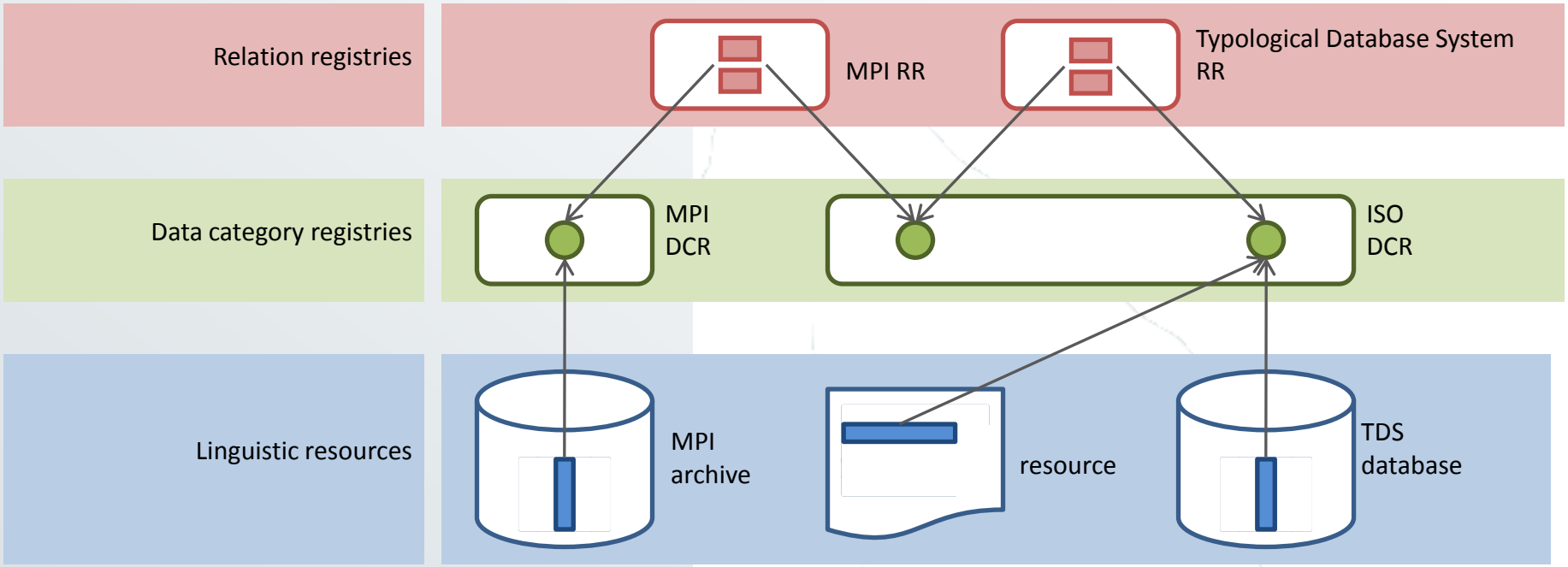
www.isocat.org

RELcat and friends

Vision


Relation registries
MPI RR
Typological Database System RR
Data category registries
MPI DCR
ISO DCR
Linguistic resources
MPI archive
resource
TDS database

How can you use a Data Category Registry?

- You can:
  - Find Data Categories relevant for your resources and embed references to them so the semantics of (parts of) your resources are made explicit
    - This can be supported by tools you use, e.g., ELAN, LEXUS and the CMDI Component Editor directly interact with ISOcat
  - Interact with Data Category owners to improve (the coverage of) their Data Categories
  - Create (together with others) new Data Categories and/or selections needed for your resources and share those
  - Submit (your) Data Categories for standardization
  - Free of charge
  - Grass roots approach

Referencing Data Categories

- Each Data Category should be uniquely identifiable
  - Ambiguity: different domains use the same term but mean different 'things'
  - Semantic rot: even in the same domain the meaning of a term changes over time
  - Persistence: for archived resources Data Category references should still be resolvable and point to the specification as it was at/close to time of creation
- Persistent IDentifiers
  - ISO 24619:2011 Language resource management -- Persistent identification and access in language technology applications
  - ISOcat uses 'cool URIs'
    - http://www.isocat.org/datcat/DC-1297 (/*grammaticalGender/*)

Where do you put these references?

