Knowledge Representation Briefing

Languages and Knowledge in Bioinformatics

Carole Goble
Information Management Group
University of Manchester, UK
Sagitus Solutions, UK

Ack: Rudi Studer, Robert Stevens, Frank van Harmelen, Ian Horrocks, Peter Karp

Roadmap

- Motivation for knowledge in bioinformatics
- What is the knowledge we are representing?
  - What is an Ontology?
- What are the representation languages
  - traditional AI/CS Frames and Logics
  - WWW metadata XML and RDF
  - WWW + AI/CS
- OIL: The Ontology Inference Layer
- Summary

Motivation 1: knowledge for knowledge sake

- "Bioinformatics is a knowledge-based discipline. Many predictions, and interpretations, of data in biology are made by comparing the data in hand against existing knowledge"

  Dr. Andy Brass

- "The technical advantages of knowledge modeling are obvious. Knowledge bases can be automatically checked for consistency; they support inference mechanisms which derive data which have not been explicitly stored; they also offer extensive request and navigation facilities. However, the most immediate benefit of knowledge base design lies in the modeling process itself, through the effort of explication, organization and structuration of the knowledge it requires."

  Editorial, Bioinformatics, July 2000

Motivation 2: The Semantic Web

"*... imagine what computers can understand when there is a vast tangle of interconnected terms and data that can automatically be followed.*" (Tim Berners-Lee, Weaving the Web, 1999)

- Interoperating resources, be it by people or systems, requires a consistent shared understanding of what the information contained means
- Ontologies are the cornerstone of encoding this understanding, BUT to be shared they they need a standard representation and exchange language "*... people [and machines] can't share knowledge if they don't speak a common language*" (Davenport)

From human interaction to machine interaction

From machine readable to machine understandable

Sharing data ⇒ Sharing meaning

- Metadata
  - Data describing the content and meaning of resources
  - Swiss-Prot schema and annotation
  - But everyone must speak the same language…
- Terminologies
  - Shared and common vocabularies
  - For search engines, agents, curators, authors and users
  - Gene Ontology, Swiss-Prot keywords
  - But everyone must mean the same thing…
- Ontologies
  - Shared and common understanding of a domain
  - Essential for search, exchange and discovery

What is an Ontology?

- Describes a formal, shared conceptualization of a domain of interest
  - Concepts that are relevant for the domain *e.g. Gene*
  - Relations between concepts *e.g. Product of Gene*
  - Axioms about these concepts and relations *e.g. Nucleic acids < 20 residues are oligonucleiotides*
  - Values *The Product of the trpA Gene is tryptophan-synthetase*
- Enforces a well-defined semantics on such a conceptualisation
- A vocabulary of terms and some specification of the meaning of the terms
- Instances
  - Nominals = instances that are part of the ontology
    - *eg. Italy, sulphur*
  - Knowledge base = ontology + instances
    - *trpA Gene, Reaction 1.1.2.4*

More on Concepts

- Primitive: properties are necessary
  - Globular protein must have hydrophobic core, but a protein with a hydrophobic core need not be a globular protein
- Defined: properties are necessary + sufficient
  - Eukaryotic cells must have a nucleus. Every cell that contains a nucleus must be Eukaryotic.

