Lexicons in an open Landscape of Registries and Repositories

Peter Wittenburg
Max-Planck-Institute for Psycholinguistics

Problem I

until now a lexicon is seen

- as one resource created by an individual or a project
- as one resource to be transferred
- as a closed resource offered by someone

- are these assumptions true for the future?
- we already have web-collaborations to create a joint lexicon
- we may want to have even more participation to get the semantics integrated

- discuss a few topics that may be of relevance for a more open infrastructure
  - where many people create lexicons and want to share/offer them
  - where many people may contribute to the content

Problem II

until now lexicons

- suffer from structural and format divergence (to be addressed later)
- suffer from semantic divergence
    - tag names are different although referring to same linguistic concept
    - tag names are the same although referring to different concepts
    - different vocabularies may be used although definition is the same

Consequences

- limited interoperability
- limited reusability

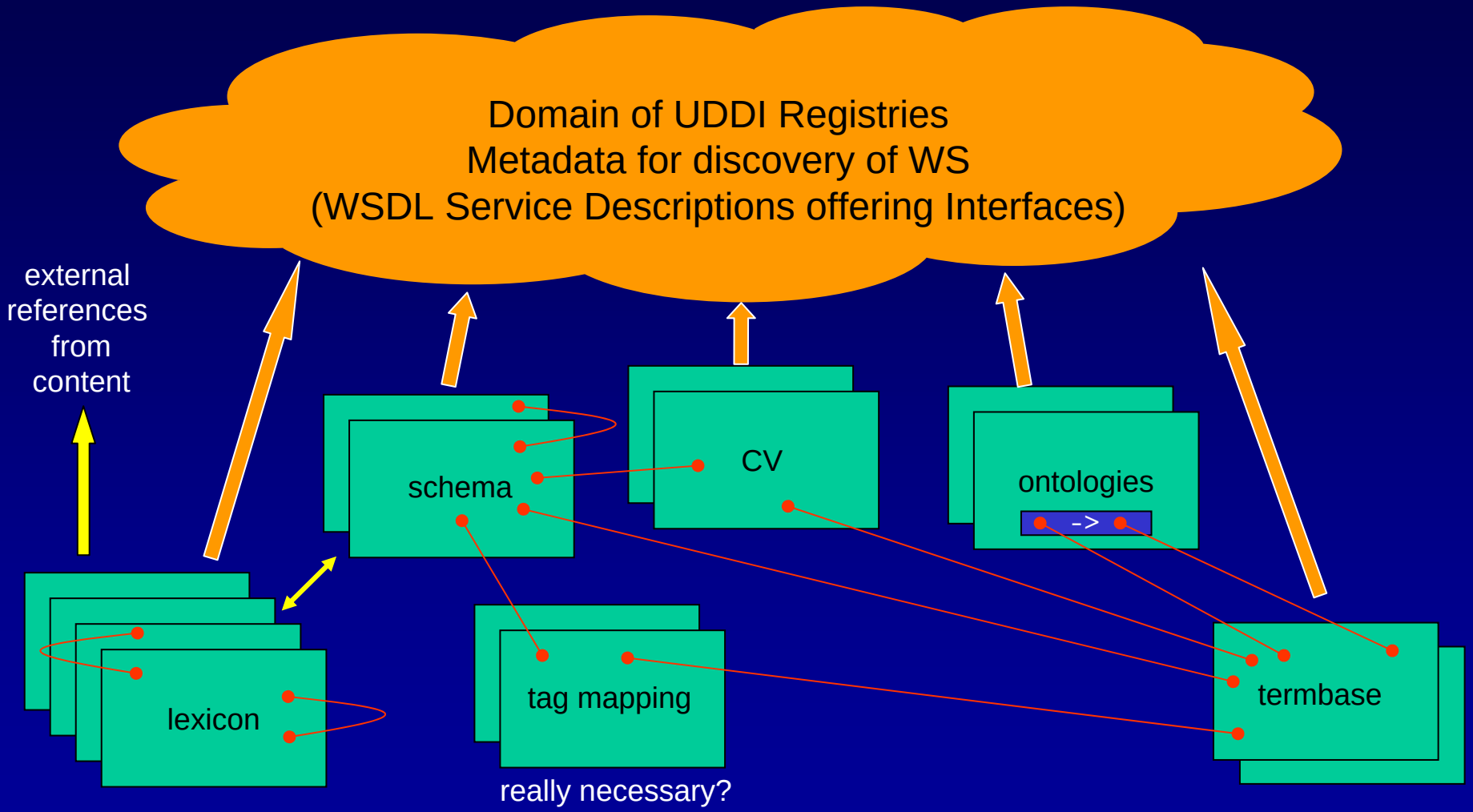
further

- how to find lexicon information
- how to register all info relevant in the context of lexicon work
- how to get access in a unified way to lexicon information

