional/collection adequacy
- Usage adequacy (type of resources)
 - Europeana users are interested in places, topics and persons
- “Quality”
 - Conceptual grain and coverage
 - Lexical coverage
 - Completeness and correctness of the semantic relationships
- Licenses

Criteria for selection

- Languages covered

French	German	English	Portuguese	Hungarian
Italian	Spanish	Polish	Swedish	Dutch
Core languages			Secondary languages	

- What is there or what can be reasonably obtained and converted

But it's getting better!

And our project is also contributing...

Contents (still being updated)

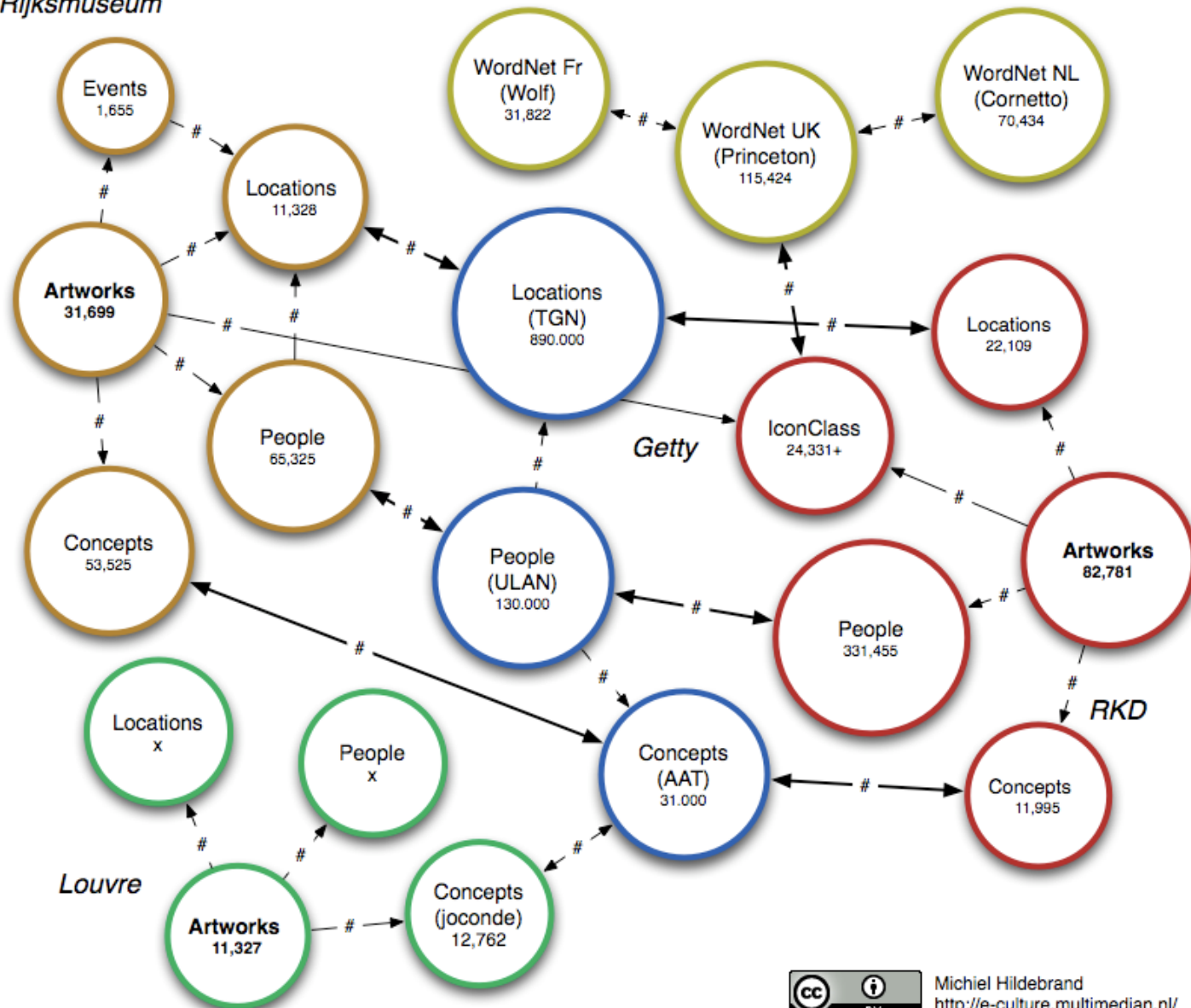
Places	Amalgame	Admin	Repository	Query	Help	Login	
------------------------	--------------------------	-----------------------	----------------------------	-----------------------	----------------------	-----------------------	--