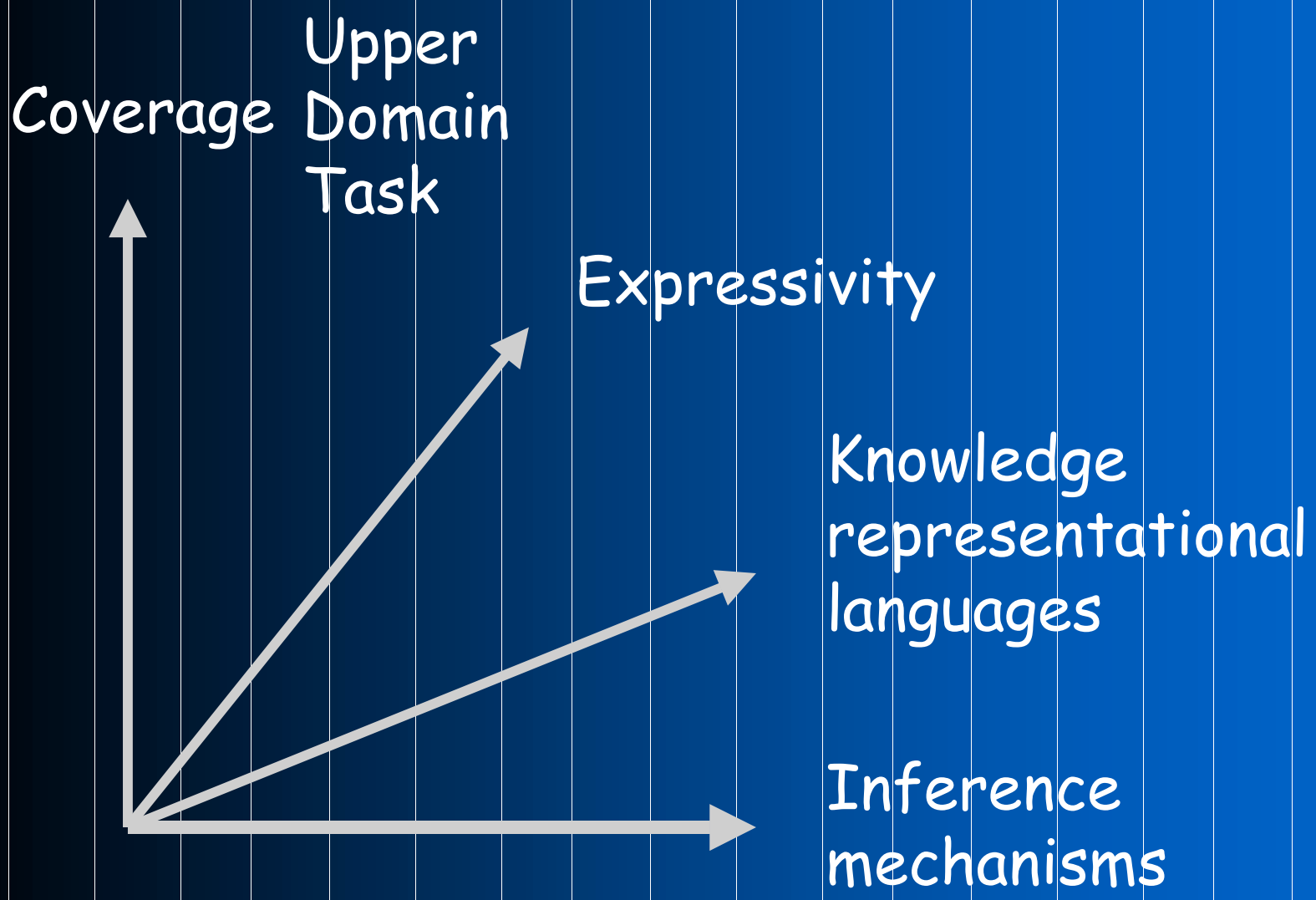
Roles of Ontologies in Bioinformatics

- Success of many biological DBs depends on
  - High fidelity ontologies
  - Clearly communicating their ontologies
  - Prevent errors on data entry, search and interpretation

- Finding, Sharing, Discovering
- Common framework for multidatabase queries - TAMBIS
- Controlled vocabularies for genome annotation - Gene Ontology
- Information-extraction applications – annotation generation
- Knowledge-acquisition tools – Ecocyc, Riboweb
- Hypothesis Generation - Ecocyc RiboWeb,

- Reuse is a core aspect of ontologies
  - Reuse of existing ontologies faster than designing new ones
  - Reuse decreases semantic heterogeneity of DBs

Ontology Description Space



Different Types of Ontologies

- Light-Weight
  - Concepts, atomic types
  - Is-a hierarchy among concepts
  - Relations between concepts

Gene Ontology

- Heavy-Weight
  - Meta classes
  - Type constraints on relations
  - Cardinality constraints
  - Taxonomy of relations
  - Reified statements
  - Axioms / semantic entailments of various flavours
    - Expressiveness (description logics, propositional, horn clauses, first order logic, higher order)
    - Inference Systems

EcoCyc
RiboWeb
Tambis Ontology

Gene Ontology -- Light Weight

Gene Ontology

Gene Ontology

molecular_function

593. uroporphyringonen-III synthase(1)
594. uroporphyrinogen decarboxylase(2
595. xanthine dehydrogenase(2)
596. xanthine oxidase(1)
597. xylulokinase
enzyme activator(25)
enzyme inhibitor(60)
apoptosis inhibitor(4)
signal transduction(834)
cell adhesion(40)
storage protein(6)
structural protein(512)
transporter(343)
ligand binding or carrier(340)
protein degradation tagging(2)
tumor suppressor(11)
molecular_function unknown(356)
ribosomal RNA(16)
small nuclear RNA(57)
small nucleolar RNA(71)
transfer RNA(510)

cellular_component

cellular_component
- 1.1. extracellular(133)
- 1.2. intracellular(4027)
- 1.3. unlocalised(70)
- 1.4. cellular_component unknow

biological_process

1. cell growth and maintenance(3411)
- 1.1.1. intracellular protein traffic(129)
- 1.1.2. metabolism(1807)
- 1.1.3. transport(209)
- 1.1.4. ion homeostasis(15)
- 1.1.5. cell volume regulation
- 1.1.6. pH regulation
- 1.1.7. autophagy
- 1.1.8. cell death(47)
  - 1.1.8.1. apoptosis(47)
    - 1.1.8.1.1. anti-apoptosis(3)
    - *1.1.8.1.2. apoptosis signalling(11)*
      - 1.1.8.1.2.1. induction of apopto
    - *1.1.8.1.3. caspase activation(2)*
    - *1.1.8.1.4. DNA fragmentation(2)*
    - *1.1.8.1.5. committment to apoptos*
    - *1.1.8.1.6. disassembly of cell struc*
    - 1.1.8.1.7. peripheral killing of activ
    - 1.1.8.1.8. killing of inflammatory ce
    - 1.1.8.1.9. killing virus-infected cell
    - 1.1.8.1.10. killing transformed cells

Current tree depth 1 | Current tree depth 1 | Current tree depth 1

1 2 3 4 5 6 7 8 | 1 2 3 4 5 6 7 8 9 | 1 2 3 4 5 6 7 8 9 10 11

Locate term:

