ISO/IEC JTC 1/SC 32 N 1290

Date: 2005-04-22

REPLACES: --

ISO/IEC JTC 1/SC 32

Data Management and Interchange

Secretariat: United States of America (ANSI)
Administered by Farance Inc. on behalf of ANSI

DOCUMENT TYPE	Other Document (Open)
TITLE	Presentation - WG 2 Metadata Registries – Next Edition
SOURCE	B. Bargmeyer
PROJECT NUMBER	
STATUS	
REFERENCES	
ACTION ID.	FYI
REQUESTED ACTION	
DUE DATE	
Number of Pages	51
LANGUAGE USED	English
DISTRIBUTION	P & L Members SC Chair WG Conveners and Secretaries

Douglas Mann, Secretary, ISO/IEC JTC 1/SC 32
Farance Inc *, 360 Pelissier Lake Road, Marquette, MI 49855-9678, United States of America
Telephone: +1 906-249-9275; E-mail: MannD@battelle.org
available from the JTC 1/SC 32 WebSite http://jtc1sc32.org/
*Farance Inc. administers the ISO/IEC JTC 1/SC 32 Secretariat on behalf of ANSI

SC 32 Tutorial Session

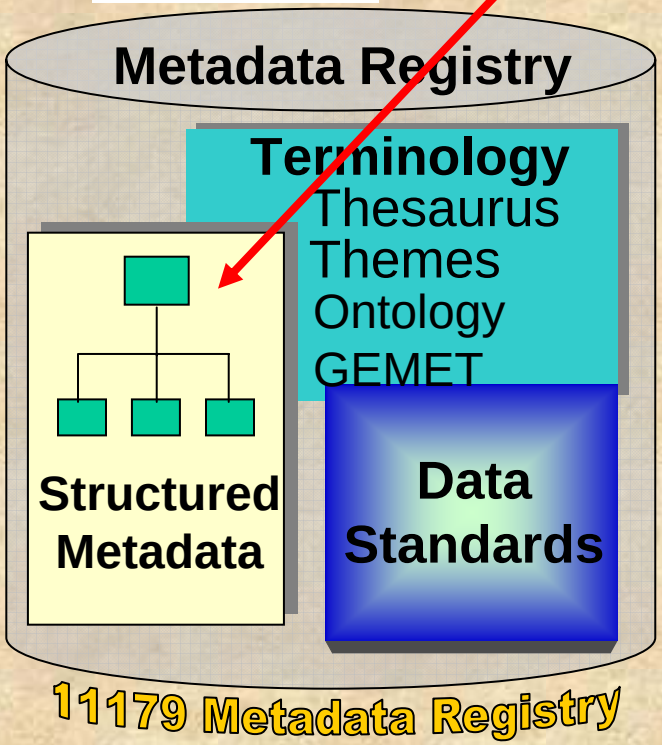
WG 2
Metadata Registries – Next Edition

April 18, 2005

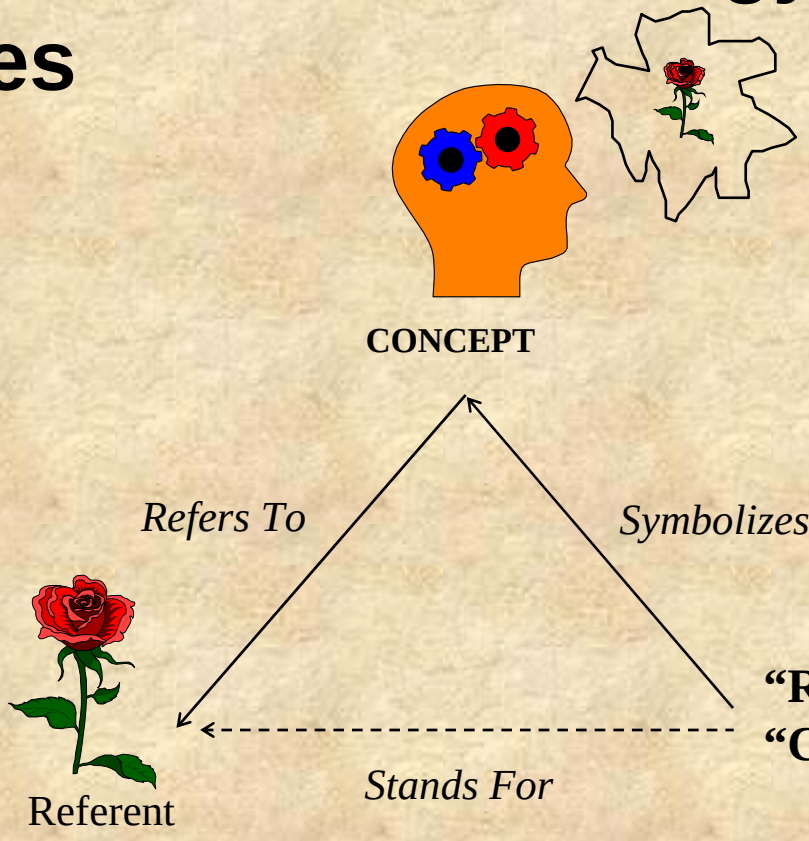
Bruce Bargmeyer,
Lawrence Berkley National Laboratory
University of California

Drawing Together

Metadata Registries



Terminology



Movement Toward Semantics Management

- Going beyond traditional Data Standards and Data Administration
- In addition to anchoring data with definitions, we want to process data and concepts based on context and relationships, possibly using inferences and rules.
- In addition to natural language, we want to capture semantics with more formal description techniques
  - FOL, DL, Common Logic, OWL
- Going beyond information system interoperability and data interchange to processing based on
  - inferences and
  - probabilistic correspondence between concepts found in natural language (in the wild) and both data in databases and concepts found in concept systems.