- Preferably in a schema
- Or in the resource itself (redundant)
- Or in the metadata of the resource (less specific)

What is a schema?

- “comes from the Greek word "σχήμα" (skhēma), which means *shape,* or more generally, *plan*.” (wikipedia)
- A collection of building blocks and rules on how to combine them into a valid resource
  - XML document:
    - DTD, XML Schema, Relax NG, …
    - easy; see http://www.isocat.org/12620
  - Text document:
    - A grammar
      - Extended Backus–Naur Form (EBNF)
      - …
    - how to embed Data Category PIDs?
  - …

XML resource

```
<lmf:lexicon xml:lang="jp" alphabet="ipa">
   <lmf:entry>
        <lmf:lemma>
              <lmf:writtenForm>nihongo</…>
              …
        </…>
        …
   </…>
   …
</…>
```

www.isocat.org

XML Relax NG schema

```
<rng:attribute name="alphabet"
    dcr:datcat="http://www.isocat.org/datcat/…">
  <rng:value
  dcr:datcat="http://www.isocat.org/datcat/…">
    ipa
  </…>
  …
</…>
```

ISOcat
www.isocat.org

CGN/DCOI grammar

tag = pos '(' feat* ')‘

pos = 'N' | ' ADJ' | 'WW' | 'TW' | 'VNW' | 'LID' | 'VZ' | 'VG' | 'BW' | 'TSW'

feat = 'NTYPE' | 'GETAL' | 'GRAAD | 'GENUS | 'NAAMVAL' | 'POSITIE' | 'BUIGING | 'GETAL-N' | 'WVORM | 'PVTIJD | 'PVAGR' | 'NUMTYPE' | 'VWTYPE' | 'PDTYPE' | 'PERSOON' | 'STATUS' | 'NPAGR' | 'LWTYPE' | 'VZTYPE’ | 'CONJTYPE' | 'SPECTYPE'

NTYPE = 'soortnaam' | 'eigennaam'

GETAL = 'enkelvoud' | 'meervoud' | 'getal'

GRAAD = 'basis' | 'comparatief' | 'superlatief' | 'diminutief'

GENUS = 'genus' | 'zijdig' | 'masculien' | 'feminien' | 'onzijdig'

NAAMVAL = 'standaard' | 'nominatief' | 'oblique' | 'bijzonder' | 'genitief' | 'datief'

POSITIE = 'prenominaal' | 'nominaal' | 'postnominaal 'vrij'

BUIGING = 'zonder' | 'met-e' | 'met-s'

GETAL-N = 'zonder-n' | 'meervoud-n'

WVORM = 'persoonsvorm' | 'buigbaar' | 'innitief' | 'onvdw' | 'voltdw‘

PVTIJD = 'tegenwoordig' | 'verleden' | 'conjunctief'

PVAGR = 'enkelvoud' | 'meervoud' | 'met-t'

NUMTUPE = 'hoofdtelwoord' | 'rangtelwoord'

VWTYPE = 'pr' | 'persoonlijk' | 'reexief' | 'reciprook' | 'bezittelijk' | 'vb' | 'vragend' | 'betrekkelijk' | 'exclamatief' | 'aanwijzend' | 'onbepaald'

PDTYPE = 'pronomen' | 'adv-pronimen' | 'determiner' | 'gradeerbaar'

PERSOON = 'persoon' | '1' | '2' | '2v' | '2b' | '3' | '3p' | '3' | '3v' | '3o'

STATUS = 'vol' | 'gereduceerd' | 'nadruk'

NPAGR = 'agr' | 'evon' | 'rest' | 'evz' | 'mv' | 'agr3' | 'evmo' | 'rest3' | 'evf' | 'mv'

LWTYPE = 'bepaald' | 'onbepaald'

VZTYPE = 'initieel' | 'versmolten' | 'naal'

CONJTYPE = 'nevenschikkend' | 'onderschikkend'