SKOS concept schemes in the RDF store

Nr	Name	# Concepts	# prefLabels	# altLabels	# not mapped	# mapped	%	Example concept
1	Thesaurus PICO 4.1	817	1634	133	514	303	(37.09%)	PICO:who
2		50732	50732	0	0	50732	(100.00%)	[Unbesetzt]
3	GTAA/Classification	99	99	0	99	0	(0.00%)	gtaa:17B beeld/geluid
4	GTAA	171990	171990	6786	146214	25776	(14.99%)	gtaa:Franko, Ivan
5	GTAA/Genre	112	112	95	112	0	(0.00%)	gtaa:interactief programma
6	GTAA/GeographicalNames	14030	14030	224	5046	8984	(64.03%)	gtaa:Amsterdam-Zuidoost
7	GTAA/Maker	20070	20070	85	20068	2	(0.01%)	gtaa:2Tall
8	GTAA/Names	27646	27646	2690	27646	0	(0.00%)	gtaa:Commissie-Posthumus II
9	GTAA/Subjects	3932	3932	1407	3932	0	(0.00%)	gtaa:luisterboeken
10	GTAA/SubjectsSandV	4404	4404	1503	4404	0	(0.00%)	gtaa:sfeerbeelden
11	GTAA/PersonNames	101697	101697	782	84907	16790	(16.51%)	gtaa:Franko, Ivan
12	SCRAN	9841	9841	0	9841	0	(0.00%)	England
13	Musées de la Ville de Lausanne - despec	2027	2027	0	2027	0	(0.00%)	a deux anses
14	Musées de la Ville de Lausanne - fonc	160	160	0	160	0	(0.00%)	accessoire-habillement
15	Musées de la Ville de Lausanne - matgen	420	420	0	420	0	(0.00%)	acacia
16	Systematik der Österreichische Mediathek	497	498	99	497	0	(0.00%)	AMT:Politik, Aktuelles
17	Fondazione Federico Zeri - Photographic Materials and Techniques	29	29	0	29	0	(0.00%)	albumina / carte

Building a **network** of semantic references

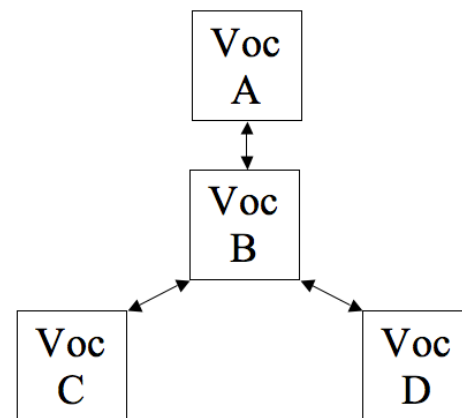
Rijksmuseum



Not everything can be aligned

- We can get links from external projects
MACS, AGROVOC...
- Sometimes in the vocabularies themselves
- But that still not does not cover everything

Pivot vocabularies?



- Looks like BS8723 backbone approach
- Hypothesis: a NKOS cloud would look the same
Not with strict pivots, but still high centrality

[...]