11179 Metadata Registries Extensions

Register (and manage) any semantics that are useful in managing data. E.g.

- Add semantic information beyond definitions
- link any concept found in concept systems or in the “wild” (or clusters of concepts and relations) to data

- This may include all of the permissible values (concepts) and the full concept systems in which the permissible values are found.
  - E.g., may want to register keywords, thesauri, taxonomies, ontologies, axiomitized ontologies….
- Lay Foundation for semantics based computing: Semantics Service Oriented Architecture, Semantic Grids, Semantics based workflows, Semantic Web …

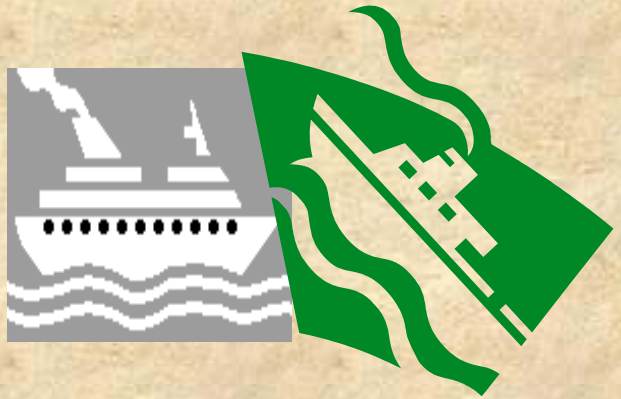
Data Base Languages

Relational data management

- Strengths
  - Underlying mathematical foundation
  - Powerful, well structured query language
- Weakness
  - Expressivity, for some concept systems
  - Performance

Other systems

- Object data management
- Graph data management



Samples of Eco & Bio Graph Data

- **Nutrient cycles in microbial ecologies** These are bipartite graphs, with two sets of nodes, *microbes and reactants (nutrients),* and directed edges indicating *input* and *output* relationships. Such nutrient cycle graphs are used to model the flow of nutrients in microbial ecologies, e.g., subsurface microbial ecologies for bioremediation.
- **Chemical structure graphs:** Here atoms are nodes, and chemical bonds are represented by undirected edges. Multi-electron bonds are often represented by multiple edges between nodes (atoms), hence these are multigraphs. Common queries include subgraph isomorphism. Chemical structure graphs are commonly used in chemoinformatics systems, such as Chem Abstracts, MDL Systems, etc.
- **Sequence data and multiple sequence alignments** . DNA/RNA/Protein sequences can be modeled as linear graphs
- **Topological adjacency** relationships also arise in anatomy. These relationships differ from partonomies in that adjacency relationships are undirected and not generally transitive.

Eco & Bio Graph Data (Continued)

- **Taxonomies of proteins, chemical compounds, and organisms, ...** These taxonomies (classification systems) are usually represented as directed acyclic graphs (partial orders or lattices). They are used when querying the pathways databases. Common queries are subsumption testing between two terms/concepts, i.e., is one concept a subset or instance of another. Note that some phylogenetic tree computations generate unrooted, i.e., undirected. trees.
- **Metabolic pathways:** chemical reactions used for energy production, synthesis of proteins, carbohydrates, etc. Note that these graphs are usually cyclic.
- **Signaling pathways:** chemical reactions for information transmission and processing. Often these reactions involve small numbers of molecules. Graph structure is similar to metabolic pathways.
- **Partonomies** are used in biological settings most often to represent common topological relationships of gross anatomy in multi-cellular organisms. They are also useful in sub-cellular anatomy, and possibly in describing protein complexes. They are comprised of *part-of* relationships (in contrast to *is-a* relationships of taxonomies). Part-of relationships are represented by directed edges and are transitive. Partonomies are directed acyclic graphs.
- **Data Provenance** relationships are used to record the source and derivation of data. Here, some nodes are used to represent either individual "facts" or "datasets" and other nodes represent "data sources" (either labs or individuals). Edges between "datasets" and "data sources" indicate "contributed by". Other edges (between datasets (or facts)) indicate derived from (e.g., via inference or computation). Data provenance graphs are usually directed acyclic graphs.

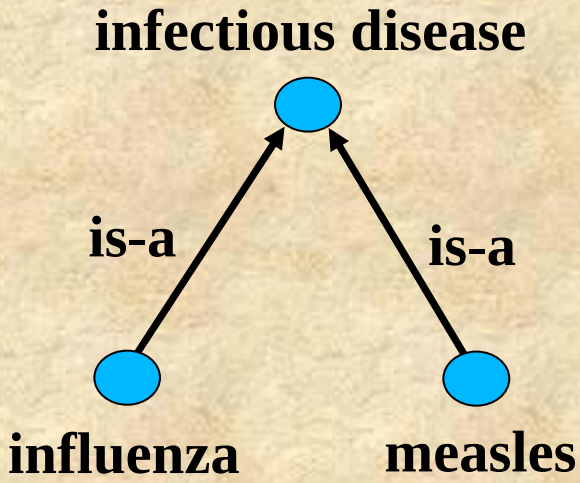
A graph theoretic characterization

- Readily comprehensible characterization of metadata structures
- Graph structure has implications for:
  - Integrity Constraint Enforcement
  - Data structures
  - Query languages
  - Combining metadata sets
  - Algorithms for query processing

Definition of a graph

- Graph = vertex (node) set + edge set
- Nodes, edges may be labeled
- Edge set = binary relation over nodes
  - cf. NIAM
- Labeled edge set
  - RDF triples (subject, predicate, object)
  - predicate = edge label
- Typically edges are directed

