

# **Patterns for thesaurus conversion to RDF/OWL**

Guus Schreiber  
Free University Amsterdam

## **Overview**

- Thesauri and thesauri standards
- Conversion process
  - Example: Union List of Artist Names
  - Example: WordNet 2.0
- SKOS model for thesauri
- Issues with respect to (adding) semantics

## Acknowledgements

- Conversion process: Mark van Assem, Jan Wielemaker, Bob Wielinga
- SKOS: Alistair Miles, Dan Brickley
- LSCOM examples: Cees Snoek, Laura Hollink
- W3C Semantic Web Best Practices & Deployment Working Group

3

## Thesauri / vocabularies

- Large bodies of domain-specific knowledge that represent consensus in particular domains
- Typically weak semantic structure
- Often lots of implicit semantics available
- Representation is typically relational database and/or XML
- Semantic Web Challenge showed that thesauri are important resources for SW applications

4

## Example thesauri

- Domain-specific vocabularies
  - Medicine: UMLS, SNOMED, Galen
  - Art history: AAT, ULAN
  - Geography: TGN
  - Food: AgroVoc,
- Generic vocabularies
  - Lexical vocabularies: WordNet, FrameNet
  - Units and dimensions,
  - Currencies, country codes, ...

5

## ISO standard for representing thesauri

- Term
  - Preferred term (USE)
  - Non-preferred term (USED FOR)
- Hierarchical relation between terms
  - Broader/narrower term (BT/NT)
    - Generic
    - Partitive
- Association between terms (RT)

6

## Conversion process


- Two steps
- Step 1: “As is” conversion
  - Keep original names
  - Make implicit semantics explicit (but this can be hard to determine)
  - Decisions on whether to keep all information
- Step 2: adding semantics
  - Separate file(s)
  - Interpretation of thesauri elements, e.g. hyponym relation as **rdfs:subClassOf**
  - May require (lots of) additional research

7


## Example thesaurus: ULAN

- 300,000 entries
- Consists of records of “Subjects” (artists and art institutions), with biographical information (place/time birth/death) and relations to other artists (student-of, ...)
- Large XML file with all data
- Basic representation:
  - association links between subjects
  - preferred/non-preferred terms relations between subjects and terms


8


Research


[Research Home](#) » [Conducting Research](#) » [Union List of Artist Names](#) » Full Record Display


Union List of Artist Names® Online  
Full Record Display

[New Search](#)
[Previous Page](#)
[Help](#)

Click the  icon to view the hierarchy.

**ID: 500000351** **Record Type: [Person](#)**

 **Koninck, Philips de** (Dutch painter and draftsman, 1619-1688)

**Note:** History and portrait painter who is today most well-known for his naturalistic panoramic bird's-eye view landscapes.

**Birth and Death Places:**  
Born: [Amsterdam \(North Holland, Netherlands\)](#) (inhabited place)  
Died: [Amsterdam \(North Holland, Netherlands\)](#) (inhabited place)

**Related People or Corporate Bodies:**  
related to (familial) .... [Koninck, Salomon](#)  
(Dutch painter, printmaker, and draftsman,  
..... 1609-1656) [500027532]  
sibling of .... [Koninck, Jacob, the elder](#)  
..... (Dutch painter and engraver, ca. 1616-1708) [500024292]  
student of .... [Rembrandt van Rijn](#)  
(Dutch painter, draftsman, and printmaker, 1606-1669)  
..... [500011051]

9

## XML fragment of ULAN

```

<Associative_Relationships>
  <Associative_Relationship>
    <Historic_Flag>NA</Historic_Flag>
    <Relationship_Type>
      1102/student of
    </Relationship_Type>
    <Related_Subject_ID>
      <VP_Subject_ID>500011051</VP_Subject_ID>
    </Related_Subject_ID>
  </Associative_Relationship>
</Associative_Relationships>

```

## Conversion issues

- XML and RDF/OWL are inherently different
  - XML = thesaurus document structure
  - RDF = thesaurus document content
- Redundant information in XML file
  - <Associative\_Relationships>**
  - <Historic\_Flag>NA</Historic\_Flag>**
- How to represent “student of”?
  - Subproperty of **Associative\_Relationship** is probably preferred
  - Needs to be derived from the data; not part of schema