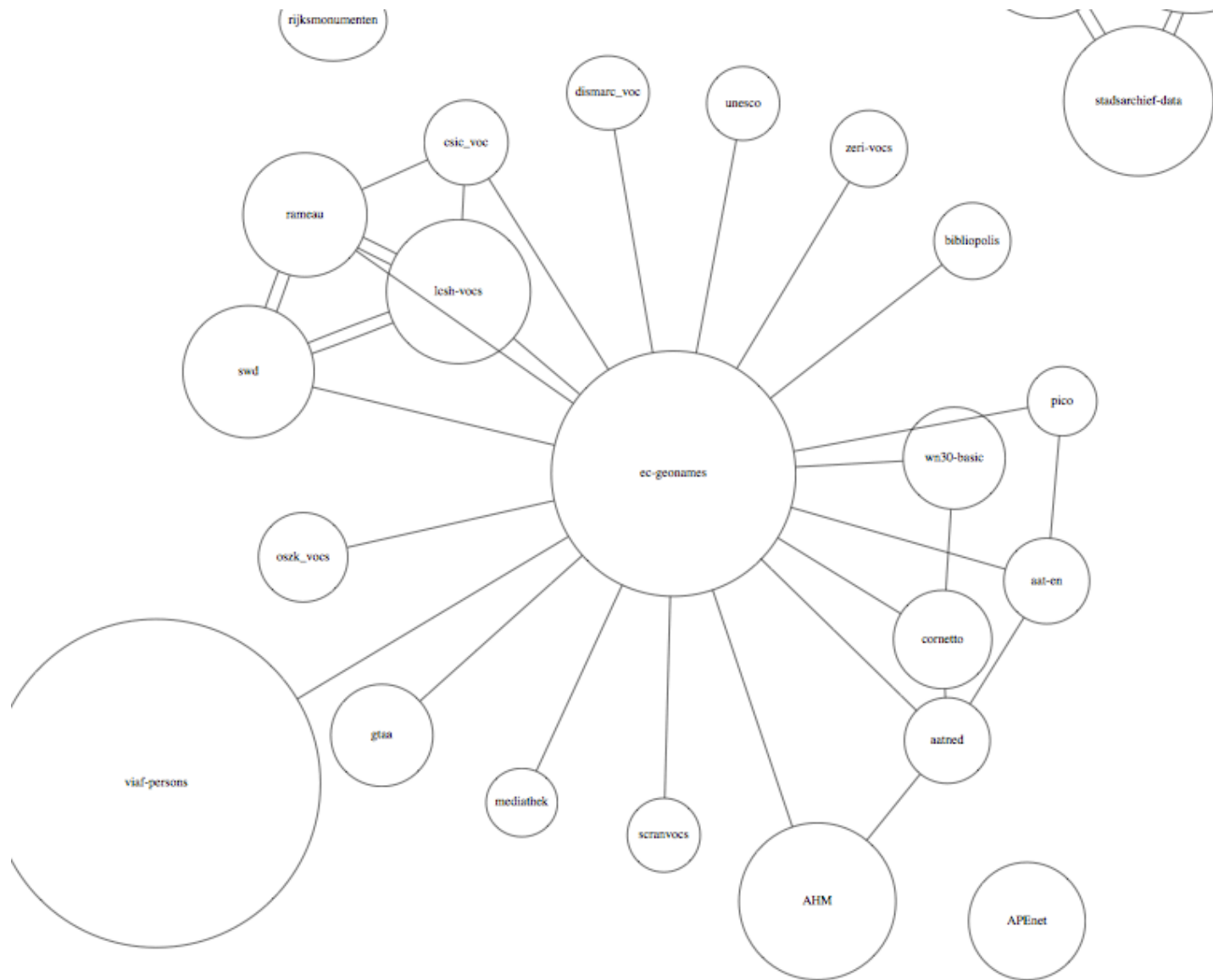
- EuropeanaConnect project will follow a (loose) pivot approach for Europeana's "semantic layer"