SPECTYPE = 'afgebroken' | 'onverstaanbaar' | 'vreemd' | 'deeleigen' | 'meta' | 'commentaar' | 'achtergrond' | 'afkorting' | 'symbool' | 'dialect'

www.isocat.org

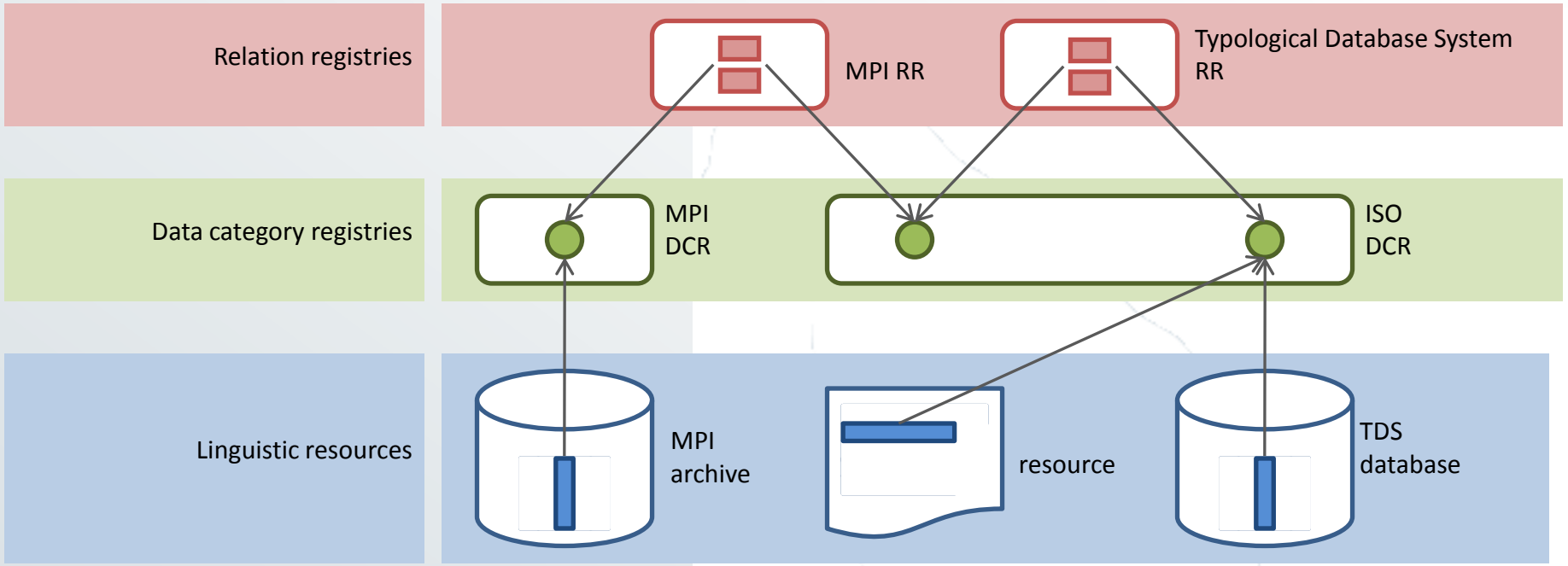
CGN/DCOI grammar with DC references

```
tag = pos '(' feat* ')'
# @datcat 'WW' http://www.isocat.org/datacat/DC-1424
# @datcat 'TW' http://www.isocat.org/datacat/DC-1334
# @datcat 'VG' http://www.isocat.org/datacat/DC-1226
# @datcat 'TSW' http://www.isocat.org/datacat/DC-2717
pos = 'N' | ' ADJ' | 'WW' | 'TW' | 'VNW' | 'LID' | 'VZ' | 'VG' | 'BW' | 'TSW'
feat = 'NTYPE' | 'GETAL' | 'GRAAD | 'GENUS | 'NAAMVAL' | 'POSITIE' | 'BUIGING | 'GETAL-N' | 'WVORM | 'PVTIJD | 'PVAGR' | 'NUMTYPE' | 'VWTYPE' | 'PDTYPE' |
      'PERSOON' | 'STATUS' | 'NPAGR' | 'LWTYPE' | 'VZTYPE' | 'CONJTYPE' | 'SPECTYPE'
NTYPE = 'soortnaam' | 'eigennaam'
GETAL = 'enkelvoud' | 'meervoud' | 'getal'
GRAAD = 'basis' | 'comparatief' | 'superlatief' | 'diminutief'
GENUS = 'genus' | 'zijdig' | 'masculien' | 'feminien' | 'onzijdig'
NAAMVAL = 'standaard' | 'nominatief' | 'oblique' | 'bijzonder' | 'genitief' | 'datief'
POSITIE = 'prenominaal' | 'nominaal' | 'postnominaal 'vrij'
BUIGING = 'zonder' | 'met-e' | 'met-s'
GETAL-N = 'zonder-n' | 'meervoud-n'
WVORM = 'persoonsvorm' | 'buigbaar' | 'innitief' | 'onvdw' | 'voltdw'
# @datcat PVTIJD http://www.isocat.org/datacat/DC-1286
# @datcat 'verleden' http://www.isocat.ord/datacat/DC-1347
# @datcat 'conjunctie' http://www.isocat.ord/datacat/DC-1843
PVTIJD = 'tegenwoordig' | 'verleden' | 'conjunctief'
PVAGR = 'enkelvoud' | 'meervoud' | 'met-t'
NUMTUPE = 'hoofdtelwoord' | 'rangtelwoord'
VWTYPE = 'pr' | 'persoonlijk' | 'reexief' | 'reciprook' | 'bezittelijk' | 'vb' | 'vragend' | 'betrekkelijk' | 'exclamatief' | 'aanwijzend' | 'onbepaald'
PDTYPE = 'pronomen' | 'adv-pronimen' | 'determiner' | 'gradeerbaar'
PERSOON = 'persoon' | '1' | '2' | '2v' | '2b' | '3' | '3p' | '3' | '3v' | '3o'
STATUS = 'vol' | 'gereduceerd' | 'nadruk'
NPAGR = 'agr' | 'evon' | 'rest' | 'evz' | 'mv' | 'agr3' | 'evmo' | 'rest3' | 'evf' | 'mv'
LWTYPE = 'bepaald' | 'onbepaald'
VZTYPE = 'initieel' | 'versmolten' | 'naal'
CONJTYPE = 'nevenschikkend' | 'onderschikkend'
SPECTYPE = 'afgebroken' | 'onverstaanbaar' | 'vreemd' | 'deeleigen' | 'meta' | 'commentaar' | 'achtergrond' | 'afkorting' | 'symbool' | 'dialect'
```

