

Towards Metadata Interoperability - from a data driven perspective -

Peter Wittenburg, Daan Broeder
Max-Planck-Institute for Psycholinguistics
Paul Buitelaar
DFKI

peter.wittenburg@mpi.nl
www.mpi.nl

- had to find practical solutions for metadata interoperability
 - INTegrated European language Resource Area (INTERA)
 - European Cultural Heritage Online (ECHO)
- content
 - introduction to the tasks
 - explanations of the problems
 - discussion of the solution
 - few words about XML vs RDF
 - some conclusions
- report about data driven work
- at the beginning lots of questions – still no answers yet
- theoretically nothing new

- increasingly large collections of language resources etc
 - at MPI 8 TB in ~ 30.000 sessions (~ 100.000 linguistic objects)
 - the IMDI domain ~ 50.000 sessions and lexicons
- want to virtually integrate these collections
 - form one searchable, browsable, manageable domain
- INTERA
 - integrate several European LR centers (in total about 40) (ELDA, BAS, MPI, PA Vienna, U Helsinki, U Lund, U Stavanger, U Uppsala, ATILF, Meertens, INL, U Stockholm, U Hamburg, ...)
 - all based on IMDI
 - gateway to OLAC/DC communities (now script)
- ECHO
 - 5 different disciplines (HoA, HoS, Ethnology, Linguistics, Philosophy)
 - 9 different repositories
 - different semantic scopes and type of concepts
 - create an interoperable metadata domain
 - full text and selective search on metadata

<http://corpus1.mpi.nl/ds/dora>

Dora - version .52.01 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

Address <http://corpus1.mpi.nl/ds/dora/DoraLayoutAction.do?active=NECEP> Go Links >>

Digital Open Resource Area

The Echo Portal

[V] Ethnology [V] History of Art [V] History of Science [V] Languages [V] Philosophy