11

## XML fragment of ULAN (2)

```
<Non-Preferred_Term>  
  <Term_Text>Koning, Philips Aerts. de</Term_Text>  
  <Term_ID>1500207734</Term_ID>  
  <Display_Order>34</Display_Order>  
  <Vernacular>Vernacular</Vernacular>  
</Non-Preferred_Term>
```

12

## Conversion issues

- Do we include all information in the conversion?
  - Display-order example
  - Source and revisions information
- Should each term have a URI?
- Making language explicit
  - “vernacular” means the string is written in the original language
  - Multi-linguality is an important issue for thesauri

13



## RDF/OWL Representation of WordNet

W3C Working Draft 19 June 2006

**This version:**

<http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/>

**Latest version:**

<http://www.w3.org/TR/wordnet-rdf/>

**Previous version:**

This is the first published version

**Editors:**

[Mark van Assem](#), Vrije Universiteit Amsterdam

[Aldo Gangemi](#), ISTC-CNR, Rome

14

## WordNet: internal representation

SynsetID    Order    LexForm    Type    SenseNum

s(108644031,1,'bed',n,3,2).  
s(108644031,2,'bottom',n,5,1).

s(102719813,1,'bed',n,1,51).  
s(102720436,1,'bed',n,2,3).

g(108644031,'(a depression forming the ground under a body of water; "he searched for treasure on the ocean bed")').  
g(102719813,'(a piece of furniture that provides a place to sleep; "he sat on the edge of the bed"; "the room had only a bed and chair")').  
g(102720436,'(a plot of ground in which plants are growing; "the gardener planted a bed of roses")').

15

## WordNet URI s

- What URIs should be chosen?
  - SynSet, WordSense, Word
- URI name:
  - ID? => difficult for human interpretation
  - Concatenated unique, human readable

**wn:synset-bank-noun-2**

First sense in synset denoted by second sense of "bank"

**wn:wordsense-bank-noun-1**

**wn:word-bank**

16



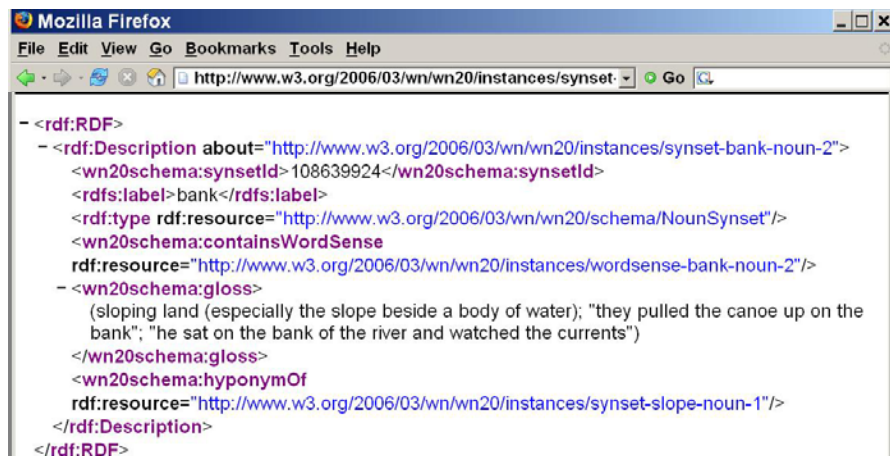
## Implicit WordNet semantics

*“The **ent** operator specifies that the second synset is an entailment of first synset. This relation only holds for verbs. “*

- Example: [breathe, inhale] entails [sneeze, exhale]
- Semantics (OWL statements):
  - Transitive property
  - Inverse property: entailedBy
  - Value restrictions for VerbSynSet (subclass of SynSet)