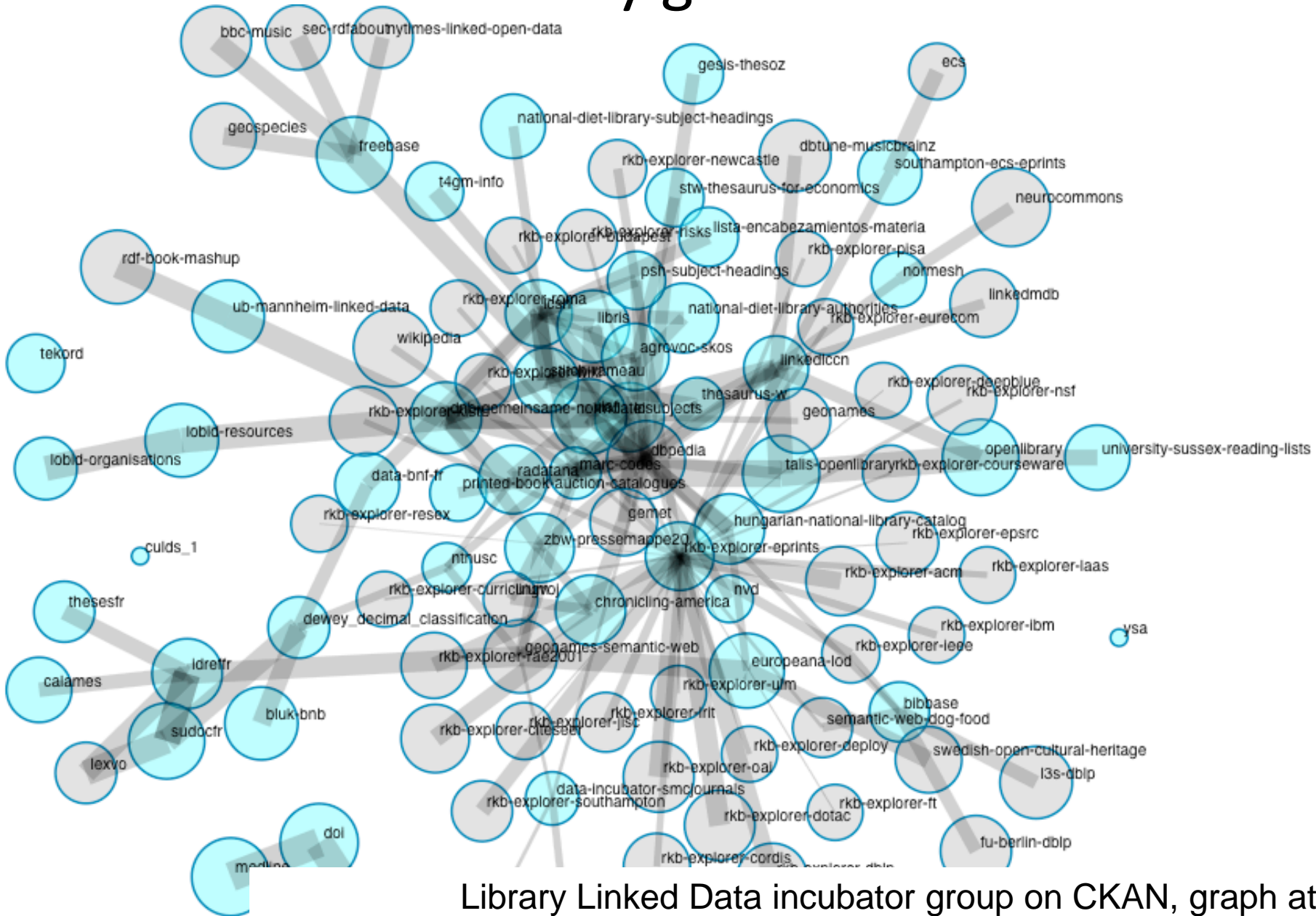
A “soft pivot” strategy?

- Matching smaller and specialized vocabularies to larger and more general pivot ones
 - Multi-lingual, wide-coverage and widely adopted vocabularies
- More specialized vocabularies may be used as anchoring points, depending on domains and existing alignments.
 - WordNet was retained as a general pivot for topics,
 - But LCSH is widely used in libraries and already (partially) mapped to other vocabularies of that sector.
- A pivot could be several vocabularies with comparable importance and complementary coverage.
 - If these vocabularies are well aligned together, it is possible to map one more specialized vocabulary to one or the other.

