

Metadata Initiative for LR

State and Perspectives

Peter Wittenburg, Michael Eichberg
Max-Planck-Institute for Psycholinguistics
Bibliotheca Hertziana

Content

- Background
- Coming Challenges
- IMDI Metadata
- IMDI Content Elements
- IMDI for Lexica
- INTERA Requirements
- ECHO Requirements
- Metadata and Web-Services
- Requirements for ISO TC37/SC4

Metadata = Catalogue type of MD describing LR

Background

- 2000 - 2002 ISLE Project (Int. Standards for Language Engineering)
 - one part was IMDI (ISLE Metadata Initiative)
 - motivation well-known (amount and complexity of resources)
 - chaos in institutions - limited reusability
- LREC 2000 - Athens: 1st Workshop on Metadata for LR
 - presentation of a White Paper - later a broad overview
- wanted to combine LE and FL (no idea about type of differences)
 - one browsable & searchable Language Resource domain
- goal: within 2 years an operational infrastructure
 - metadata set, controlled vocabularies, editor, browser, bridge
 - framework for all linguistic datatypes
 - strategy: bottom up, i.e. understand community requirements
 - so: many discussions with LE and FL
- now version 3.x; distributed repository with ~ 15.000 objects, 3000 h

Metadata for LR still relatively new - still to convince / still dynamics

Future Challenges

- new projects funded by EC and others
 - DOBES: 20 internat. teams documenting endangered languages
 - INTERA (Integrated European language Resource Area)
 - integration of holdings of European data centers and more
 - integration of resource and tool repositories
 - *nice* formulation (Data Categories, RDF, open repositories)
 - ECHO (European Cultural Heritage Online)
 - complicated project covering networks, content, technology
 - 4 disciplines now (arts hist, science hist, ethnology, languages)
 - goal: common technological framework for CH type of data
 - many scientists in ECHO not familiar with concept of MD
 - strong pillars: institutes with high ambitions
 - Virtual Wisdom
 - *corporate memory* for Max-Planck-Society
 - 85 institutes operating in different disciplines - complex semantics
 - multidisciplinary integration & interoperability; persistence

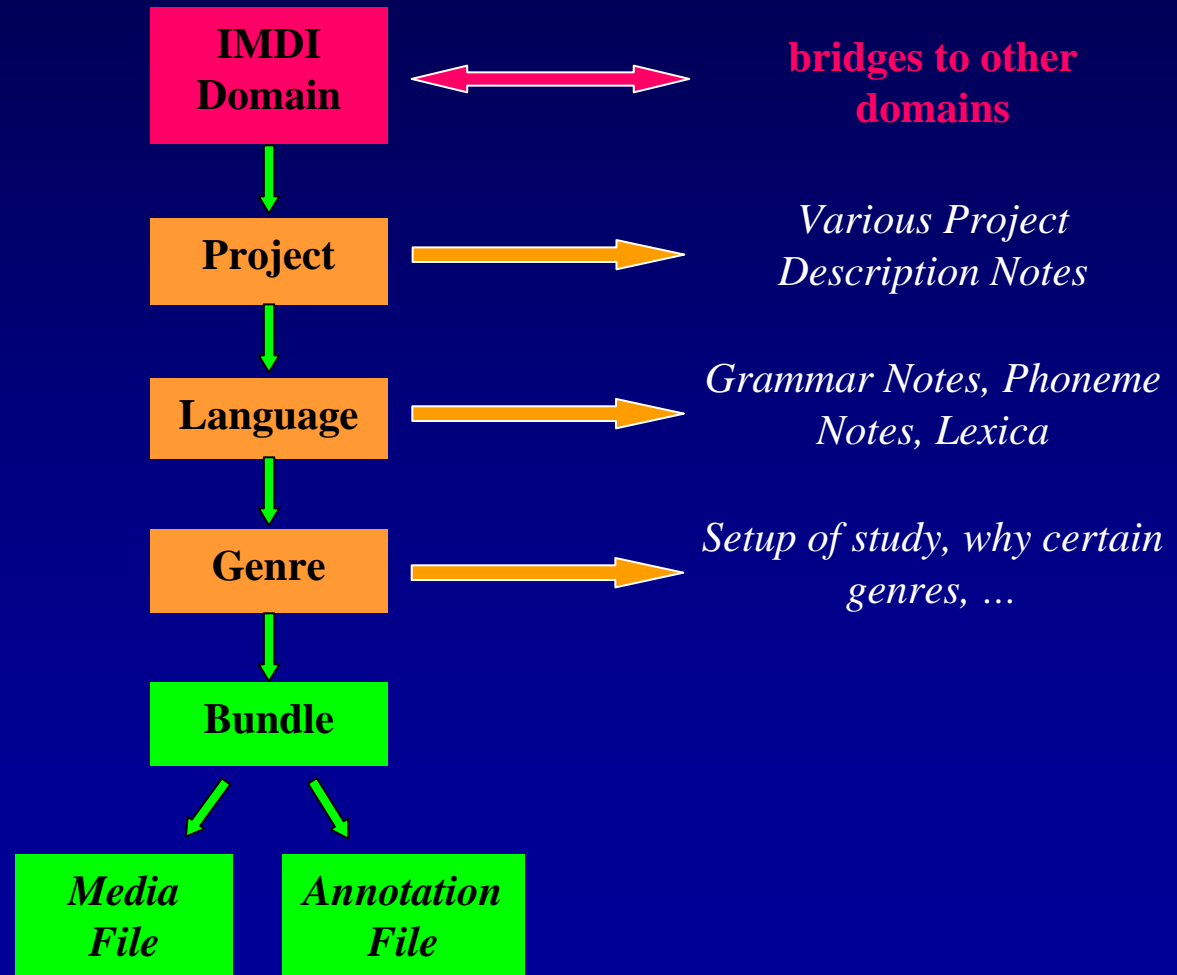
our MD work will remain project driven
need to retain information - no pidginization

IMDI Keypoints

- Canonical domain organization, but users can define own org

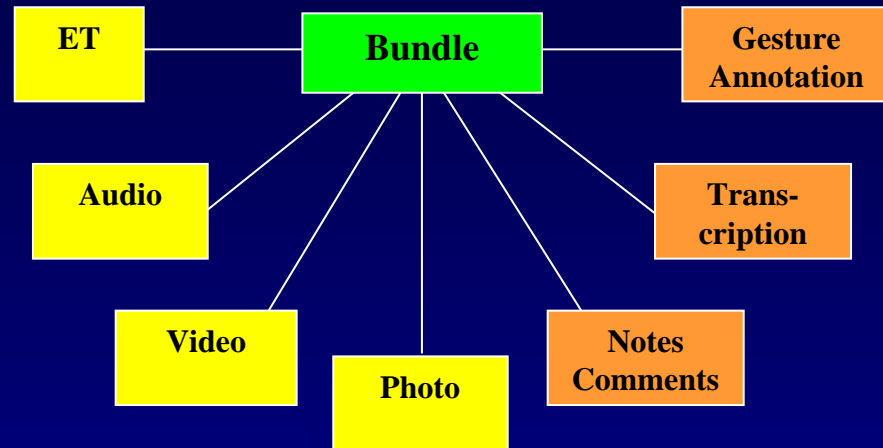
- browsable & searchable domain
- fully distributed
- local support
- manageable structure
- all based on XML
- all open
- immediate start of tools
- simple registration

- editor
- browser
- search
- management scripts



IMDI Keypoints

- Concept of Bundling
 - bundle is glue
 - several annotation files
 - tools must know relations



- Type of queries

Give me all resources that include a female speaker of 6 years old speaking Yaminyung and where an English translation exists.