17

## Query for WordNet URI returns “concept-bounded description”

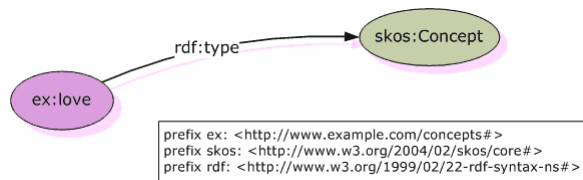


```
- <rdf:RDF>
- <rdf:Description about="http://www.w3.org/2006/03/wn/wn20/instances/synset-bank-noun-2">
  <wn20schema:synsetId>108639924</wn20schema:synsetId>
  <rdfs:label>bank</rdfs:label>
  <rdf:type rdf:resource="http://www.w3.org/2006/03/wn/wn20/schema/NounSynset"/>
  <wn20schema:containsWordSense
    rdf:resource="http://www.w3.org/2006/03/wn/wn20/instances/wordsense-bank-noun-2"/>
- <wn20schema:gloss>
  (sloping land (especially the slope beside a body of water); "they pulled the canoe up on the
  bank"; "he sat on the bank of the river and watched the currents")
</wn20schema:gloss>
  <wn20schema:hyponymOf
    rdf:resource="http://www.w3.org/2006/03/wn/wn20/instances/synset-slope-noun-1"/>
</rdf:Description>
</rdf:RDF>
```

18

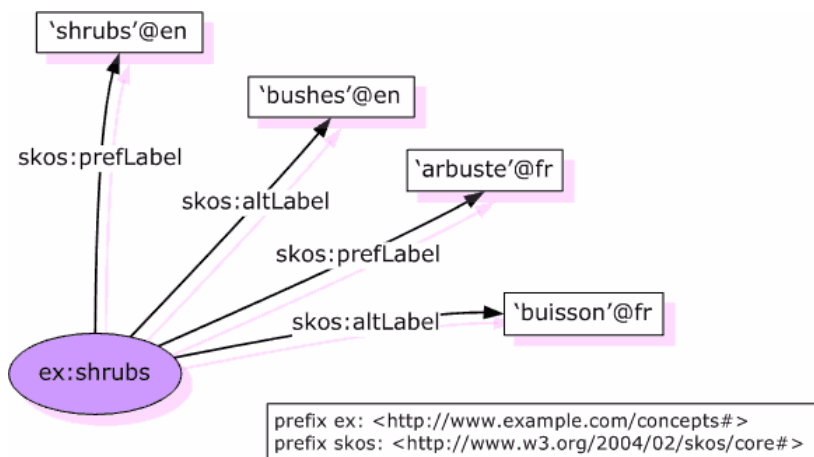
## SKOS: pattern for thesaurus modeling

- Based on ISO standard
- RDF representation
- Documentation:  
<http://www.w3.org/TR/swbp-skos-core-guide/>
- Base class: SKOS **Concept**