Use the '' as a wild card.?*

Locating terms is case insensitive

List gene products for selected terms | Only current | or | and | and not

EcoCyc - Heavy Weight

Relationships Editor

Application Knowledge-Base Frame View Preferences

KB: ECOBASE Package: ECOCYC Slot-Value Relationships Viewer

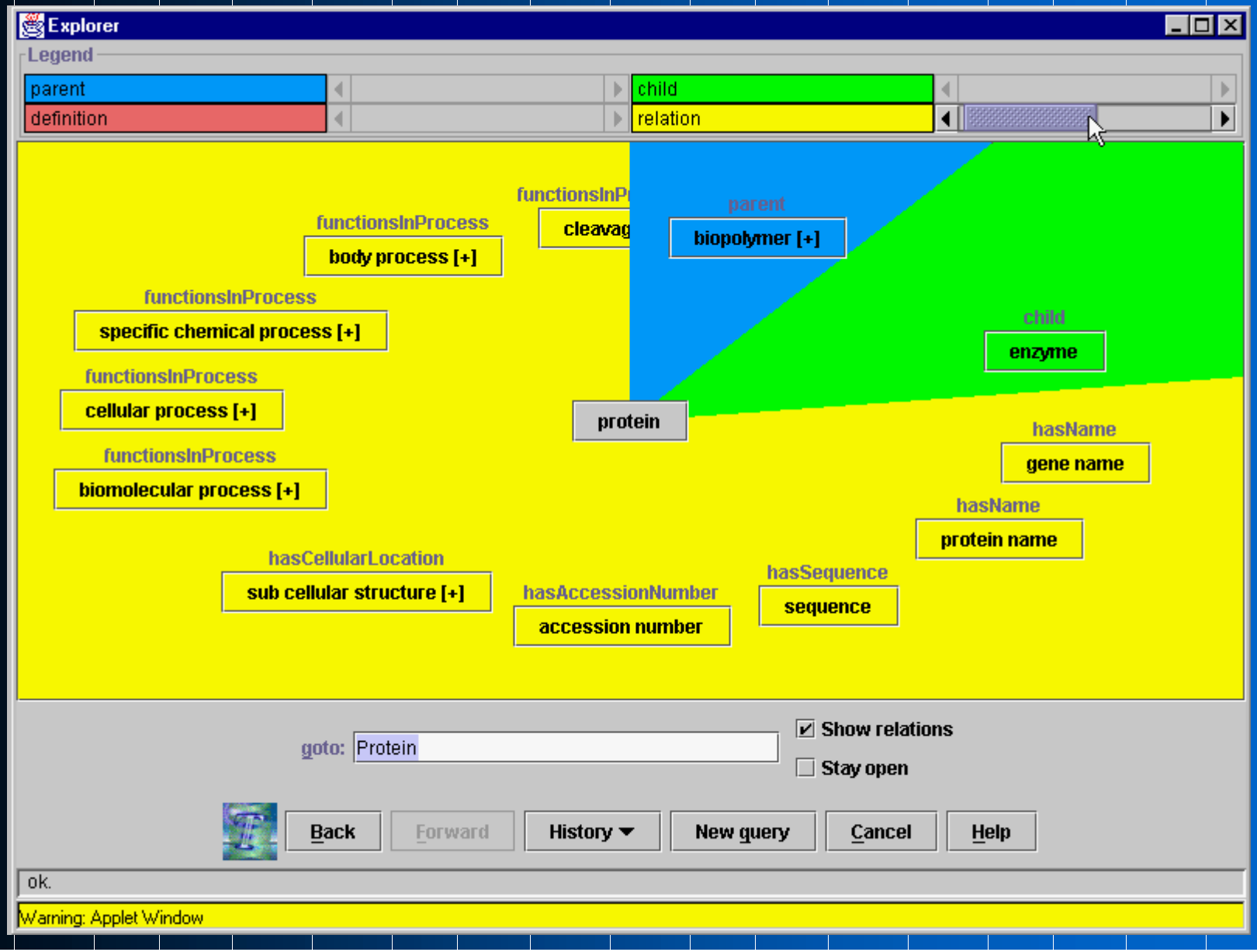
NAD phosphorylation and dephosphorylation
- REACTION-LIST → NADPPHOSPHAT-RXN —ENZYMATIC-REACTION→ NADP phosphatase —ENZYME→ NADPPHOSPHAT-MONOMER
- REACTION-LIST → NAD-KIN-RXN —ENZYMATIC-REACTION→ NAD kinase —ENZYME→ NAD-KIN-MONOMER
- REACTION-LIST → 1.6.1.2-RXN —ENZYMATIC-REACTION→ pyridine nucleotide transhydrogenase —ENZYME→ PYRNUTRANSHYDROGEN-CPLX

NAD phosphorylation and dephosphorylation
- REACTION-LIST → NADPPHOSPHAT-RXN —ENZYMATIC-REACTION→ NADP phosphatase
- REACTION-LIST → NAD-KIN-RXN —ENZYMATIC-REACTION→ NAD kinase —ENZYME
- REACTION-LIST → 1.6.1.2-RXN —ENZYMATIC-REACTION→ pyridine nucleotide transhydroge

```
Command:
Command: Deselect
Command: ▮
```

L: Deselect; R: Menu of completions.

TAMBIS - Heavyweight ontology



Representation Languages for Ontologies

- Controlled vocabulary
  - List of terms possibly in a taxonomy
- Data modelling and Knowledge Representation
  - Object oriented models- EER, UML
  - Frames- Ontolingua, OKBC, OCML, Protégé, Ocelot/GKB
  - Logics– Description Logics, Conceptual Graphs, FLogic
- WWW standards
  - XML & XML Schema
  - RDF & RDF Schema
- Web-based ontology specification languages
  - XOL, SHOE - XML based
  - OIL - RDF(S) based

Corcho and Gomez-Perez
A Roadmap to Ontology Specfication Languages
EKAW 2000, LNAI 1937

Frame based languages

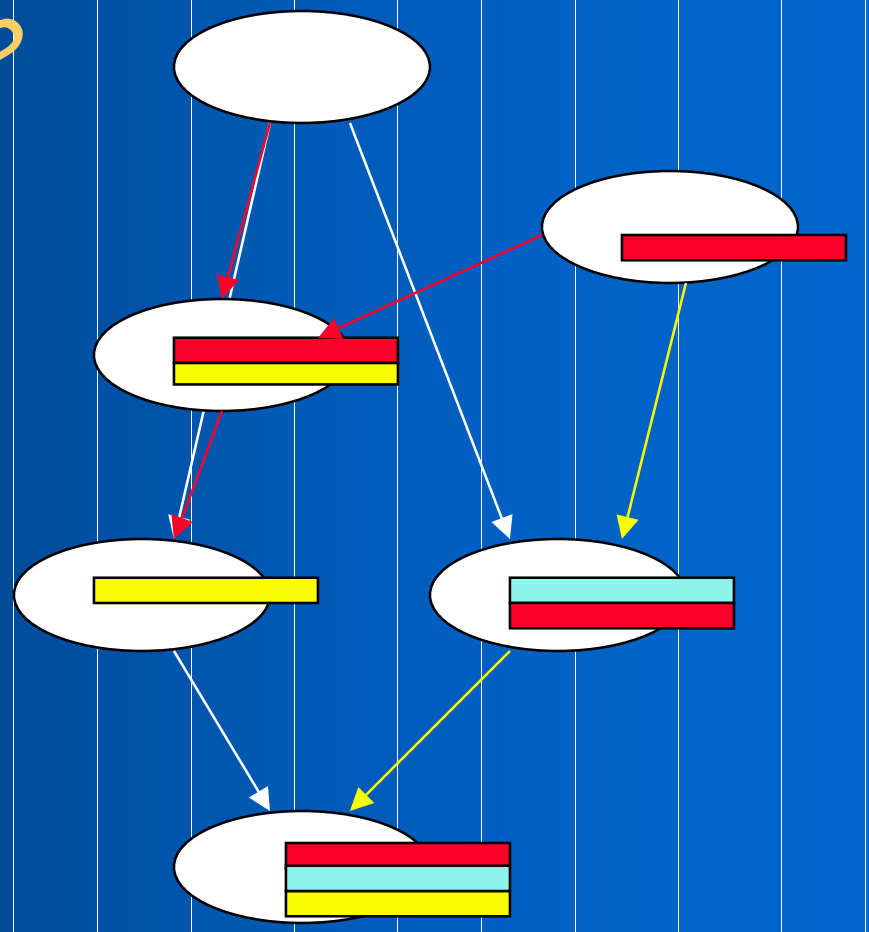
- Intuitive for many users
- Extensive set of modelling primitives
  - Frames, slots, facets, defaults
  - Classes are primitives
- OKBC, OKBC-Lite, XOL
- Lots of tools
- Little reasoning support

(Description) Logic-based languages

- Defined classes
- Negation and disjunction
- Properties for slots/relations
  - E.g. transitivity for *contained-in*
- Formal semantics
- Reasoning support
  - Inconsistency detection
  - Implicit superclass detection
  - Optimised inference engines for tractable reasoning

Why Reasoning support?