Give me all resources from the SmartCom project, that includes speech and gestures, where the topic was "Route Direction" and where a transcript and a gesture annotation exists.

General IMDI Structure

Bundle	(Name, Title, Date)
Location	(Continent, Country, Region, Address, <u>Description+</u> , <u>Keys</u>)
Project	(Name, Title, ID, <u>Contact</u> , <u>Description+</u> , <u>Keys</u>)
Content	(StructureGenre, ContentGenre, CommContext, Task, Subject, Modalities, <u>Description+</u> , <u>Keys</u>)
Languages	(<u>Language+</u> , <u>Description+</u>)
Actors	(<u>Description</u>)
Actor+	(Type, Name+, FullName, Code, Role, <u>Language+</u> , EthnicGroup, Age, Sex, Education, <u>Description+</u> , <u>Keys</u>)
Resources	
MediaFile+	(ResourceLink, Size, Type, Format, Quality, RecordingCondition, Position, <u>Access</u> , <u>Description+</u> , <u>Keys</u>)
AnnotationUnit+	(ResourceLink, Annotator, Date, Type, Format, ContentEncoding, CharacterEncoding, <u>Access</u> , LanguageID, <u>Description</u>)
WrittenR+	(ResourceLink, Type, SubType, Format, Size, Validation, ContentEncoding, CharacterEncoding, <u>Access</u> , LanguageID, <u>Description</u> , <u>Keys</u>)
LexiconR+	(LexicalEntry, MetaLanguages, Format, Size, ...)
Source+	(ID, Format, Quality, Position, <u>Access</u> , <u>Description+</u>)
References	(<u>Description+</u>)

IMDI Editor

The screenshot shows the 'IMDI Editor' window with the following sections:

- General Content Data:**
 - Description: EPISODE: 06 Stage Direction; the ashtray experiment ;
 - Keys: A table with columns 'Name' and 'Value'. The first row contains 'KEYWORDS' and 'EXPERIMENT, ASHTRAY/ LEXICON, REFERENCE TO SPACE/'.
 - Buttons: 'Add a Key'.
 - Infofile: 'ladhk23.prt' with a 'Details' button.
- Languages:**
 - 1. Name: Dutch, Description: Target Langage, Infofile: 'Details' button.
 - 2. Name: Arabic, Description: Source Langage, Infofile: 'Details' button.
 - 3. Name: (empty), Infofile: 'Details' button.
 - Button: 'Add a language...'.

- Java
- CV update
- CV caching
- constraints
- XML generation
- sub-blocks
- continuously optimized
- download - JNLP
- free usage
- OpenSource