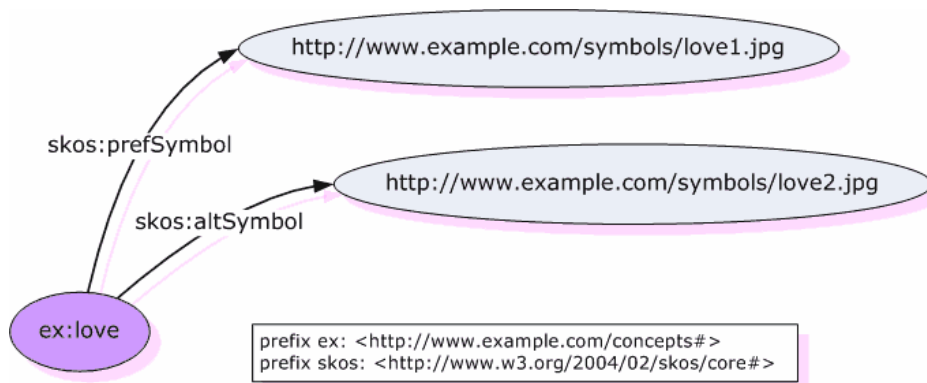
19

## Multi-lingual labels for concepts



20

## Visualizations of concepts



21

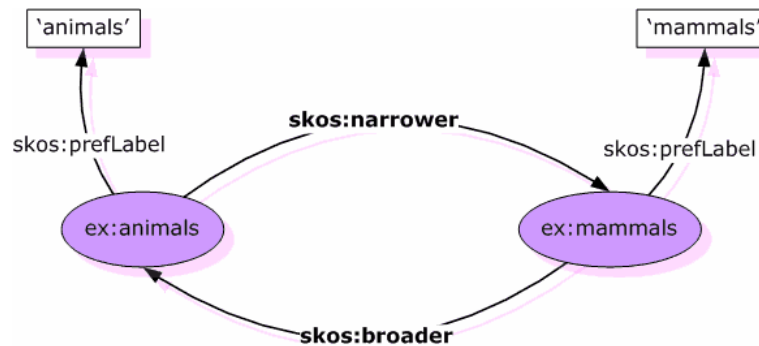
## Documenting concepts

```
skos:note
|
+-- skos:definition
|
+-- skos:scopeNote
|
+-- skos:example
|
+-- skos:historyNote
|
+-- skos:editorialNote
|
+-- skos:changeNote
```

22

## Semantic relation: broader and narrower

- No subclass semantics assumed!

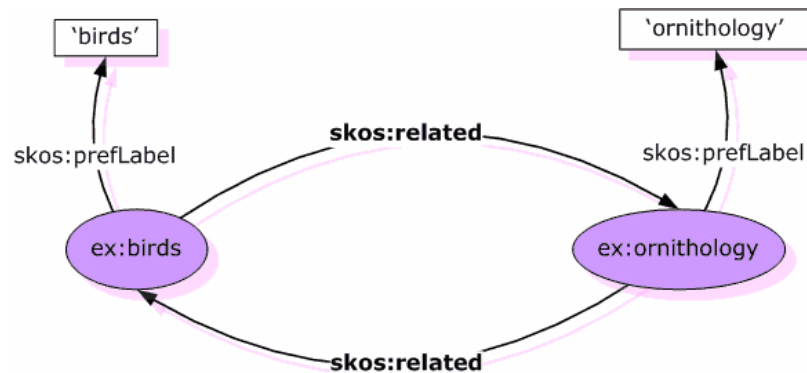


prefix ex: <http://www.example.com/concepts#>  
prefix skos: <http://www.w3.org/2004/02/skos/core#>

23

## Semantic relations: related

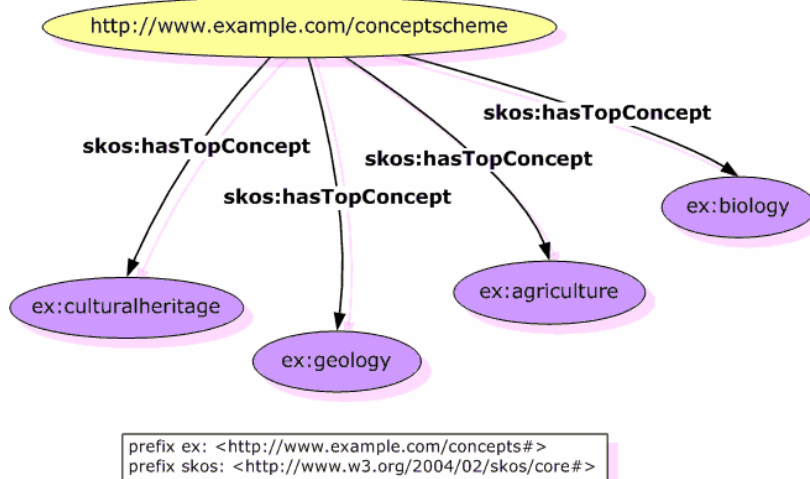
- Symmetry is issue (OWL use)



prefix ex: <http://www.example.com/concepts#>  
prefix skos: <http://www.w3.org/2004/02/skos/core#>