Or use:

- an XML representation for the schema
- an XML schema for an XML representation of the content

Vision


Relation registries
MPI RR
Typological Database System RR
Data category registries
MPI DCR
ISO DCR
Linguistic resources
MPI archive
resource
TDS database

Multiple DCRs?

- Actually we don't need multiple DCRs to have overlapping subsets
  - Overlaps are created due to
    - Data categories are typed, and might not have the type you need
    - External sets are imported just as they are
      - NKJP, GOLD, STTS, MDF, …
      - Only some take the effort to also provide mappings
    - There might be very fine differences between your data category and an existing one, and the owner doesn't want to adapt
- Still we would like to know that these data categories are the same or almost the same

Which data category types?

- TDGs give DC types based on some reference model:
  - Metadata: CMDI
  - Morphosyntax: LMF
  - Terminology: TBX
  - *POS field (closed DC) of the lexical entry "walk" gets the value 'verb' (simple DC)*
    - *PoS = 'verb'*
- If the DC type doesn't fit your needs:
  - *Verb (open DC) feature of a feature structure gets the value "walk"*
    - *Verb = 'walk'*
  - Unfortunately the DCR data model hasn't facilities yet to share a semantic core between various types
  - Create your own and add a sameAs relationship to RELcat

Relation Registry - RELcat

- The Relation Registry is basically a triple store:
  - Subject: resource 1
  - Predicate: relationship
  - Object: resource 2
- Example:

  # */first plural exclusive vernacular/* is-a */vernacular/*

  isocat:DC-1234 rel:subClassOf isocat:DC-4 .

