The Role of XML Databases in Intelligent Search and Case-Based Reasoning

Larry Kerschberg

E-Center for E-Business; http://eceb.gmu.edu/

Department of Computer Science and

Information and Software Engineering

Volgenau School of Information Technology and Engineering

George Mason University, Fairfax, Virginia, USA

Copyright 2007 by Larry Kerschberg, All Rights Reserved

Topic Outline

- Databases - An Historical Perspective
- CBR at 30,000 feet
- CBR Life-Cycle Models
- XML/RDF/Semantic Web
- Role of XML/RDF/Semantic Web in CBR
- XML Databases to Support CBR
- CBR Framework for Collaborative Semantic Search - Knowledge Sifter
- CBR Meets 2.0 Challenge
- Conclusions

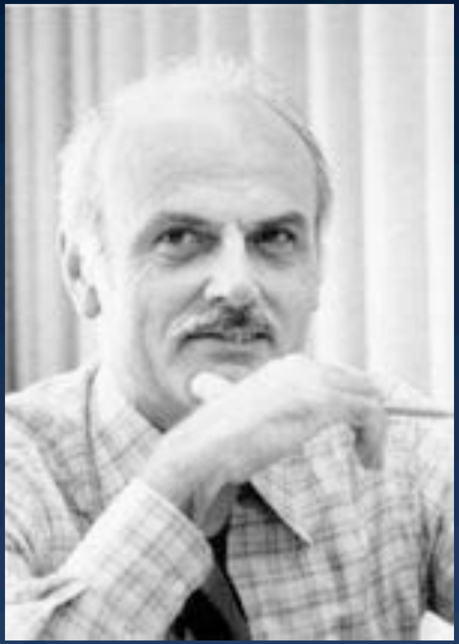
Databases - An Historical Perspective

Years	Models	Systems	Players	Conferences Journals
1960s	Network (CODASYL) Hierarchical	IDMS - GE/Honeywell IMS from IBM System 2000	**Charles Bachman*** Edgar Sibley Mike Senko	SIGFIDET
1970s	Relational Model Entity/Relationship Functional Model	Ingres Oracle (Prototypes)	**E.F (Ted) Codd*** Stonebraker Larry Ellison, Peter Chen Kerschberg, Shipman	SIGMOD ACM TODS IEEE TKDE
1980s	**SQL Standard - 1986** Semantic Model Normalization Theory Object-Oriented Transaction Management Active Databases Information Security	Ingres, Sybase, Imformix DB2, Oracle (commercial)	Jim Gray David Dewitt Mike Stonebraker David Mayer Dennis McLeod Stefan Ceri Jennifer Widom	PODS EDBT (European) Expert Database Systems (EDS) VLDB Journal
1990s	Semi-structured Data Mining Scientific DB	Oracle, DB2, Informix Tamino (XML)	Many authors	JIIS SIGKDD Data Mining & KD
2000s	Bioinformatics Stream Processing Event-Driven DBs RDF Databases	Aurora (Brown Univ.) Streambase	Zdonik Stonbraker	Journal of Data Semantics

*ACM Turing Award Winners

Hall of Luminaries

Charles Bachman



Ted Codd

Michael Stonebraker

Larry Ellison

Peter Chen

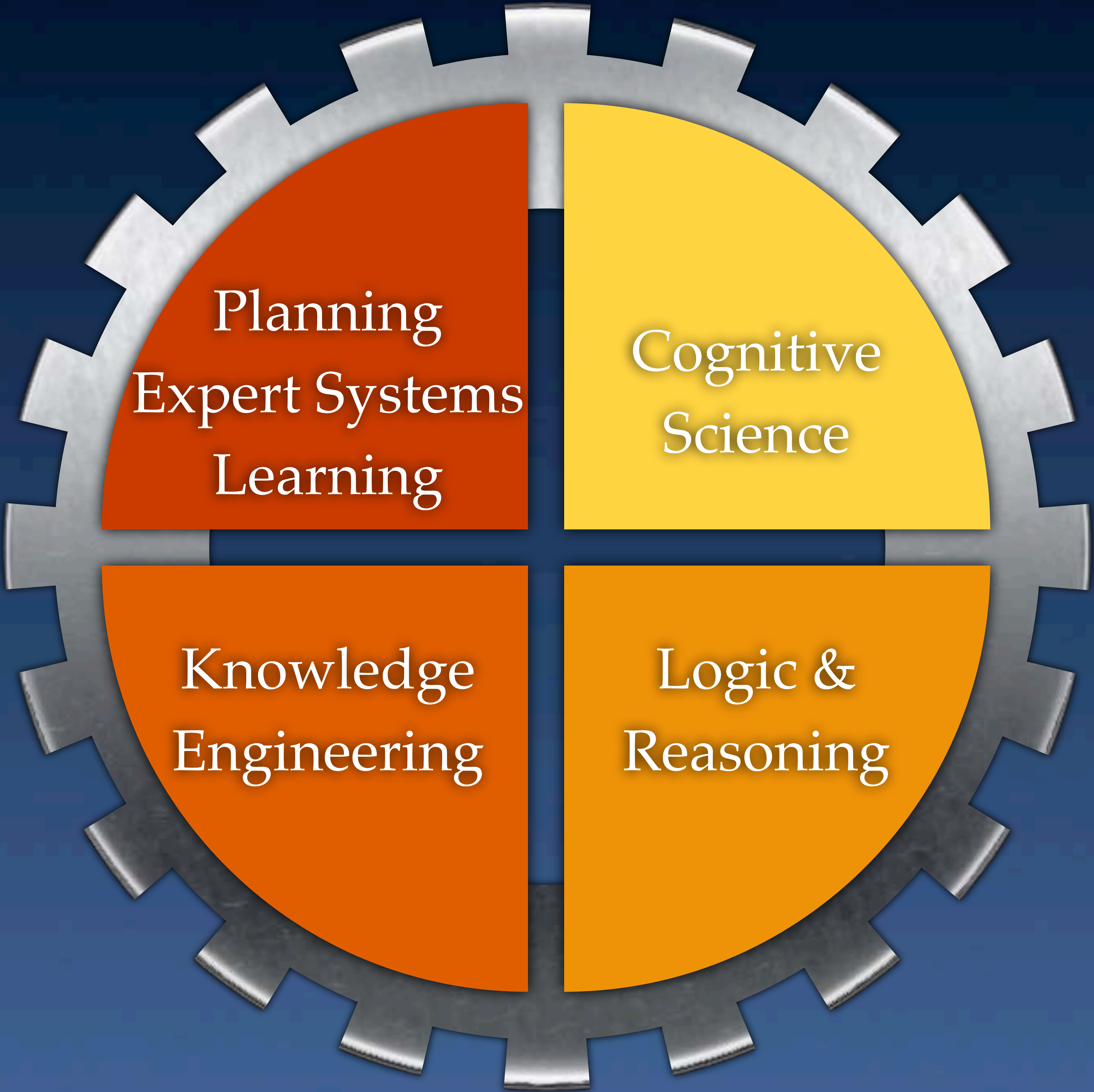
Stefano Ceri

Jennifer Widom

News
Steve Jobs and The Must-Have Music Player Everyone Is Talking About
iPod, Therefore i Am