Please Select Views To Build Query

RMV

[X] NECEP

society name

alternative name

language name

country

continent

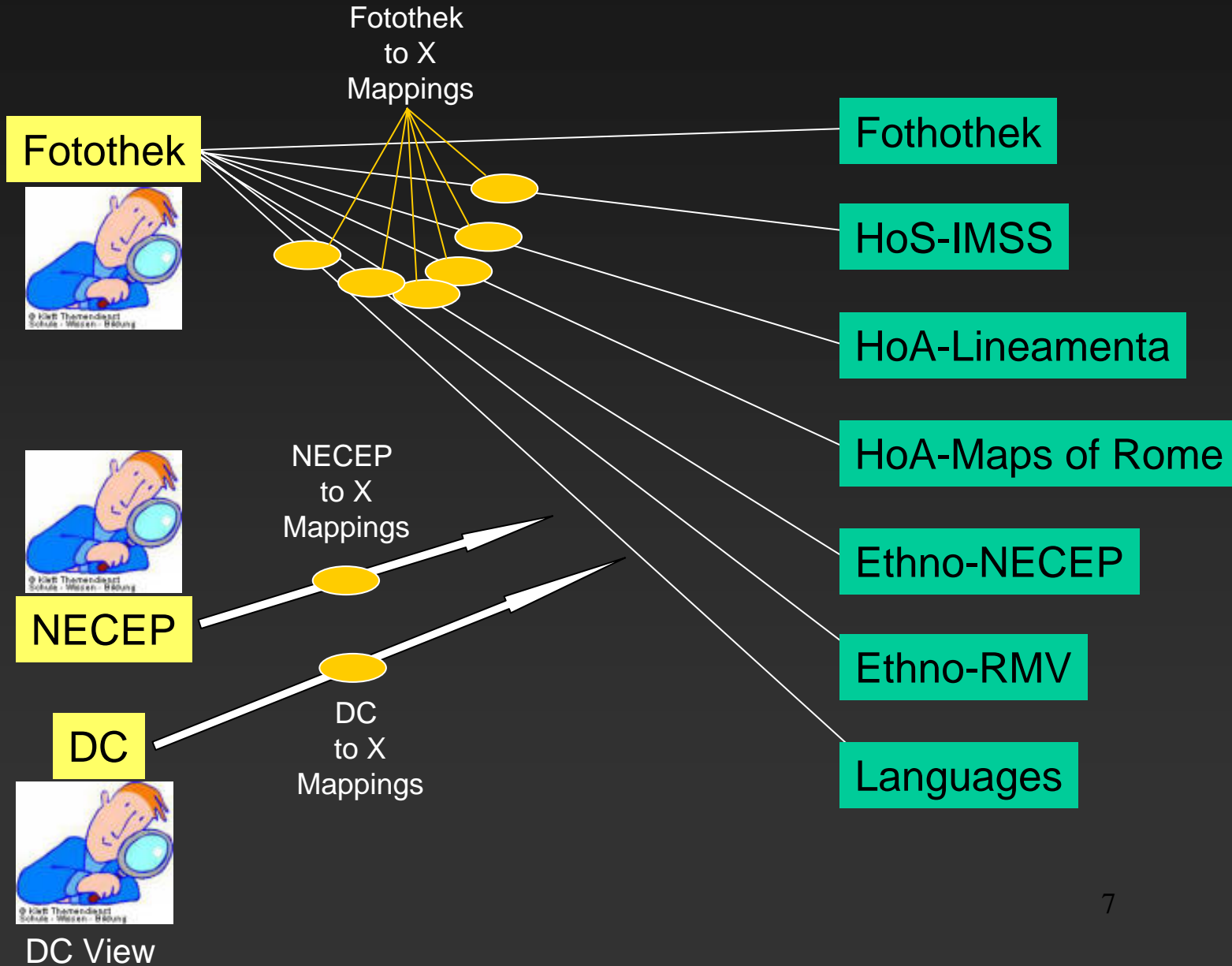
ethnic region

Simple Search

Done Internet

Domain	Exp Nr. Records	Nr Records	Set	Thesauri	Harvesting Type
Fotothek	very large	~ 65.000	MIDAS exhaustive	IconClass	XML -offline
Lineamenta	small	~ 140	new – DC like	-	XML -offline
Maps of Rome	small	~ 300	self-defined	-	XML -offline
HoS various	pot large	-	new - DC like	-	XML -offline
IMSS	large	~ 1.000	DC	-	OAI
NECEP	small	11	new - exhaustive	-	XML - online
RMV	very large	1000	OMV	OMV (AAT) GeoOMV	OAI
Language	large	~ 30.000	IMDI	Geo	XML online OAI
Philosophy	small	~ 40	self defined	-	XML - online
		> 100.000			

- Harvesting
 - OAI not yet standard – responsibility at provider side
 - configuration driven harvesting only partly applicable
- Encoding
 - UTF-8 specification not true
 - etc
- Structure
 - extraction from databases – no validation
 - not well-formed XML
 - continuous changes
- Semantic
 - remaining slides