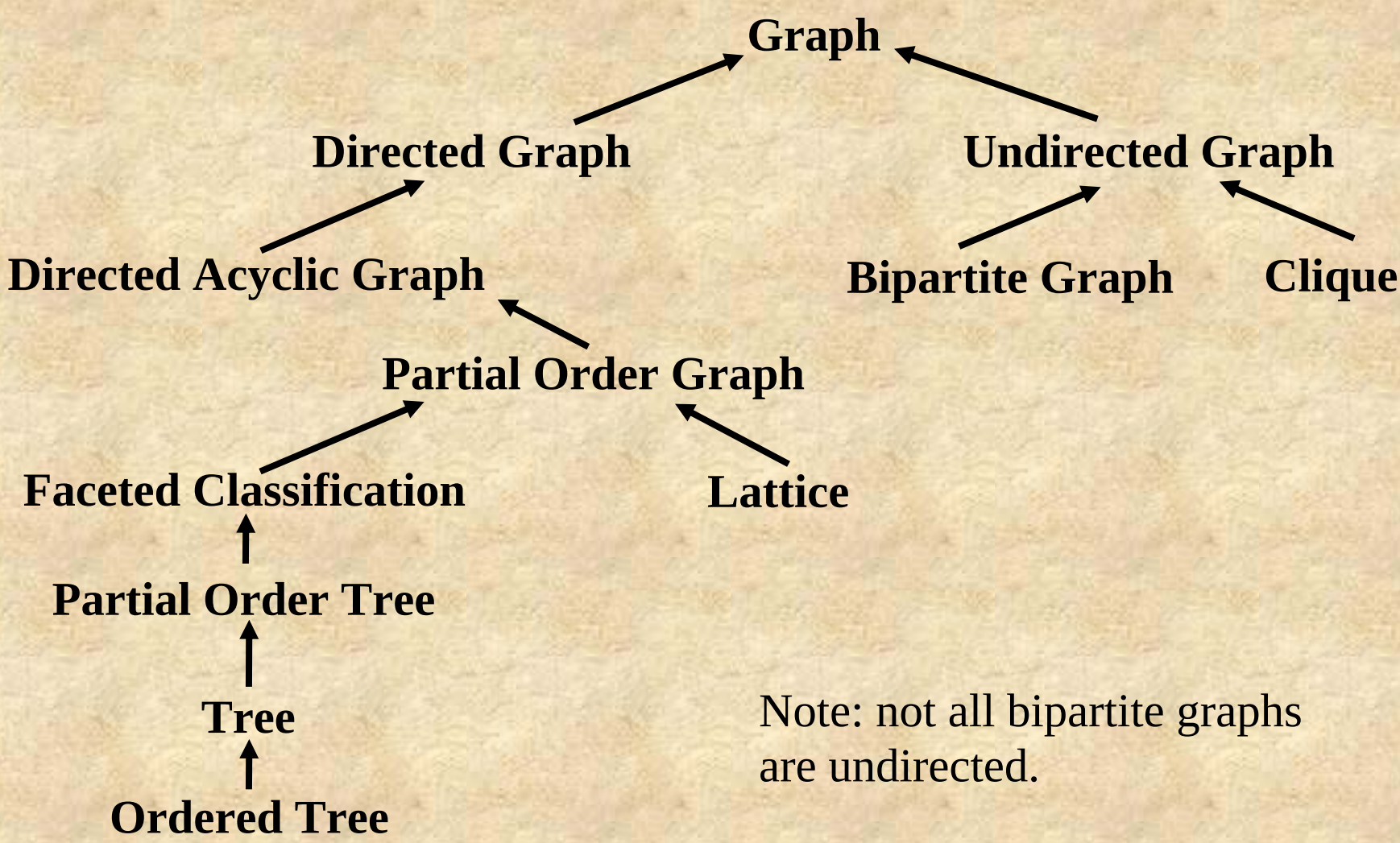
Example of a graph



Types of Metadata Graph Structures

- Trees
- Partially Ordered Trees
- Ordered Trees
- Faceted Classifications
- Directed Acyclic Graphs
- Partially Ordered Graphs
- Lattices
- Bipartite Graphs
- Directed Graphs
- Cliques
- Compound Graphs