  # */first person/* part-of */first plural exclusive vernacular/*

  isocat:DC-1 rel:partOf isocat:DC-1234 .

  # */plural/* part-of */first plural exclusive vernacular/*

  isocat:DC-2 rel:partOf isocat:DC-1234 .

  # */exclusive/* part-of */first plural exclusive vernacular/*

  isocat:DC-3 rel:partOf isocat:DC-1234 .

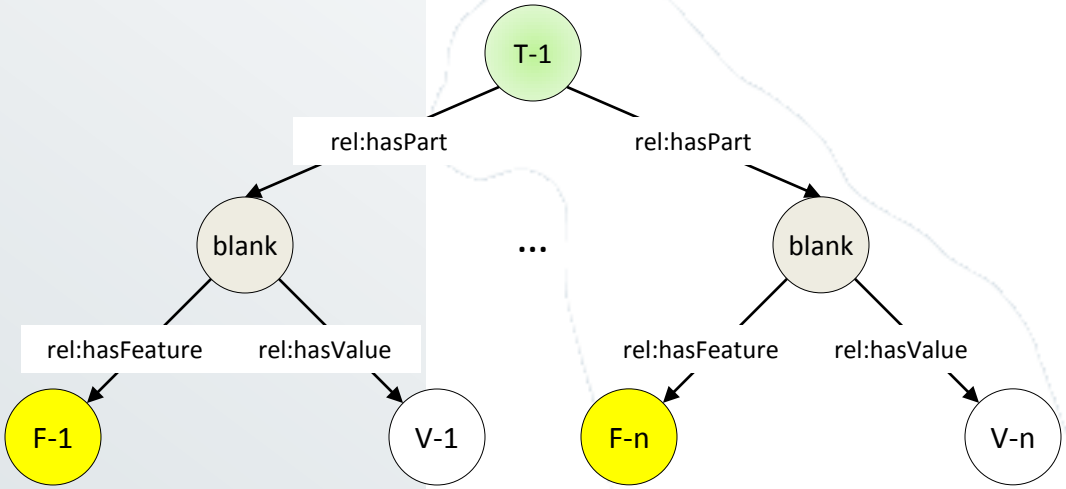
Relation type taxonomy

- rel:related
  - rel:sameAs (symmetric)
  - rel:almostSameAs (symmetric)
  - rel:narrower (inverse of rel:broader)
    - rel:superClassOf (inverse of rel:subClassOf)
  - rel:broader (inverse of rel:narrower)
    - rel:subClassOf (inverse of rel:superClassOf)
    - rel:partOf
      - rel:directPartOf
      - rel:indirectPartOf

Extensible taxonomy inspired by OWL and SKOS, which other relation types might be needed or other standard sets should be included?

! To cover *n*-ary relationships more advanced RDF constructs are probably needed.

n-ary relationships



RELcat status

- A first alpha version is running on the ISOcat test server
  - Only a web services interface, i.e., no user interface yet
  - Import of sets relations done by the system administrator
  - Initial sets are imported to support the CMDI semantic search
    - http://lux13.mpi.nl/relcat/rest/set/dc
    - http://lux13.mpi.nl/relcat/rest/set/cmdi
      - http://lux13.mpi.nl/relcat/rest/set/cmdi/relations?relation=sameAs&resource=dc:language
- For now your relations can be send to the system administrator in a simple spreadsheet with three columns
  - Subject Data Category
  - Relationship (see previous slide or suggest a new one)
  - Object Data Category

A new kitten: SCHEMAcat

- Resource schema's should be stored somewhere persistently
  - Get a PID, i.e., a handle
- These schema's can/should be annotated with data categories
  - SCHEMAcat → ISOcat
- These data categories will have (typed) relationships among eachother
  - SCHEMAcat → RELcat

Schema types

- SCHEMAcat should support any schema format
  - XML:
    - XML Schema
    - Relax NG
    - DTD
    - …
  - Text (how?):
    - EBNF
    - …
  - …