- different “dimensions” and vocabularies along them
- one typical example

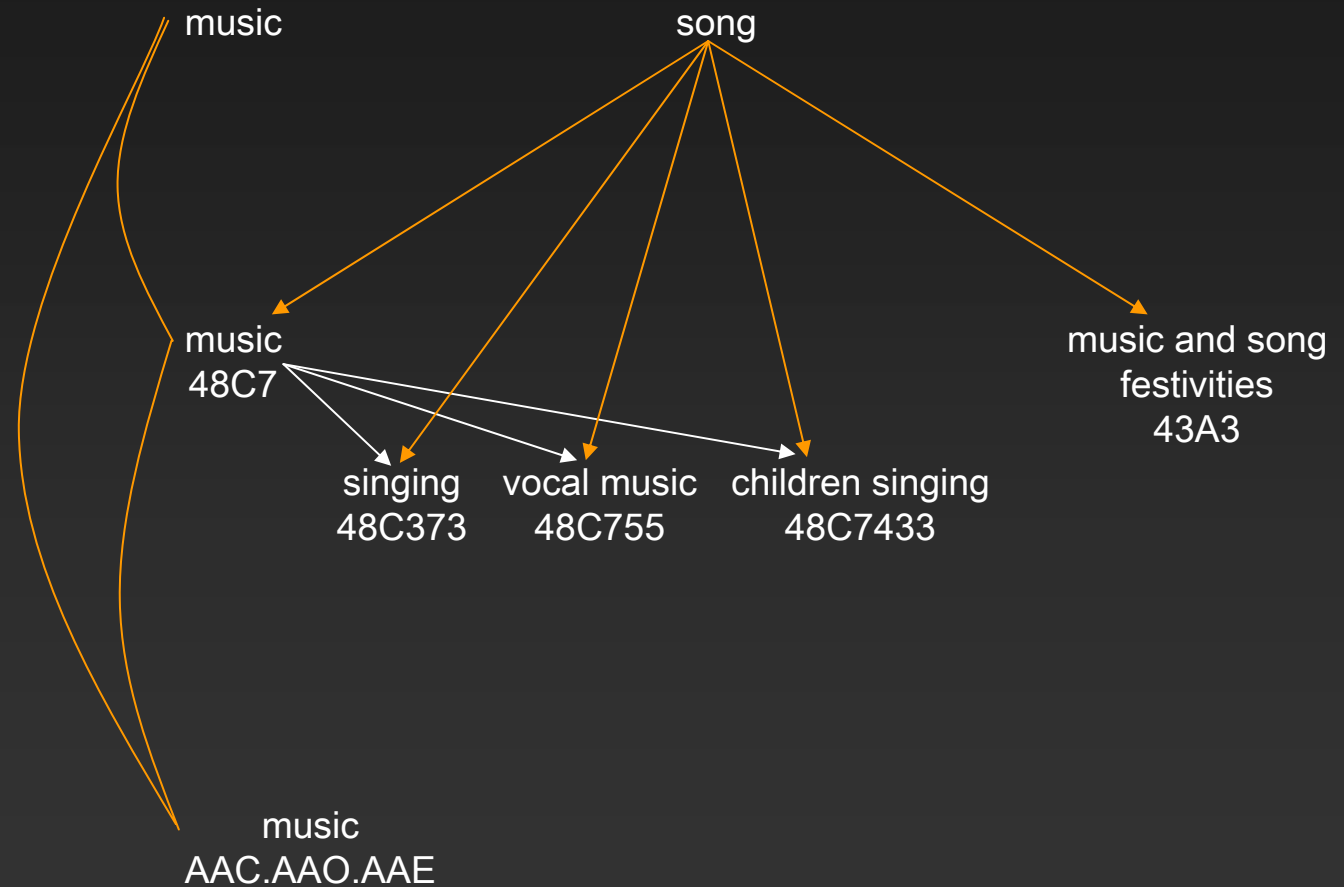
HoA	IMDI
Name artist	Participant
Date	Date
Period	Date
Location of creation	Continent/country/region
Object title	Title
Title of building	Title
Object type	Content
Prim iconography	Content
Sec iconography	Content
Place of content	Continent/country/region

- location of creation according to which vocabulary?
- Object type ?= content
- Iconography according to the IconClass thesaurus (complex system to associate content keywords) vs. Content (task, subject, genre, modality)