Graph Taxonomy

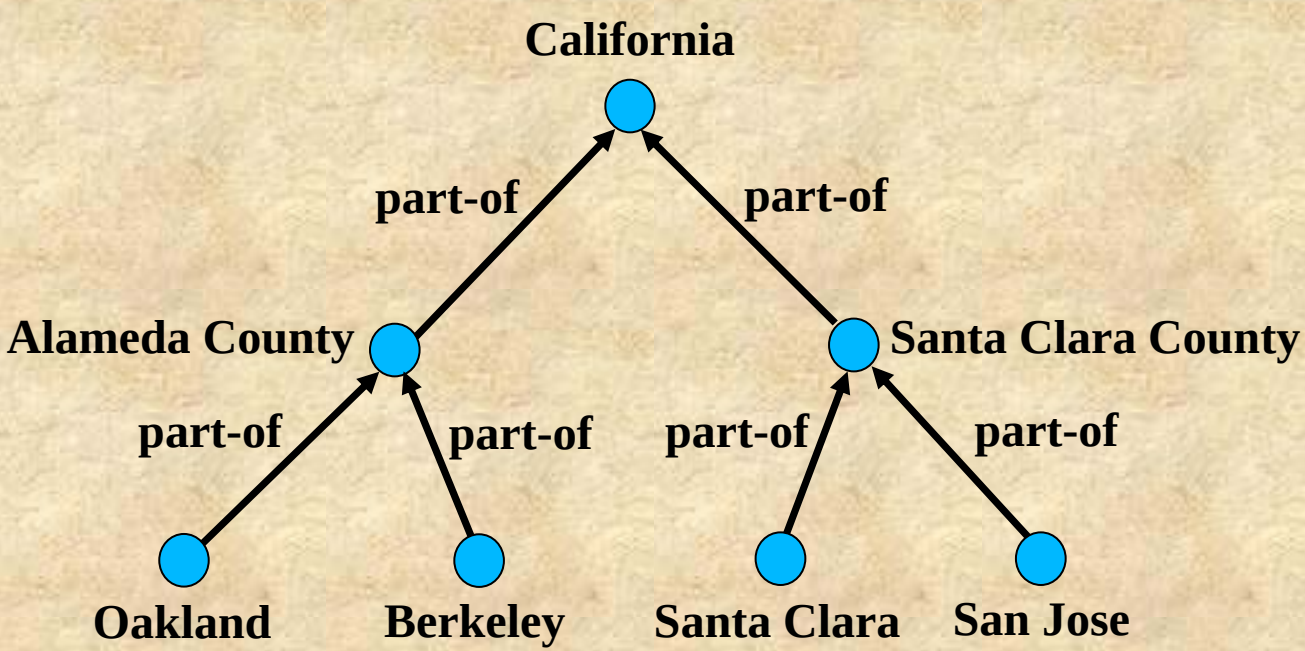


Note: not all bipartite graphs are undirected.

Trees

- In metadata settings trees are almost most often directed
  - edges indicate direction
- In metadata settings trees are usually partial orders
  - Transtivity is implied (see next slide)
  - Not true for some trees with mixed edge types.
  - Not always true for all partonomies

Example: Tree



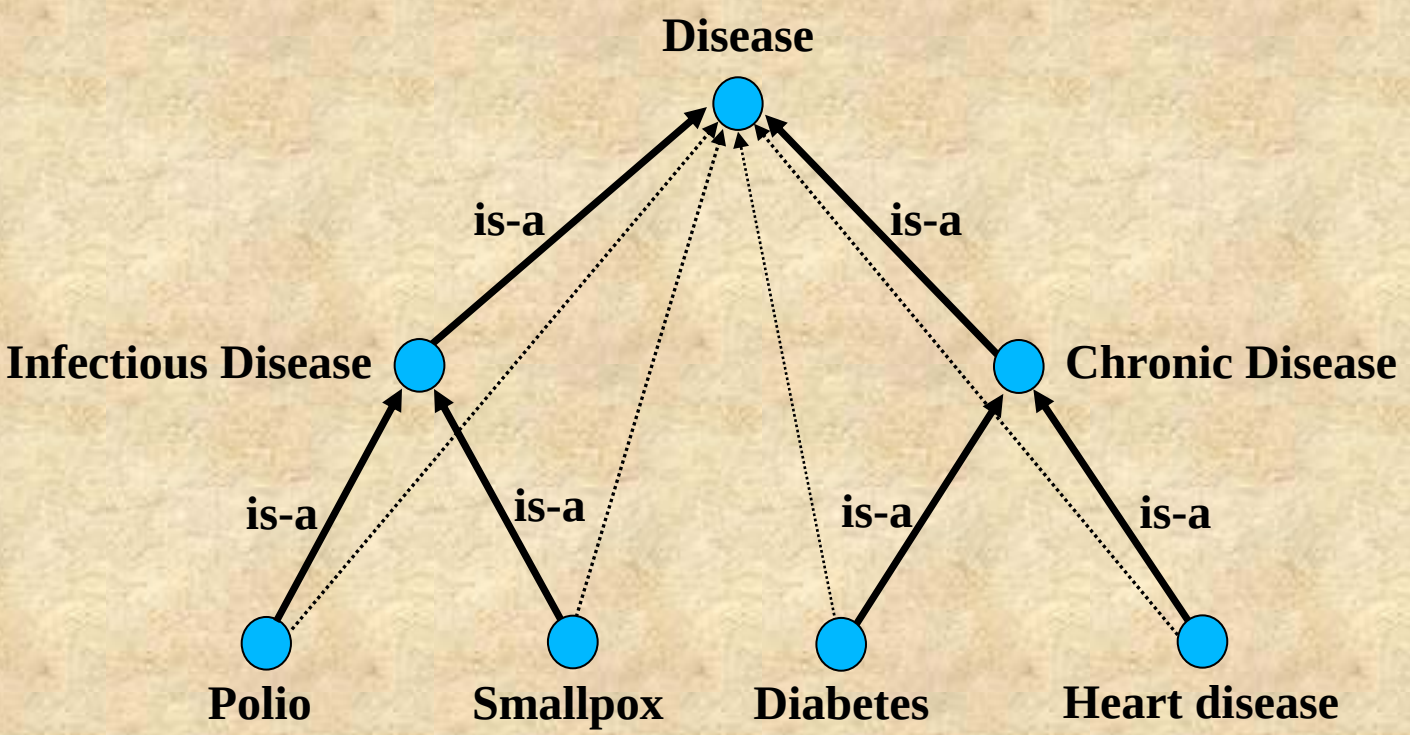
Trees - cont.

- Uniform vs. non-uniform height subtrees
- Uniform height subtrees
  - fixed number of levels
  - common in dimensions of multi-dimensional data models
- Non-uniform height subtrees
  - common terminologies

Partially Ordered Trees

- A conventional directed tree
- Plus, assumption of transitivity
- Usually only show immediate ancestors (transitive reduction)
- Edges of transitive closure are implied
- Classic Example:
  - Simple Taxonomy, “is-a” relationship