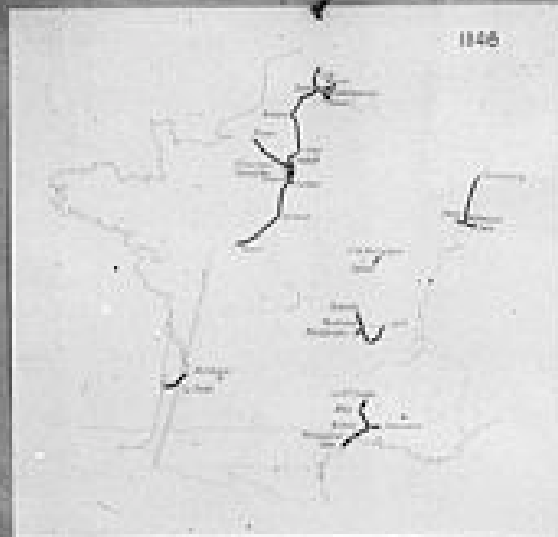
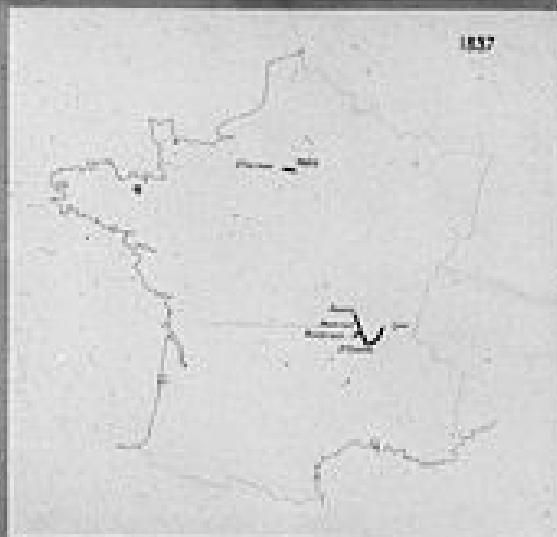
Datacloud in the making



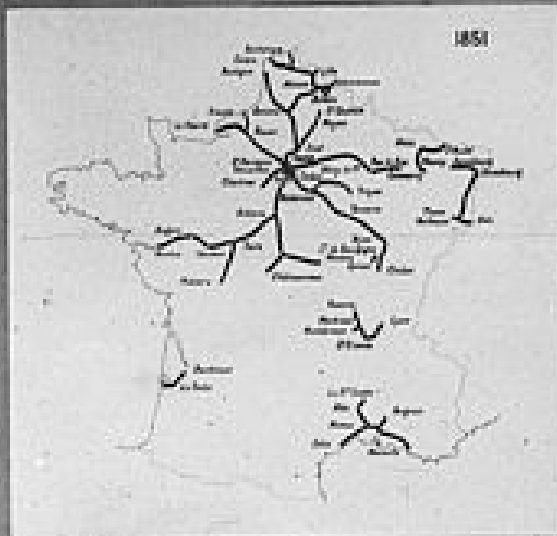
Community guidance?



Library Linked Data incubator group on CKAN, graph at <http://semantic.ckan.net/group/?group=http://ckan.net/group/lld>



Community strategy?