Steve Jobs

CBR at 30,000 Feet



Planning
Expert Systems
Learning

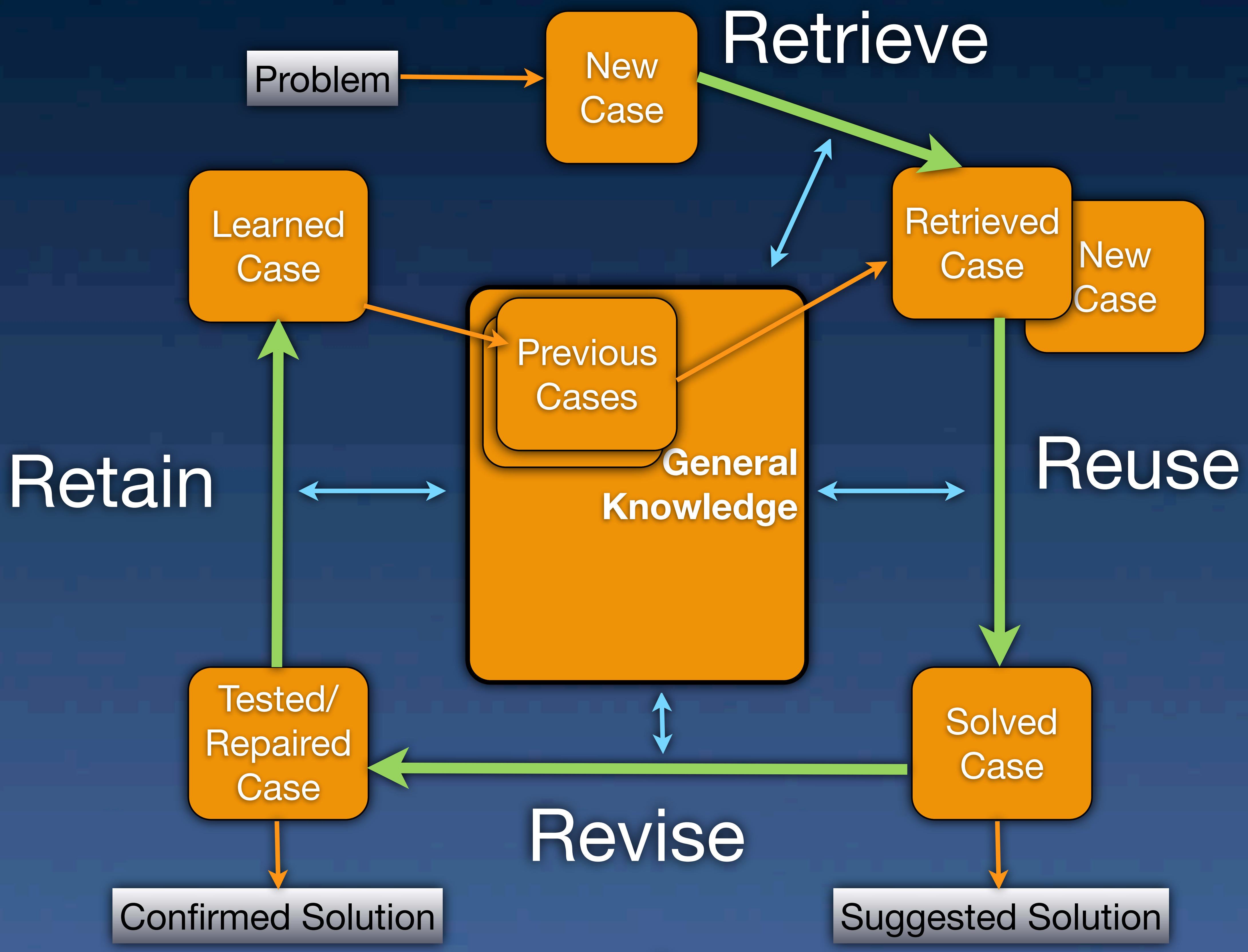
Cognitive
Science

Knowledge
Engineering

Logic &
Reasoning

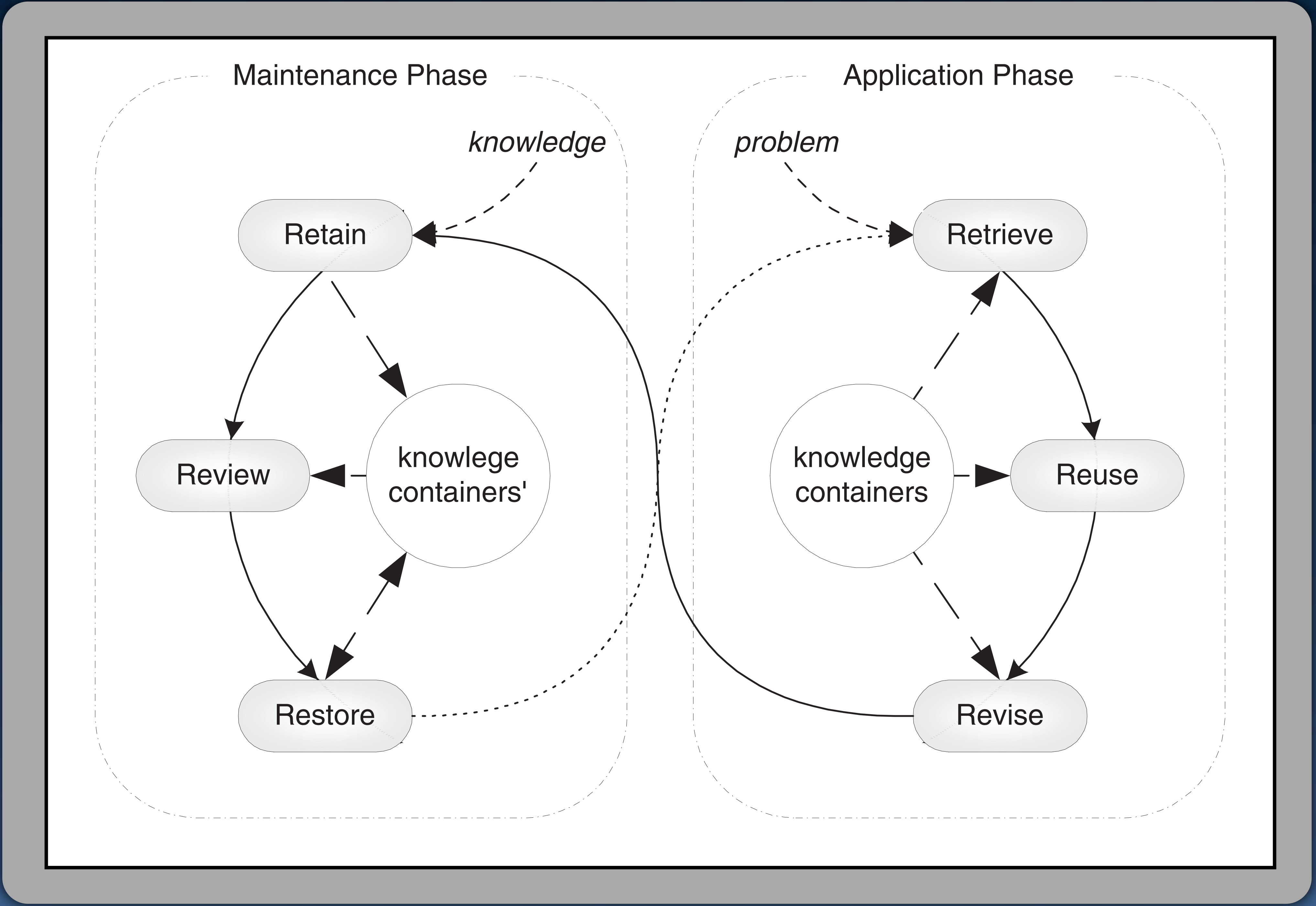
CBR Life-Cycle Models

CBR Life Cycle Model



From Aamodt & Plaza, 1994

CBR Model with Maintenance



From Roth-Berghofer and Iglezakis, GWCBR 2001

XML, RDF, and Semantic Web

Role of Metadata

- Metadata is data about data - describes the data and how it should be interpreted
  - Text, Numbers;
  - Class, Property, Task, etc.
- Metadata may be embedded within a document (e.g., tags) or external to the document (e.g., Relational DB Schema or a shared ontology).
- Embedded metadata provides the context and meaning for the data.
- Data DNA - Data knows everything that will possibly happen to it.