Example: Partial Order Tree

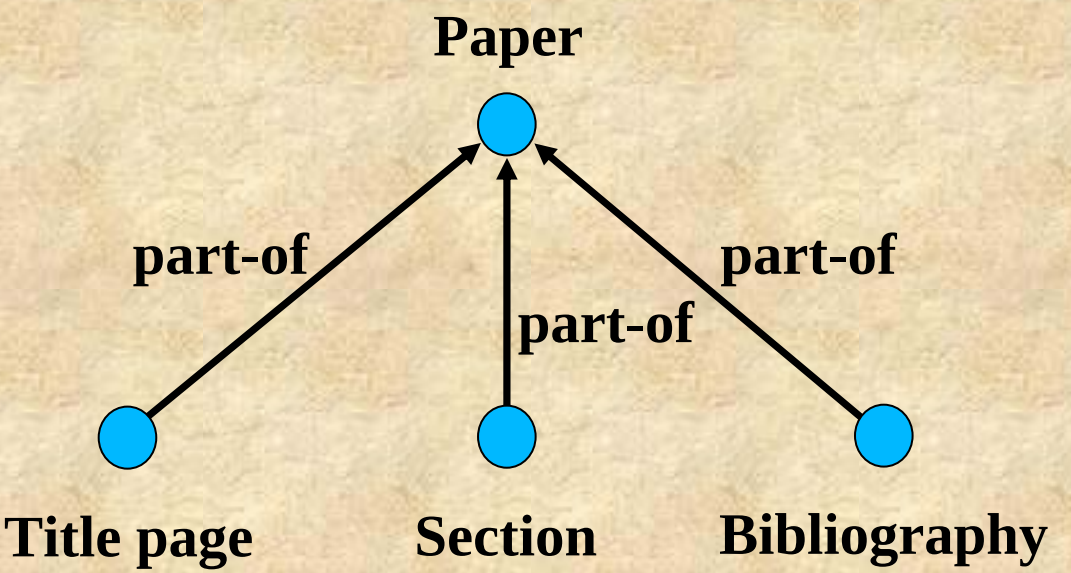


Signifies inferred is-a relationship

Ordered Trees

- Order here refers to order among sibling nodes (not related to partial order discussed elsewhere)
- XML documents are ordered trees
  - Ordering of “sub-elements” is to support classic linear encoding of documents

Example: Ordered Tree

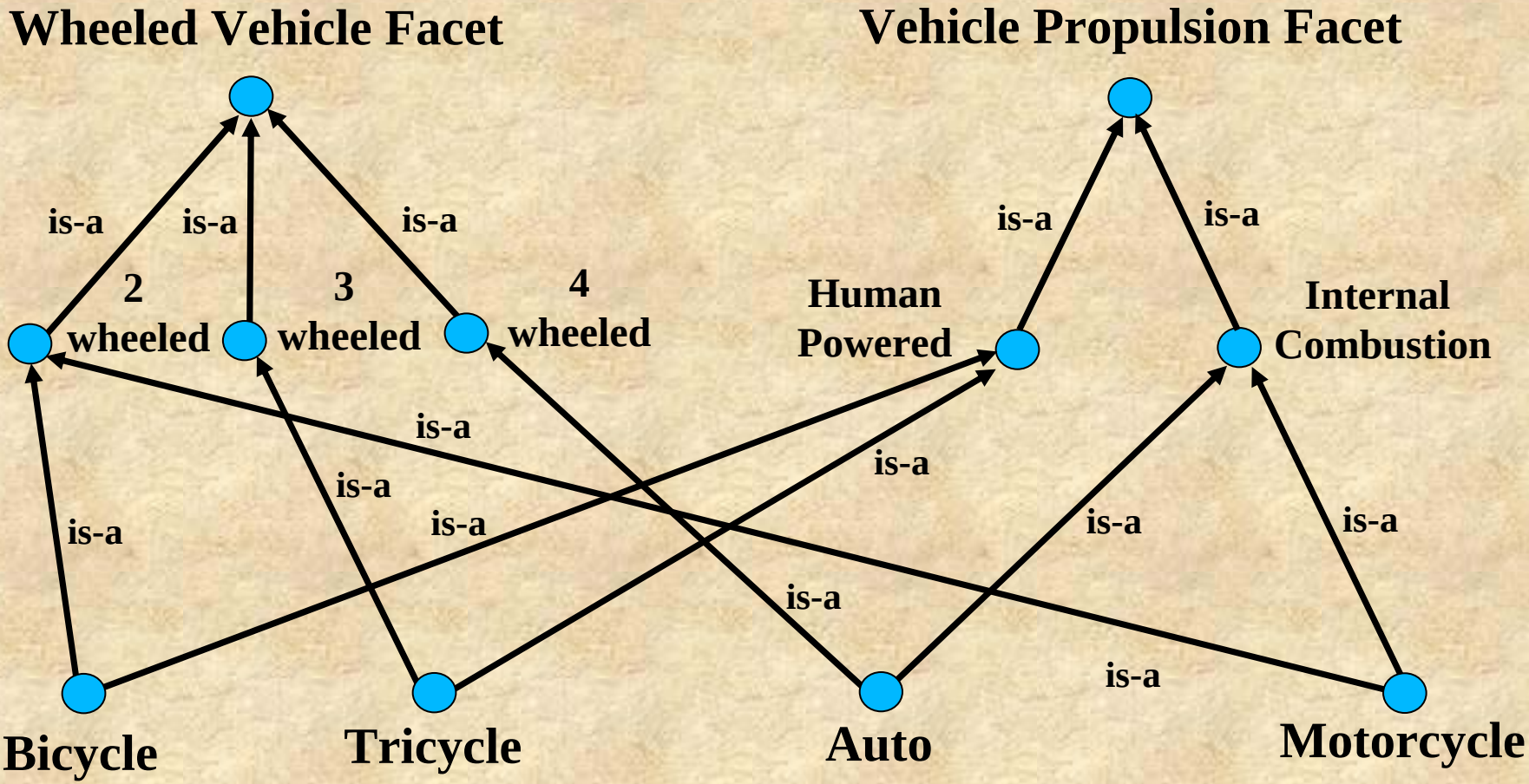


Note: implicit ordering relation among parts of paper.