- Useful for
  - As design support tool
  - For large ontologies
  - Multiple authors
  - Integrating and sharing
  - Evolving ontologies
- Allows to
  - Establish inter-ontology relationships
  - Check for consistency
  - Check for unexpected implied relationships

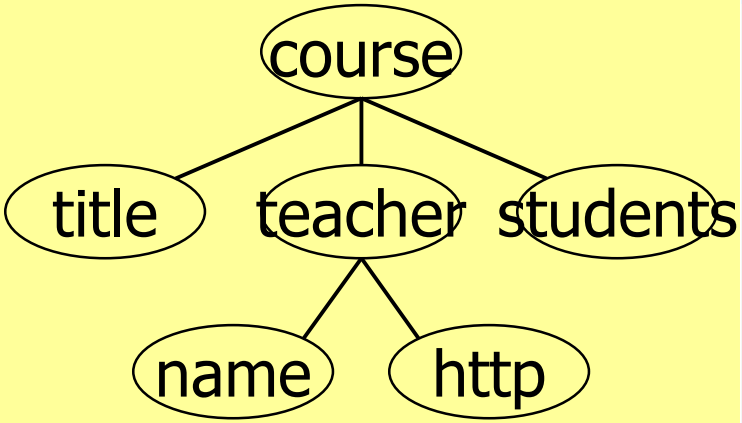


XML is not good for describing ontologies

- XML defines grammars to verify and structure documents
- The grammar enforces constraints on tags
- Different grammars define the same content
- XML lacks a semantic model – it only has a *surface* model which is a tree.

• node = label + attr/values + contents

```
<course date="...">
      <title>...</title>
      <teacher>...</teacher>
            <name>...</name>
            <http>...</http>
      <students>...</students>
</course>
```



XML is not good for describing ontologies

- Meaning of XML documents is intuitively clear
  - "*semantic*" markup tags are domain terms
- But computers do not have intuition
  - Tag names *per se* do not provide semantics
  - The semantics are encoded outside the XML specification
- XML makes no commitment on:

❶ Domain specific ontological **vocabulary**

❷ Ontological **modelling primitives**

⇨ requires pre-arranged agreement on ❶ & ❷

Feasible for closed collaboration

- agents in a small & stable community
- pages on a small & stable intranet

XML DTDs and XML Schema

- DTD does not distinguish between objects and relations
- XML Schema's type extension mechanism is a red herring – it can't be used to model ontological subtypes
- XML has been used as a serialisation syntax for other markup languages – e.g. SMIL, XOL

```
<class>
  <name> person </name>
</class>
<slot>
  <name>year-of-birth</name>
  <domain.person</domain>
  <slot-cardinality>1</slot-cardinality>
</slot>
```

RDF and RDF Schema is a bit better

- RDF a small set of modelling primitives + syntax
- RDF *does not* commit to a domain vocabulary

- RDF Schema
  - Extended set of modeling primitives
    - resource, class, subClassOf, type
    - property, subPropertyOf
    - domain, range, constraintResource, constraintProperty
  - Define a domain vocabulary for RDF
  - Organise this vocabulary in a typed hierarchy
  - BUT expressive inadequacy and poorly defined semantics
- RDFS is the framework for a vocabulary

RDF Schema syntax in XML

```
<rdf:Description ID="MotorVehicle">
    <rdf:type resource="http://www.w3.org/...#Class"/>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/...#Resource"/>
   </rdf:Description>
<rdf:Description ID="Truck">
    <rdf:type resource="http://www.w3.org/...#Class"/>
    <rdfs:subClassOf rdf:resource="#MotorVehicle"/>
</rdf:Description>
<rdf:Description ID="registeredTo">
    <rdf:type resource="http://www.w3.org/...#Property"/>
    <rdfs:domain rdf:resource="#MotorVehicle"/>
    <rdfs:range rdf:resource="#Person"/>
</rdf:Description>
<rdf:Description ID="ownedBy">
    <rdf:type resource="http://www.w3.org/...#Property"/>
    <rdfs:subPropertyOf rdf:resource="#registeredTo"/>
</rdf:Description>
```

Describing and Exchanging Ontologies

- To reuse an ontology we need to share it with others in the community
- Exchanging ontologies requires a language with:
  - common syntax
  - clear and explicit shared meaning
- Tools for parsing, delivery, visualising etc

Requirements for an Ontology-language

- Well designed
  - Useful and proven *modelling primitives*
  - Intuitive to human users
  - Can say simple things simply
  - *Expressive* enough to capture many ontologies
  - Efficient, sound and complete *reasoning support*
- Well defined
  - clear *syntax* - read ontologies
  - Formal *semantics* – understand (process) ontologies - to facilitate machine interpretation of that semantics;
  - *Expressive* enough to capture many ontologies
- Compatible
  - Easy *mapping* to/from other ontology languages
  - Maximum *compatibility* with XML and RDF(S);