24

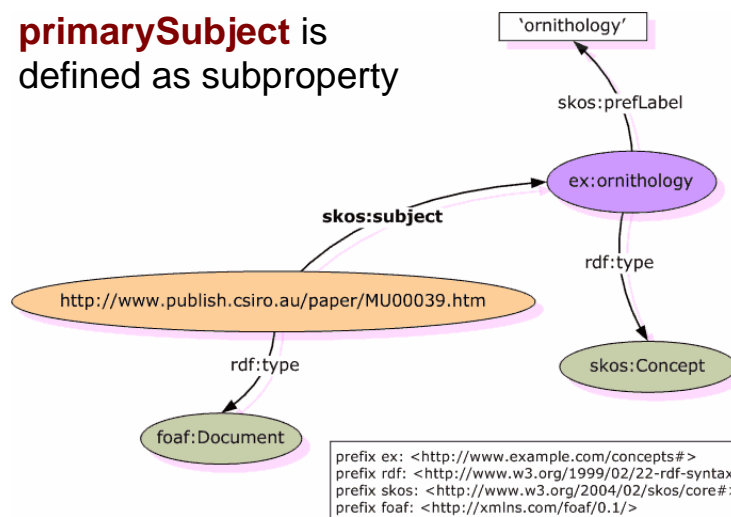
## Defining the top level of the hierarchy



25

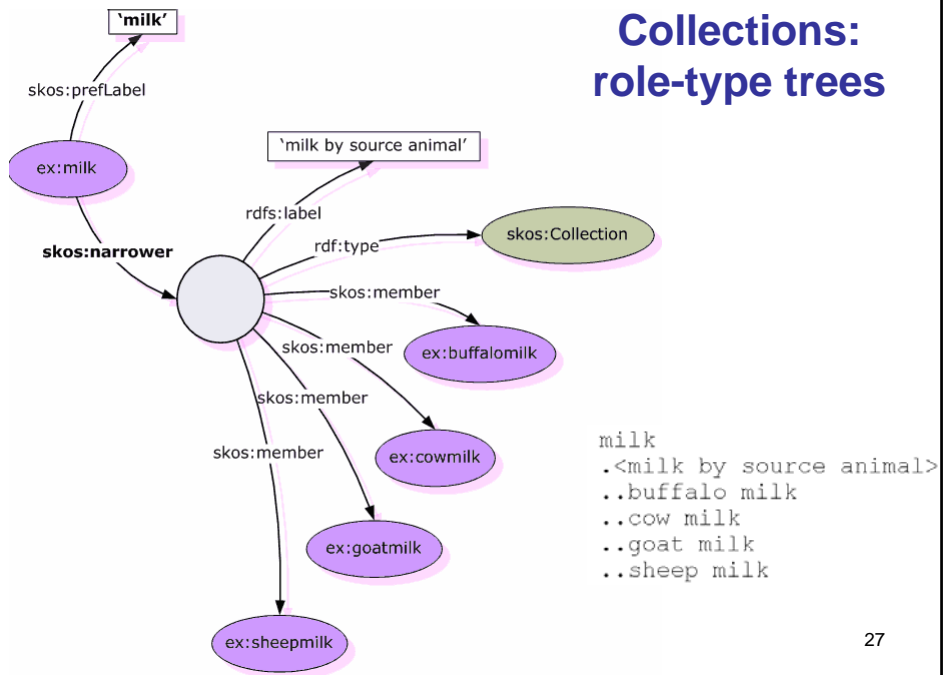
## Indexing a resource with a SKOS concept