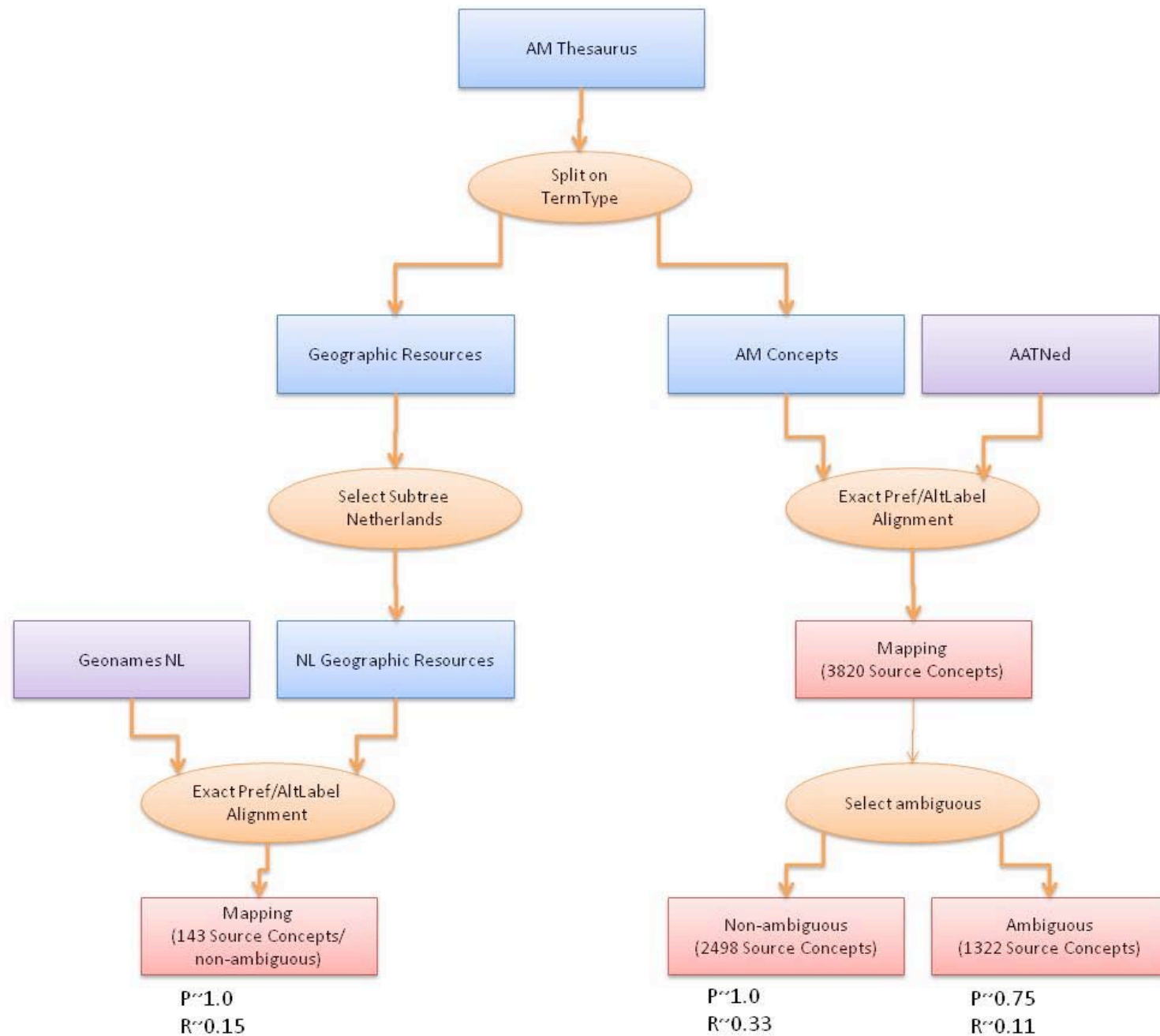
- Hmmmm, not there yet
- And maybe it's alright to have no strategy now?
- Anyway, we had to find something that would bring hope
E.g., an alignment tool to support a loose strategy, with trials and errors
Different from the less-agile state-of-the art ontology alignment tools

AMALGAME

Presented at TPDL

AMALGAME

- Do not try to make the tool smart
 - use simple atomic matching components
 - fast, scalable, with predictable and understandable results
- Empower domain experts by interactive iteration loop
 - **try** most promising component(s)
 - **analyze** the results
 - **decide** on next step:
 - discard results from last attempt
 - refine current results to improve precision
 - look for more results to improve recall



Thanks!

aisaac@few.vu.nl

Europeana.eu team

EuropeanaConnect WP1 @ HUB

Steffen Hennicke, Marlies Olenski, Stefan Gradmann

EuropeanaConnect project

<http://www.europeanaconnect.eu/>