Others

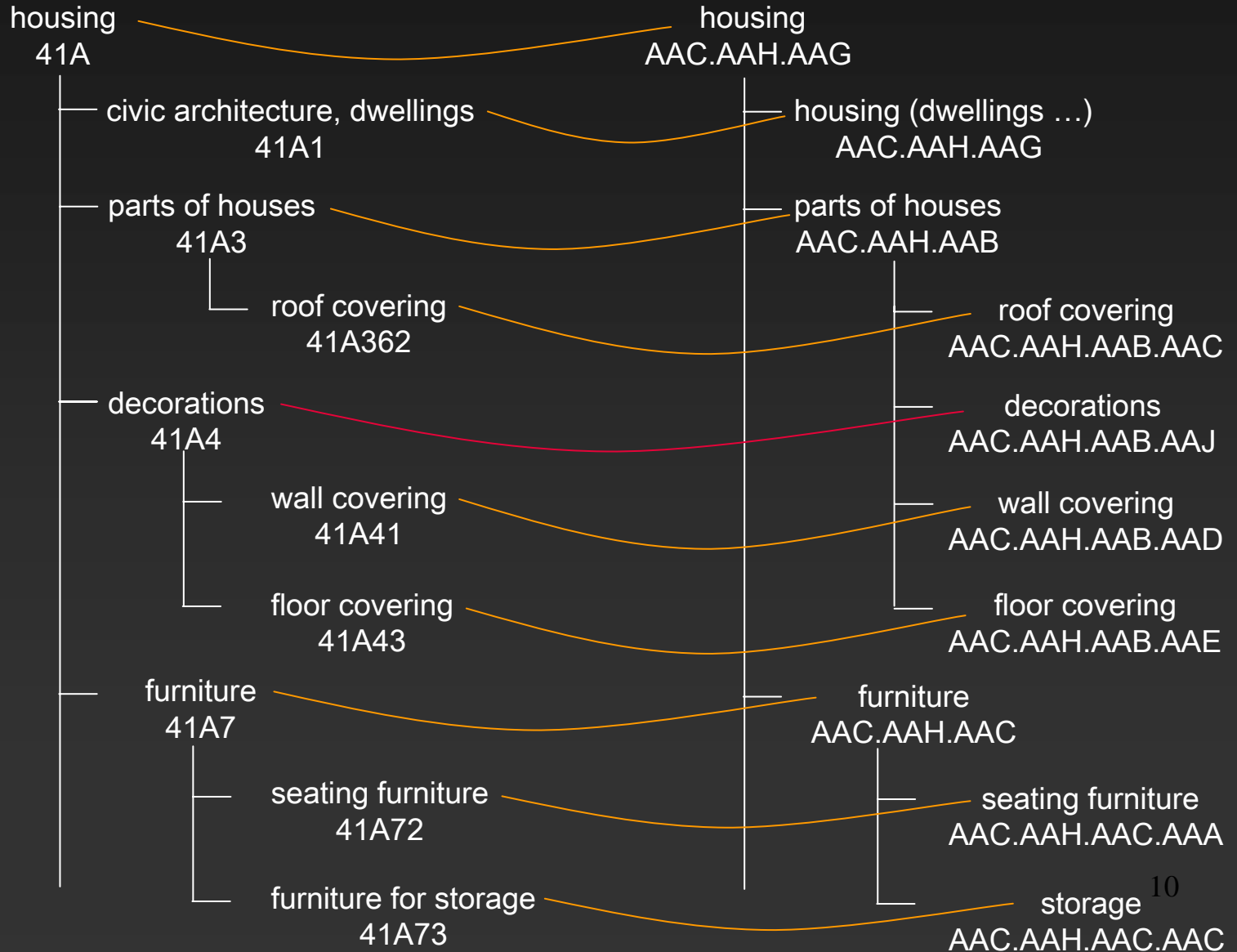
IconClass

OVM



IconClass

OVM



- one directional usage only
- 4 types of mappings
 - isEqualTo synonym
Example: “HoS:language” isEqualTo “NECEP:language name”
 - isSubclassOf
Example: “RMV:coverage spatial” is SubclassOf “IMDI:continent”
 - isSuperclassOf inversion
 - mapsTo fuzzy type of relation
logically not exploitable
Examples: “IMDI:Genre” mapsTo “Fotothek:Iconography”
“IMDI:Task” mapsTo “Fotothek:Iconography”
“IMDI:Subject” mapsTo “Fotothek:Iconography”
“RMV:Category” mapsTo “Fotothek:Iconography”

Example 1

Simple Search “dogon”

1 match was found: NECEP: 1

Complex Search “dogon”

View NECEP - society name: 1 in NECEP

View IMSS - language: 1 in NECEP

View DC - language: 1 in NECEP

View Language - language: 1 in NECEP

Complex Search “mali”

View Language - country: 1 in NECEP

language mapsTo society name
Dogon isSubClassOf Mali

Example 1

Simple Search “agriculture”

75 matches are found: Language: 73, Fotothek: 2

Complex Search “agriculture”