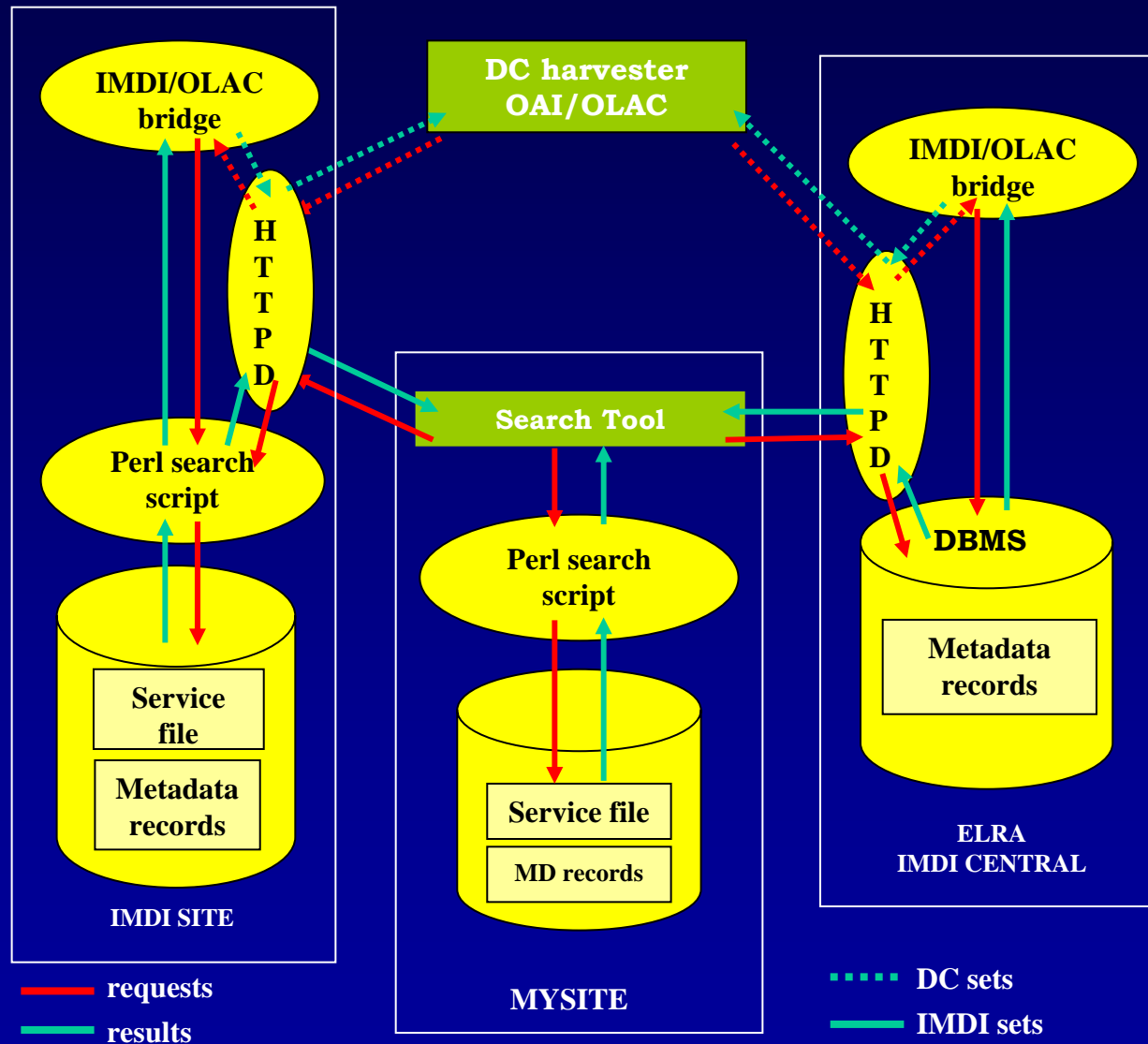
IMDI Browser

The screenshot displays the IMDI Browser interface. On the left, there is a 'Browser Action' panel with buttons for 'Exit', 'Options', 'About', and 'MD Search'. Below it, the 'History' panel lists recent actions: 'MPI corpora (UNIX)', 'World map', 'MPI Corpora (HTTP)', and 'Search results'. The 'Status' is 'Ok'. The main area is divided into two sections: 'Meta Descriptions' and 'Info/Content'. The 'Meta Descriptions' section shows a hierarchical tree view of metadata, with the root URL 'http://www-server.mpi.nl/topics/BC/mpi-unix.imdi' at the bottom. The 'Info/Content' section displays 'Participant information' for a performer named 'LC', including details like 'Type: Performer (shaman)', 'Sex: m', 'Age: ca. 70', and 'EthnicGroup: Mayan'. It also lists 'Additional keys' and 'Languages Info'.

- Java
- XML processing
- bookmarks
- info file rendering
- MD search
- continuously optimized
- node creation to come
- **immediate tool start - own tools**

download - JNLP
free usage

Search Infrastructure



Content Description

Languages		the languages the resource is about
Description+		descriptions of the languages assembly as a whole
Language+	ccv	substructure to include as many languages as needed
Description+		substructure to include descriptions of the content part as a whole
Keys+		substructure to define private/project specific key-value pairs (CGN, Management)
Task	ov	typical name for the task the subject has to carry out
Modalities	ovl	the modalities that are included in the recording
Communication Context		linguistic features concerning the recording context
Interactivity	ccv	degree of interactivity between participants
Planning Type	ccv	degree of planning by the consultant
Involvement	ccv	degree of involvement of researcher
Channel	ov	Face-to-face, experimental, broadcasting, telephone, Human-machine-dialogue, other
Social Context	ov	Family, Private, Public, ControlledEnvironment, other
Structure	ov	Monologue, Dialogue, Conversation, other
StructuralGenre	ov	Literature, Poetry, Song, ...
ContentGenre	ov	Narrative, Procedural, Oratory, ...

currently in discussion

Basics for Lexicon MD

- no changes to first proposal until now (need experience with set)
- lexica appear in different flavors (wordlists, dictionaries, ...)
- restrict at first instance to those including headwords and descriptions
- no concept oriented lexical databases (thesauri, ontologies)
- characteristic: no unified structure, very heterogeneous domain (language, theory)
- therefore in L-MD no detailed structure dependent description
- content by the major lexicon categories and indication of included information
 - examples Orthography Spelling, Syllabification, Hyphenation
 Morphosyntax POS, Inflection, Gender
 - no information how POS is encoded and how it is embedded
- for detailed information the schema could be associated with the MD

Re-used Blocks

Administrative Info (if possible conform to other resources)

- Name no change
- Title no change
- Date just one date,
others in history description
- Location no change
- Project (Name, ID, Contact, Description) no change
- Version as possible key-value pair or in version description

Content Info (if possible conform to other resources)

- Languages (Description, Language +) languages that are included as
subject languages (not comment, etc)
- CommCont n.a.
- Genre n.a. special domain lexica (kinship wordlist)?
- Subgenre n.a. ?
- Subject n.a. new, topic of resource
- Modality useful for sign language, gestures, ...
- Task n.a. only useful for corpora?

Re-used Blocks

Actor Info (if possible conform to other resources, to be used for creator etc)

• Type	creator, validator, contributor, ...
• Name	n.a.
• FullName	no change
• Code	n.a.
• Role	n.a.
• Language	could be used
• EthnicGroup	n.a.
• Age	n.a.
• Sex	n.a.
• Education	n.a.

Editor should not show not applicable fields of course.

Lexicon Blocks

Lexicon Info I

- LexicalEntry
 - HeadwordType Sentence, Phrase, Wordform, Lemma, ...
 - Orthography Hyphenated Spelling, Syllabified Spelling, ...
 - Morphology Stem, Stem Allomorphy, Segmentation, ...
 - Morphosyntax POS, Inflection, Countability, ...
 - Syntax Complementation, Alternation, Modification, ...
 - Phonology Transcription, IPA Transcription, CV pattern, ...
 - Semantics Sense distinction, Ontological classification, ...
 - Etymology
 - Usage Region, Style, ...
 - Frequency

(vocabularies to be open to offer flexibility at least in the first years)

Lexicon Blocks

Lexicon Info II

- MetaLanguages
- Format
- Size
- NoEntries
- CharacterEncoding+
- SchemaRef
- Access
- Keys
- Description+

Languages used in comments,
sense definitions, ...

indication of format (text/x-sbx)

size in bytes

size in no of entries

special fonts used etc

reference to the schema of the lexicon

will cover history, version, media included, ...

INTERA

- *nice* formulation of all
 - describe all terms used (element & vocabulary names) as DatCats
 - integrate them into open repositories (?)
 - add designations in various languages
 - re-use where possible existing DatCats (?)
 - define sub-blocks explicitly (complex DatCats)
 - describe all relations with RDF and make RDF descriptions also open
 - mapping to DC also to be described by RDF
- integration of Resource and Tool Repositories (TR)
 - once found a (number of) resource user wants to execute operation in distributed scenario
 - browser sends application profile to TR (resources, environment, function)
 - TR matches profile with tool (metadata) descriptions
 - TR responds with tool list (names, references, descriptions, advice)
 - user selects a tool and wants to execute it

TR Service to be implemented as Webservice (re-usage)

ECHO

- interoperability between 4 disciplines at MD (and resource) level
 - query: find resources about Venetian ship building (arts, science, language)
 - not interested in MD pidginization (lose all interesting data)
 - what else?
 - make MD schemas explicit and understand usage
 - start with individual mappings and apply them
 - offer different views at query interface
 - later same as in INTERA: define DatCats, make them open & make all relations explicit with RDF
 - in fact creation of practical ontologies (by a few)
 - not at all clear what kind of relations are needed
- others will never agree
 - so:
 - offer all details
 - let people define their own service by re-using & modifying

ECHO Example

- interop layer between History of Arts and Languages
 - query: find resources about Venetian ship building (arts, language)
 - one side: IMDI records as described
 - History of Arts: 186.000 objects linking to 320.000 online photos
 - description according to MIDAS (ICONCLASS) rules
 - all descriptions in hierarchical HIDA database
 - **are in the phase of *fishing in the dark* / MD made for (sub) community**

Object (Building, Type, Category, ...)

