UNL Universal Words

Should we go in details about UWs?

Is this the important point??

Or other more general issues?

Nicoletta Calzolari
ILC – CNR
glottolo@ilc.cnr.it

5 Topics

a. What is to be considered a “Universal Word”?

b. Which named entities should be introduced in the dictionary of UW’s, if any?

c. UW’s must correspond to roots, to stems or to word forms?

d. Antonyms should be represented as a single UW or as different UW’s?

e. When a multiword expression must be represented as a UW?

5 Starting questions

1. How many UW's should be recognized in the sentence below?

   "Charles Dickens is generally regarded as the most important English novelist of the Victorian period"

2. "Charles Dickens" should be represented as a permanent UW or as a temporary UW?
3. "hunger" (= "a physiological need for food"), "hungry" (= "feeling hunger"), "hungrily" (= "in the manner of someone who is very hungry") and "hunger" (= "to cause to experience hunger") should be represented as simple, compound or complex UW's?
4. Antonyms such as "mortal" and "immortal", "hot" and "cold", and "son" and "father" should be represented as a single UW (and the corresponding attributes) or as different UW's?
5. "Farbfernsehgerät" ("color television set", in German) should be represented as a simple or complex UW?

Request

But also

- **Asked to "Suggest some general procedures"**

I'll go more in this direction

But first ...

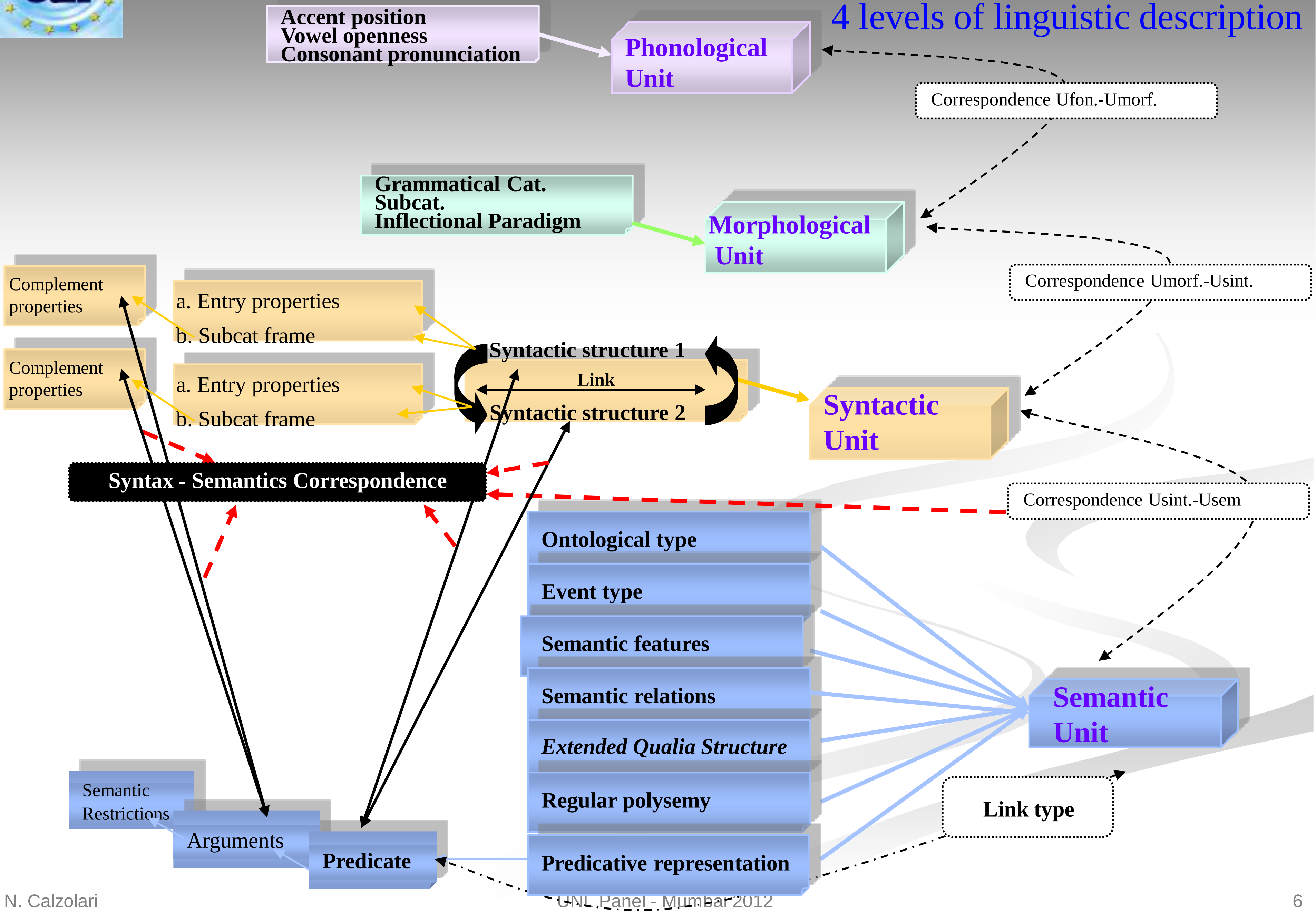
a. What is to be considered a "Universal Word"?

- They are universal in the sense that they are uniform identifiers to the entities defined in the UNL Knowledge Base, which is expected to map everything that we know about the world, and that is used to assign translatability to any concept

  - Nodes in a Semantic Network
  - Nodes of an Ontology?
  - Look at other examples

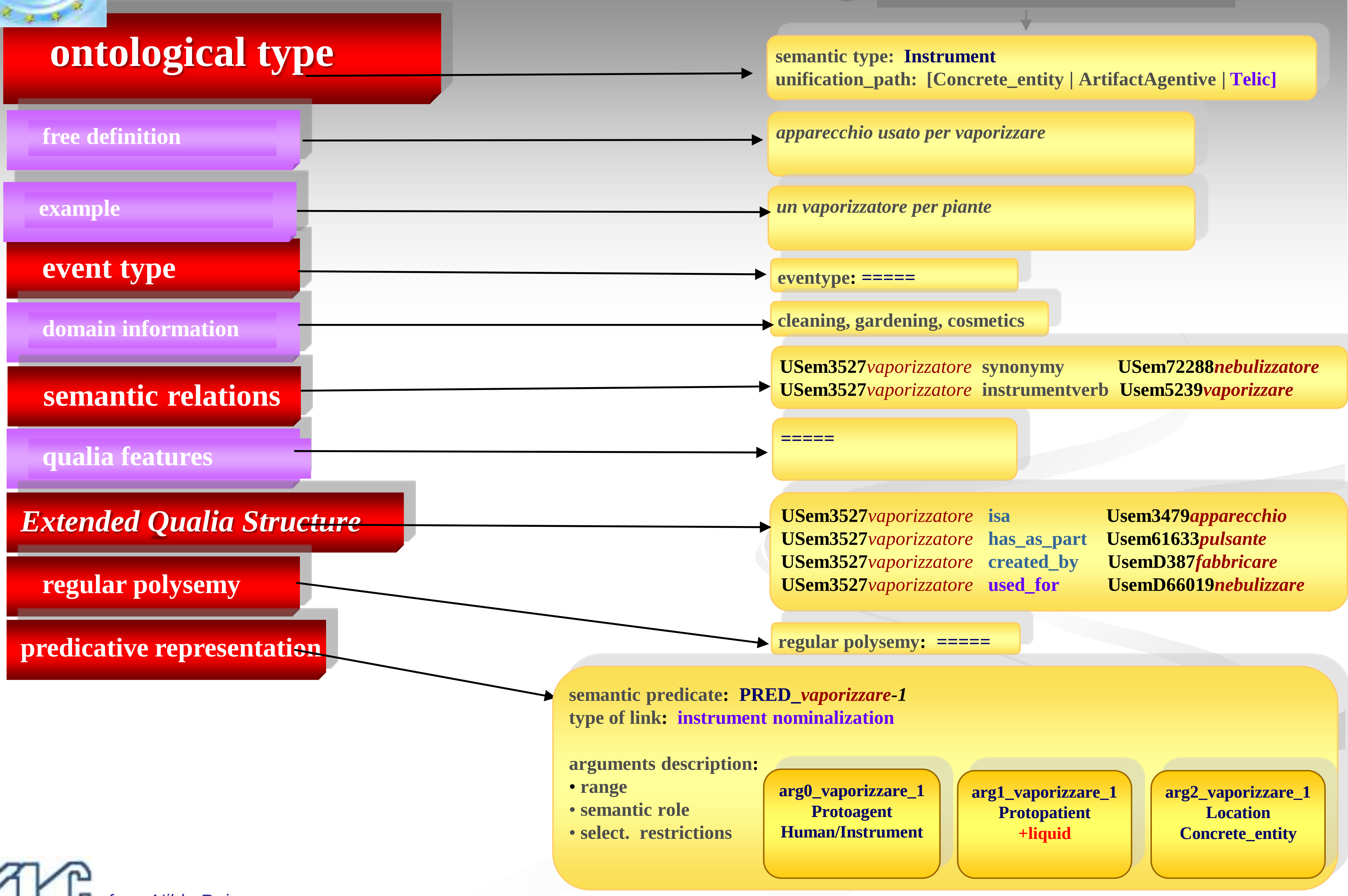
PAROLE-SIMPLE:

4 levels of linguistic description



Semantic entry

USem3527**vaporizzatore**



ontological type

semantic type: **Instrument**
unification_path: [Concrete_entity | ArtifactAgentive | Telic]

free definition

apparecchio usato per vaporizzare

example

un vaporizzatore per piante

event type

eventype: =====

domain information

cleaning, gardening, cosmetics

semantic relations

USem3527*vaporizzatore*	synonymy	USem72288*nebulizzatore*
USem3527*vaporizzatore*	instrumentverb	Usem5239*vaporizzare*

qualia features

=====

Extended Qualia Structure

USem3527*vaporizzatore*	isa	Usem3479*apparecchio*
USem3527*vaporizzatore*	has_as_part	Usem61633*pulsante*
USem3527*vaporizzatore*	created_by	UsemD387*fabbricare*
USem3527*vaporizzatore*	used_for	UsemD66019*nebulizzare*

regular polysemy

regular polysemy: =====

predicative representation

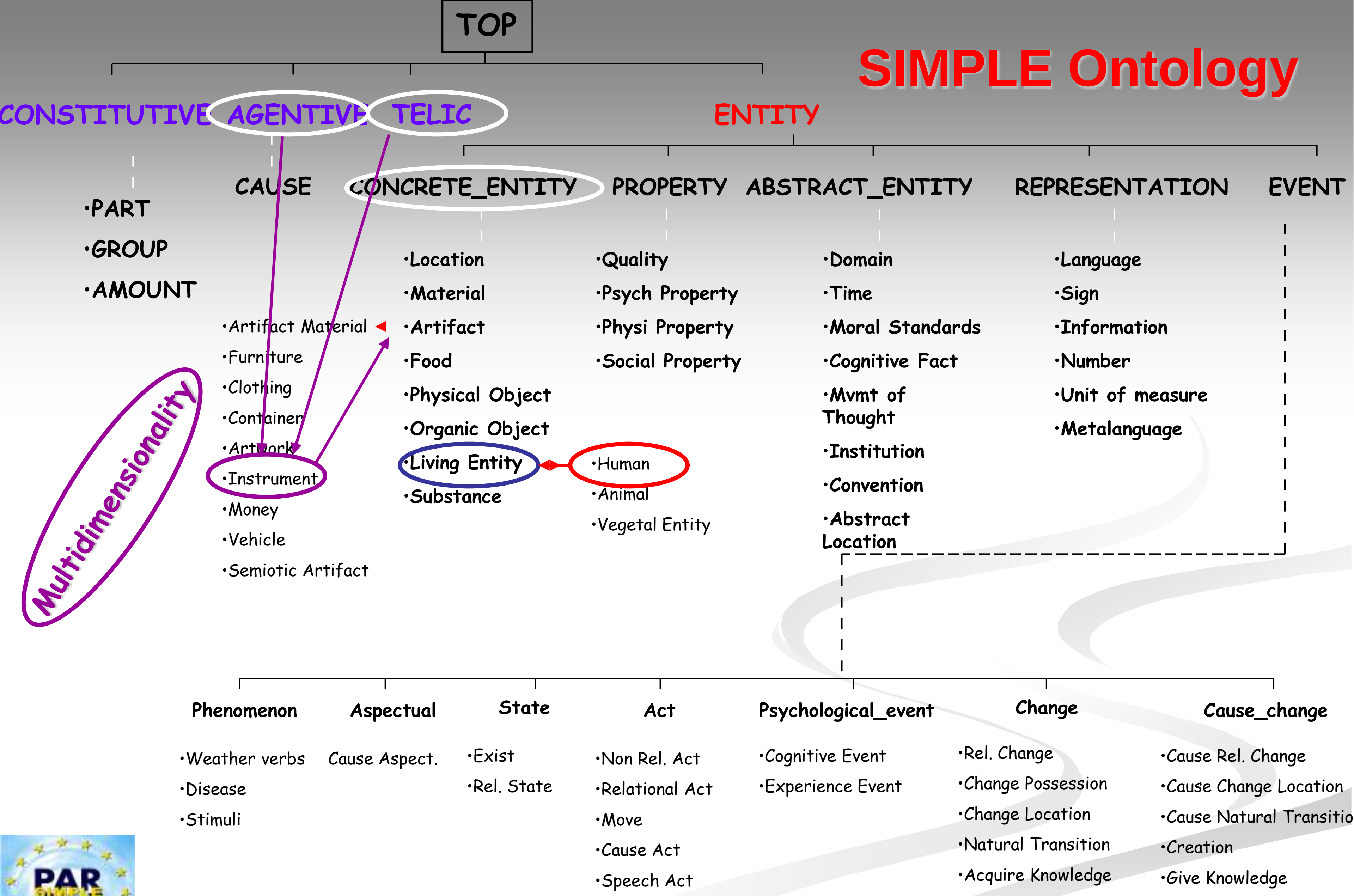
semantic predicate: **PRED_*vaporizzare*-1**
type of link: **instrument nominalization**