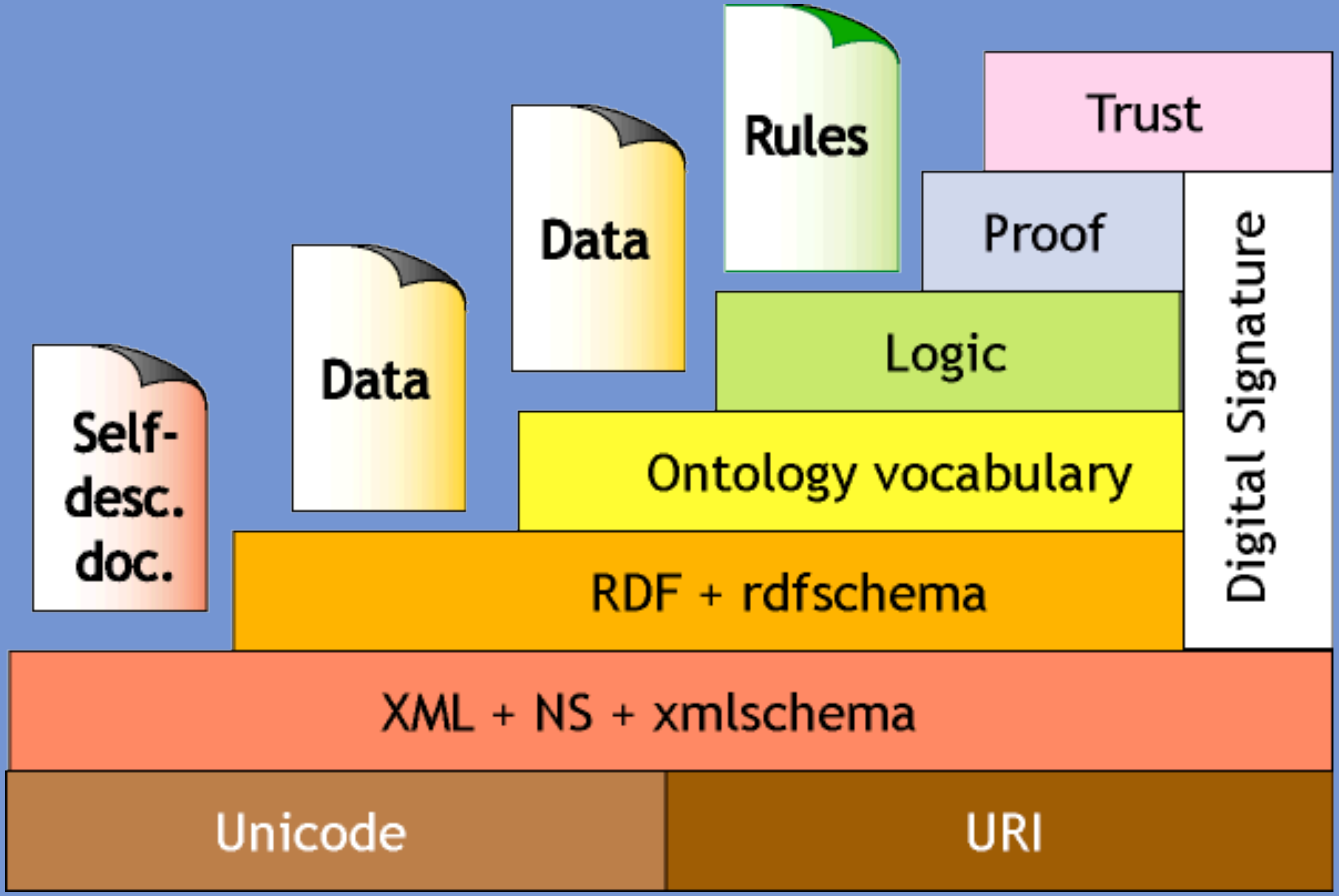
Metadata Standards Initiatives

- Dublin Core for library and Intellectual Property - hosted by OCLC in Dublin, Ohio
- XML - Extensible Markup Language
  - Provides the syntax for tagging document
  - XML Schema, XSLT, XML Protocol (SOAP)
- RDF - Resource Description Framework
  - Markup of Web resources, binary relations.
- Web Services and the Semantic Web
  - View the Web as a distributed information space
  - Allow computers, programs and agents to communicate in peer-to-peer using standard protocols.

Dublin Core Metadata Types

Content	Intellectual Property	Instantiation
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

The layered Semantic Web will have successive layers of knowledge, reasoning, learning, and trust.



Semantic Web Layers

XML, RDF and Relatives

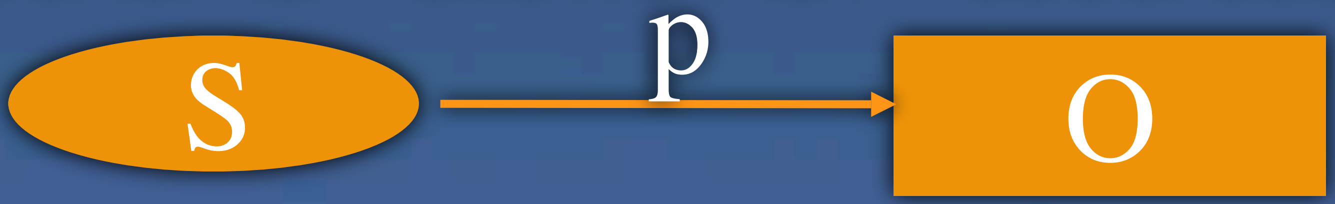
- XML (eXtensible Markup Language) is a markup language which indicates the meaning of the marked-up text.
  - Differs from HTML which deals with the presentation of information.
  - XML is really a *meta-language*, a mechanism for representing other languages in a standardized way.
  - The *interpretation, i.e., the meaning,* of the tags is left to the community which uses that markup language.

RDF - Resource Description Framework

- RDF is a meta-model to describe “things” on the Web.
  - Things are *resources* in the RDF vocabulary.
  - RDF model deals with:
    - Resources - a thing on the Web
    - Properties - a specific aspect, characteristic, attribute or relation the describes a resource.
    - Statements - consists of a specific resource, with a named property together with that property’s value.
    - The value can be either a resource or a literal (free text).