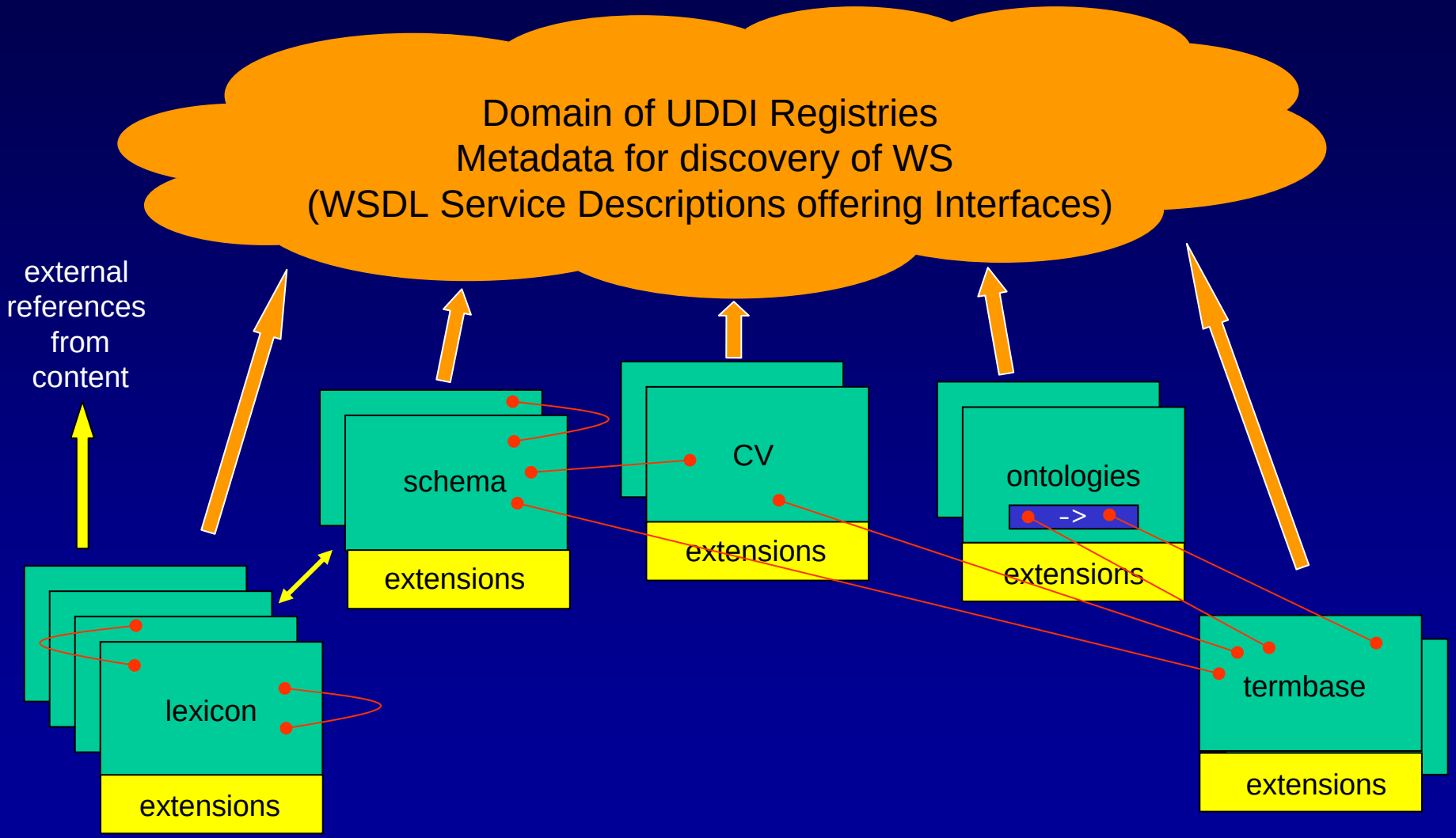
Various attempts to understand needs (MILE, Genelex, …)

The architecture presented may help to overcome part of these problems.