arguments description:
- range
- semantic role
- select. restrictions

arg0_vaporizzare_1	arg1_vaporizzare_1	arg2_vaporizzare_1
Protoagent	Protopatient	Location
Human/Instrument	+liquid	Concrete_entity

from Nilda Ruimy

SIMPLE Ontology



TOP

CONSTITUTIVE

- PART
- GROUP
- AMOUNT

AGENTIVE

CAUSE

TELIC

ENTITY

CONCRETE_ENTITY

- Location
- Material
- Artifact
  - Artifact Material
  - Furniture
  - Clothing
  - Container
  - Artwork
  - Instrument
  - Money
  - Vehicle
  - Semiotic Artifact
- Food
- Physical Object
- Organic Object
- Living Entity
  - Human
  - Animal
  - Vegetal Entity
- Substance

PROPERTY

- Quality
- Psych Property
- Physi Property
- Social Property

ABSTRACT_ENTITY

- Domain
- Time
- Moral Standards
- Cognitive Fact
- Mvmt of Thought
- Institution
- Convention
- Abstract Location

REPRESENTATION

- Language
- Sign
- Information
- Number
- Unit of measure
- Metalanguage

EVENT

Phenomenon

- Weather verbs
- Disease
- Stimuli

Aspectual

Cause Aspect.

State

- Exist
- Rel. State

Act

- Non Rel. Act
- Relational Act
- Move
- Cause Act
- Speech Act

Psychological_event

- Cognitive Event
- Experience Event

Change

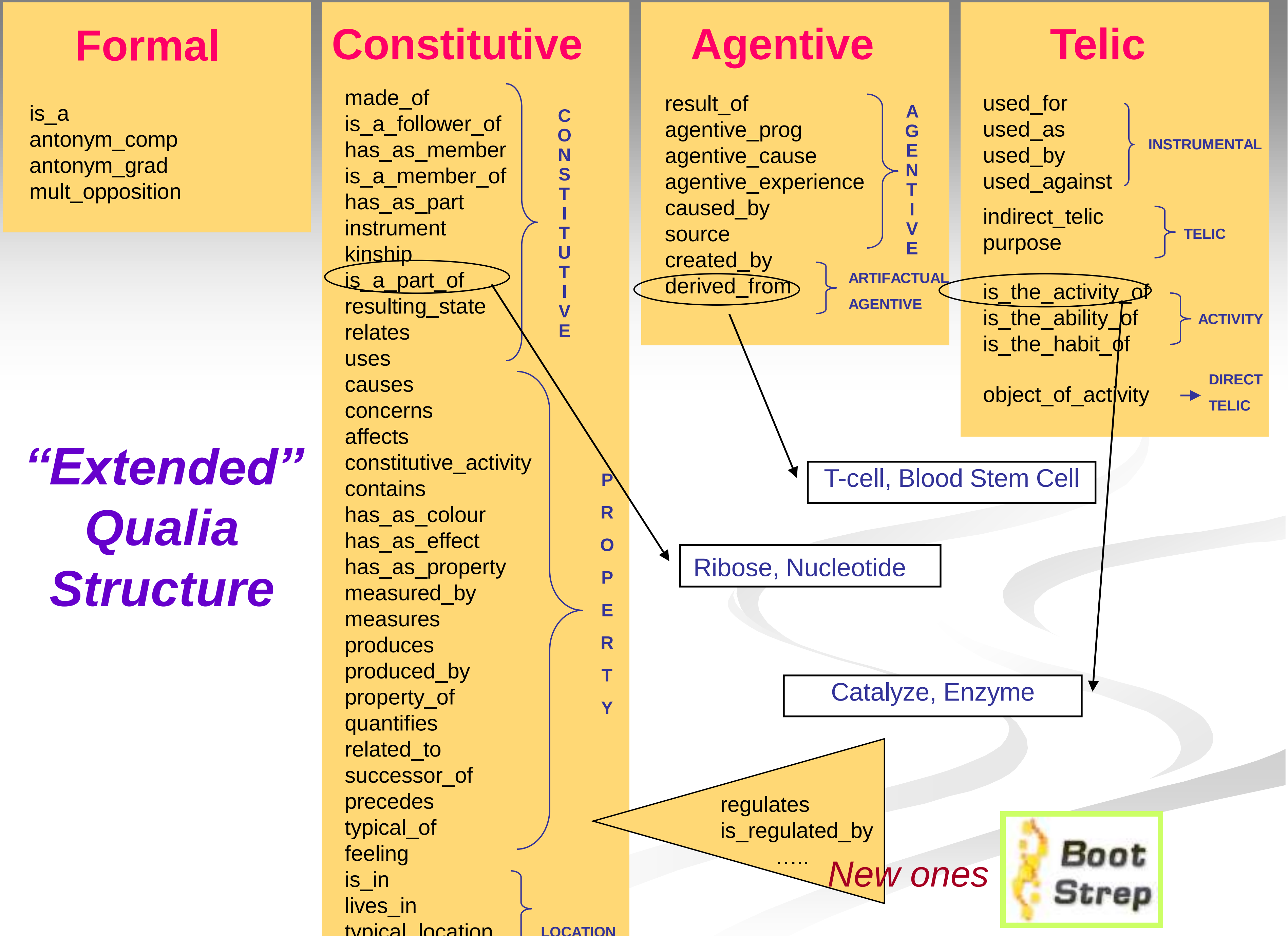
- Rel. Change
- Change Possession
- Change Location
- Natural Transition
- Acquire Knowledge

Cause_change

- Cause Rel. Change
- Cause Change Location
- Cause Natural Transitio
- Creation
- Give Knowledge

Multidimensionality

from Nilda Ruimy



"Extended" Qualia Structure

Formal

is_a
antonym_comp
antonym_grad
mult_opposition

Constitutive

made_of
is_a_follower_of
has_as_member
is_a_member_of
has_as_part
instrument
kinship
is_a_part_of
resulting_state
relates
uses

CONSTITUTIVE

causes
concerns
affects
constitutive_activity
contains
has_as_colour
has_as_effect
has_as_property
measured_by
measures
produces
produced_by
property_of
quantifies
related_to
successor_of
precedes
typical_of
feeling

PROPERTY

is_in
lives_in
typical_location

LOCATION

Agentive

result_of
agentive_prog
agentive_cause
agentive_experience
caused_by
source

AGENTIVE

created_by
derived_from

ARTIFACTUAL AGENTIVE

Telic

used_for
used_as
used_by
used_against

INSTRUMENTAL

indirect_telic
purpose

TELIC

is_the_activity_of
is_the_ability_of
is_the_habit_of

ACTIVITY

object_of_activity → DIRECT TELIC

T-cell, Blood Stem Cell

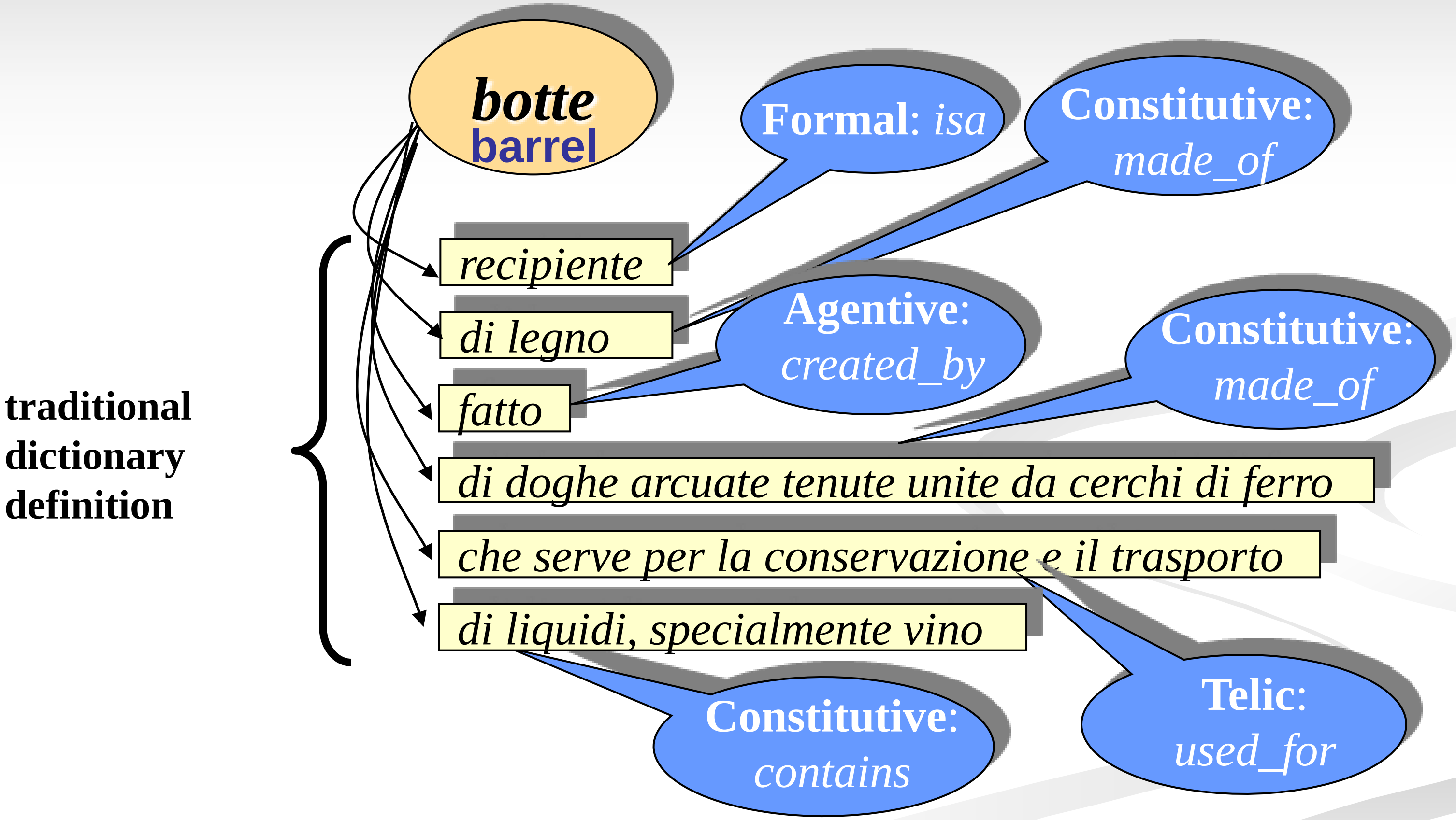
Ribose, Nucleotide

Catalyze, Enzyme

regulates
is_regulated_by
…..