- **primarySubject** is defined as subproperty



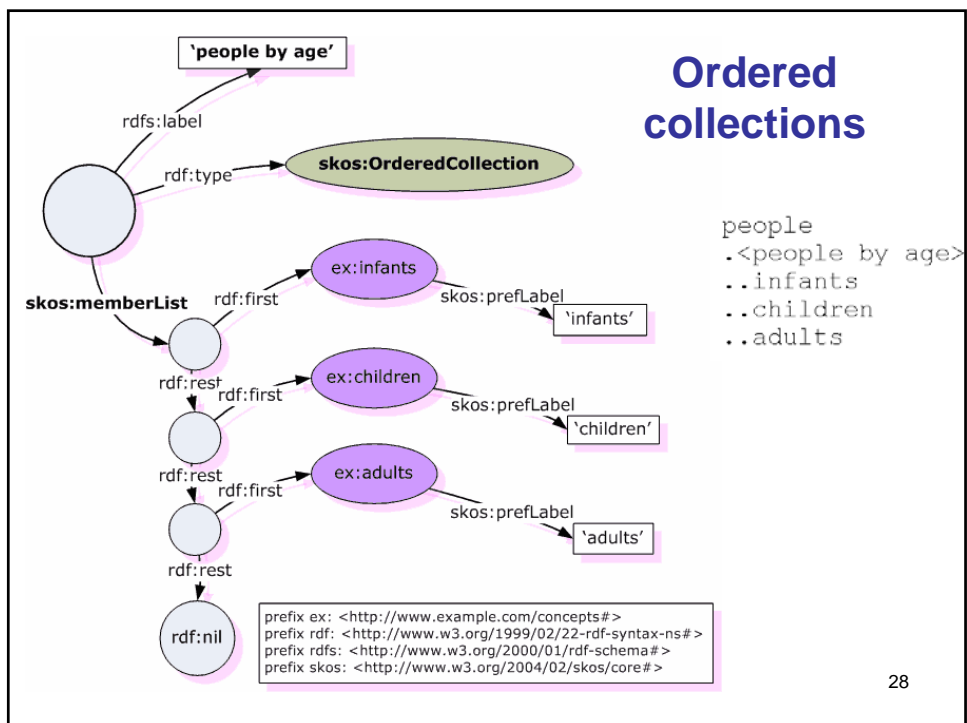
26

## Collections: role-type trees



27

## Ordered collections



28

## Recipes for vocabulary URIs

- Simplified rule:
  - Use “hash” variant” for vocabularies that are relatively small and require frequent access  
<http://www.w3.org/2004/02/skos/core#Concept>
  - Use “slash” variant for large vocabularies, where you do not want always the whole vocabulary to be retrieved  
<http://xmlns.com/foaf/0.1/Person>
- For more information and other recipes, see:  
<http://www.w3.org/TR/swbp-vocab-pub/>

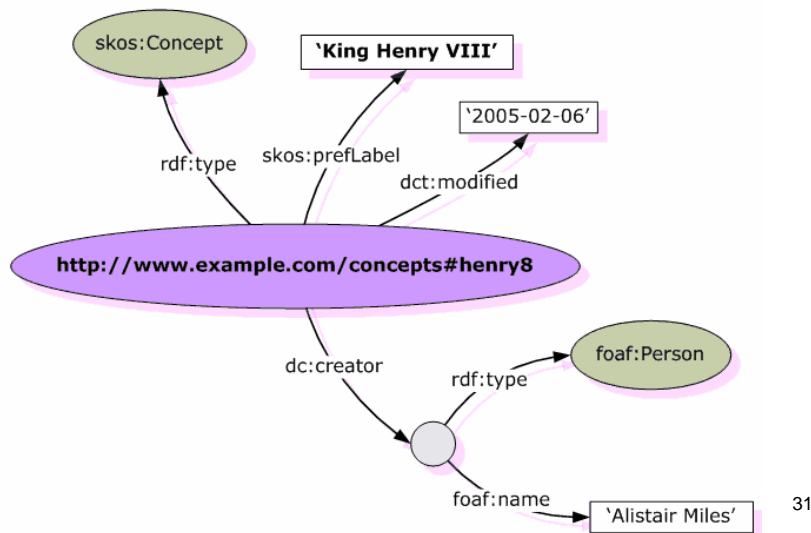
29

## Adding semantics

- Adding OWL statements
- Interpretations of thesaurus relations such as **narrower** as subclass-of are often imprecise (but can still be useful)
- Learning relations between thesauri is important form of additional semantics
  - Example: AAT contains styles; ULAN contains artists, but there is no link
  - Availability of this kind of alignment knowledge is extremely useful

30

## SKOS semantics: concepts are not the real things



## SKOS semantics inference rules (1)

- Collection membership rule  
 $(?i \text{ skos:subject } ?x) (?x \text{ skos:broader } ?y)$   
 $\rightarrow (?i \text{ skos:subject } ?y)$
- If a painting of Van Gogh has as **subject SunFlowers** and if **Flowers** is a **broader** term of **SunFlowers**, then **Flowers** is also the **subject** of the painting.