Scenario


Domain of UDDI Registries
Metadata for discovery of WS
(WSDL Service Descriptions offering Interfaces)
external references from content
schema
CV
ontologies
->
lexicon
tag mapping
really necessary?
termbase

Dynamic Scenario



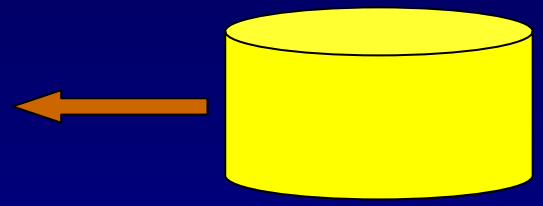
How to make this world flexible? How to connect added contributions?

Terminology + Ontologies

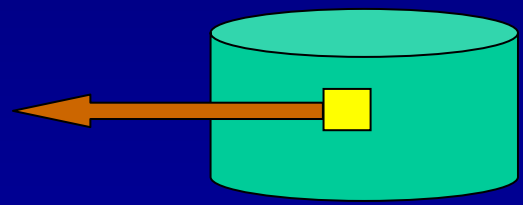
Laurent

Information to be offered

- services from institutions
- resources as a whole
  - ontology addresses
  - termbase addresses
  - lexicon schemas
  - lexicon addresses
  - controlled vocabulary addresses



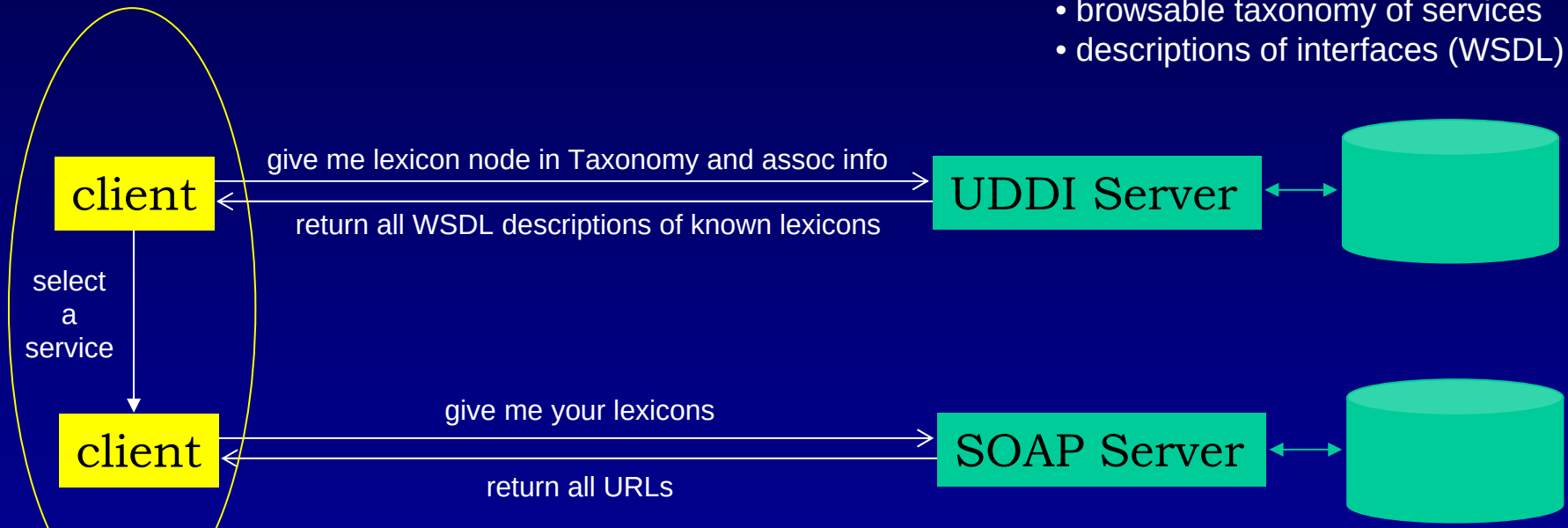
- content of resources
  - specific ontology content
  - concept definition records
  - specific lexicon content
  - CV themselves



- which granularity of services? which Import/Export Types?