Date

Period

Location

Title

ICONCLASS

LocalRelation

Building (Location, Category, ...)

Creator (Name, ...)

Relation (Link, Creator, ...)

Photo+ (Nr, Owner, Date, Person,

ResourceLink, View, ...)

Bundle

Date

Date / Description / Keys

Location.xxx / Content.Language

Name / Title

Content.Subject / .Language / .Keys / .Description

Location.xxx

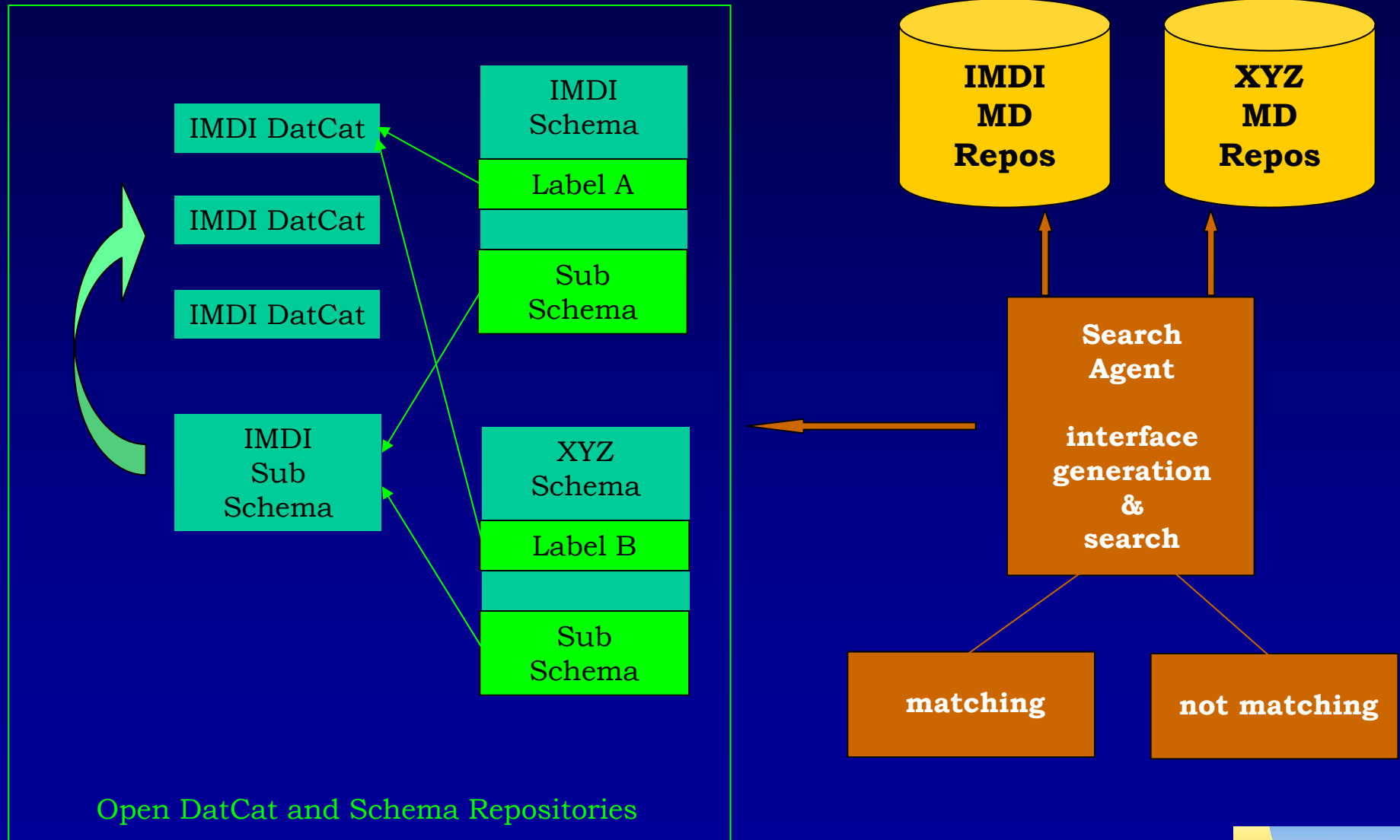
not relevant

Actor.Type=Creator (Name, ...) *mostly not relevant*

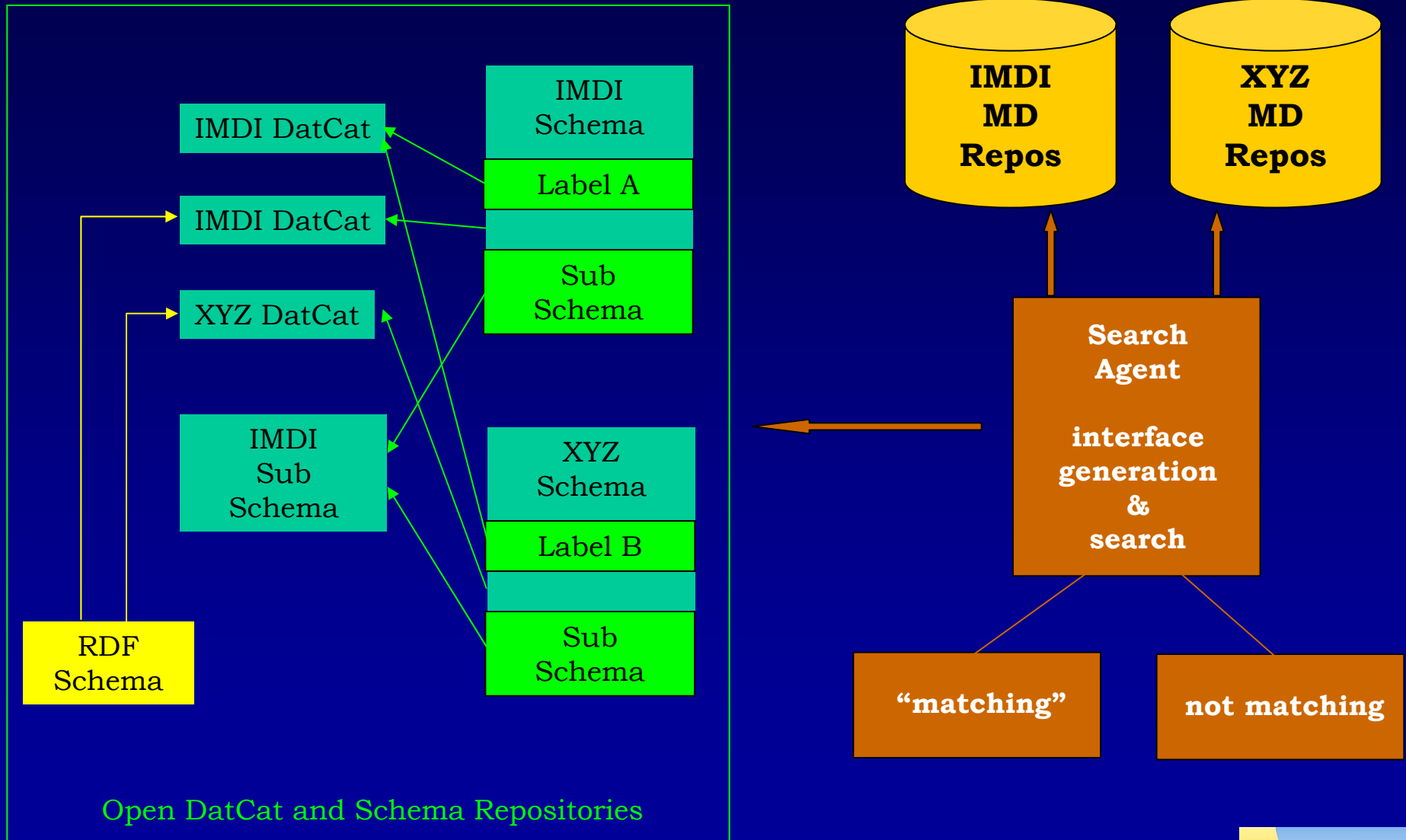
could be investigated??