32



## SKOS semantics inference rules (2)

- Collectable property rule  
 $(?x \ ?p \ ?c) \ (?c \ skos:member \ ?m)$   
 $(?p \ rdf:type \ skos:collectableProperty)$   
 $\rightarrow (?x \ ?p \ ?m)$
- If **GoatMilk** is a member of the collection **<milk by source animal>**, and the latter is a **narrower** concept for **Milk**, and **narrower** is a **collectableProperty**, then **GoatMilk** is also a **narrower** concept of **Milk**
- **broader** and **related** are also collectable

33

## Metamodelling for thesauri: should terms be classes or instances?

- Many thesauri have a inherent metamodeling aspects:
  - The structure of the thesaurus: concepts, relations
  - The actual terms also have a class flavor
- Engineers feel compelled to choose which level to represent as classes
  - Treating terms as classes looses the semantics of the structure-level model  
**Sneeze is an instance of Verb**
  - Treating terms as instances loses the semantics of term relations  
**Bank is a subclass of FinancialInstitution**

34

## Metamodelling

- OWL DL requires strict separation of classes and instances
- But on the Semantic Web my instances may be your classes!
- Metamodelling features especially required in vocabulary/ontology mapping and/or interpretation
- Cf. Protégé metamodelling facilities
- OWL 1.1 (not standardized) allows limited metamodelling within OWL DL scope

35

## Example: WordNet

```
Class(LexicalConcept)
Class(Noun subClassOf(LexicalConcept))
Property(hyponymOf
  domain(LexicalConcept)
  range(LexicalConcept))
Individual(1000768 type(LexicalConcept)
  wordForm(Human))
```

Problem: how to use the hyponym hierarchy as a subclass hierarchy?

36

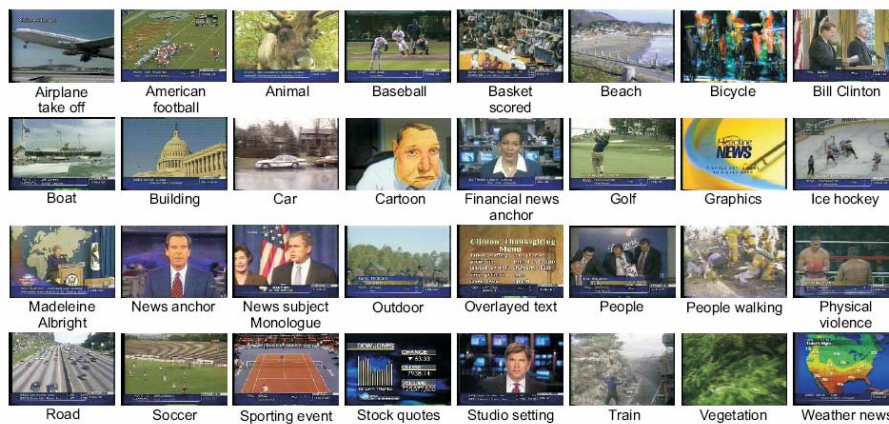
## RDF solution: use metamodelling

**subClassOf(LexicalConcept Class)**  
**subPropertyOf(hyponymOf subClassOf)**  
**subPropertyOf(wordForm rdfs:label)**

- Corresponds to our intuition that WordNet model is a metamodel

37

## Concepts for video detectors (Snoek et al)



38

## LSCOM lexicon: 110 – Female Anchor

- Composite concept
- Alignment needed with general resource to understand semantics



- ♦ S: (n) anchor, anchorman, anchorperson (a television reporter who coordinates a broadcast to which several correspondents contribute)
  - direct hypernym / inherited hypernym / sister term
    - ♦ S: (n) television reporter, television newscaster, TV reporter, TV newsmen (someone who reports news stories via television)
      - ♦ S: (n) reporter, newsmen, newsperson (a person who investigates and reports or edits news stories)
        - ♦ S: (n) communicator (a person who communicates with others)
          - ♦ S: (n) person, individual, someone, somebody, mortal, soul (a human)

## Issues

- Many thesauri do not have a rich semantic structure like WordNet
- Need for learning additional semantic relations between thesaurus concepts
- Result: “ontologizing thesauri”

## **New W3C work: Semantic Web Deployment Working Group**

- Mission to help in vocabulary deployment
- Chartered to standardize SKOS  
Pattern for RDF/OWL representation of (ISO-compliant)  
thesauri
- Guidelines for adding semantics to existing  
vocabularies

<http://www.w3.org/2006/07/SWD/>