New ones

Meaning dimensions expressed by Qualia relations

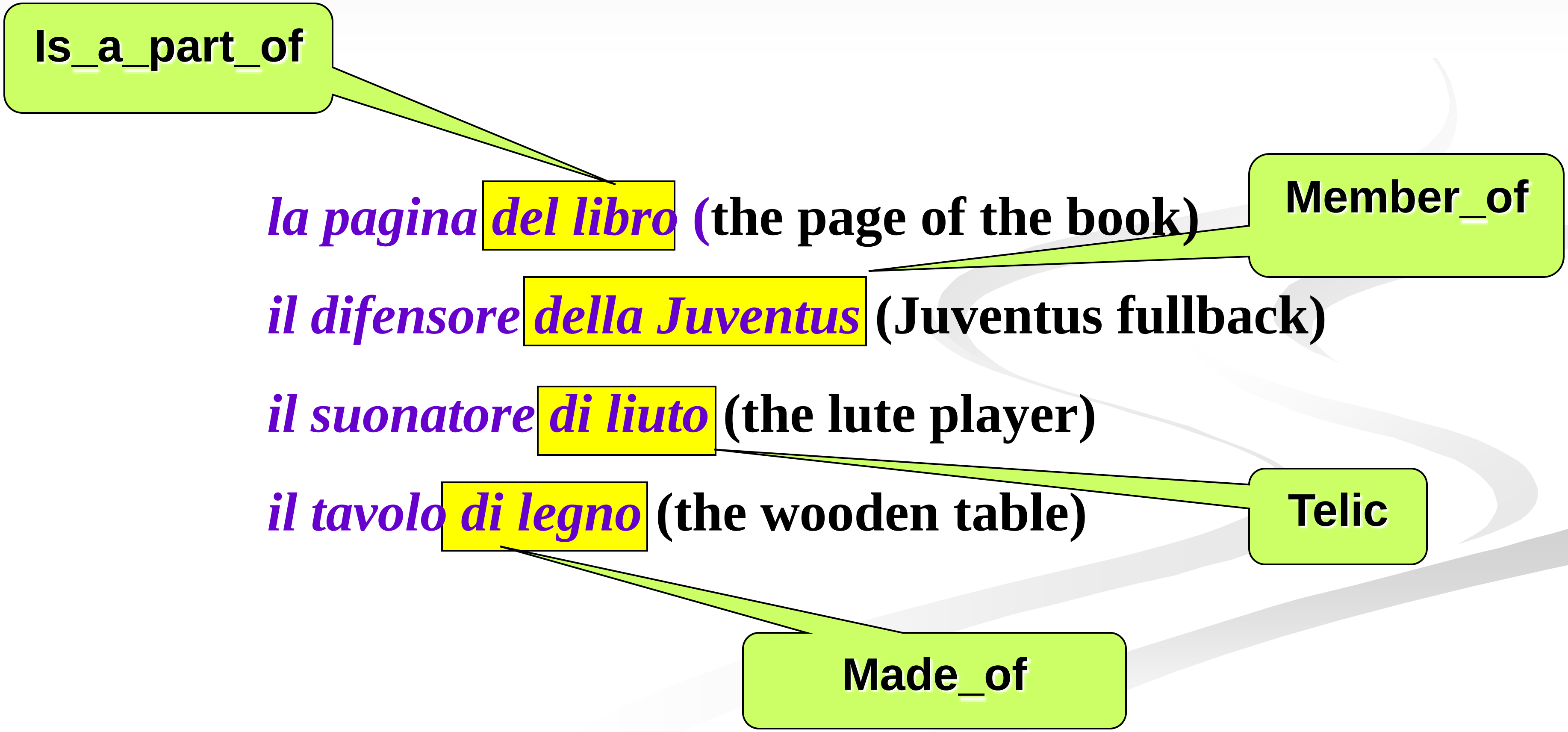

botte
barrel
Formal: isa
Constitutive: made_of
recipiente
di legno
Agentive: created_by
Constitutive: made_of
fatto
di doghe arcuate tenute unite da cerchi di ferro
traditional dictionary definition
che serve per la conservazione e il trasporto
di liquidi, specialmente vino
Telic: used_for
Constitutive: contains

from Nilda Ruimy

Semantic Multidimensionality & NLP

NLP tasks (IE, WSD, NP Recognition, etc.) need to access **multidimensional aspects of word meaning**:

Extended Qualia Relations



Is_a_part_of

la pagina ***del libro*** **(the page of the book)**

Member_of

il difensore ***della Juventus*** **(Juventus fullback)**

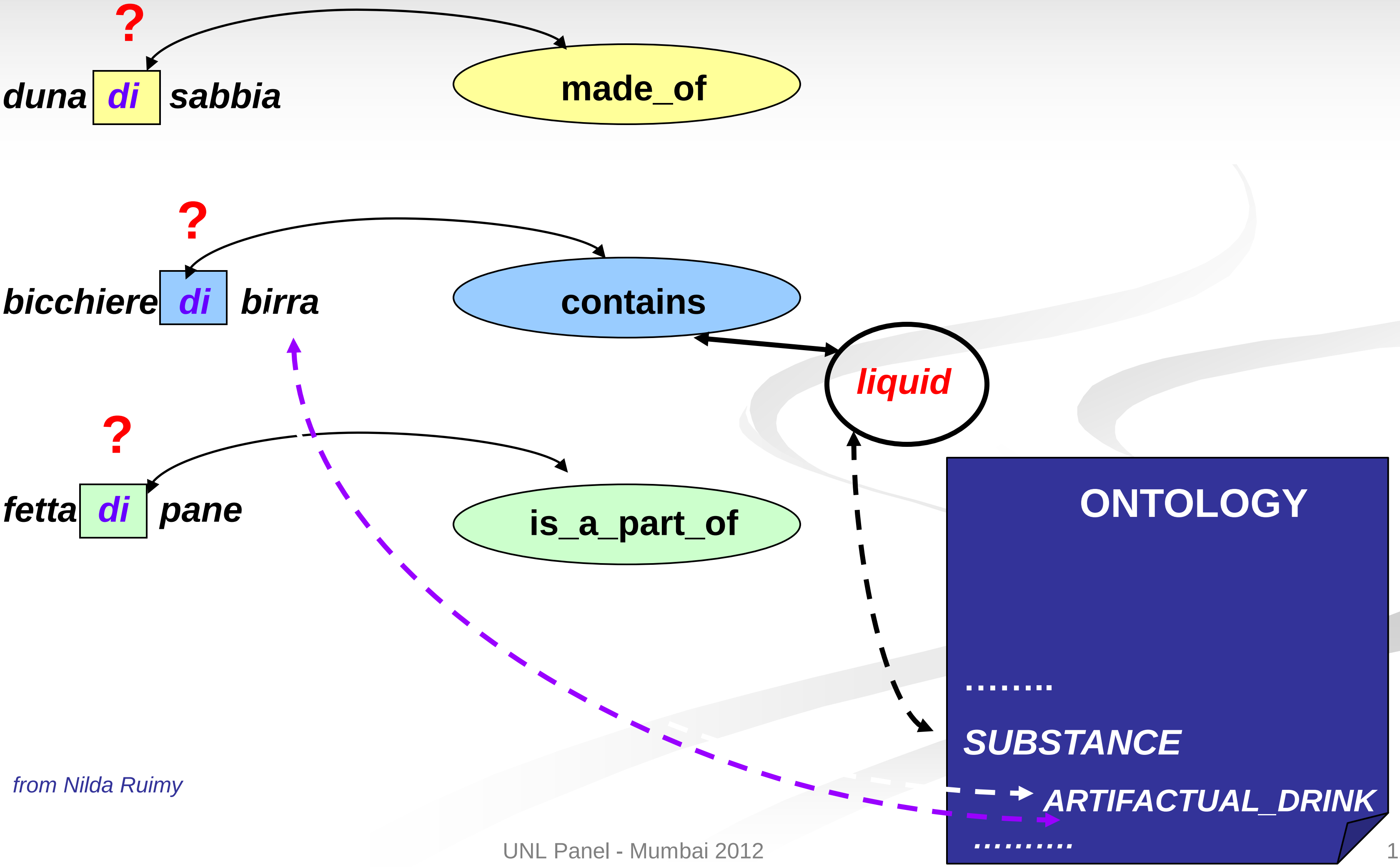
il suonatore ***di liuto*** **(the lute player)**

Telic

il tavolo ***di legno*** **(the wooden table)**

Made_of

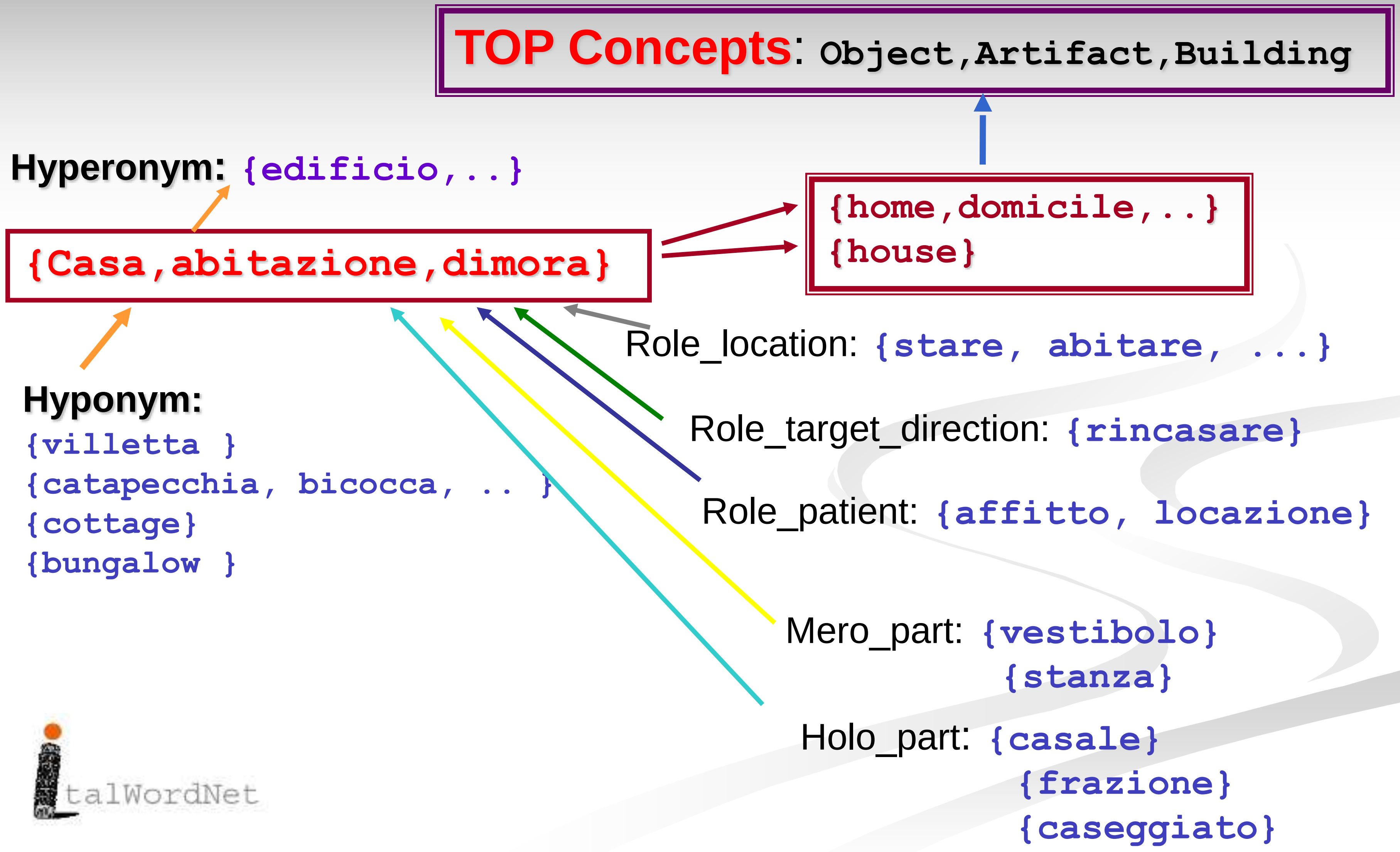
Disambiguation = Interpretation of conceptual relations in context