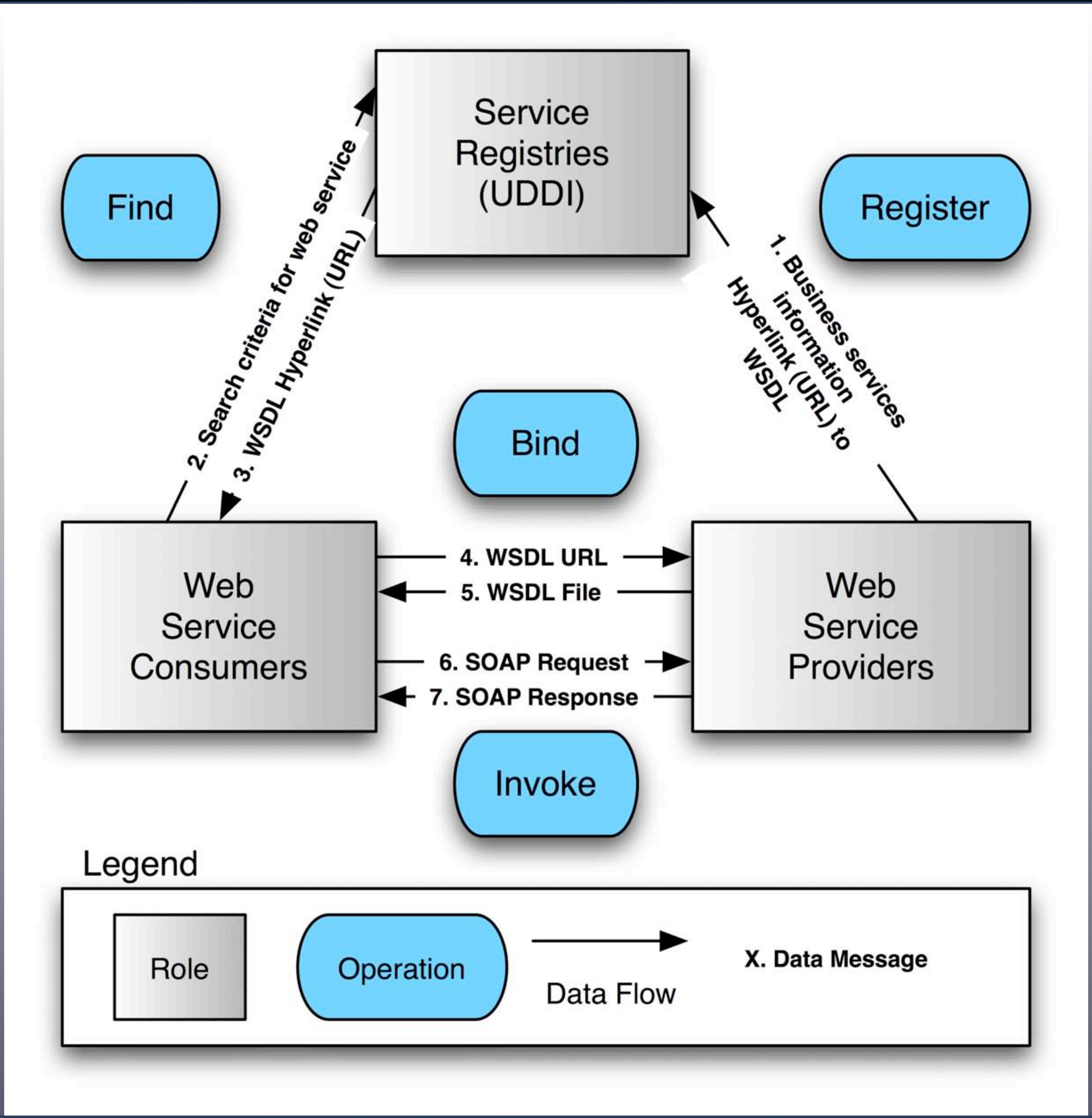
RDF Data Model

- The RDF data model is defined as follows:
  - There is a set of Resources.
  - There is a set of Literals.
  - There is a subset of Resources called Properties.
  - There is a set of Statements, each element of which is a triple of the form:
    - {pred, sub, obj}
      Where pred is a property (member of Properties),
      sub is resource (member of Resources) and
      obj is either a resource or a literal (member of Literals).
- RDF Schema - allows RDF resources to be typed.



Ontology

- **Ontology** is defined as the “*science or study of being*”, *Oxford English Dictionary*
- Ontology building involves identifying the domain objects, their relationship to one another.
- ***Semantic Web*** researchers consider an ontology to consist of:
  - A set of knowledge terms, which includes the vocabulary,
  - the semantic interconnections, and
  - some rules of inference and logic for some particular domain of discourse.



Web Services Protocols

UDDI, WSDL and SOAP

Role of XML, RDF, and Semantic Web in CBR

XML Specification of a Case

From Coyle, Hayes, Cunningham, Representing Cases for CBR in XML, ICCBR, 1999

```
<case name="DUB-OSL #34">
  <features>
    <username>Coyle</username>
    <traveloffer>
      <origin>DUB</origin>
      <destination>OSL</destination>
      <departuretime>Mon, 2 Dec 2002 at 6:45 GMT</departuretime>
      <arrivaltime>Mon, 2 Dec 2002 at 12:00 CET</arrivaltime>
      <distance>1051</distance>
      <flighttime>255</flighttime>
      <hops>
        <numberofhops>2</numberofhops>
        <hop>
          <origin>DUB</origin>
          <destination>AMS</destination>
          <carrier>KLM</carrier>
          <departuretime>Mon, 2 Dec 2002 at 6:45 GMT</departuretime>
          <arrivaltime>Mon, 2 Dec 2002 at 9:20 CET</arrivaltime>
          <class>Coach</class>
        </hop>
<hop>
          <origin>AMS</origin>
          <destination>OSL</destination>
          <carrier>KLM</carrier>
          <departuretime>Mon, 2 Dec 2002 at 10:10 CET</departuretime>
          <arrivaltime>Mon, 2 Dec 2002 at 12:00 CET</arrivaltime>
          <class>Coach</class>
        </hop>
      </hops>
    </traveloffer>
    <recommendation>5</recommendation>
  </features>
</case>
```