Faceted Classification

- Classification scheme has mulitple facets
- Each facet = partial order tree
- Categories = conjunction of facet values (often written as [facet1, facet2, facet3])
- Faceted classification = a simplified partial order graph
- Introduced by Ranganathan in 19th century, as Colon Classification scheme
- Faceted classification can be descirbed with Description Logc, e.g., OWL-DL

Example: Faceted Classification



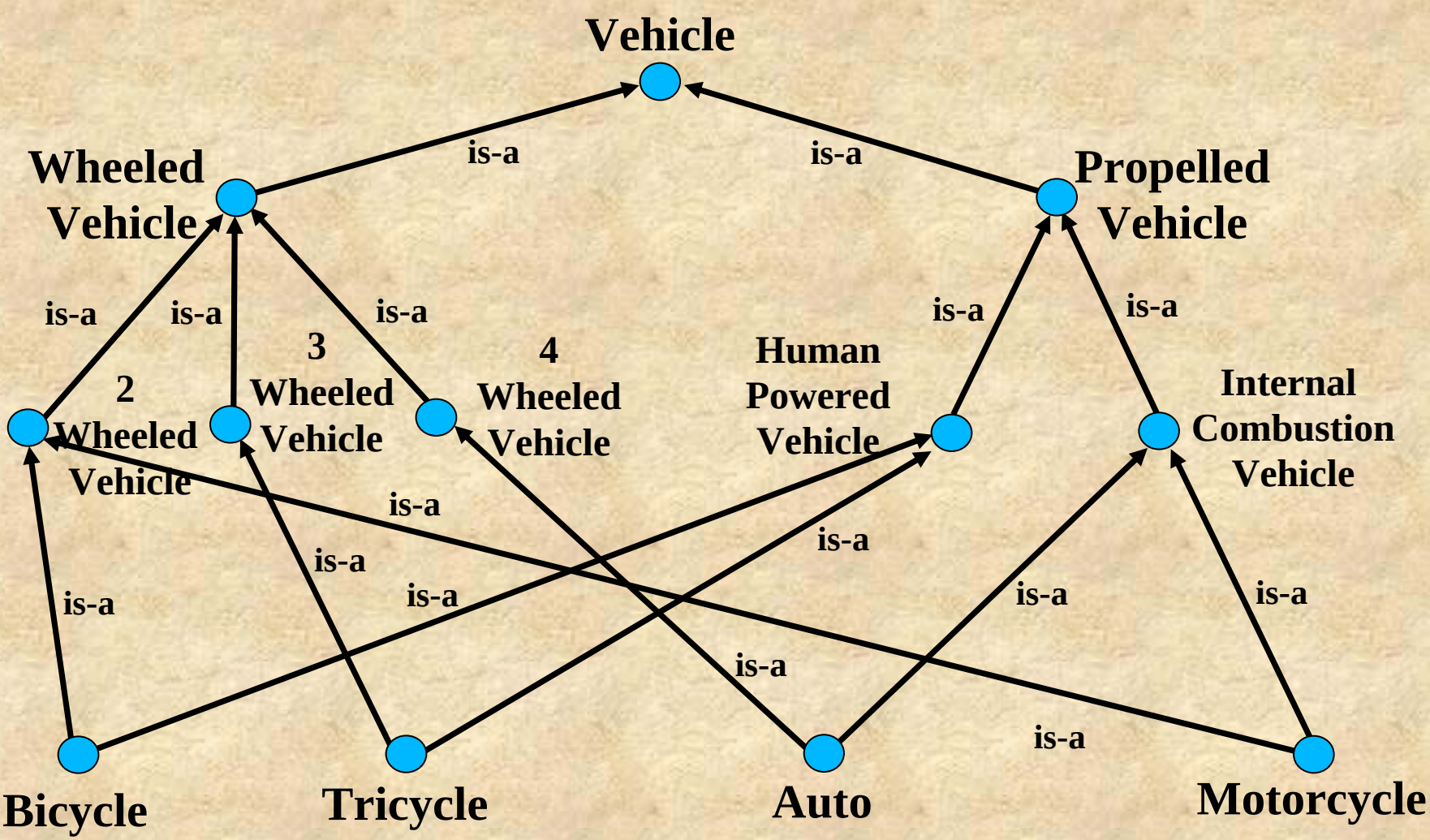
Faceted Classifications and Multi-dimensional Data Model

- MDM – a.k.a. OLAP data model
  - Online Analytical Processing data model
  - Star / Snowflake schemas
- Fact Tables
  - fact = function over Cartesian product of dimensions
  - dimensions = facets
    - geographic region, product category, year, ...

Directed Acylic Graphs

- Graph:
  - Directed edges
  - No cycles
  - No assumptions about transitivity (e.g., mixed edge types, some partonomies)
  - Nodes may have multiple parents
- Examples:
  - Partonomies (“part-of”) - transitivity is not always true