GO EcoCyc TAMBIS

OIL: Ontology Inference Layer

Frames:
modelling primitives,
OKBC-Lite

Description Logics:
formal semantics &
automated reasoning
support

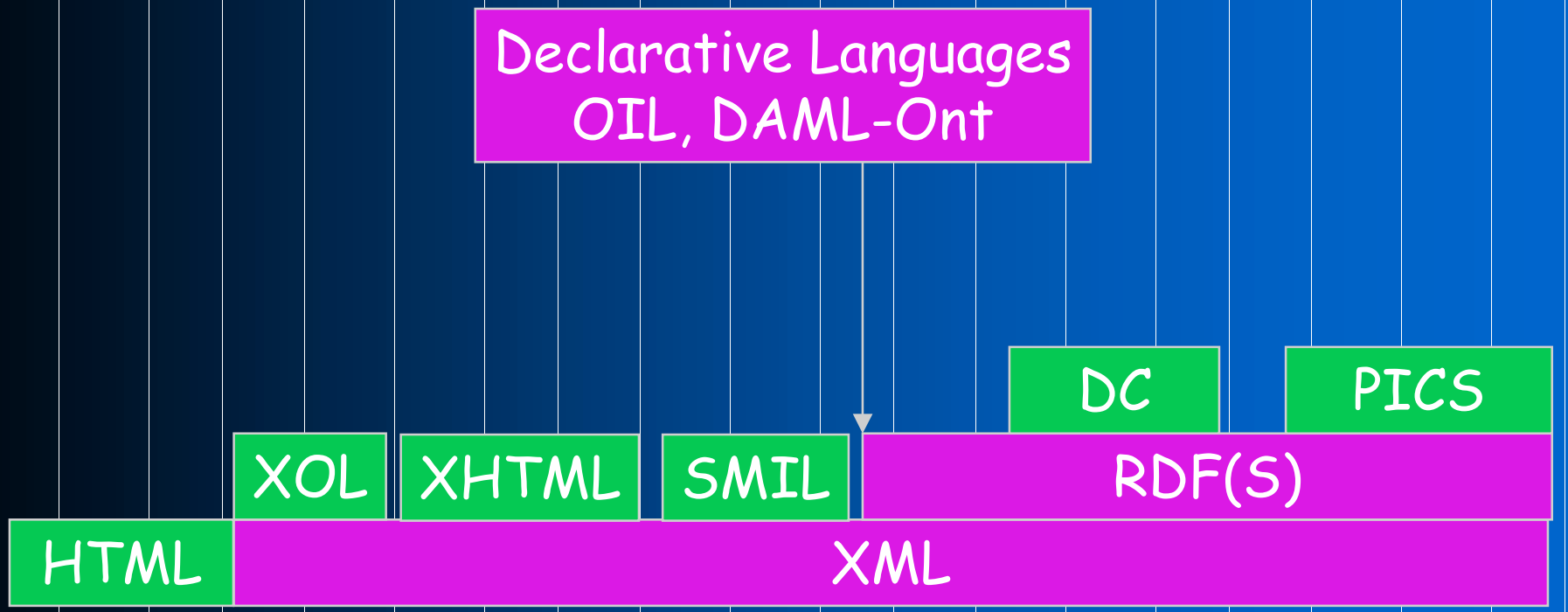
Originally based on XOL from the BioOntology Working Group

A knowledge representation language and inference mechanism for the web

Web languages:
XML & RDF based syntax, RDFS mapping

W3C vision of ontologies on the Web

- The W3C hierarchy of languages



OIL by example

Slot-def part-of	• part-of is a slot
subslot-of structural-relation	• sub-slot of structural-relation
inverse has-part	• inverse is has-part
properties transitive	• it is transitive
Class-def defined herbivore	• herbivore exactly defined as:
subclass-of animal	• sub-class of animal
slot-constraint eats	• that eats
value-type plant OR	• only plants
slot-constraint part-of	• or parts of
has-value plant	• plants
min-cardinality 2 vegetable	• and >= 2 types of vegetable
Disjoint herbivore carnivore	• herbivore and carnivore disjoint

Semantics via translation to SHIQ DL

Slot-def part-of	
subslot-of structural-relation	• part-of ② structural-relation
inverse has-part	• has-part ≡ part-of$^-$
properties transitive	• part-of ∈ R_+
Class-def defined herbivore	• herbivore ≡
subclass-of animal	• animal ↑
slot-constraint eats	
value-type plant OR	• ∀eats.(plant ➔
slot-constraint part-of	
has-value plant	• ∃part-of.plant) ↑
min-cardinality 2 vegetable	• >= 2.*eats*vegetable
Disjoint herbivore carnivore	• herbivore ② ¬carnivore

RDFS delivery syntax

```
<rdfs:Class rdf:ID = "herbivore">
  <rdf:type
   rdf:resource=http://www.ontoknowledge.org/#DefinedClass/>
   <rdfs:subClassOf rdf:resource="#animal"/>
   <rdfs:subClassOf>
     <oil:NOT>
       <oil:hasOperand rdf:resource="resource="#carnivore"/>
     </oil:NOT>
   </rdf:subClassOf>
</rdfs:Class>
```

OIL as RDF(S) extension

```
<rdfs:Class rdf:ID = "herbivore">


  <rdfs:subClassOf rdf:resource="#animal"/>
  <rdfs:subClassOf>




   </rdf:subClassOf>
</rdfs:Class>
```

GO in OIL

```
slot-def part-of
    inverse has-part
    properties transitive

class-def nucleolus-organiser
    documentation "go:0005731"
    subclass-of cellular_component
    slot-constraint part-of has-value nucleolus
```

OIL: Ontology Metadata (Dublin Core)

Ontology-container
 title "macromolecule fragment"
  creator "robert stevens"
  subject "macromolecule generic ontology"
  description "example for a tutorial"
  description.release "1.0"
  publisher "R Stevens"
  type "ontology"
  formal "pseudo-xml"
  identifier "http://www.ontoknowledge.org/oil/oil.pdf"
  source "http://img.cs.man.ac.uk/ismb00/mmexample.pdf"
  language "OIL"
  language "en-uk"
  relation.haspart "http://www.ontoRus.com/bio/mmole.onto"

OIL primitive ontology definitions

```
slot-def has-backbone
   inverse is-backbone-of
slot-def part-of
  inverse is -part-of
  properties transitive

class-def rna
  subclass-of nucleic-acid
  slot-constraint has-backbone
     value-type ribophosphate

class-def  ribophosphate
class-def  deoxyribophosphate
disjoint ribophosphate,
         deoxyribophosphate
```

```
class-def defined dna
   subclass-of nucleic-acid
   slot-constraint has-backbone
     value-type deoxyribophosphate

disjoint rna, dna

class-def defined catalyst
   subclass-of macromolecule
   slot-constraint promotes
     has-value reaction

class-def defined enzyme
    subclass-of protein, catalyst
```