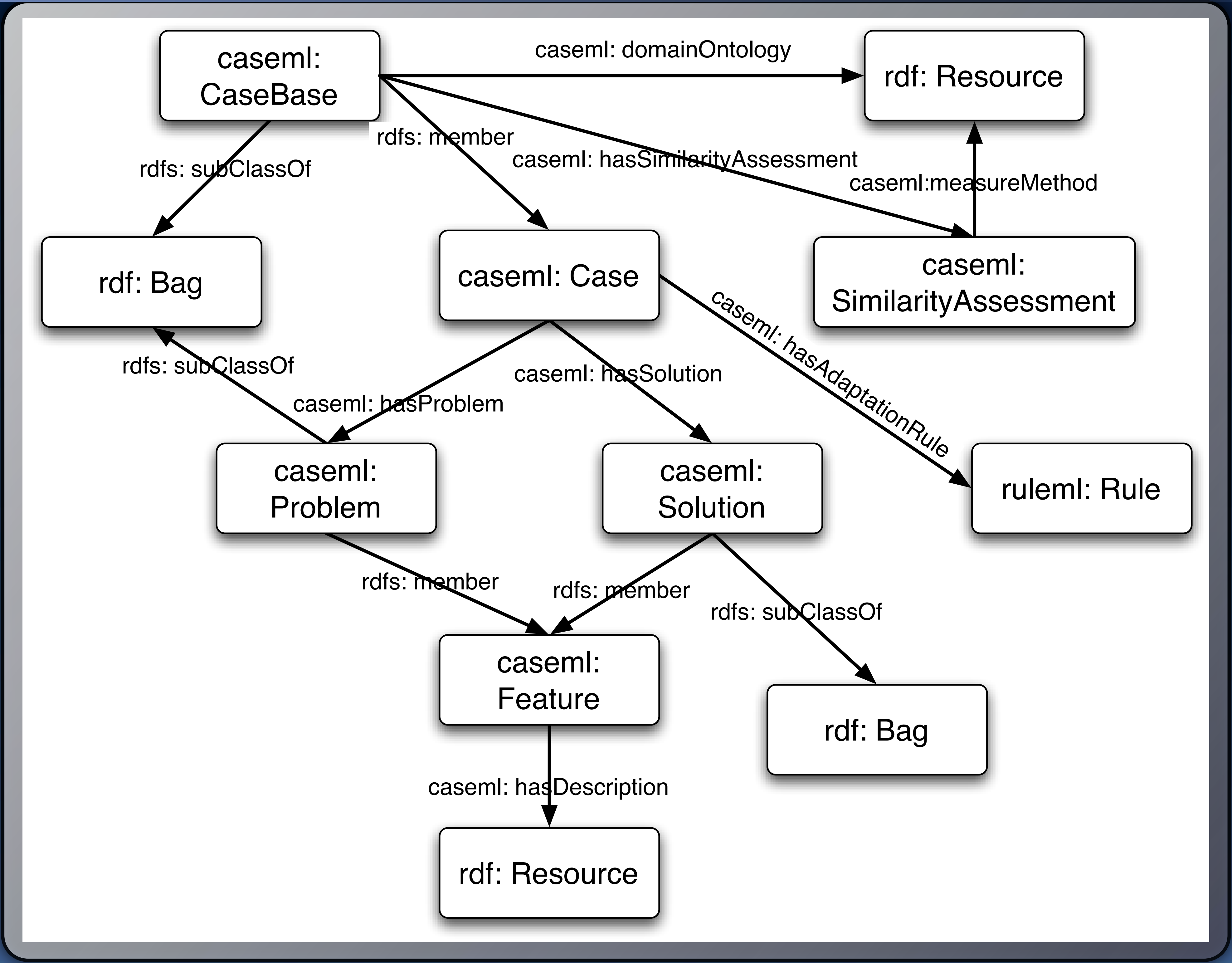
Definition of a Case in CaseML

- CaseML - a Case Markup Language (See Chen and Wu, ICCBR03, 2006)
- Classes in CaseML
  - **CaseBase** - class which acts as a container for cases;
  - **Case** - has one problem description and one solution description;
  - **Problem** - One problem has one or more features;
  - **Feature** - Feature contains *attribute-value pairs* or objects that are described by ***domain ontologies***.
  - **Solution** - One solution has one or more features
  - **Similarity Assessment** - The class encapsulates the detail about hos the case contained in this CaseBase would be assessed.

Definition of a Case in CaseML

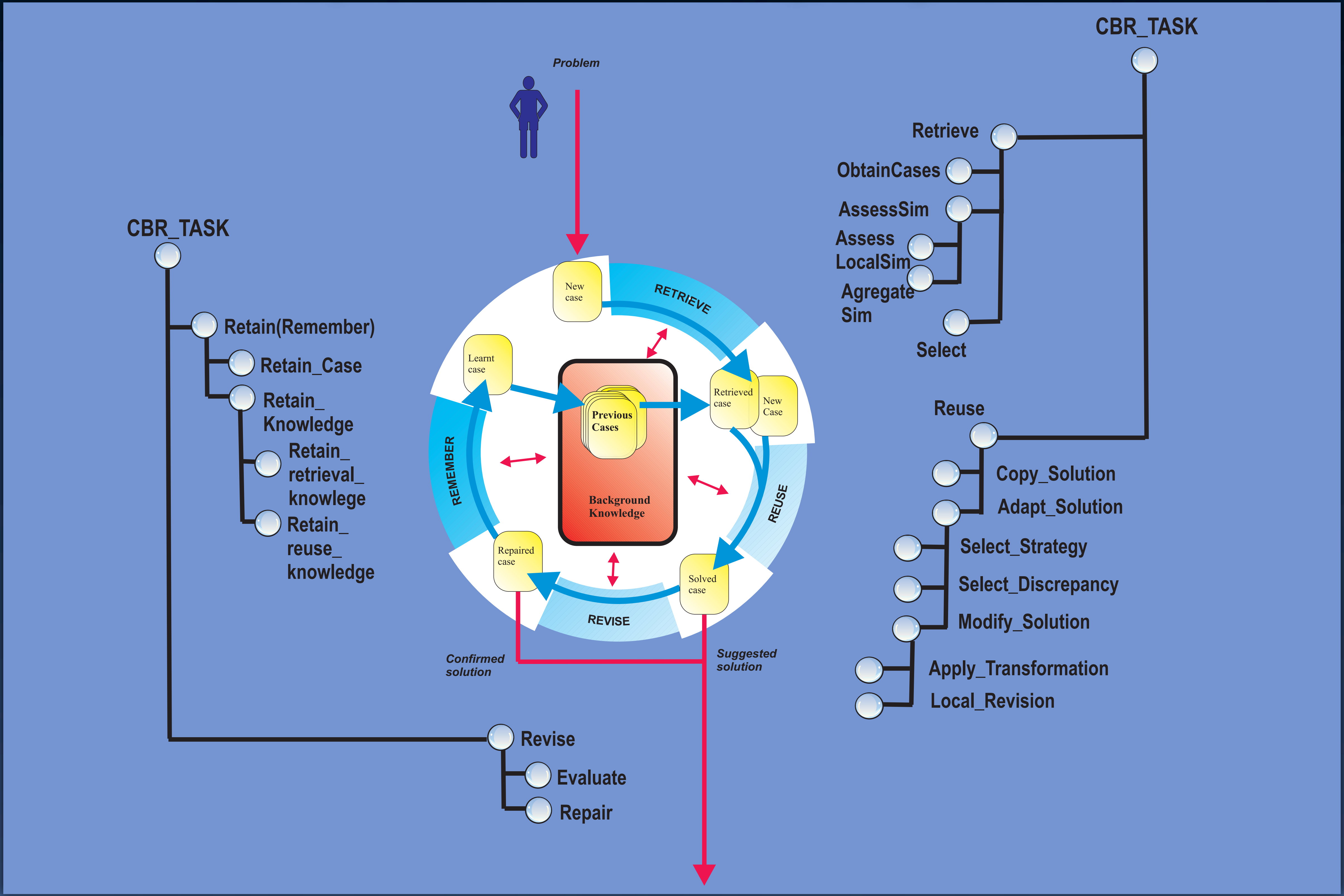
- Properties in CaseML
  - **domainOntology** - one case base belongs to one domain which has a URL that points to its definition;
  - **hasProblem** - this property establishes the relationship between the Case and Problem classes;
  - **hasDescription** - relates Feature class to domain ontology;
  - **hasSolution** - relationship between the Case and Solution classes.
  - **hasSimilarityAssessment** - property points to multiple Similarity Assessment classes, indicating multiple assessment algorithms;
  - **hasMeasureMethod** - relates the SimilarityAssessment class to specific measurement method identified as a Web resource.
  - **hasAdaptationRule** - relates Feature class with RuleML-specified adaptation rule.