View Fotothek - iconography: 2 in Fotothek

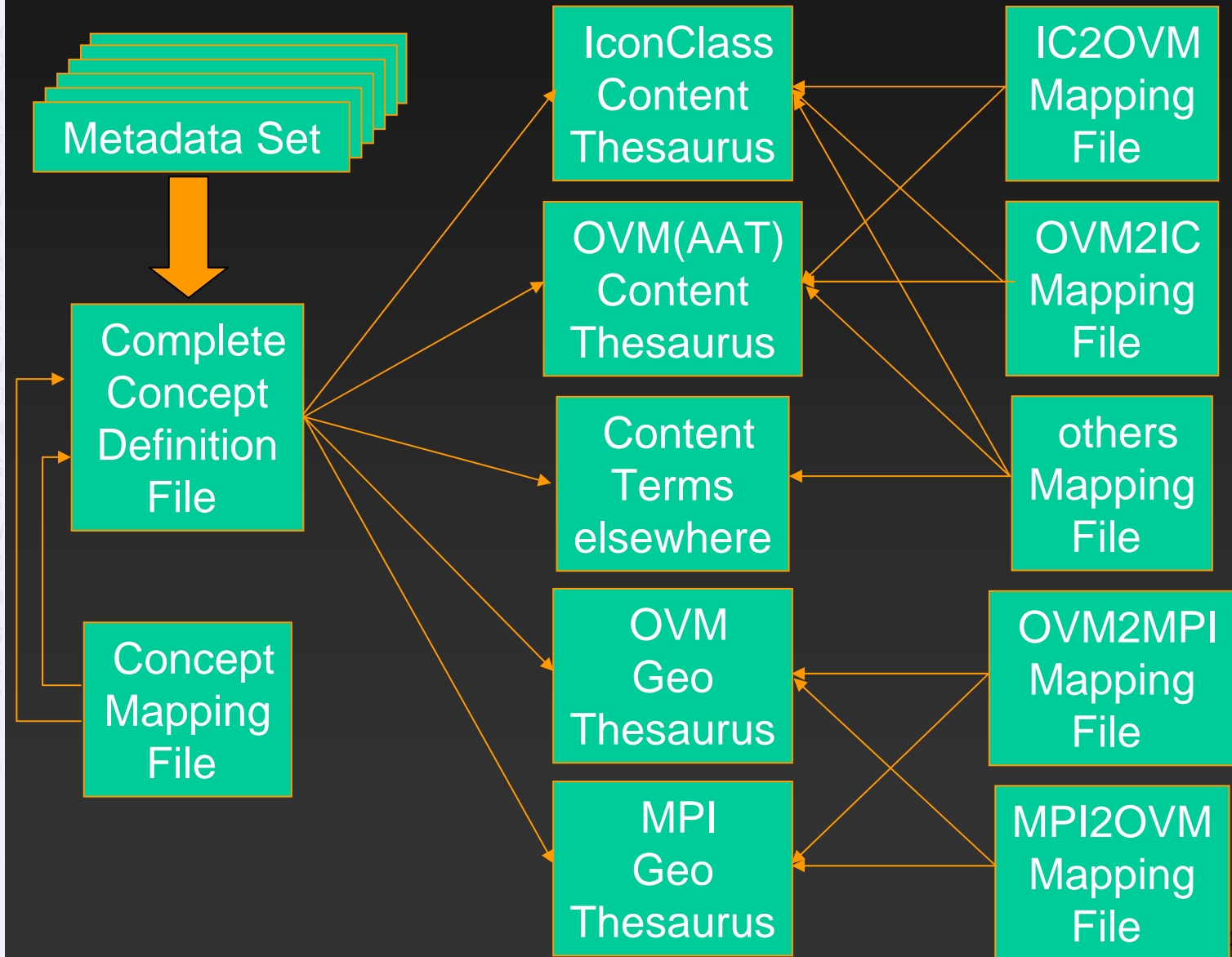
View RMV – content: 2 in Fotothek

View IMDI – content: 2 in Fotothek

Recording Place = Southern Agricultural Kindergarten

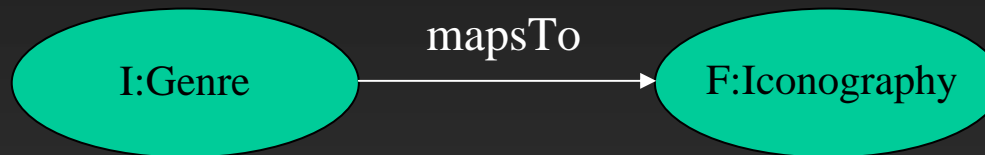
Actor working at Ministry of Agriculture