from Nilda Ruimy

WordNets

Synsets linked by semantic relations



TOP Concepts: `Object,Artifact,Building`

Hyperonym: {edificio,..}

{Casa,abitazione,dimora}

{home,domicile,..}
{house}

Role_location: {stare, abitare, ...}

Role_target_direction: {rincasare}

Role_patient: {affitto, locazione}

Hyponym:
{villetta }
{catapecchia, bicocca, .. }
{cottage}
{bungalow }

Mero_part: {vestibolo}
{stanza}

Holo_part: {casale}
{frazione}
{caseggiato}

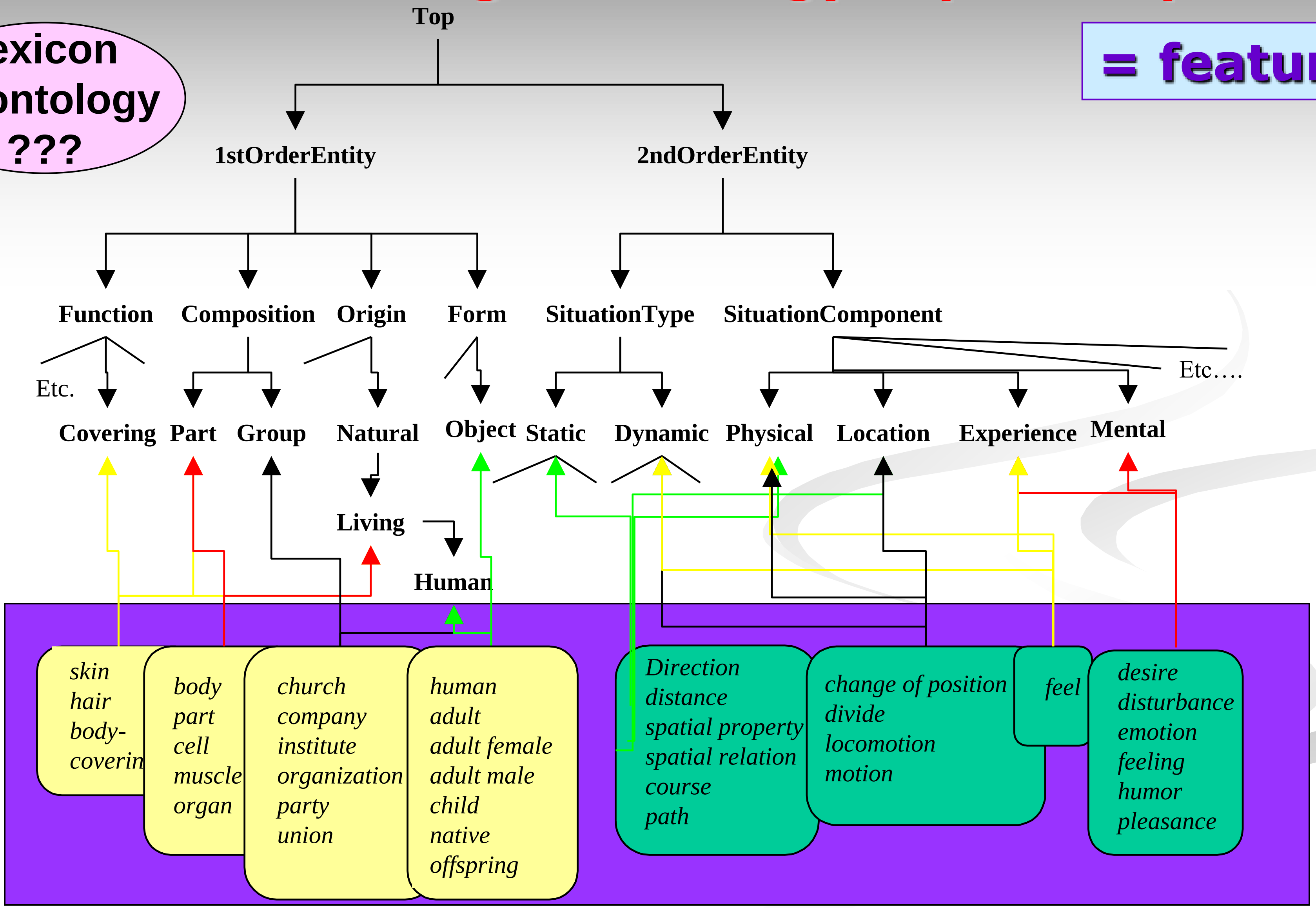
ItalWordNet

EuroWordNet: Clusters of "Base Concepts" classified according to Ontology Top Concepts

= words

= features

Lexicon or ontology ???



Top

1stOrderEntity

2ndOrderEntity

Function Composition Origin Form SituationType SituationComponent

Etc. Etc….

Covering Part Group Natural Object Static Dynamic Physical Location Experience Mental

Living

Human

skin
hair
body-coverin

body
part
cell
muscle
organ

church
company
institute
organization
party
union

human
adult
adult female
adult male
child
native
offspring

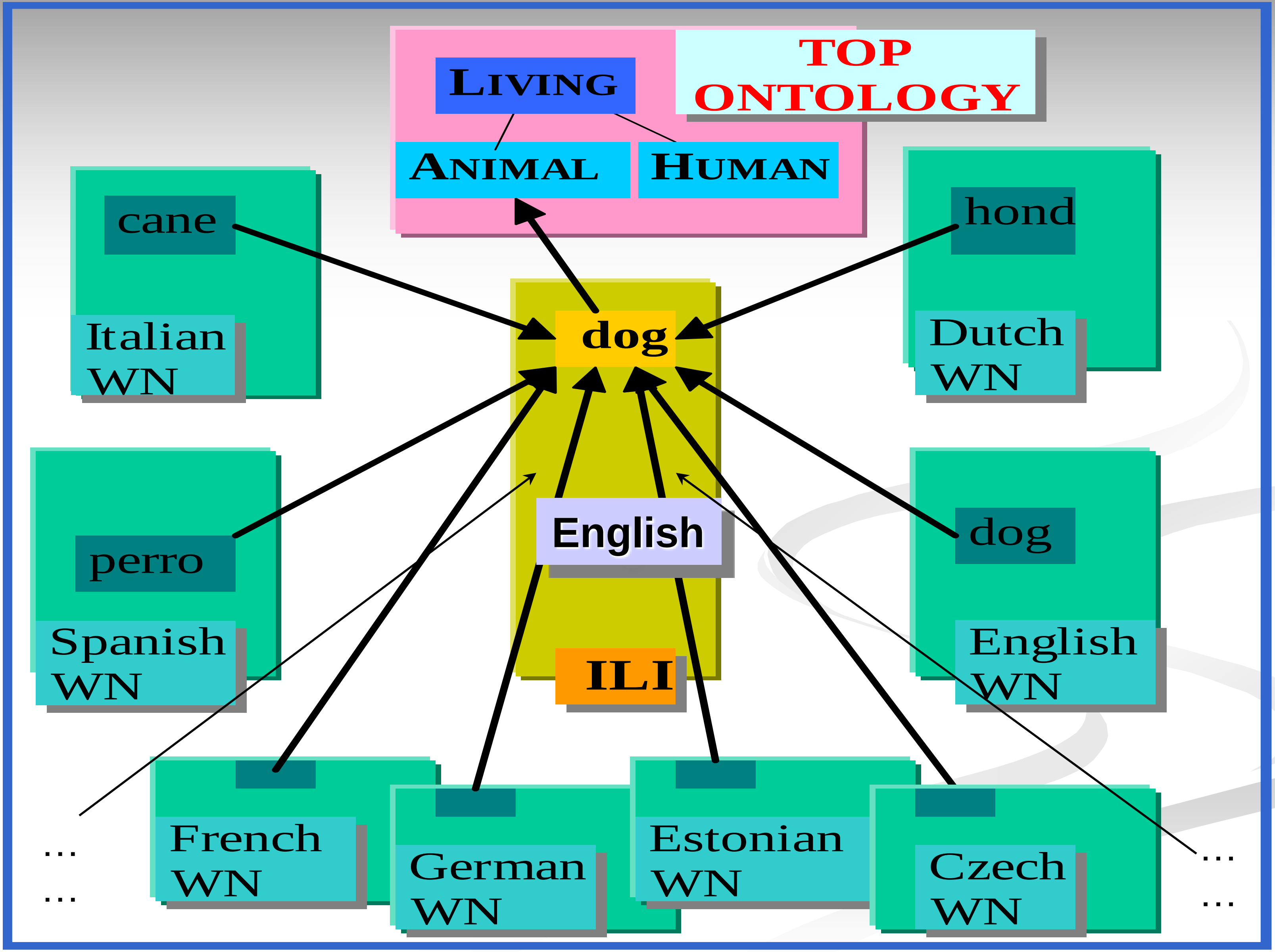
Direction
distance
spatial property
spatial relation
course
path

change of position
divide
locomotion
motion

feel

desire
disturbance
emotion
feeling
humor
pleasance

EuroWordNet Multilingual Data Structure



Reusability

- **Interesting to map UWs to WordNet(s)?**

Suggestion

Interoperability

- E.g**. linking to ILI**
- And through this **to many WordNets** in many languages

Population

- Also to **facilitate populating the NL Dictionaries**

1. How many UW's should be recognized in the sentence

- No unique & no “right” answer
  - 8 Nodes? Less? More?
- It depends on the theoretical framework
- Otherwise we would have solved many of our problems ….

2. "Charles Dickens" should be represented as a permanent or temporary UW?

- Named Entity &
  - As such different from e.g. “writer”
- In UNL they are “Temporary UWs”:
  - **Fine if consistent**

- Most named entities (names of people, places, …) are represented as temporary UW's... Nevertheless, some named entities of widespread use (such as "England” …) have been included in the UNL Dictionary and are treated as permanent UW's**. Our current criteria is the Wikipedia.** If a proper name is defined as an entry in the Wikipedia, then it should be defined as a permanent UW and included.

- Right criterion? Wikipedia has a different purpose
- Introduces the possibility of different representations for same type of unit:
  - **Consistency problem??**

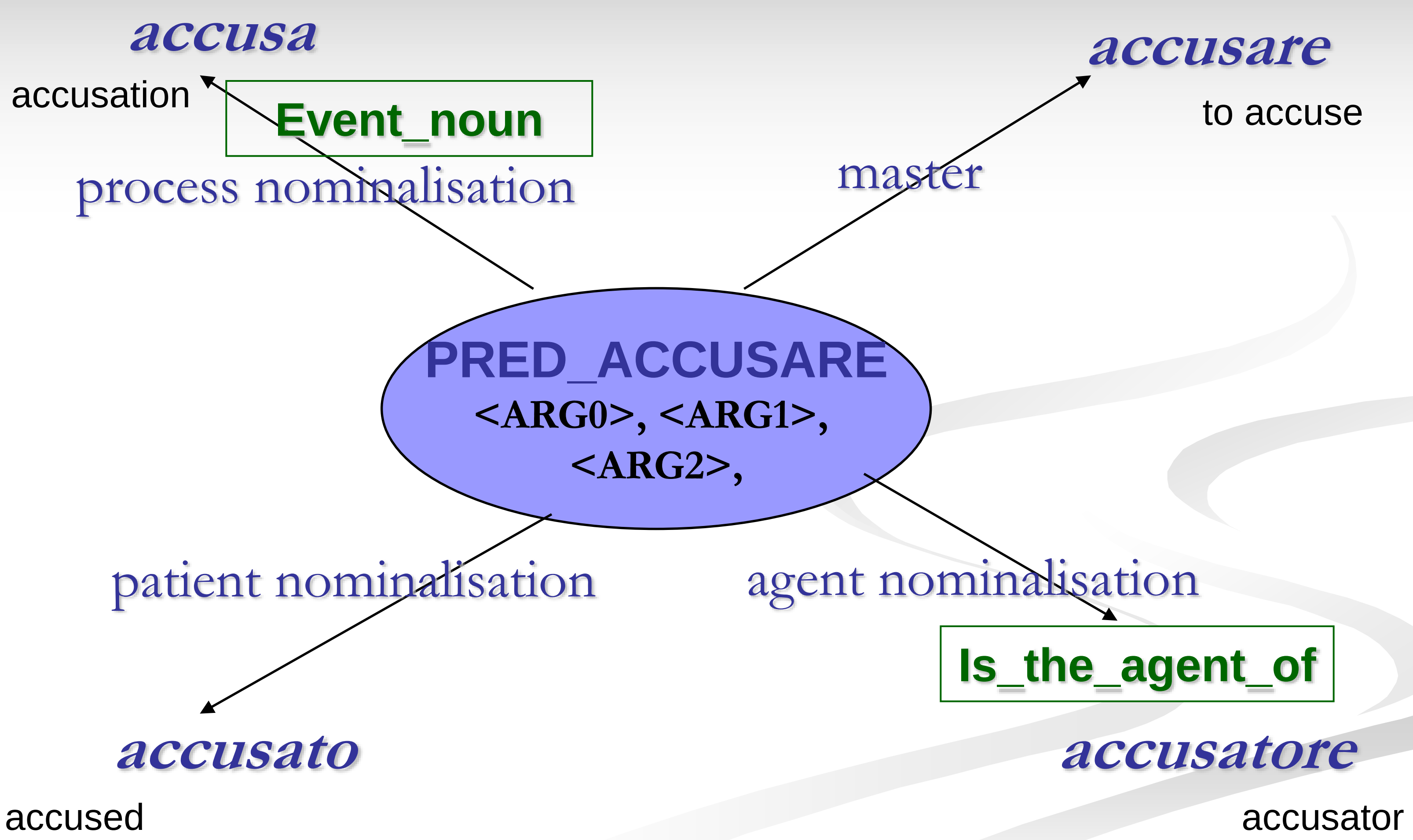
3. "hunger”, "hungry”, "hungrily”, "hunger” should be represented as simple, compound or complex UW's?