Type of Interaction

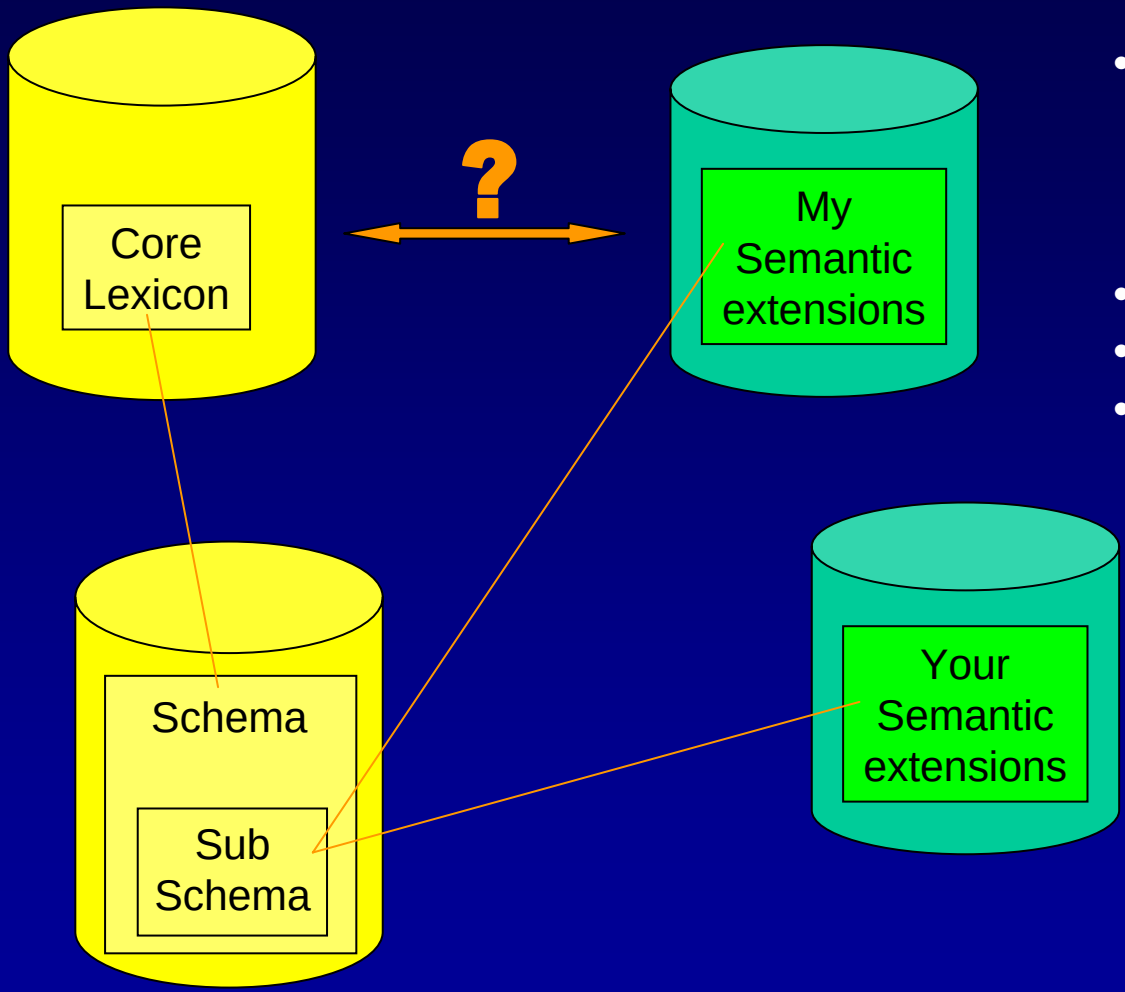
Metadata about Content
- MD descriptions of services (OLAC/IMDI/…)
- browsable taxonomy of services
- descriptions of interfaces (WSDL)



User Defined Applications of any complexity could even exploit several different services in one application

Content
- the content (can be on different levels)
- SOAP supports offered interfaces

Repository Aspects



- extensions can be at any level
  - added entries
  - added sense descriptions
  - …
- how to register them?
- what about URIDs?
- what about link stability?

Lexicon Services Needed

- tools/services needed
  - need editors that allow browsing in termbase
    - while creating a schema
    - while creating an ontology
    - while creating a CV
  - need editors that allow browsing in other lexicons
    - while creating a lexicon entry
  - search tools that can exploit the link structure
  - flexible visualization of assembled content
  - ….

ISO TC37/SC4 Tasks

- what is needed (without termbase and onto details)?
  - setup UDDI registries for LR services
  - give advice of how to register
  - create a taxonomy for LR domain to integrate services
  - support suitable metadata descriptions for discovering services (see MILE)
  - give advice about type and granularity of WS
  - specify the types of import and export of WS

  - define lexical elements as datcats
  - define controlled vocabularies for elements
    ("values" are DatCats, defines the conceptual domain)
  - recommend exchange schema for vocabulary definitions

  - define requirements for tools