Harvesting is node in Agriculture Subtree



- miss standards for the registration of concepts
 - ISO TC37/SC4 Data Category Registry coming for LR
 - ISO 11179 and ISO 12620 compliant
 - formulated in XML – why? ask Laurent 😊
 - contains hierarchical relations as part of concept definitions
 - IMDI and OLAC definitions to be included
 - allows reference in schemas to DCR entries
 - allows people to create their own MD sets without losing interoperability
- their will be “personal” DCRs since people disagree and science needs flexibility → how to prevent proliferation?
- yet no infrastructure to register registries

- miss standards for the registration of relations
 - on the way to defined relation types (RDF(S) – OWL)
 - little agreement on relations (ECHO)
 - intention-dependent relations (optimal data mining)



- “transitive verb” isSubClassOf “verb” is different
- maintain relations outside of concept definitions – of course
- emergence of many “practical ontologies”
 - yet no registry standards (or?)
 - where to find useful practical ontologies?
 - how to combine practical ontologies?
 - how to resolve conflicts?

- in our projects XML is chosen to represent all knowledge for short term arguments
- special search engine makes use of all structural information
- ISO has chosen XML as representation format for DCRs
 - XML sufficient since all information is of structural type
 - allows to represent trees and references
 - references are structural not semantic
- RDF(S) and OWL come with language to describe semantics
- RDF designed to represent any type of relation
- simple RDF assertions allow machines to combine relations from different repositories
- for us no direct use due to project specific approach and have fuzzy relation (mapsTo)
- but relations not easy to modify, to adapt and to combine

- one reviewer argued that problems are well-known and solved
- for data-driven people like us almost nothing is solved
- need well-established standards and ready to use infrastructures
 - for describing concepts (ISO DCR is excellent start)
 - for describing relations (RDF(S) + OWL, what with mapsTo)
 - for registering DCRs and Relation Registries
 - for selecting and combining such components
- our repositories and the results of all intensive work will not be directly re-usable
- still interdisciplinary work as in ECHO is difficult

XML:

Surface syntax, no semantics

XML Schema:

Describes structure of XML documents

RDF:

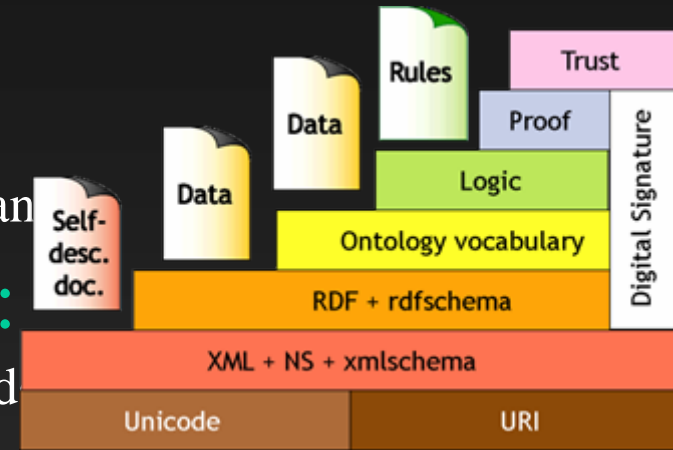
Datamodel for “relations” between “things”

RDF Schema:

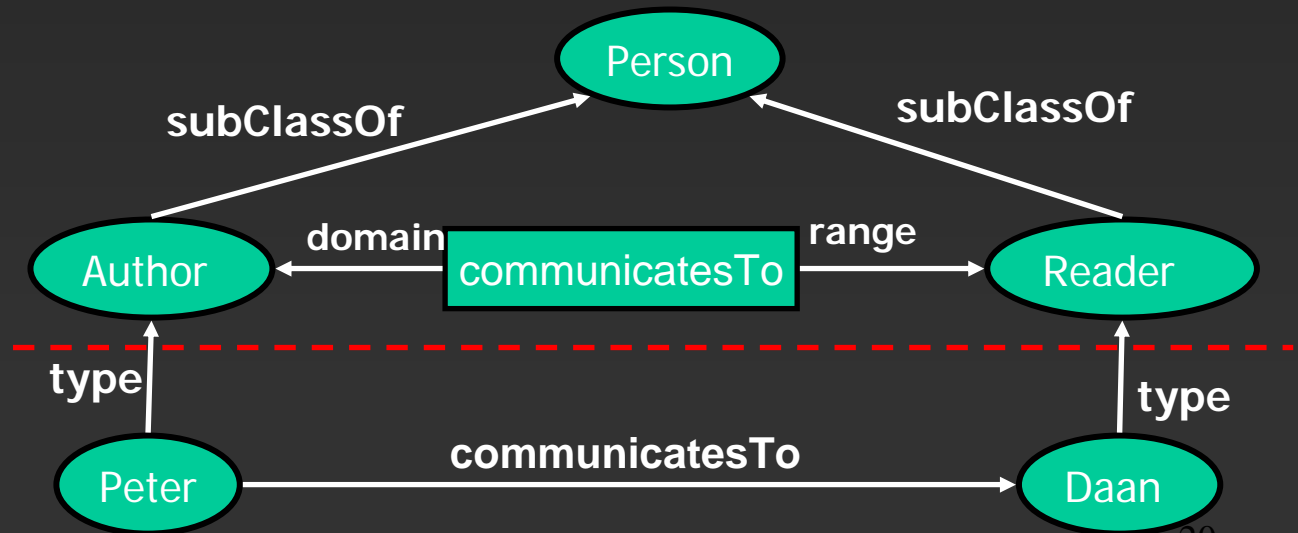
RDF Vocabulary Definition Language

OWL:

A more expressive
Vocabulary Definition Language



- Defines **vocabulary** for RDF
- Organizes this vocabulary in a **typed hierarchy**
 - Class, subClassOf, type
 - Property, subPropertyOf
 - domain, range



OWL Lite:

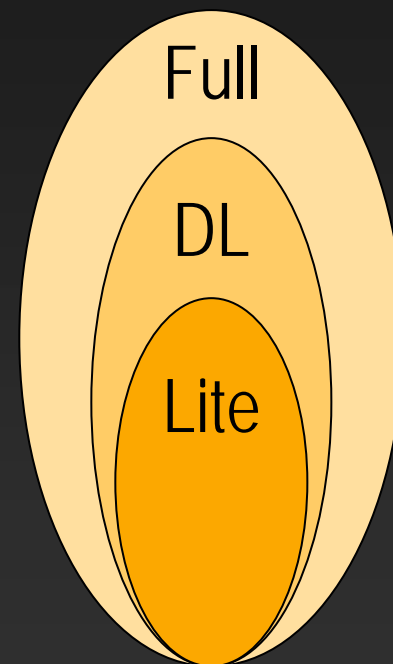
Classification hierarchy
Simple constraints

OWL DL:

Maximal expressiveness
While maintaining tractability
Standard formalisation

OWL Full:

Very high expressiveness
Loosing tractability
Non-standard formalisation
All syntactic freedom of RDF
(self-modifying)



Syntactic layering
Semantic layering

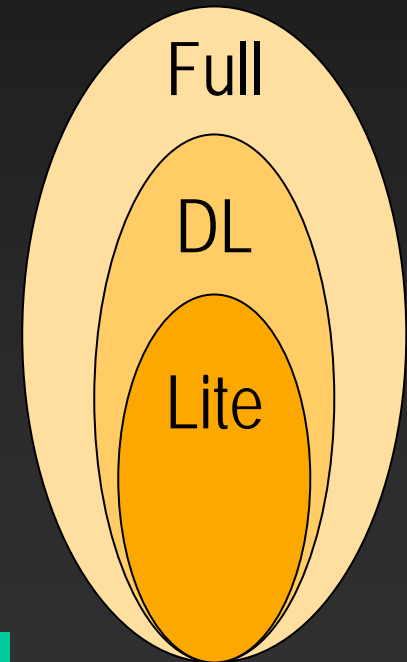
OWL Light

- (sub)classes, individuals
- (sub)properties, domain, range
- conjunction
- (in)equality
- cardinality 0/1
- datatypes
- inverse, transitive, symmetric
- hasValue
- someValuesFrom
- allValuesFrom

OWL DL

- Negation
- Disjunction
- Full Cardinality
- Enumerated types

RDF Schema



OWL Full

- Allow meta-classes etc