OIL defined ontology definitions

```
class-def defined mitochondrial
        subclass of location
        slot-constraint cellularlocation
            has-value ( mitochondrion OR
                (slot-constraint  part-of
                    has-value mitochondrion)
            cardinality 1

class-def defined succinate-dehydrogenase
        subclass of  enzyme
        slot-constraint promotes
                value-type oxidation
        slot-constraint cellularlocation
                has-value
                        (slot-constraint part-of
                                has-value mitochondrion)
                cardinality 1
```

OIL extends frames

- Classes can be primitive (necessary conditions)
  - Elephant ⇒ animal that has-colour grey
- Or defined (necessary and sufficient conditions)
  - Vegetarian ⇔ person who eats neither meat nor fish
- Classes allowed in slot constraints
  - Slot-constraint eats has-value meat (eats some meat)
  - Slot-constraint eats value-type meat (eats only meat)
- Can use arbitrary class expressions instead of only class names
  - Slot-constraint eats value-type NOT (OR meat fish)
- Cardinality constraints can include value-types
  - Slot-constraint eats max-cardinality 1 plant
- Sub-slot relation -- Daughter-of subslot-of child-of
- Slot properties - Transitive (e.g. part-of), Symmetrical …
- Concrete data types
  - Integers and strings, with min, max, ranges..
- Additional axioms
  - Disjointness, disjoint-coverings, equivalence …

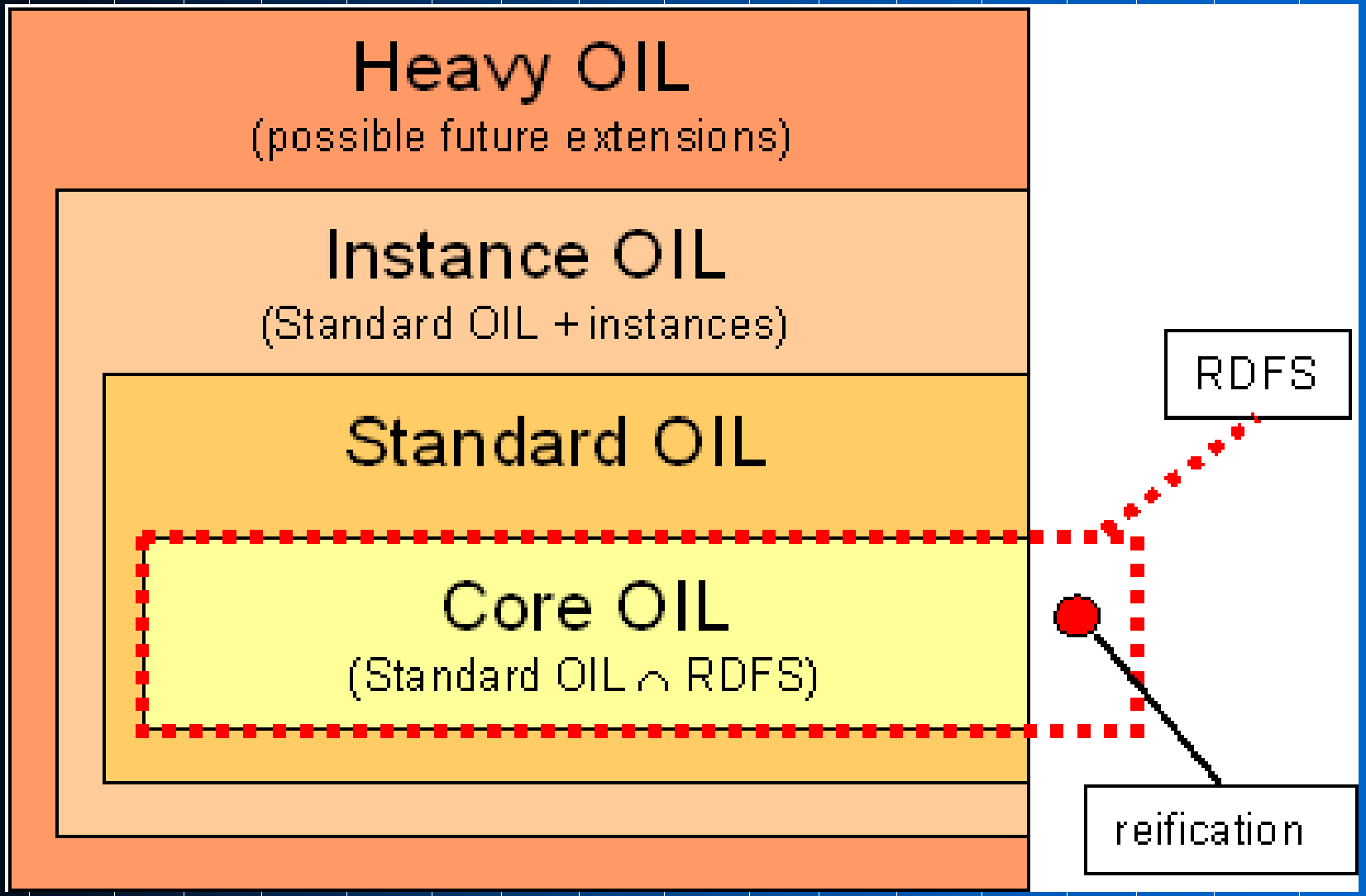
OIL Core restricts frames

- No defaults
- Limited axioms / rules
- Second order expressivity
- Ontology only (limited form of individuals)
- No lists