- They are not compounds, but derived: different
  - They are in some relation (simple or complex) with “hunger”

Another possibility:

- They can be simple UW
- And in addition have the relation marked

“Predicate - semantic unit(s)” link & Relations



accusa

accusation

Event_noun

process nominalisation

accusare

to accuse

master

PRED_ACCUSARE

<ARG0>, <ARG1>, <ARG2>,

patient nominalisation

agent nominalisation

Is_the_agent_of

accusato

accused

accusatore

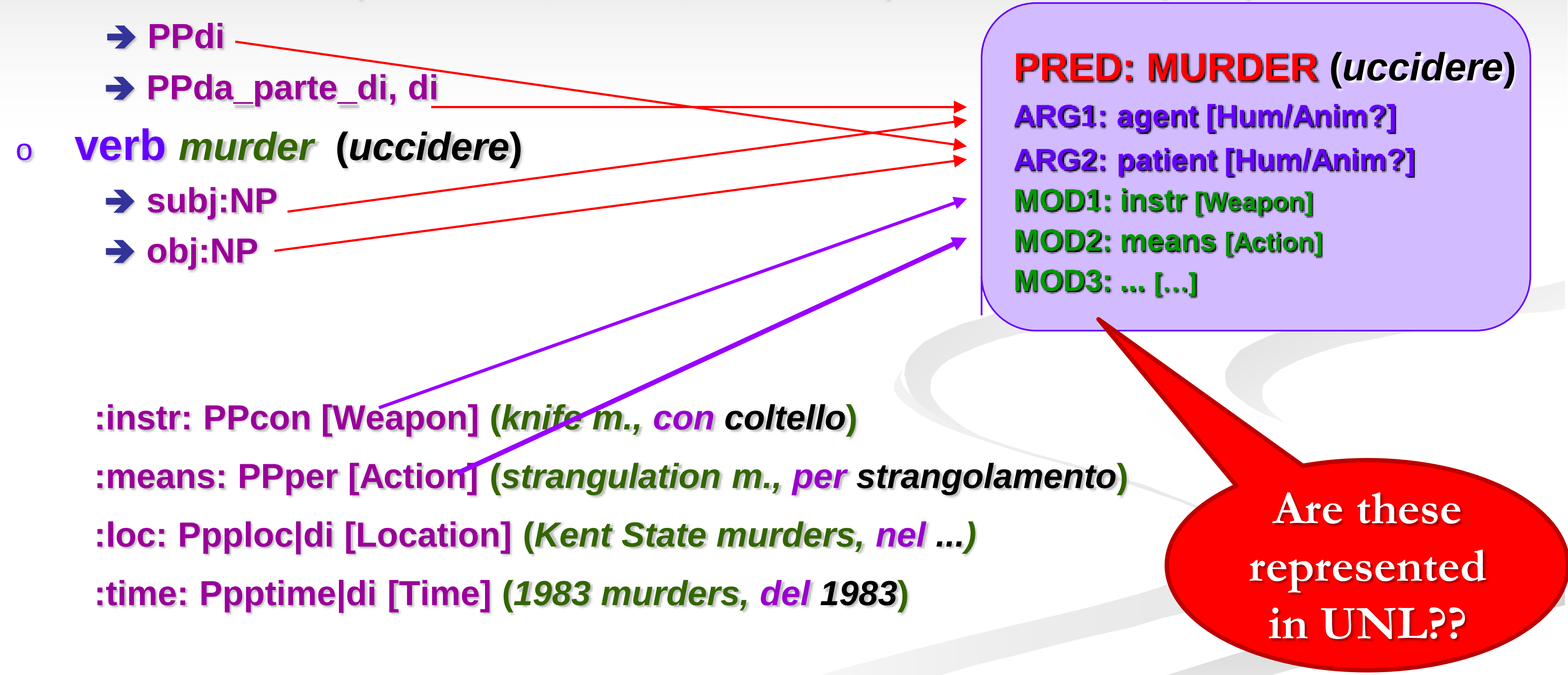
accusator

from Nilda Ruimy

SIMPLE:

Deverbal nominalisation:

- noun *murder* (*uccisione, delitto, omicidio* (different sem. pref.))
  - ➔ PPdi
  - ➔ PPda_parte_di, di
- verb *murder* (*uccidere*)
  - ➔ subj:NP
  - ➔ obj:NP



PRED: MURDER (*uccidere*)

ARG1: agent [Hum/Anim?]

ARG2: patient [Hum/Anim?]

MOD1: instr [Weapon]

MOD2: means [Action]

MOD3: ... [...]

:instr: PPcon [Weapon] (*knife m., con* coltello)

:means: PPper [Action] (*strangulation m., per* strangolamento)

:loc: Ppploc|di [Location] (*Kent State murders, nel* ...)

:time: Ppptime|di [Time] (*1983 murders, del* 1983)

Are these represented in UNL??

4. Antonyms such as "mortal" - "immortal", "hot" - "cold", "son" - "father" should be represented as a single UW (and the corresponding attributes) or as different UW's?

Similar as above:

- They can be simple UW
- And in addition have the relation marked

- Or is a minimal set of UWs needed??

5. "Farbfernsehgerät" ("color television set", in German) should be represented as a simple or complex UW?

Given the principle:

- ❑ The UNL must be independent from any particular natural language

- It should be a complex UW?

But

- In some language it may be expressed in one word
- It denotes a specific entity, and it has a specific meaning .....
- See "ferro da stiro" (iron)

➧ **See Interannotator agreement**

Suggestion

Compounds & Idioms

Locutions & Figurative usages

- **Where is the boundary of the MWE?**
  - "*andare_a_piedi*" vs. *andare* (Pos V) *a_piedi* (Pos Adv.loc).?

- *due lavoratori su tre* ***sono a casa*** (= essere disoccupato)

  [the collocation with '*lavoratori*' disambiguates the expression]
- *uomo [di polso]*

- If annotation of individual components, loss of the semantic contribution of the MWE
  - *acquistare un oggetto a buon* (Pos A) *mercato* (Pos S) !!

Noun Compounds/Complex Nominals …are pervasive

- There is a motivation in most N+N construction:
  - the context provides it

Theory based approaches

- The **FrameNet** (**SIMPLE**) way
  - appeal to **specific frame structures** (**qualia structures**) **associated with the head noun,**
  - determine from corpus attestations **which frame elements** (**qualia**) can get instantiated **as a modifier word**

- ***"container":*** complex nominals can specify:
  - **material** *(aluminium c., glass c., …)*
  - **contents** *(food c., trash c., …)*
  - **size** *(3 quart c., …)*
  - **function** *(shipping c., storage c., …)*
  - **…**

Noun Compounds/Complex Nominals & multidimensional semantic approaches

a. FrameNet

"***Container***" Frame Structure: **Frame Elements**:

- Material: *aluminum container, glass c., metal c., tin c.*
- Contents: *food container, beverage c., trash c., water c., milk c., fuel c.*
- Size: *3 quart container*
- Function: *shipping container, storage c.*

b. SIMPLE

Qualia Relations of "***container***" as used in compounds:

- Constitutive: *made_of* [MATERIAL] *aluminum container, glass c., metal c., tin c.*
- Telic: *contains* [ENTITY] *food container, beverage c., trash c., water c., milk c., fuel c.*
- Constitutive:*size* [QUANTITY] *3 quart container*
- Telic:*is_used_for* [EVENT] *shipping container, storage c.*

Complex Nominals

E.g. *knife (coltello)* triggers:

- a "cutting frame" (FrameNet)
- specific (SIMPLE) dimensions of meaning

SIMPLE Extended Qualia structure

for the interpretation of the semantic relation betw. Ns

(internal relational structure of MWE)

butcher's knife	***(coltello da macellaio)***	➔ TELIC	(used_by)	Y [Human]	➔ PPda
plastic knife	***(coltello di plastica)***	➔ CONST	(made_of)	X [Material]	➔ PPdi
table knife	***(coltello da tavola)***	➔ TELIC	(used_in)	Z [Location]	➔ PPda
hunting knife	***(coltello da caccia)***	➔ TELIC	(used_in_activity)	E[Activity]	➔ Ppda

piatto di legno	➔ CONST	(made_of)	X	[Material]	➔ PPdi
piatto di pasta	➔ CONST	(contains)	X	[Food]	➔ PPdi

PP disambig.

Difficult task to answer too specific issues/questions

- **If** we have to leave the principles untouched, the model & general approach as given,
- We only can speak about implementation details …
- Difficult to change details
- So I prefer to touch the issues in a different way

And

- In a moment to hint at some general principles & recommendations for LRs & lexicons

Other reflections

- Present some other examples
- To see if some lesson can be learnt
  - ➧… Some small suggestions

    Mapping UWs – Individual Languages words:

    Mapping e.g. to WordNet, or other Ontologies?

Some questions:

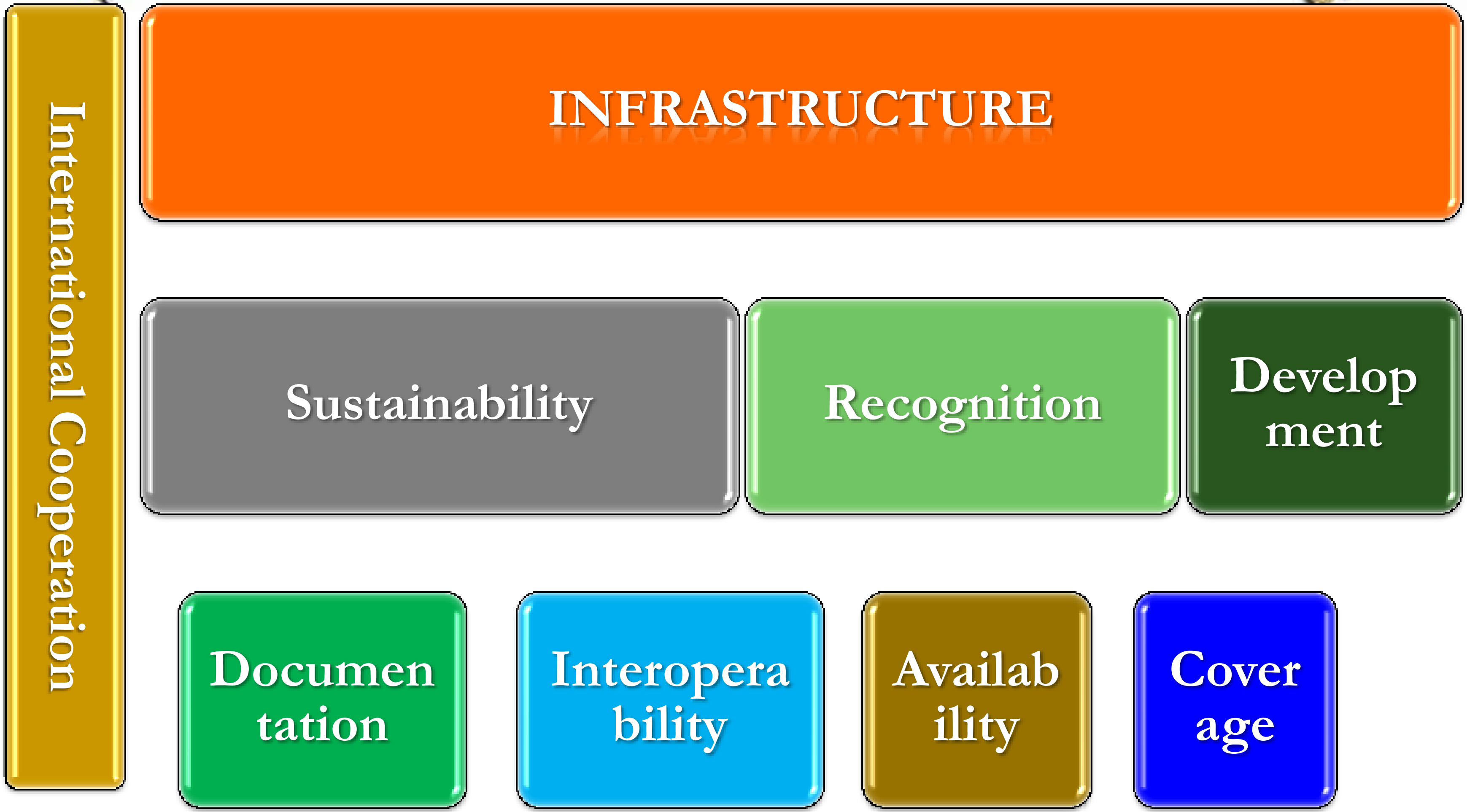
- **Is there a model behind?**
- **Has it grown in a "principled" way?**
- **Are specs clear enough?**
- **Interannotator agreement?** Lexicon encoders agreement?
- **Consistency?**