for access methods only

Query Landscape



Query Landscape



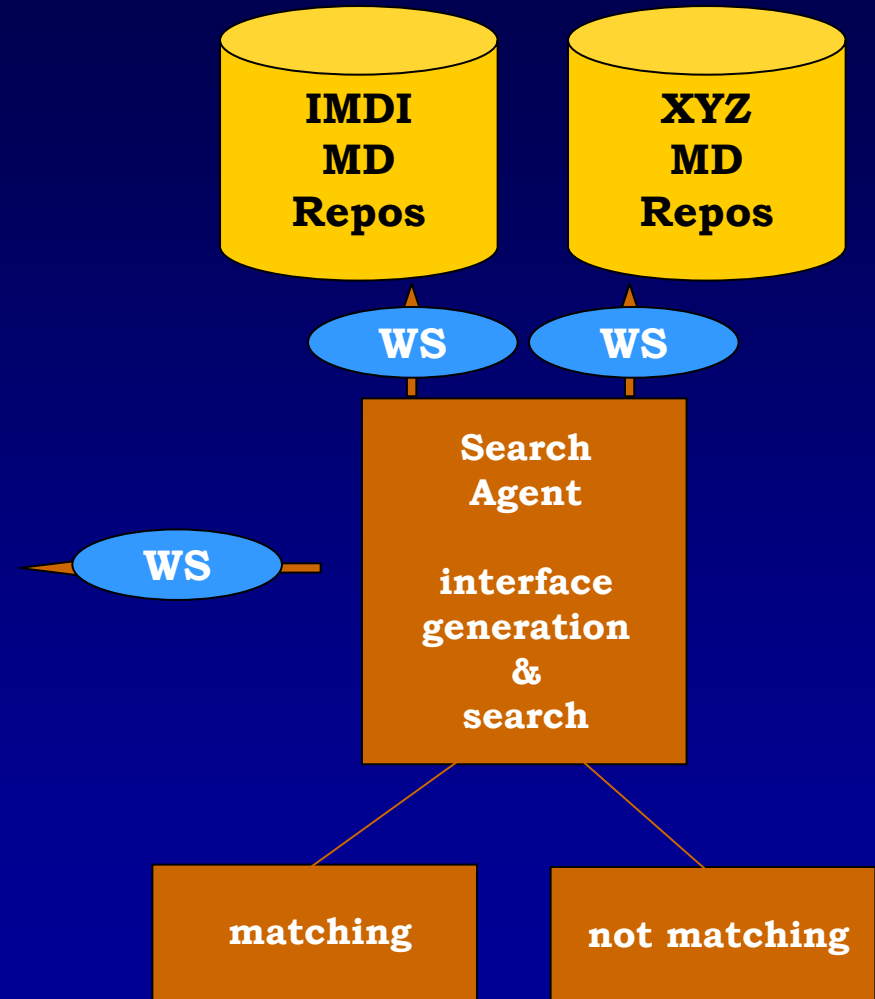
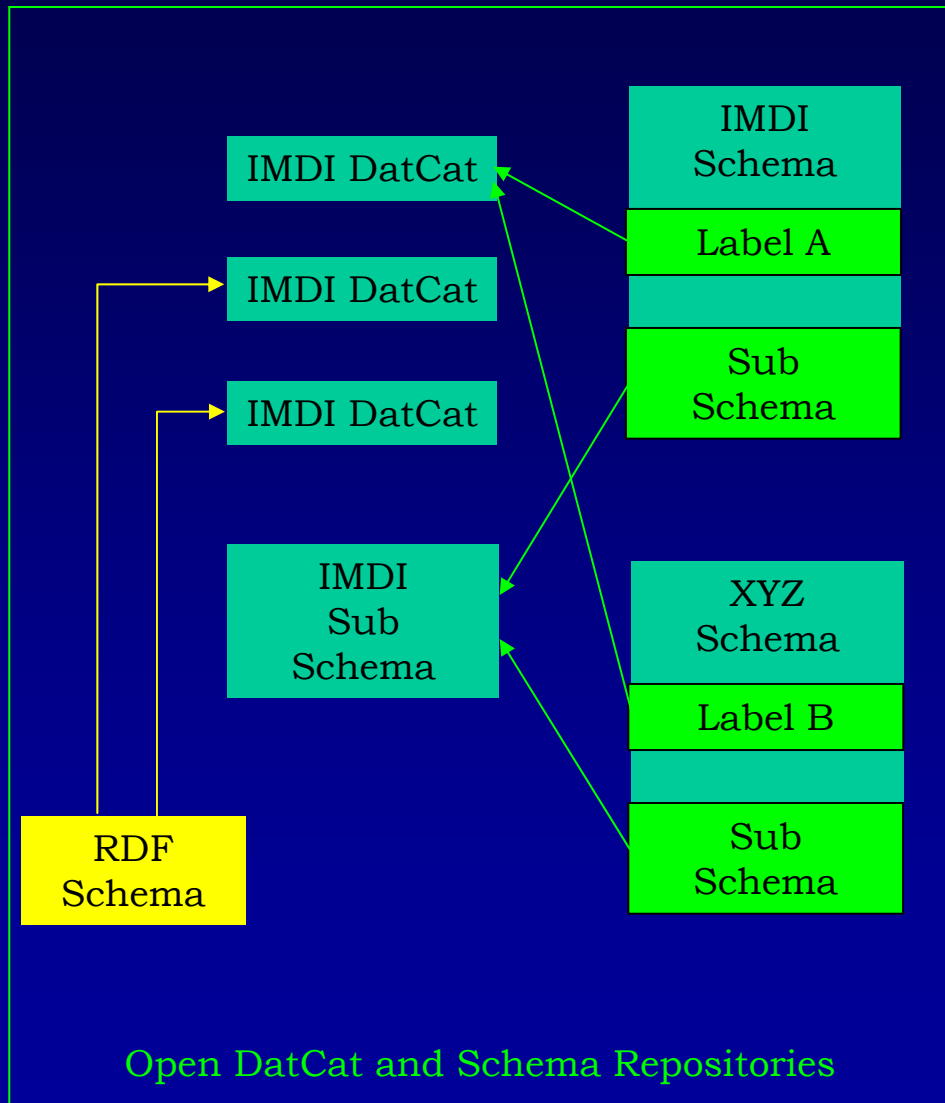
Metadata & Web-Services