XML Relax NG schema

```
<rng:attribute name="alphabet"
    dcr:datcat="http://www.isocat.org/datcat/…">
  <rng:value
  dcr:datcat="http://www.isocat.org/datcat/…">
    ipa
  </…>
  …
</…>
```

CGN/DCOI grammar with DC references

www.isocat.org

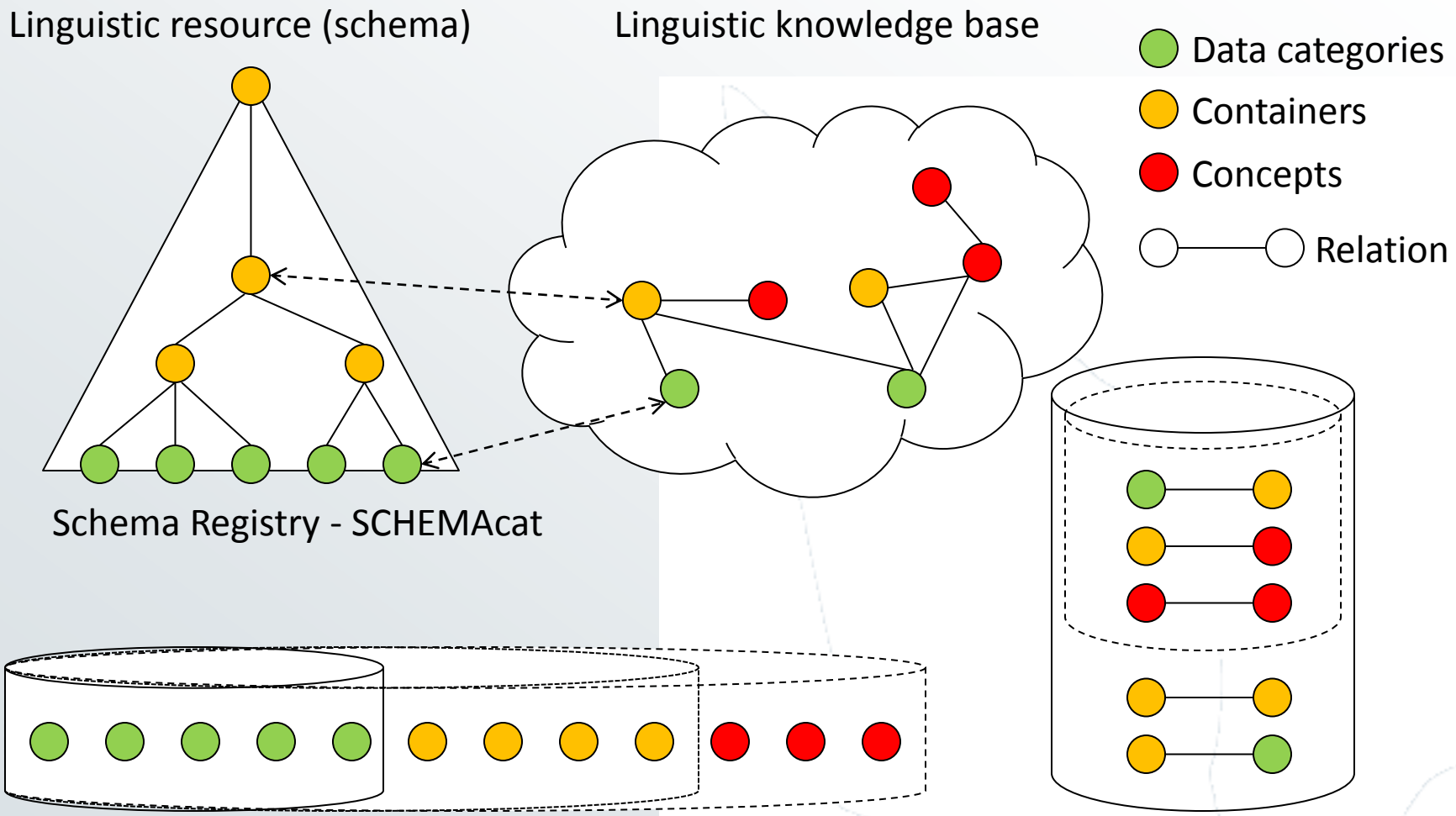
```
tag = pos '(' feat* ')'
# @datcat 'WW' http://www.isocat.org/datacat/DC-1424
# @datcat 'TW' http://www.isocat.org/datacat/DC-1334
# @datcat 'VG' http://www.isocat.org/datacat/DC-1226
# @datcat 'TSW' http://www.isocat.org/datacat/DC-2717
pos = 'N' | ' ADJ' | 'WW' | 'TW' | 'VNW' | 'LID' | 'VZ' | 'VG' | 'BW' | 'TSW'
feat = 'NTYPE' | 'GETAL' | 'GRAAD | 'GENUS | 'NAAMVAL' | 'POSITIE' | 'BUIGING | 'GETAL-N' | 'WVORM | 'PVTIJD | 'PVAGR' | 'NUMTYPE' | 'VWTYPE' | 'PDTYPE' |
      'PERSOON' | 'STATUS' | 'NPAGR' | 'LWTYPE' | 'VZTYPE' | 'CONJTYPE' | 'SPECTYPE'
NTYPE = 'soortnaam' | 'eigennaam'
GETAL = 'enkelvoud' | 'meervoud' | 'getal'
GRAAD = 'basis' | 'comparatief' | 'superlatief' | 'diminutief'
GENUS = 'genus' | 'zijdig' | 'masculien' | 'feminien' | 'onzijdig'
NAAMVAL = 'standaard' | 'nominatief' | 'oblique' | 'bijzonder' | 'genitief' | 'datief'
POSITIE = 'prenominaal' | 'nominaal' | 'postnominaal 'vrij'
BUIGING = 'zonder' | 'met-e' | 'met-s'
GETAL-N = 'zonder-n' | 'meervoud-n'
WVORM = 'persoonsvorm' | 'buigbaar' | 'innitief' | 'onvdw' | 'voltdw'