Structure of a Case



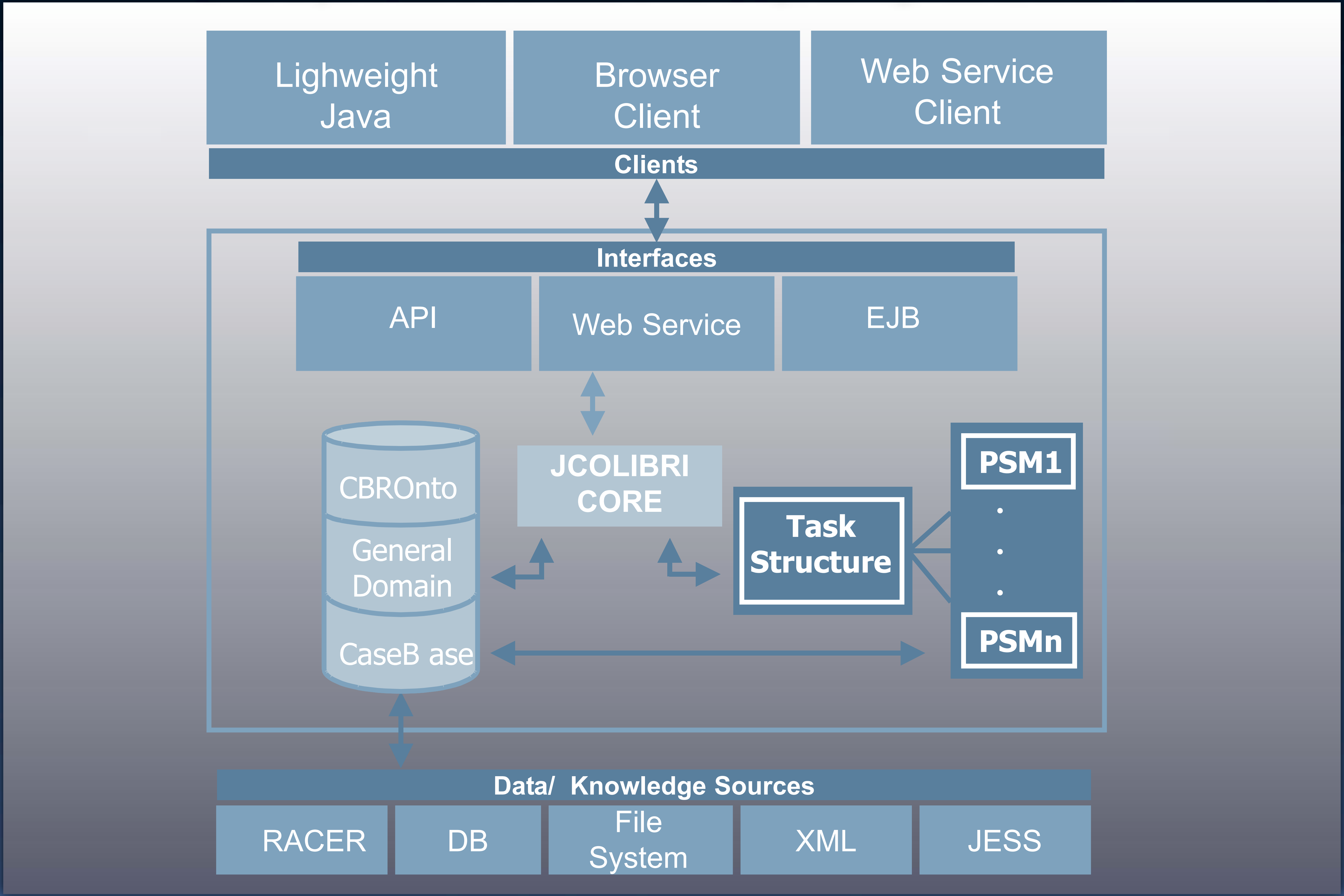
Namespaces:
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w4.org/2000/01/rdf-schema#"
xmlns:caseml="http://grid.zju.edu.cn/caseml#"

From Diaz-Agudo, et al, On Developing a Distributed CBR Framework through Semantic Web Services, OWL: Experiences and Directions Workshop, Galway, Ireland, 2005.



CBR Life Cycle and CBROnto Task Structure

From Diaz-Agudo, et al, On Developing a Distributed CBR Framework through Semantic Web Services, OWL: Experiences and Directions Workshop, Galway, Ireland, 2005.



jCOLIBRI Architecture

Role of XML & RDF Databases in Support of CBR

Role of Databases in CBR

- Direct support for two main life-cycle tasks:
  - ***Retrieve***
    - Use query languages such as XML-based XQuery, SQL/XML, or RDF-Based SPARQL, RDQL, RQL, etc.
  - ***Retain***
    - Store large collection of "case" instances in formats such as: relational, native-XML, or RDF.
    - Create indexes to allow fast retrieval of cases based on features, context, etc.

XML Databases

- Native XML Databases
  - Defines a logical model for an XML document, versus defining just the data in the document.
  - Model must include elements, attributes, PCDATA, and document order. Examples include the XPath DM and Document Object Model (DOM).
  - Document-based storage - entire document can be stored and retrieved
  - Node-based storage - individual nodes of the document stored and retrieved.
- Vendors
  - Berkeley DB from Oracle,
  - Tamino by Software AG

RDF Databases

- Oracle Spatial 10g includes an open, scalable, secure and reliable RDF management platform. Based on a graph data model, RDF triples are persisted, indexed and queried, similar to other object-relational data types.
- IBM's Web Ontology Manager is a lightweight, Web-based tool for managing ontologies expressed in Web Ontology Language (OWL).
- IBM's IODT, IBM's toolkit for ontology-driven development.
- IBM Semantic Layered Research Platform - IBM SLRP is a family of open-source Semantic Web software components including an enterprise RDF store, query engine, web application framework, RCP development libraries, etc.
- SemWeb for .NET supports persistent storage in MySQL, Postgre, and Sqlite; has been tested with 10-50 million triples; supports SPARQL.

CBR Meets Web 2.0 Challenge

Distributed Heterogeneous Collaborative Filtering for Case Discovery and Learning

CBR-2.0-Distributed Heterogeneous Collaborative Filtering

- Combine web-based authoritative (recommender, collaborative) sources:
  - Amazon (Books); iTunes (Music); Netflix and IMDB for Movies; ...
- Access Web 2.0 Collaborative Markup Applications
  - Wikipedia - Collaborative Encyclopedia
  - Delicious for tagged URLs
  - Flikr for pictures
  - Social Netorking Sites: FaceBook, MySpace, and Linkedin.
- The entire Internet and Web constitute the Case Base.
- Search for emergent case patterns by querying the markup tags across heterogeneous domains.
- Create a semantic web of concepts from a domain model

Conclusions

Conclusions

- Databases have not played major role in CBR, partly because the case bases have been small.
- However, XML, RDF, and the Semantic Web will change this and the CBR community should explore the use of DBMS to support the REs Live Cycle.
- Extend CBR to resource discovery in Web 2.0 - your new Case Base
- Invitation to the CBR community for a Special Issue of the Journal of Intelligent Information Systems.