- Web-Services
 - transparent offering of all available resources at each level of detail
 - can export data or offer components to be executed
 - SOAP implements access to resources via HTTP
 - WSDL describes the type of service in detail
 - UDDI allows to register service (distributed yellow page system)
 - distributed UDDI require taxonomy - does not exist yet (for LR)
 - all XML based
 - with transformation also non-XML data can be served
 - to be processed by machines

Metadata & Web-Services

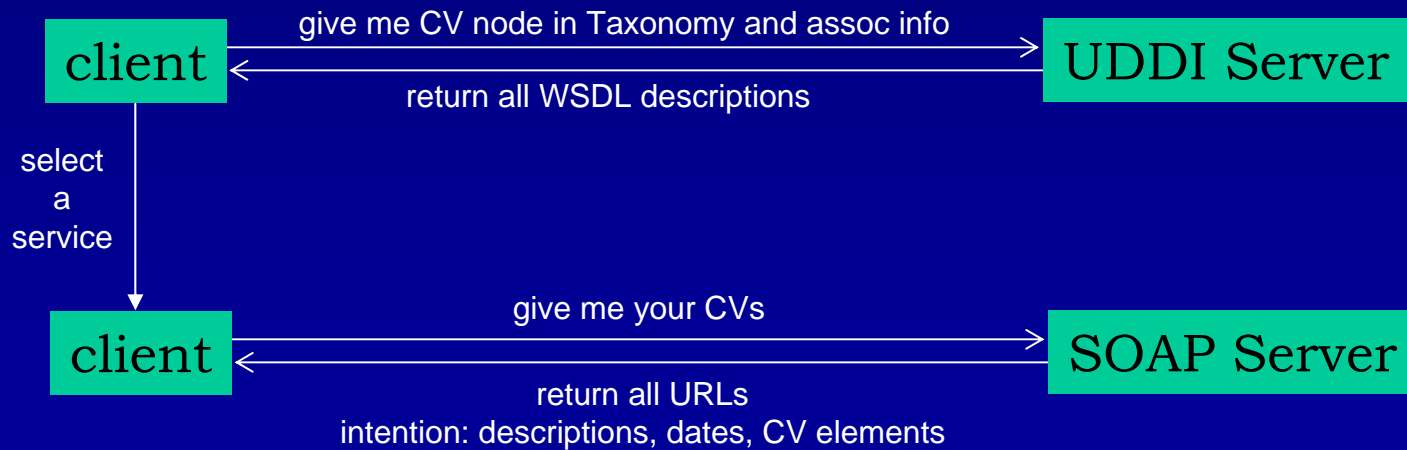
- Metadata Web-Services
 - what do we want to serve?
 - controlled vocabularies (re-usage)
 - building blocks (re-usage)
 - element defs (they are defined elsewhere - how are they served?)
 - mapping files (IMDI to DC, ...)
 - search for metadata records and serve elements of them
 - MD editing as a component
 - which granularity?
 - do we offer the CVs itself or do we offer the URLs
 - do we offer all CVs or an individual CV as a service
 - just metadata records or even element values as functions
 - where to register?
 - are there reliable and open places that all know to register/find services
 - can we agree on a taxonomy and naming to allow machines to find things

Query Landscape



CV Web-Service

- recently a Web-Service to offer IMDI's controlled vocabularies
 - just a test to implement complete chain (UDDI, WSDL, SOAP)
 - service delivers URLs of IMDI CVs
 - service implemented with Java (JAXR, JAXM, JAX-RPC, JWSDP, ...)
 - service registered at MS UDDI server
- not useful so far, but learned how to do it