# @datcat PVTIJD http://www.isocat.org/datacat/DC-1286
# @datcat 'verleden' http://www.isocat.ord/datacat/DC-1347
# @datcat 'conjunctie' http://www.isocat.ord/datacat/DC-1843
PVTIJD = 'tegenwoordig' | 'verleden' | 'conjunctief'
PVAGR = 'enkelvoud' | 'meervoud' | 'met-t'
NUMTUPE = 'hoofdtelwoord' | 'rangtelwoord'
VWTYPE = 'pr' | 'persoonlijk' | 'reexief' | 'reciprook' | 'bezittelijk' | 'vb' | 'vragend' | 'betrekkelijk' | 'exclamatief' | 'aanwijzend' | 'onbepaald'
PDTYPE = 'pronomen' | 'adv-pronimen' | 'determiner' | 'gradeerbaar'
PERSOON = 'persoon' | '1' | '2' | '2v' | '2b' | '3' | '3p' | '3' | '3v' | '3o'
STATUS = 'vol' | 'gereduceerd' | 'nadruk'
NPAGR = 'agr' | 'evon' | 'rest' | 'evz' | 'mv' | 'agr3' | 'evmo' | 'rest3' | 'evf' | 'mv'
LWTYPE = 'bepaald' | 'onbepaald'
VZTYPE = 'initieel' | 'versmolten' | 'naal'
CONJTYPE = 'nevenschikkend' | 'onderschikkend'
SPECTYPE = 'afgebroken' | 'onverstaanbaar' | 'vreemd' | 'deeleigen' | 'meta' | 'commentaar' | 'achtergrond' | 'afkorting' | 'symbool' | 'dialect'
```

SCHEMAcat status

- In preparation …

ISOcat
www.isocat.org

A whole litter!



ISOcat status

- Standardization is slowly taking off
  - The Metadata TDG is starting next week
  - Terminology is planning the same
  - Their workflow and progress will be reported on the TC37 plenary and will hopefully spark some action of the other TDGs
- ISOcat keeps improving in general, although development effort should be moving more towards RELcat and SCHEMAcat …