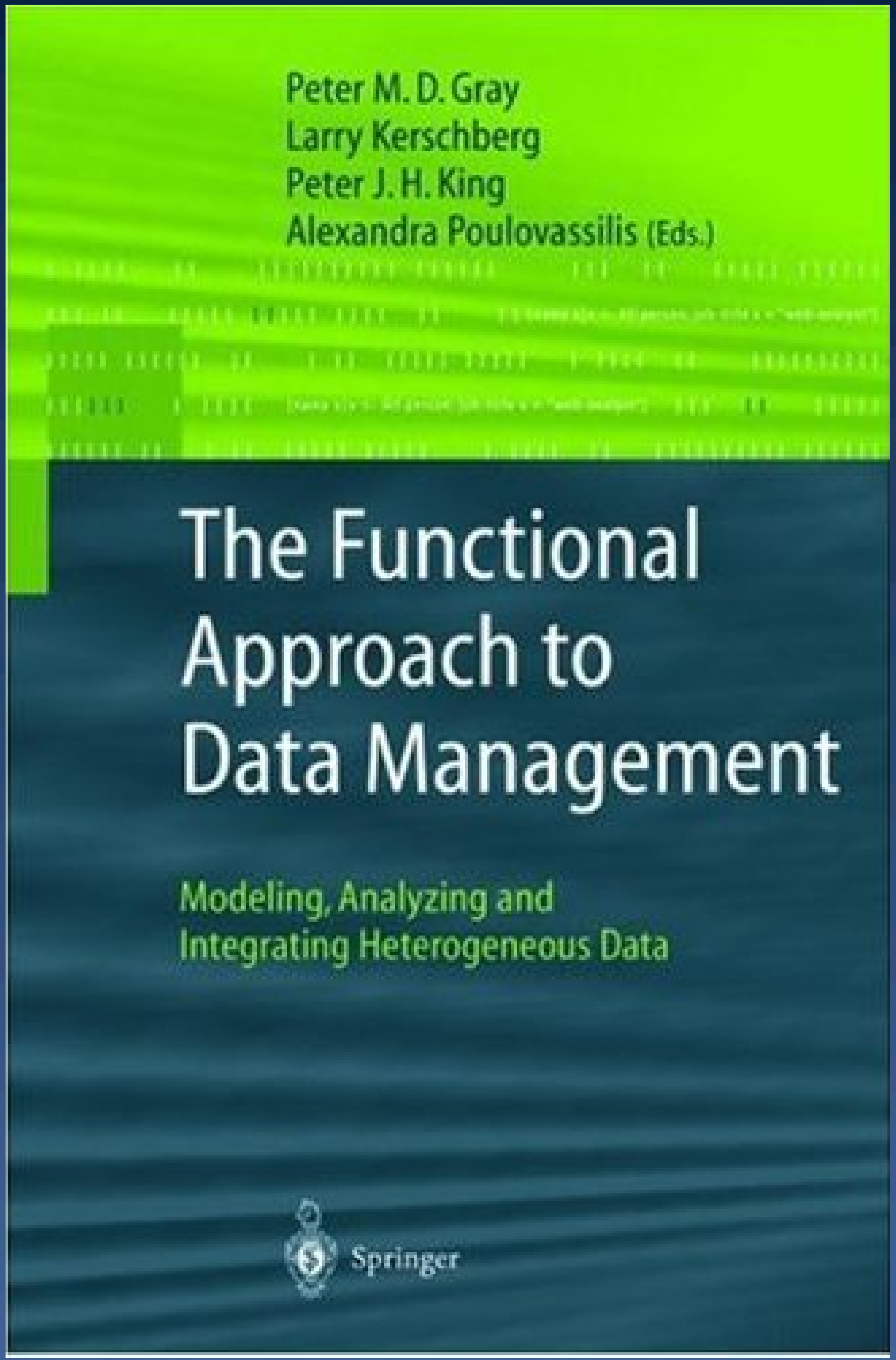
Unabashed Book Plug

The Functional Approach to Data Management:

Modeling, Analyzing and Integrating Heterogeneous Data

Peter Gray
Larry Kerschberg
Peter King
Alex Poulovassilis

Springer



Peter M. D. Gray
Larry Kerschberg
Peter J. H. King
Alexandra Poulovassilis (Eds.)