Comparison with statement from Indian national program @ LREC Workshop

- A lot of **attention to infrastructural and policy issues**, coordination, **standards & interoperability**
- **Before starting building, in the planning phase**
  - Also because of the complexity
  - Use of de facto standards, e.g. WordNet
  - Common platforms
  - Evaluation
- Establishing a model that could be reused more globally

Good model

FLaReNet Recommendations
A comprehensive perspective



International Cooperation

INFRASTRUCTURE

Sustainability

Recognition

Development

Documentation

Interoperability

Availability

Coverage

Resource Interoperability

"Design and set up an interoperability framework for LRT"

- Facts
  - The **lack** of interoperability and compliance with standards **costs a fortune**
  - "Why should I care?"
  - An essential **prerequisite for successful data exploitation** of the enormous amount of data
- Actions to be taken
  - **Encourage/enforce use of best practices** or standards in LR production projects
  - Make **standards operational** and put them in use
  - **Invest** in standardisation activities
  - Identify new **mature areas** for standardisation and promote joint efforts between **R&D and industry**
  - **→ RDF for LLOD**

Suggestion

LMF - ISO

- Specifically designed to accommodate as many models of lexical representation as possible
- Its pros:
  - **Meta-model**: a high-level specification ISO24613
  - **Data Category Registry**: low-level specifications ISO12620
- Not a *monolithic* model, rather a *modular* framework
  - LMF library provides the hierarchy of lexical objects (with structural relations among them)
  - Data Category Registry provides a library of descriptors to encode linguistic information associated to lexical objects (N.B. Data Categories can be also user-defined)

The field is mature

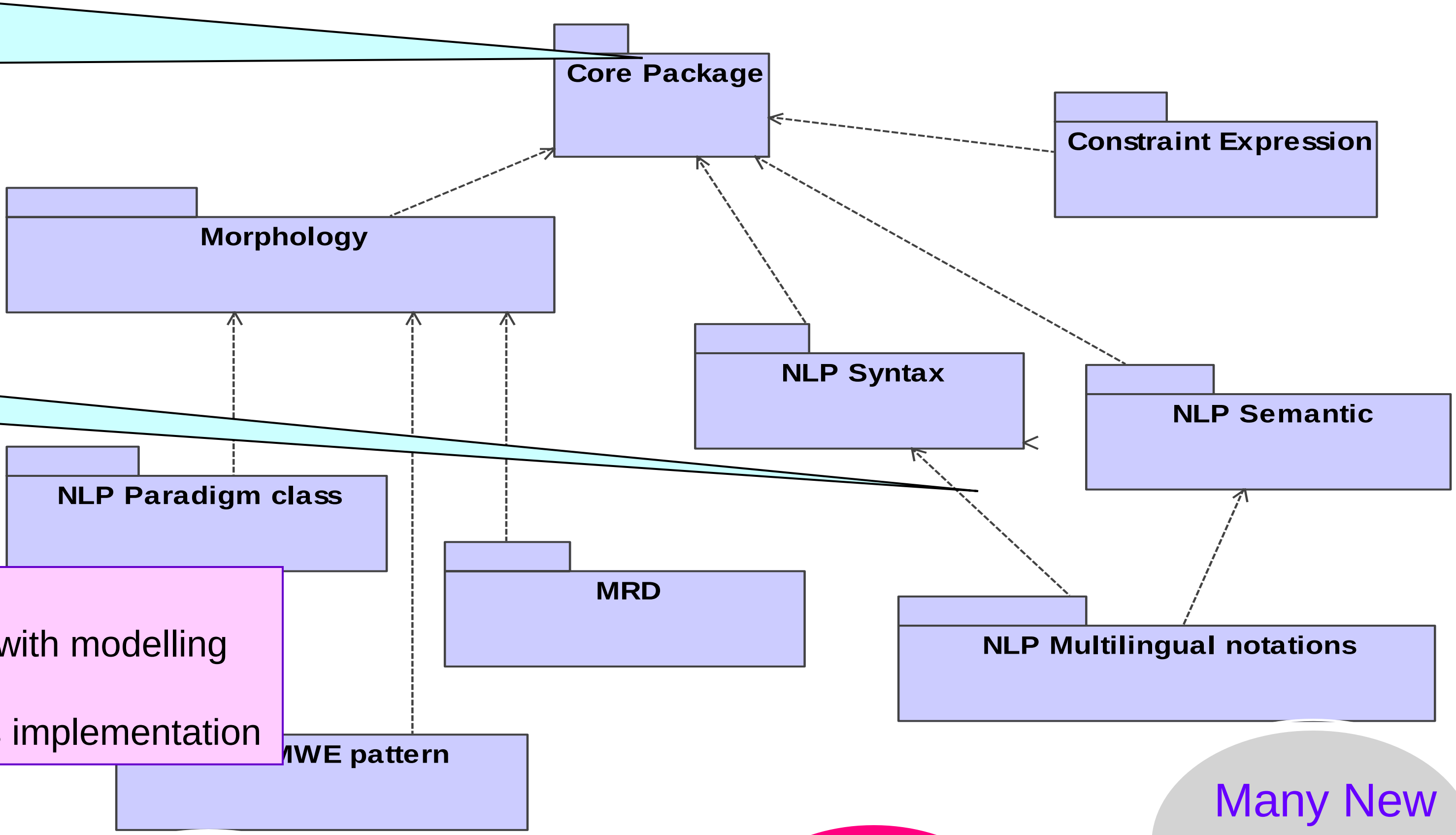
ISO LMF
Lexical Markup Framework

The field is mature

Builds on EAGLES/ISLE

Structural skeleton, with the basic hierarchy of information in a lexical entry

+ various extensions



- Modular framework
- LMF specs comply with modelling UML principles
- an XML DTD allows implementation

LIRICS

eContent

NEDO Asian Lang.uages

NICT Language-Grid Service Ontology

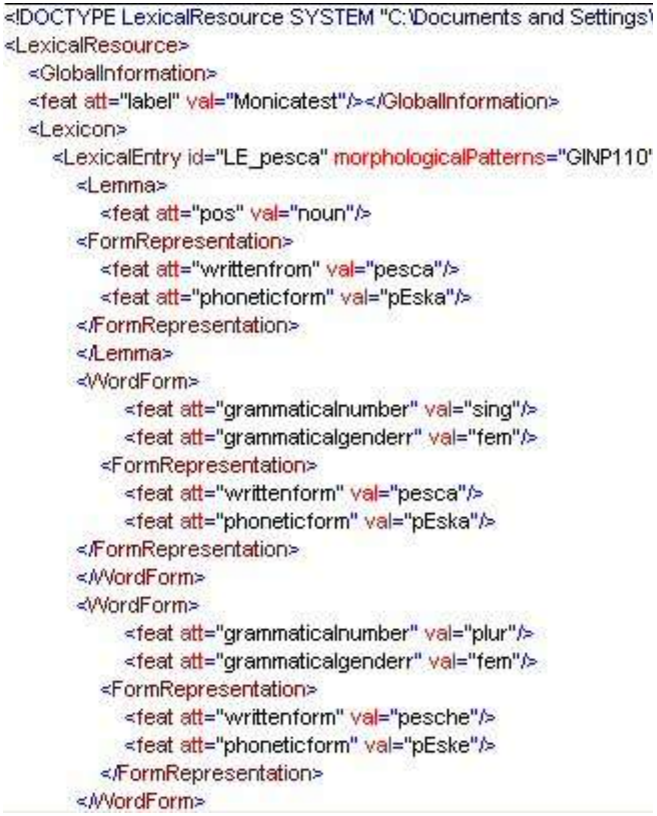
ICT KYOTO

LexInfo

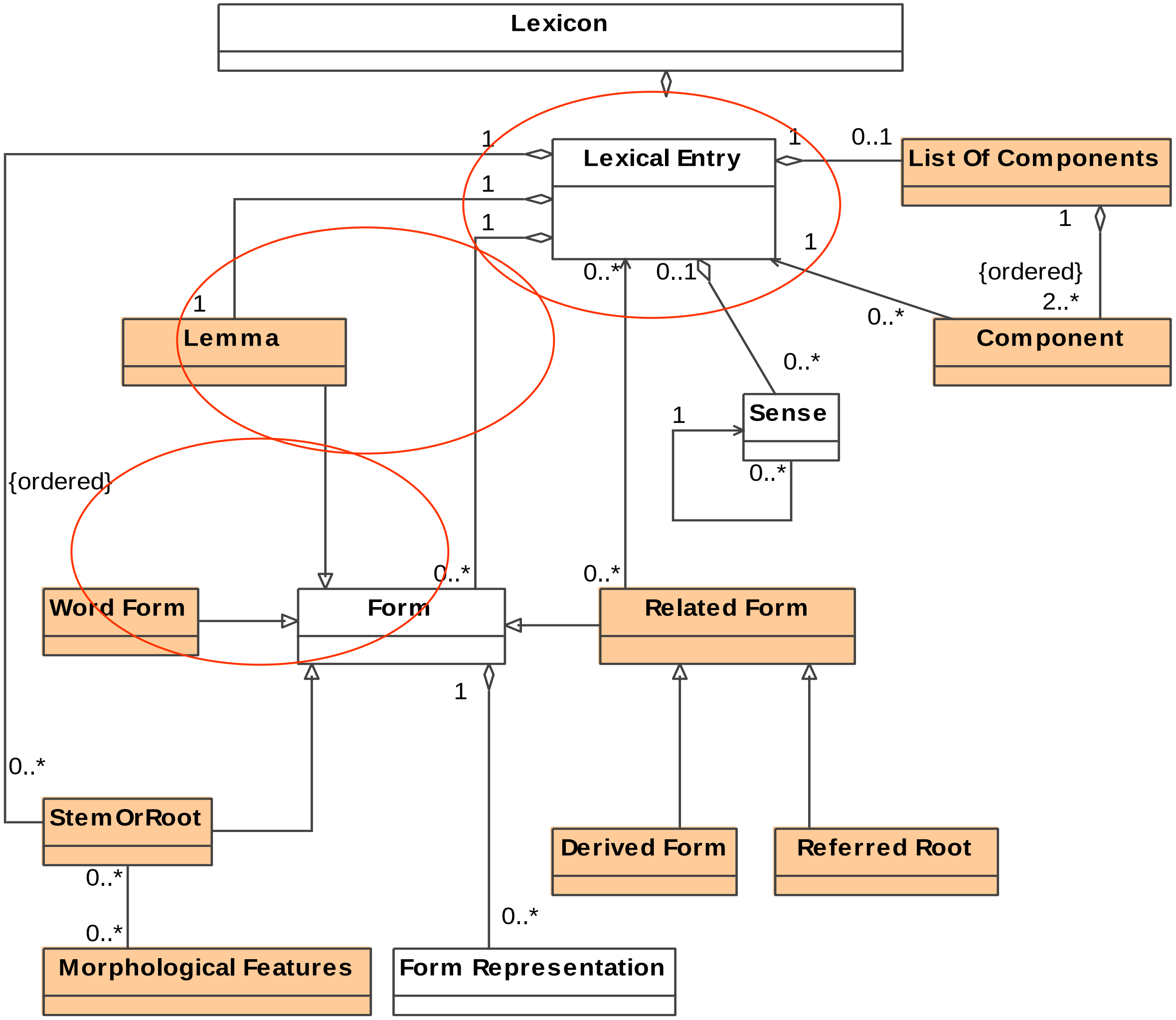
Many New initiatives …

Principles of LMF:

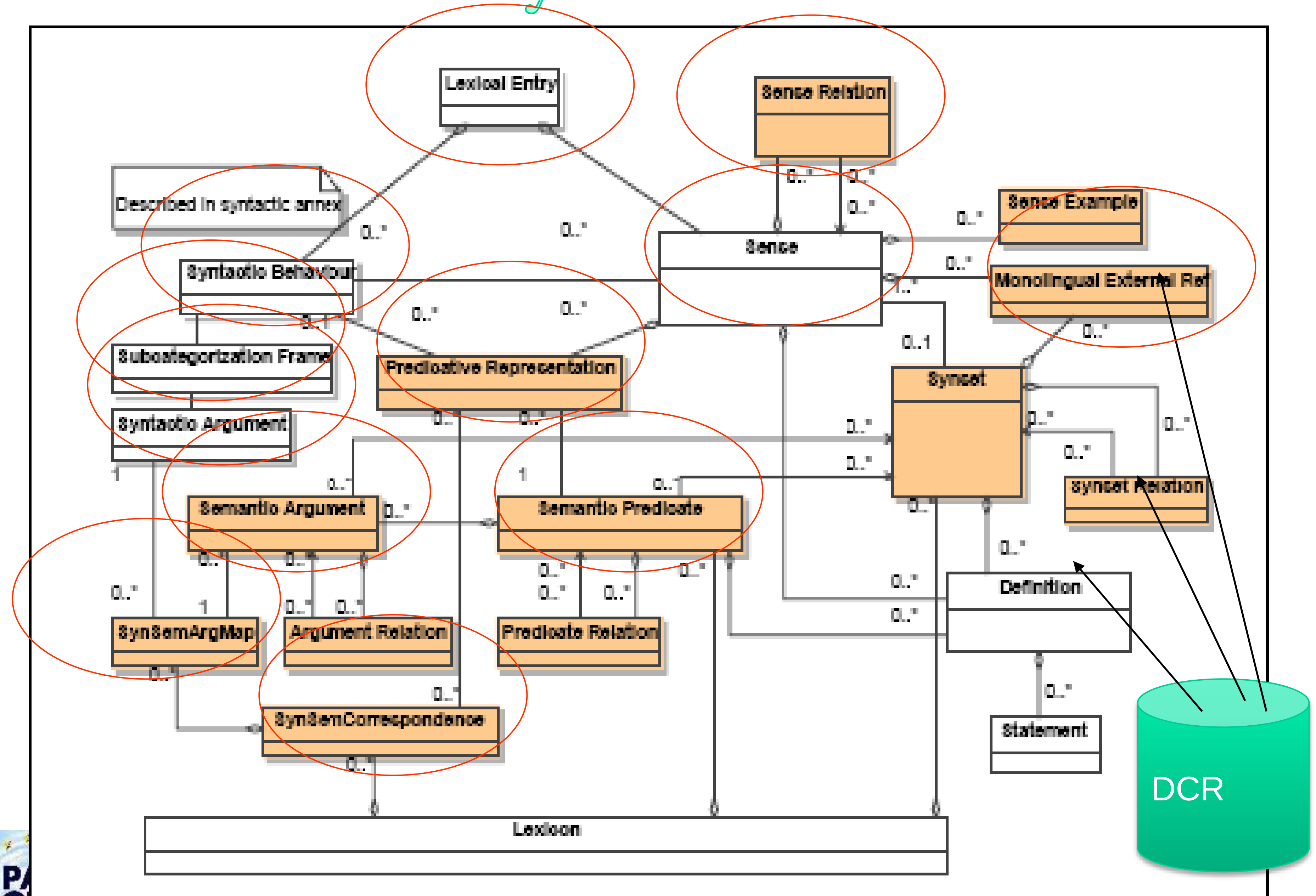
from very simple lexicons …



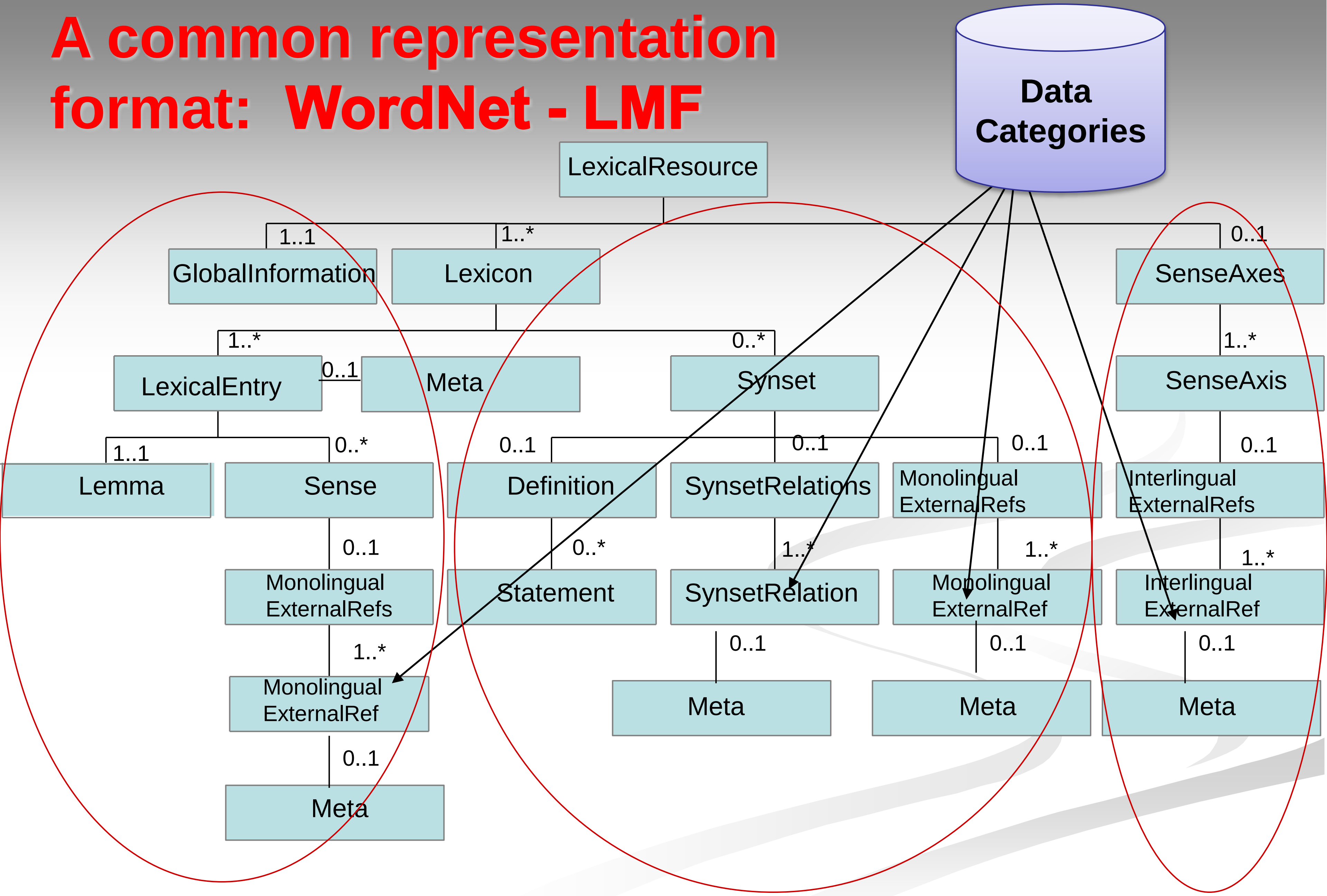
```
<!DOCTYPE LexicalResource SYSTEM "C:\Documents and Settings\
<LexicalResource>
  <GlobalInformation>
  <feat att="label" val="Monicatest"/></GlobalInformation>
  <Lexicon>
    <LexicalEntry id="LE_pesca" morphologicalPatterns="GINP110'
      <Lemma>
        <feat att="pos" val="noun"/>
      <FormRepresentation>
        <feat att="writtenfrom" val="pesca"/>
        <feat att="phoneticform" val="pEska"/>
      </FormRepresentation>
      </Lemma>
      <WordForm>
          <feat att="grammaticalnumber" val="sing"/>
          <feat att="grammaticalgenderr" val="fem"/>
        <FormRepresentation>
          <feat att="writtenform" val="pesca"/>
          <feat att="phoneticform" val="pEska"/>
      </FormRepresentation>
      </WordForm>
      <WordForm>
          <feat att="grammaticalnumber" val="plur"/>
          <feat att="grammaticalgenderr" val="fem"/>
        <FormRepresentation>
          <feat att="writtenform" val="pesche"/>
          <feat att="phoneticform" val="pEske"/>
        </FormRepresentation>
      </WordForm>
```



to very rich ones …



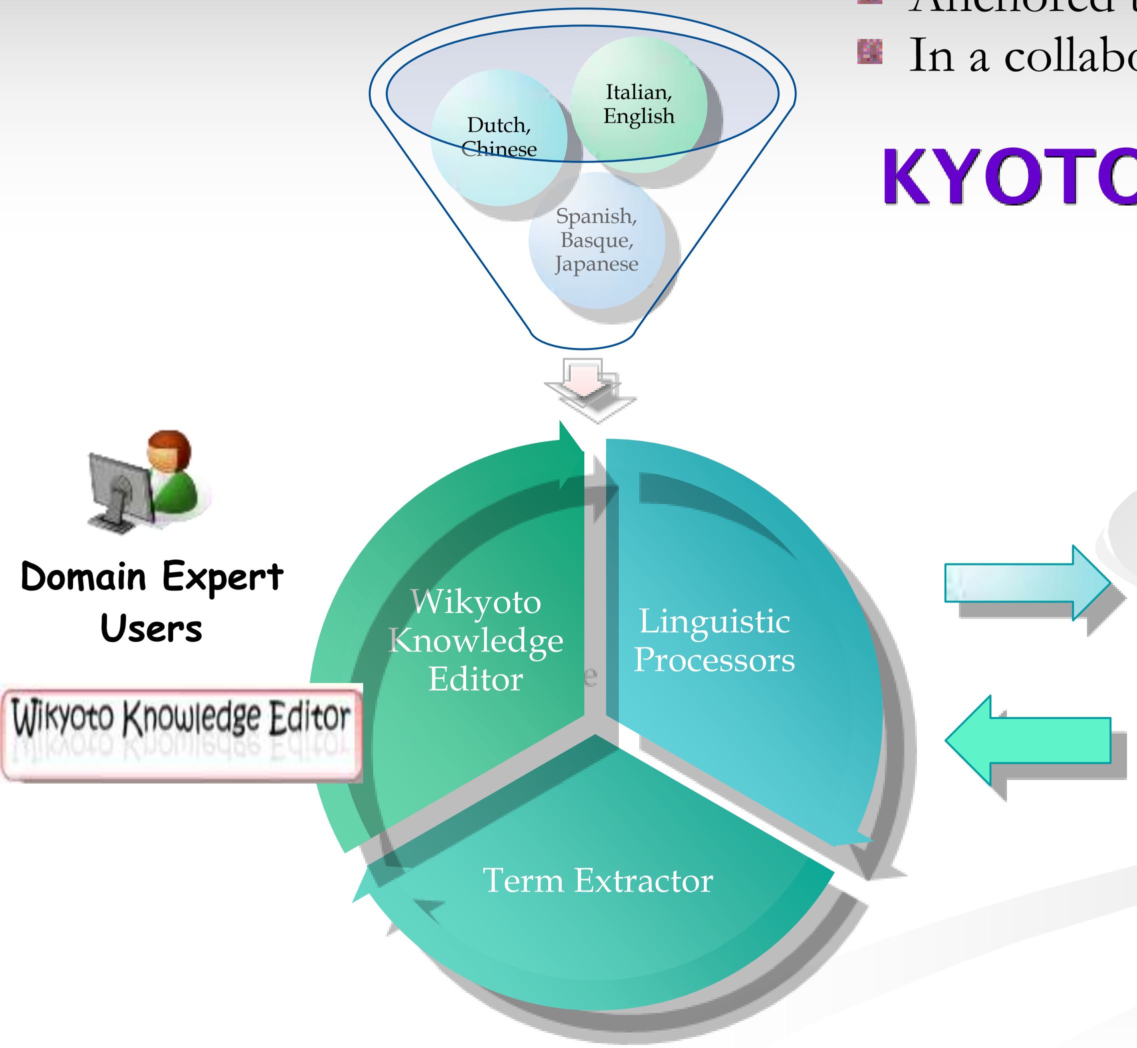
A common representation format: WordNet - LMF