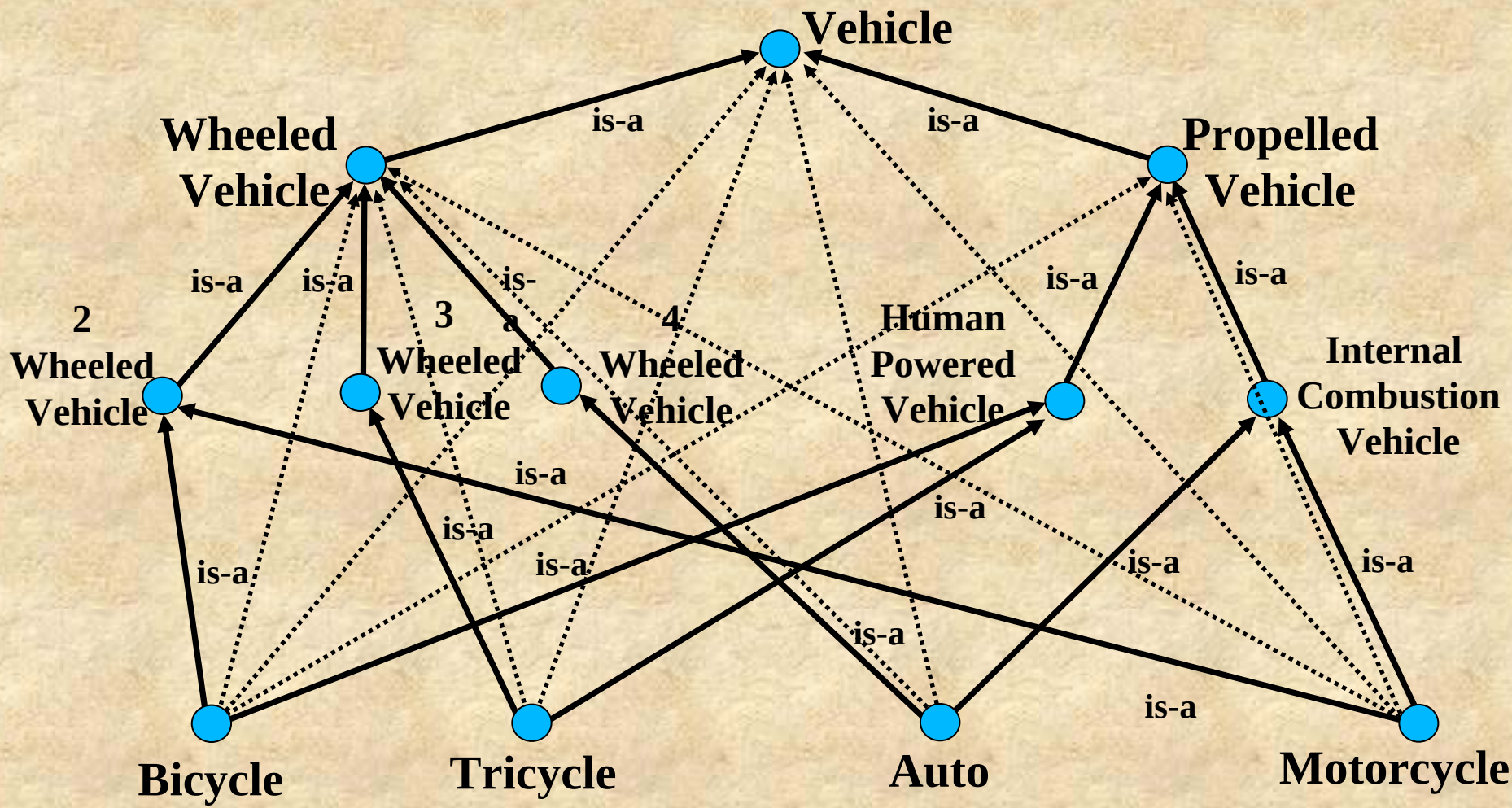
Example: Directed Acyclic Graph



Partial Order Graphs

- Directed acyclic graphs + inferred transitivity
- Nodes may have multiple parents
- Most taxonomies drawn as transitive reduction, transitive closure edges are implied.
- Examples:
  - all taxonomies
  - most partonomies
  - multiple inheritance
- POGs can be described in Description Logic, e.g., OWL-DL

Example: Partial Order Graph



Dashed line = inferred is-a (transitive closure)