- Because
  - Reasoning support
  - Semantics

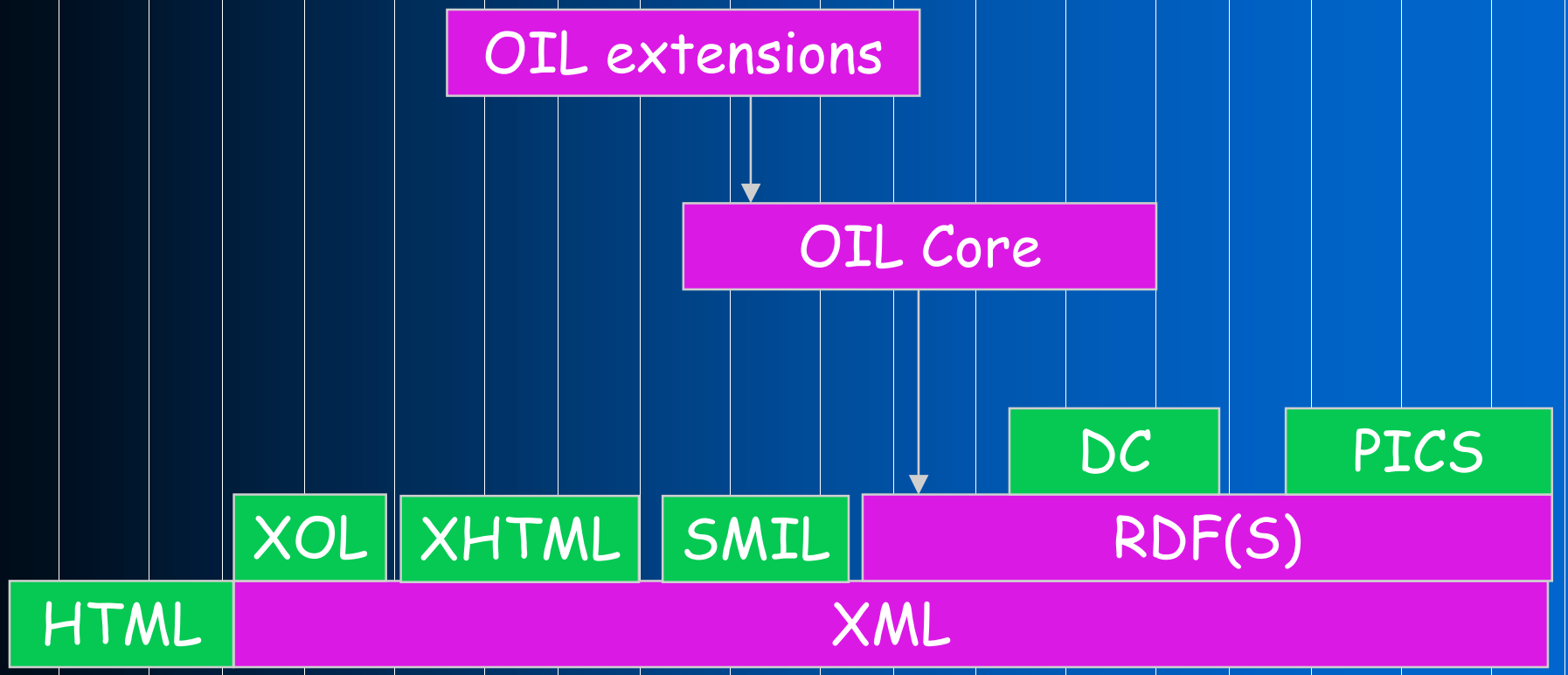
- Instance OIL
  - Instances can be described but they won't be reasoned about.

One size does not fit all
OIL's Layered Architecture



Layered OIL extensions

- Build on top of OIL Core
- Use RDFS and OIL Core modelling primitives where possible



What OIL Doesn't do

- It doesn't interoperate between ontologies
- It doesn't merge or align ontologies

- The inferencing helps tools that do this but it doesn't do this *per se*

Its easier to integrate two programs written in Java than one in C and one in Java., but moving both programs into the same language doesn't integrate them.

DARPA

- Darpa Agent Markup Language http://www.daml.org/
  - Developing a RDFS based ontology language
- Joint EU/US ad hoc Agent Markup Language Committee
  - OIL Consortium
  - W3C RDF-Logic
  - DARPA DAML
  - EU commission and DARPA observers
  - http://www.daml.org/committee

- Some differences between OIL and DAML
- DAML Ont 2nd release will be DAML-OIL
- Aim: DAML-OIL based W3C standardisation proposal

W3C RDF WG

Acrobat Reader - [W3C Activity Prop.pdf]

File Edit Document View Window Help

W3C

Activity Proposal: Semantic Web

Summary · Context · Proposal · Timeline/Resources
Acknowledgements

1. Executive Summary

Facilities to put machine-understandable data on the Web are becoming a high priority for many communities. The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently. This must be an area in which W3C maintains a leadership role, both in developing open specifications for those technologies that are ready for large scale deployment and identifying through open source advanced development the infrastructure components that will be necessary to scale in the Web in the future. We propose a *Semantic Web Activity* to leverage the existing results obtained in the W3C Metadata Activity:

A. **Continue the work of the RDF Interest Group**
The RDF Interest Group will coordinate implementation and deployment of RDF and will provide liaison with new work in the W3C and the wider XML community on matters relating to RDF.
B. **Charter a Working Group to consider revisions to the RDF Model and Syntax Recommendation**
C. **Complete work on the RDF Schema specification**
This Working Group group will incorporate the results of the on-going RDF implementation experience and consider directic
Recommendation.
D. **Coordinate with selected nc**
on Semantic Web technolog
This coordination includes, but is
E. **Charter additional working**
supporting the above objec
F. **Create an advanced develo**
XML and RDF technologies
The development project is intended to facilitate distributed collaboration with a specific intent to increase the level of automation of the W3C Web site and to develop open-source RDF infrastructure support modules.

tiatives and individual activitie

to, DCMI, DAML, OIL, and SHOE,

groups identified as neccessar

2. Context: Market Analysis, Related W3C Work

100% 1 of 5 8.26 x 11.69 in

OIL Tools and Infrastructure

- Reasoning services provided by CORBA FaCT system
  - Currently via OIL <-> FaCT translators XSLT
  - CORBA OIL soon http://www.cs.man.ac.uk/FaCT/
- Frame ontology editors being build/adapted
  - Protégé editor (Stanford)
  - ODE (Madrid)
  - OntoEdit (Karlesruhe)
  - OilEd (Manchester) http://img.cs.man.ac.uk/oil/
- Demonstrator web resource annotation tools
  - COHSE (Manchester)
  - OntoAnnotate (Karlesruhe)
- Case Studies
  - Tambis II, Gene Ontology

OilEd 1.0

Oiled

File Log Export Reasoner Help

Classes | Slots | Individuals | Axioms | Container

Classes

- company
- dog
- dog_liker
- dog_owner
- driver
- girl
- grownup
- haulage_company
- haulage_truck_driver
- haulage_worker
- kid
- lorry
- lorry_driver
- magazine
- man
- newspaper
- old_lady
- person
- pet_owner
- publication
- quality_broadsheet