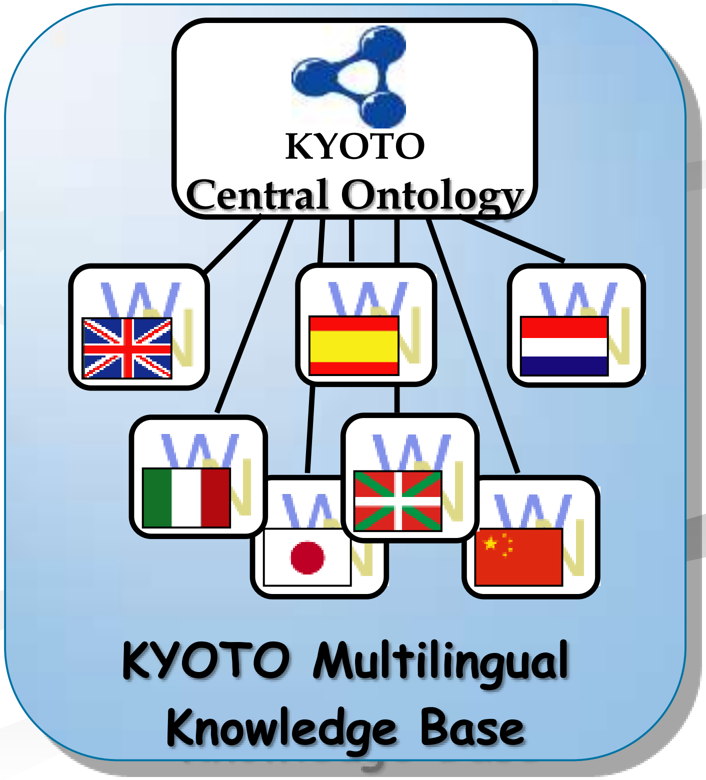
from Monica Monachini

Collaborative Platform

- Automatic extraction of concepts
- Validation & enrichment of domain WordNets
- Anchored to the ontology by domain experts
- In a collaborative platform

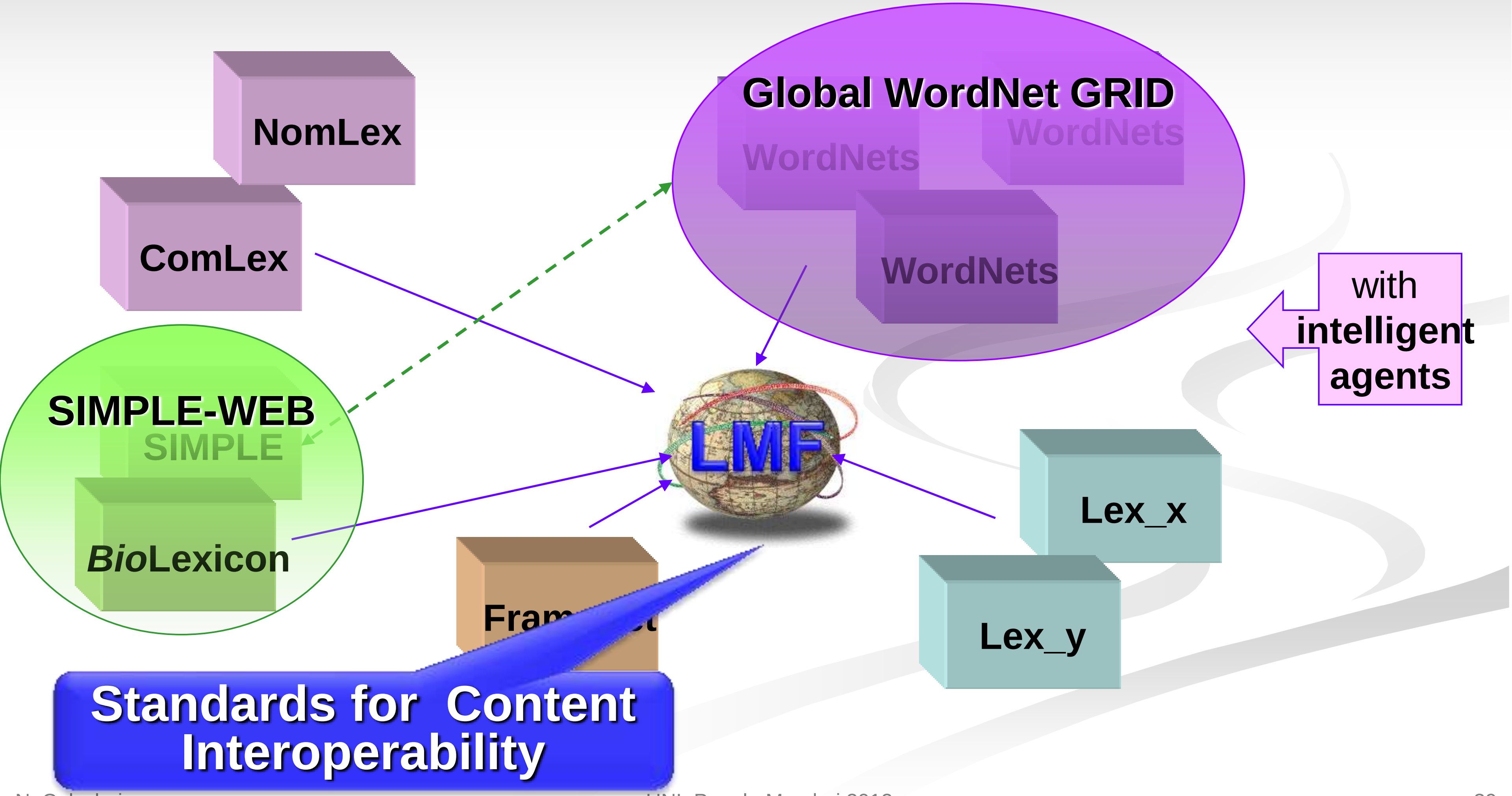


KYOTO Knowledge Base



Standards Lexical WEB

As a critical step for **semantic mark-up** in the Semantic Web



Standards for Content Interoperability

Resource Development

"Define a reference model for future Language Resource development"

- Facts
  - Lack of a model for proper and effective development of new resources
  - Tendency to start from scratch
- **Actions to be taken**
  - Ensure **strong public and community support** to definition and dissemination of **resource production best practices**
  - **Go Green**: enforce recycling, reusing and repurposing
  - Work towards the **full automation** of LR data production
  - **Invest in Web 2.0/3.0 methods for collaborative creation and extension** of high-quality resources, also as a means to achieve better coverage

Story about BIG DATA

Open Data

i.e. the backstage
Not in the forefront wrt applications

Keywords:

- LR sharing/linking/integrating/reusing/ …
- "Content" interoperability → towards Knowledge Resources
- Paradigm of accumulation of knowledge so successful in more mature disciplines

Collaborative building of LRs

- A Unified Framework for (future) LRs & (old?) SW (LLOD)?
  - Cross- fertilisation
  - New methodology of work
  - Interoperability acquires even more value

Infrastructural issues

Rationale

Accumulation of **massive amounts** of

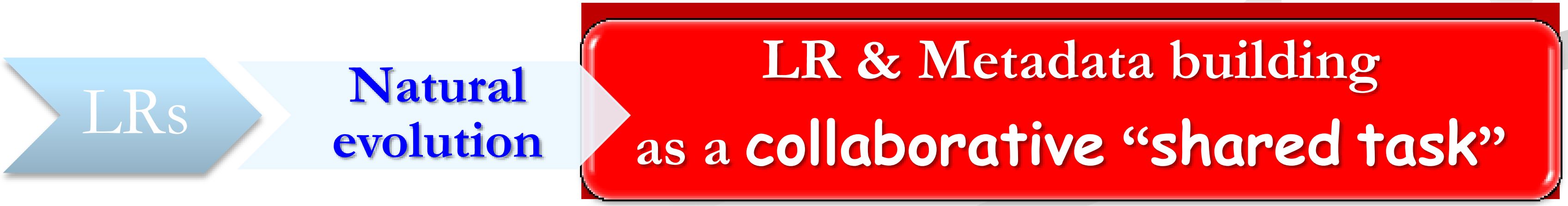
- **multi-dimensional data** &
- **meta-data**

is a key to foster advancement

The **history of LRs** brings us through concepts such as

- **Reusability**
- **Integration**
- **Standards and Interoperability**
- **Cooperative projects**
- **Subsidiarity**
- **Infrastructural role of LRs**
- **Sharing**
- …

How these fit in UNL?



LRs

Natural evolution

LR & Metadata building

as a collaborative "shared task"

Distributed Language Services

A scenario implying:

- content interoperability standards
- supra-national cooperation
- architectures enabling accessibility

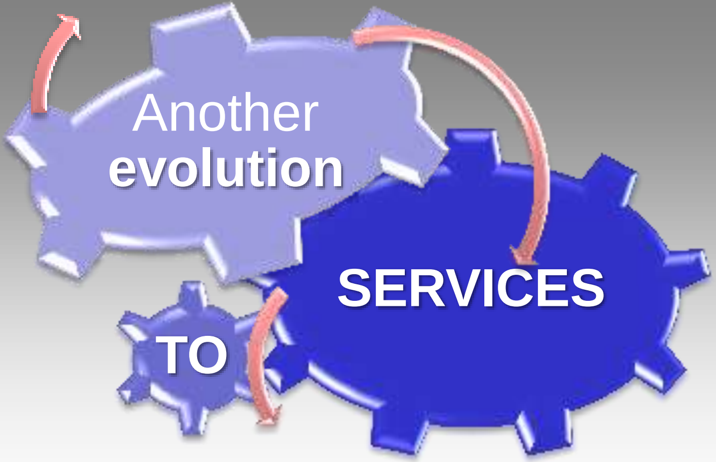
Enabling:

- Create new resources on the basis of existing
- Exchange & integrate information across repositories
- Compose new services on demand

Collaborative & collective/social development & validation,

cross-resource integration & exchange of information

USE



LRs as services &

Services around LRs

- **LRs as services**
  - Composite access
  - Web-services for Visualisation, Analysis, …
  - Extracting, Adapting, Merging, Linking, …
  - …

- **Services around LRs**
  - Describing with MD
  - Sharing: Authentication, …
  - Legal: licensing, …
  - Web-services for Collecting, Crawling, Cleaning, Linking, Integrating, Clustering, …
  - Inventorying
  - Converting (around Interoperability)
  - Annotating , (Content) Analysing, Acquiring info, …
  - Adapting , Repurposing, Evaluating,
  - Crowdsourcing
  - Translating, Localising, …
  - Summarising , Mining, …
  - Understanding, …

Resource Infrastructure

- **Facts**
  - Need for facilities supporting seamless access, use, re-use and trust of data
  - Coordination among infrastructural initiatives is needed
- **Actions to be taken**
  - Build a **sustainable facility for discovering, accessing and sharing data and tools**
  - Establish **international hub of resources and technologies** for speech and language **services**, **– Pooling of services**, **L-Apps**

International Cooperation

“Promote synergies among initiatives at international level”

And communities!

- **Facts**
  - Cooperation among countries and programs is essential to drive the field forward in a coordinated way and avoid duplication of efforts and fragmentation
- **Actions to be taken**
  - Establish an **International Forum** to share information, discuss **future policies and priorities on a global scale**
  - **Share** the **effort** for **production** of LRs between international bodies and individual countries
  - Maintain a **public survey** on the LT and LR situation **worldwide**