The Functional Approach to Data Management

Modeling, Analyzing and Integrating Heterogeneous Data

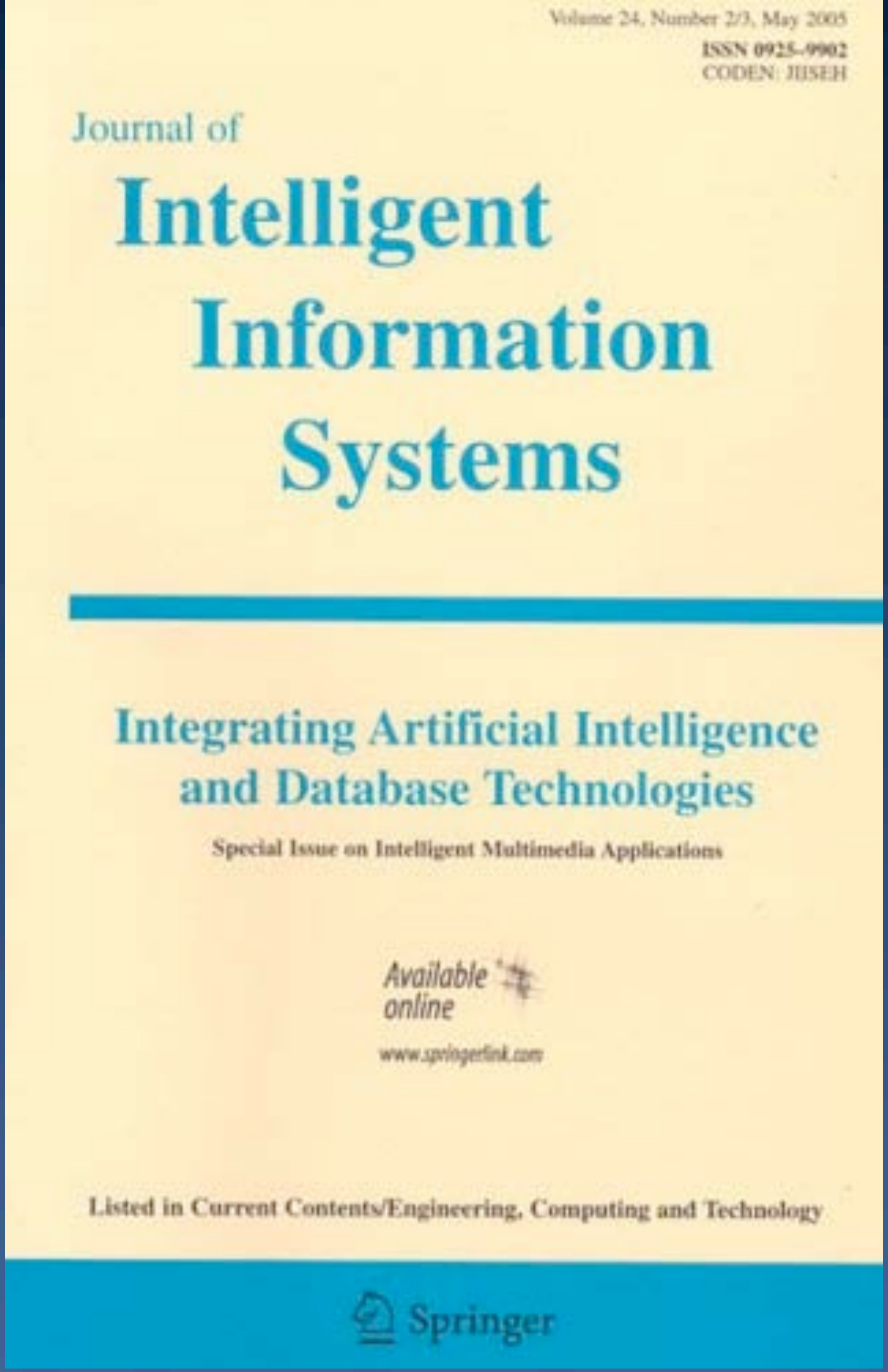
Springer

Unabashed Journal Plug

Journal of Intelligent Information Systems

Integrating Artificial Intelligence and Database Technologies

Editors-in-Chief
Larry Kerschberg
Zbigniew Ras
Maria Zemankova



Volume 24, Number 2/3, May 2005
ISSN 0925-9902
CODEN: JIISEH

Journal of
Intelligent Information Systems

Integrating Artificial Intelligence and Database Technologies

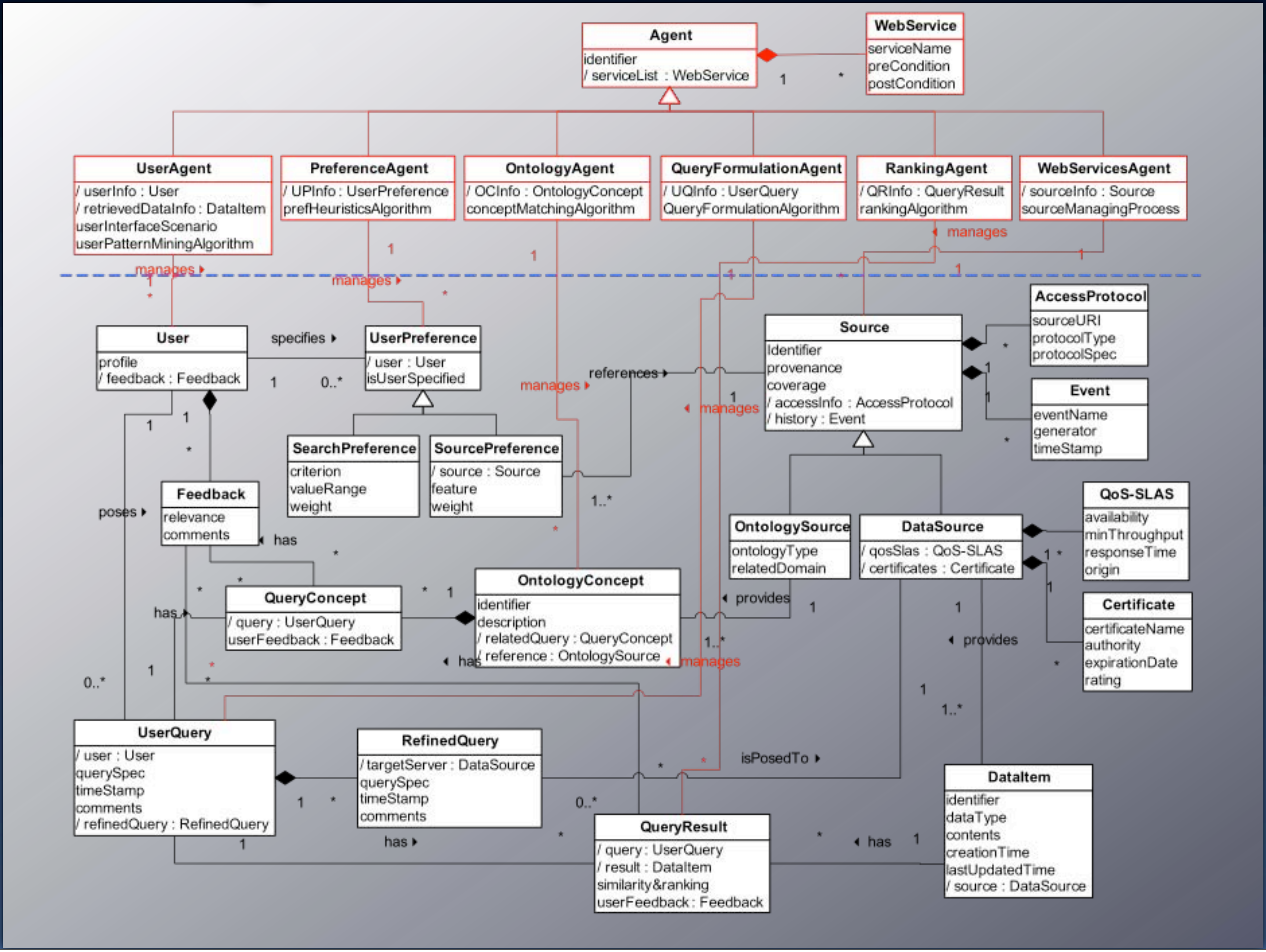
Special Issue on Intelligent Multimedia Applications

Available online
www.springerlink.com

Listed in Current Contents/Engineering, Computing and Technology

Springer

Knowledge Sifter Meta-Model



Agent
identifier
/ serviceList : WebService

WebService
serviceName
preCondition
postCondition

UserAgent
/ userInfo : User
/ retrievedDataInfo : DataItem
userInterfaceScenario
userPatternMiningAlgorithm

PreferenceAgent
/ UPInfo : UserPreference
prefHeuristicsAlgorithm

OntologyAgent
/ OCInfo : OntologyConcept
conceptMatchingAlgorithm

QueryFormulationAgent
/ UQInfo : UserQuery
QueryFormulationAlgorithm

RankingAgent
/ QRInfo : QueryResult
rankingAlgorithm

WebServicesAgent
/ sourceInfo : Source
sourceManagingProcess

User
profile
/ feedback : Feedback

UserPreference
/ user : User
isUserSpecified

SearchPreference
criterion
valueRange
weight

SourcePreference
/ source : Source
feature
weight

Source
Identifier
provenance
coverage
/ accessInfo : AccessProtocol
/ history : Event

AccessProtocol
sourceURI
protocolType
protocolSpec

Event
eventName
generator
timeStamp

Feedback
relevance
comments

OntologySource
ontologyType
relatedDomain

DataSource
/ qosSlas : QoS-SLAS
/ certificates : Certificate

QoS-SLAS
availability
minThroughput
responseTime
origin

Certificate
certificateName
authority
expirationDate
rating

QueryConcept
/ query : UserQuery
userFeedback : Feedback

OntologyConcept
identifier
description
/ relatedQuery : QueryConcept
/ reference : OntologySource

UserQuery
/ user : User
querySpec
timeStamp
comments
/ refinedQuery : RefinedQuery

RefinedQuery
/ targetServer : DataSource
querySpec
timeStamp
comments

QueryResult
/ query : UserQuery
/ result : DataItem
similarity&ranking
userFeedback : Feedback

DataItem
identifier
dataType
contents
creationTime
lastUpdatedTime
/ source : DataSource