ISO TC37/SC4 Tasks

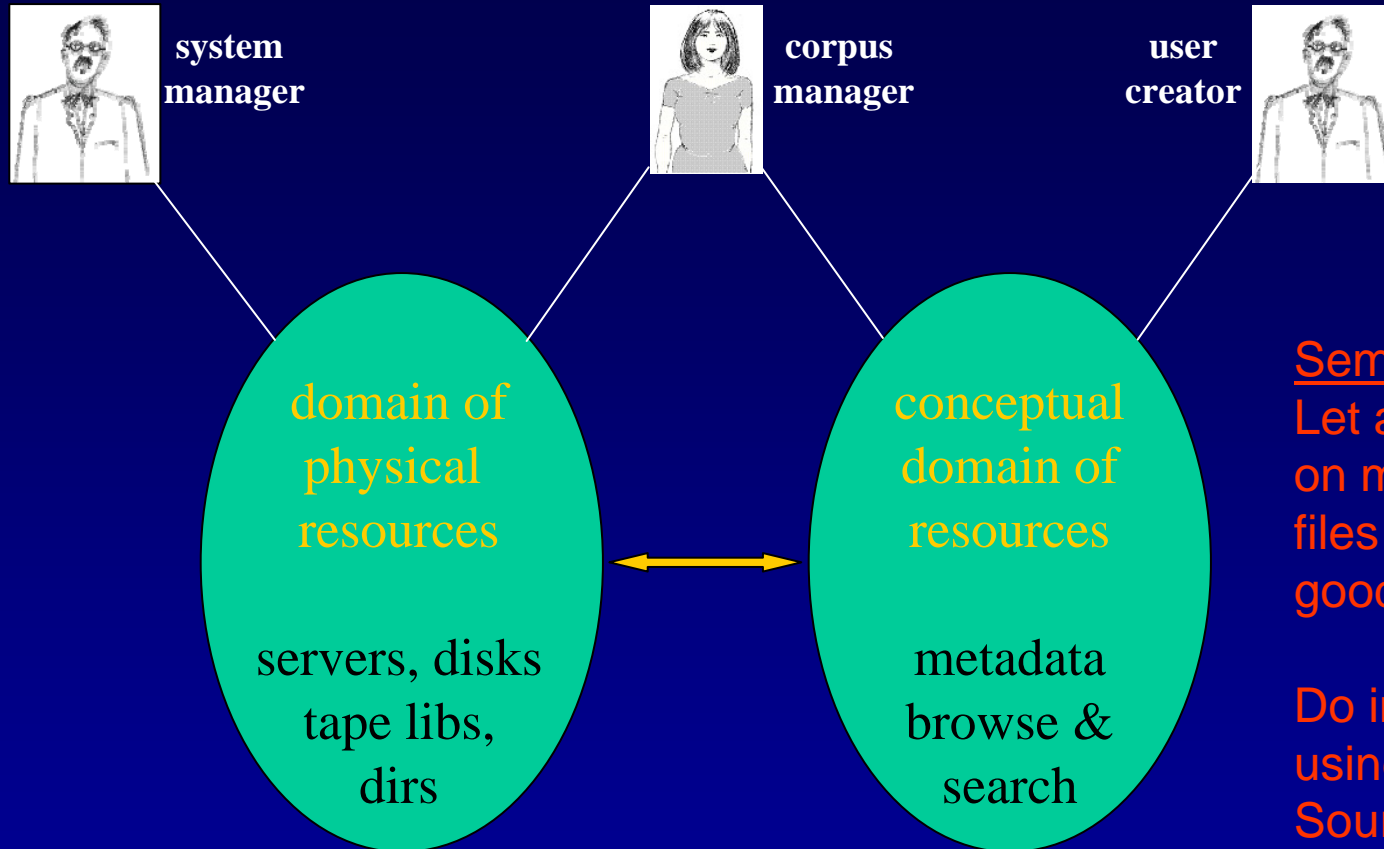
- what do we need?
 - repository infra for element definitions (primitive and constructed) (11179 ?)
 - define general controlled vocabularies
 - (“values” are DatCats, defines the conceptual domain)
 - recommend exchange schema for vocabulary definitions
 - (given the termbase framework)
 - recommend framework for re-usage of elements and schemas
 - recommend framework for making all relations explicit (RDFS repositories)
 - make suitable UDDI registry facility and create a taxonomy for LR domain
 - give advice about type and granularity of WS
- have to take care
 - 11179 seems to be excellent ref and framework
 - have to keep it simple and tractable
 - narrow down complexity for LR

End

IMDI	www.mpi.nl/ISLE papers at LREC 2002 conference
DOBES	www.mpi.nl/DOBES workshop at LREC 2002 conference
ECHO	echo.mpiwg.mpg.de www.mpi.nl/echo
INTERA	paper about LREP protocol at LREC 2002 conference
Tools	www.mpi.nl/tools (free for usage, open source soon)
Corpus	www.mpi.nl/corpus (much content not free)

Thanks for your attention

The Vision



Semantic Web

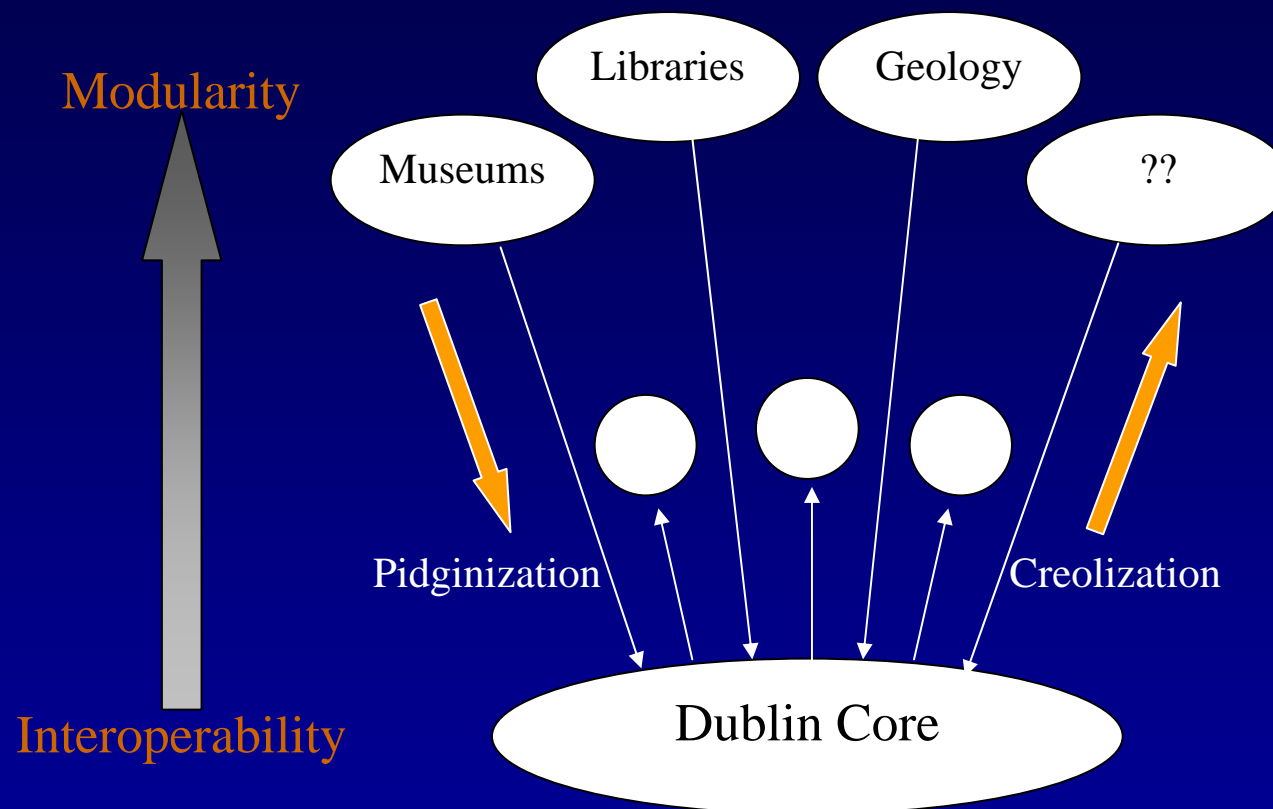
Let agents operate on machine readable files to do something good for us.