Name

haulage_truck_driver

Documentation

Subclass of

person

Slot Constraints

type	slot	
has-value	works_for	[par
has-value	drives	truc

D:\OIL\OilEd\ontologies\people

Class Hierarchy

Hierarchy

- top
  - animal
  - colour
  - company
  - haulage_worker
    - haulage_truck_driver
  - person
    - cat_liker
    - dog_liker
    - driver
      - bus_driver
      - haulage_truck_driver
      - lorry_driver
      - van_driver
        - white_van_man
    - grownup
    - kid
    - pet_owner
  - publication
  - vehicle

Supers

- driver
- haulage_worker

Done

Oil Collaborators and further info

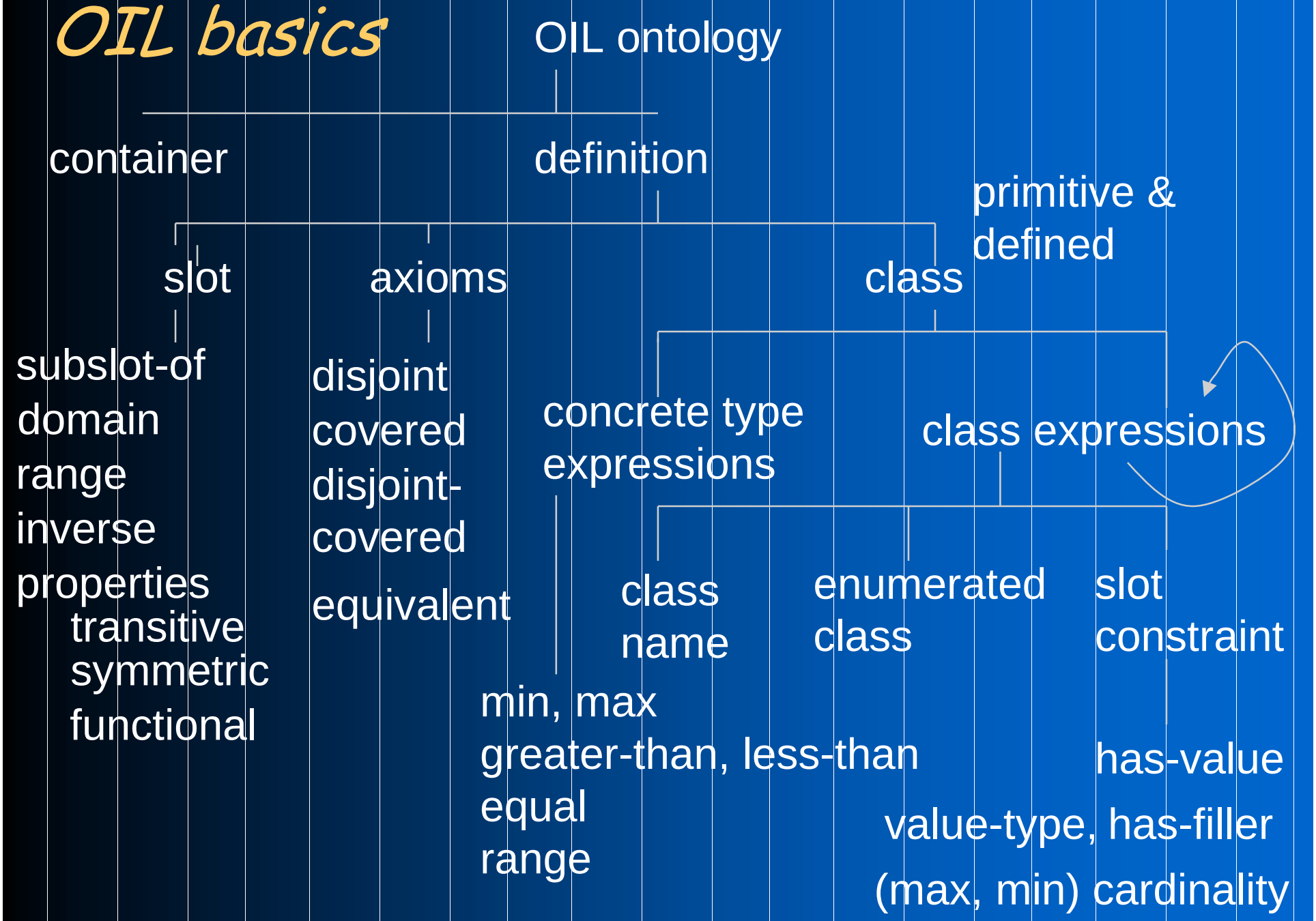
- Projects
  - OntoWeb Network of Excellence
  - On-To-Knowledge
  - IBROW
  - Comma
  - SemanticWeb IST EU Proposal
- EU academics
  - University of Manchester, UK
  - Free University of Amsterdam, , Nederland
  - University of Amsterdam, Nederland
  - Knowledge Media Institute, Open University, UK
  - University of Karlsruhe, Germany
  - University of Bremen, Germany
- EU Industrialists
  - Aidministrator, Nederland
  - Swiss Life, BT, CogNT
- US Academics
  - Stanford University, USA (DB, KSL, Med Inf)
  - University of Maryland
  - SRI
- Outside Academia
  - W3C (RDF working group)
  - DARPA (DAML initiative)
- Bio sign ups
  - BioWisdom, Biovista,
  - Ingenuity
  - SmithKlineBeecham
  - GlaxoWellcome, Merck

Summary

- We should mix the best of frames and logics together with web technologies
- We believe that something OIL-like will be adopted as the ontology language of the web
- Because of its extension of RDFS, any RDFS aware agent can use an OIL ontology
- A layered approach maintains compatibility and ensures careful understanding of the semantics and reasoning involved

- The Semantic Web
- http://www.ontoknowledge.org/oil
- http:/www.semanticweb.org/

OIL basics


OIL ontology
container
definition
slot
axioms
class
primitive &
defined
subslot-of
domain
range
inverse
properties
transitive
symmetric
functional
disjoint
covered
disjoint-
covered
equivalent
concrete type
expressions
class expressions
class
name
enumerated
class
slot
constraint
min, max
greater-than, less-than
equal
range
has-value
value-type, has-filler
(max, min) cardinality