Do intelligent searches using other Knowledge Sources as well.

Users want to search & browse in a conceptual space and immediately execute some useful tool that is suggested to operate on the format and on the platform he is working on

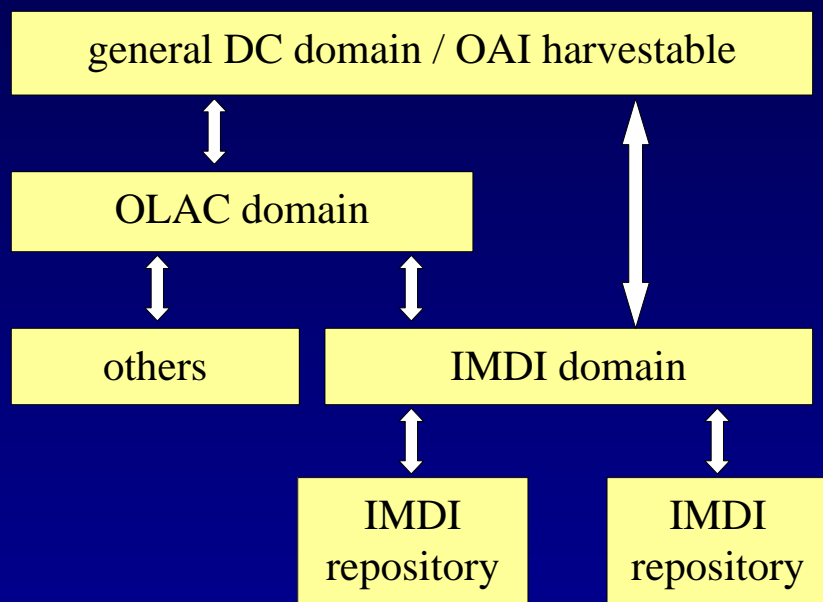
Managers want to control whole resource workflow in this conceptual space.

Metadata Genesis



- have seen the Pidginization - necessary for librarians
- experience a phase of Creolization - necessary for domain
- modularization already foreseen in the Warwick model

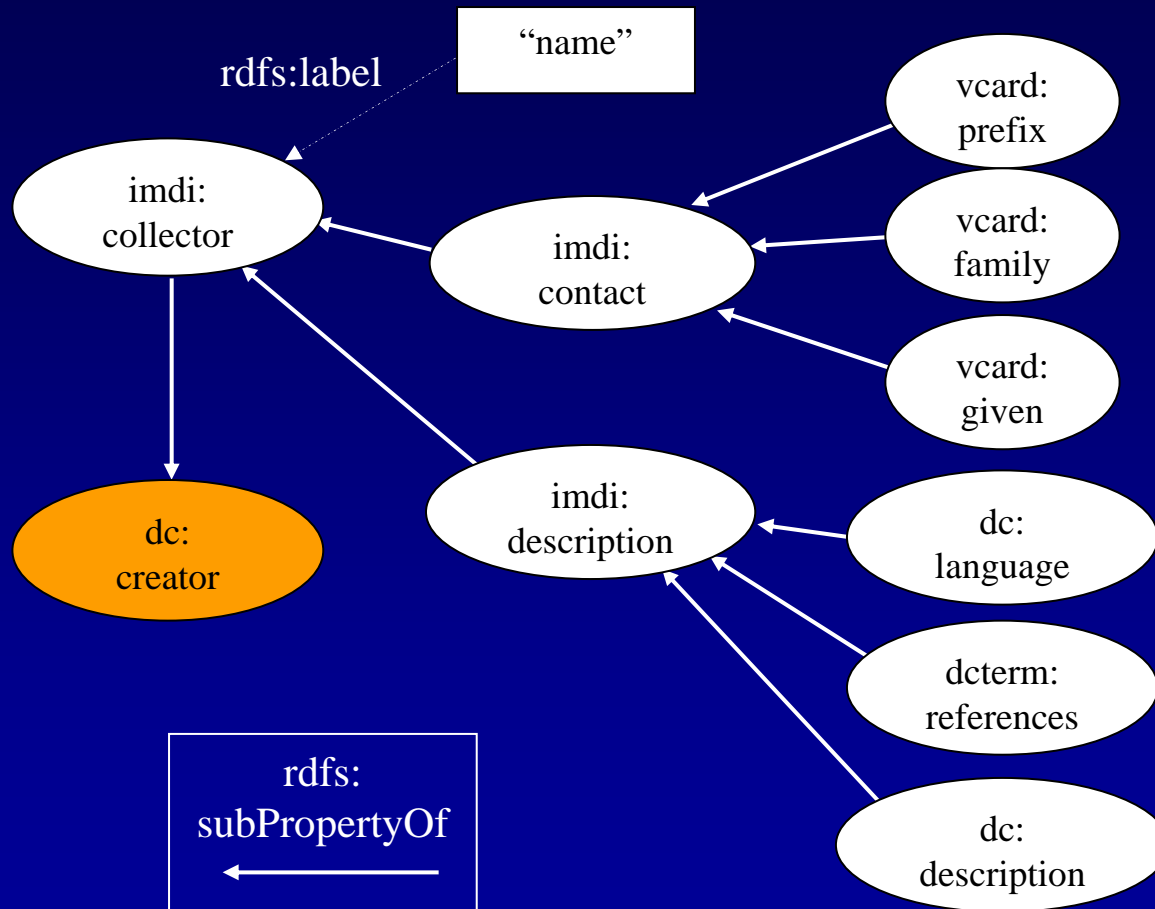
Metadata Interoperability



- mapping document on the web
- conversion and OLAC harvesting ready
- suffering from dynamics
- some problems
 - DC:Creator vs. IMDI:Collector
 - session resource bundle
media files, annotation files, sources
become individual resources
copying other elements
 - substructure in IMDI
non in DC

so: for interoperability there is a price to be paid
mapping is now implicit in scripts - not a nice solution

Future Metadata



- define all terms or refer to ontology concepts
- make all relations explicit internal and external
- re-use well-defined terms
- re-use whole sub-blocks
- need ontology
- need term repositories
- apply RDF/OWL
- ISO can play great role

General IMDI Structure

- **General Info** Name, Date, Creator, Project, etc
- **Content Info** Languages, Genre, Modality, Lexical Entry, etc
- **Participants Info** information about various agents, ...
- **Media File Info** typical info describing media files
- **Annotation File Info** typical info describing annotations
- **Written Resource Info** typical info describing written resources
- **Lexicon Resource Info** typical info describing lexica
- **Source Info** typical info referring to sources
- **References** references of all sort