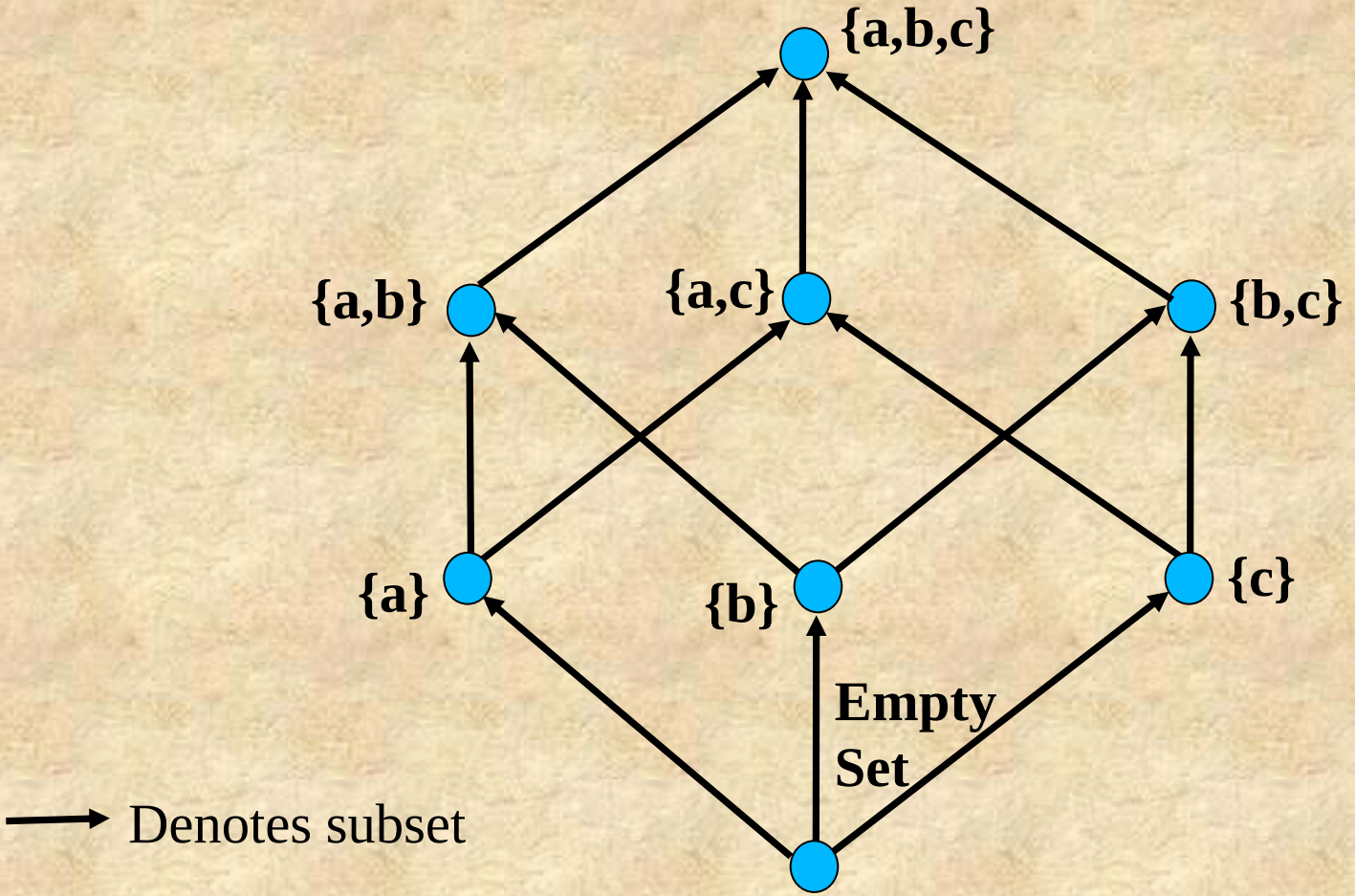
Directed Graph

- Generalization of DAG (directed acyclic graph)
- Cycles are allowed
- Arises when many edge types allowed
- Example: UMLS

Lattices

- A partial order
- For every pair of elements A and B
  - There exists a least upper bound
  - There exists a greatest lower bound
- Example:
  - The power set (all possible subsets) of a finite set
  - LUB(A,B) = union of two sets A, B
  - GLB(A,B) = intersect of two sets A,B

Example Lattice: Powerset of 3 element set



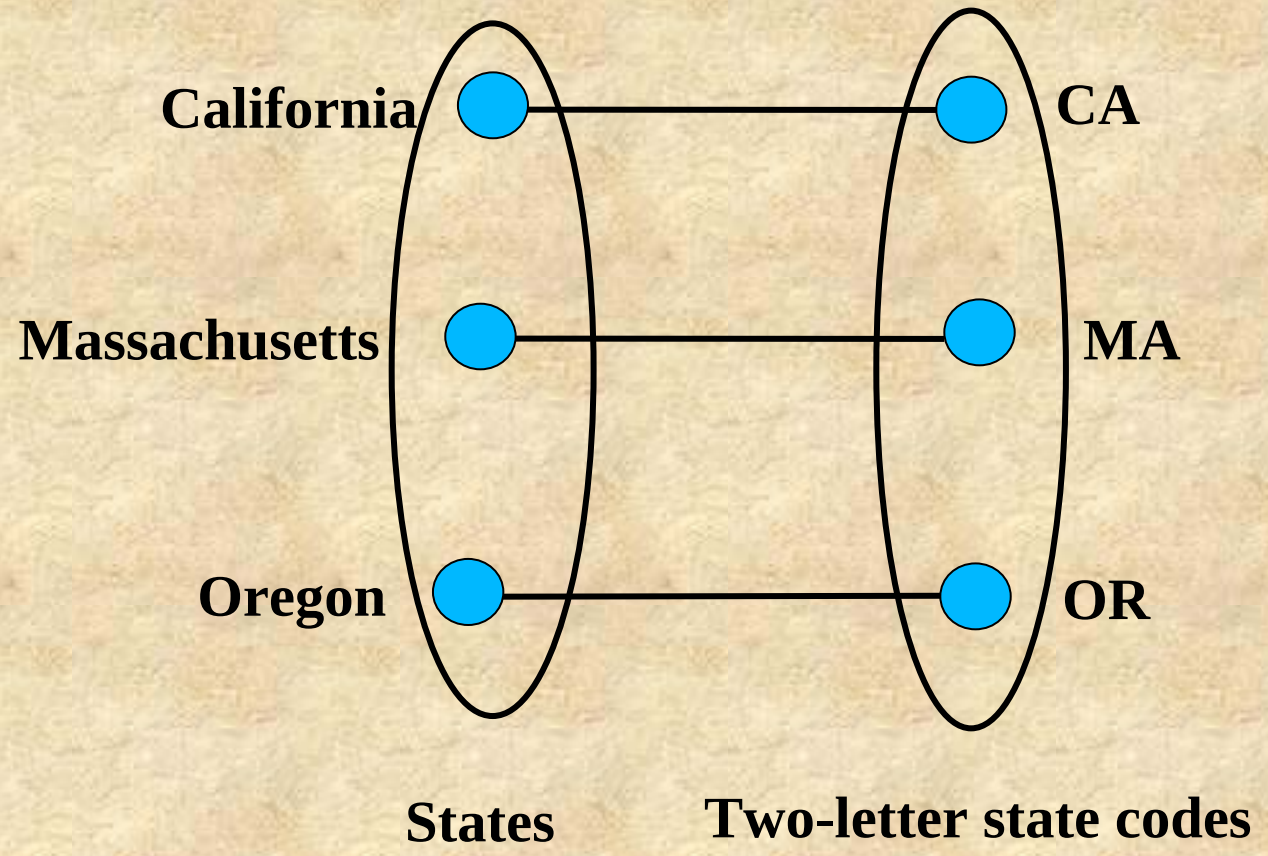
Lattices - Applications

- Formal Concept Analysis
  - synthesizing taxonomies
- Machine Learning
  - concept learning

Bipartite Graphs

- Vertices = two disjoint sets, V and W
- All edges connect one vertex from V and one vertex from W
- Examples:
  - mappings among value representations
  - mappings among schemas
  - (entity/attribute, relationship) nodes in Conceptual Graphs

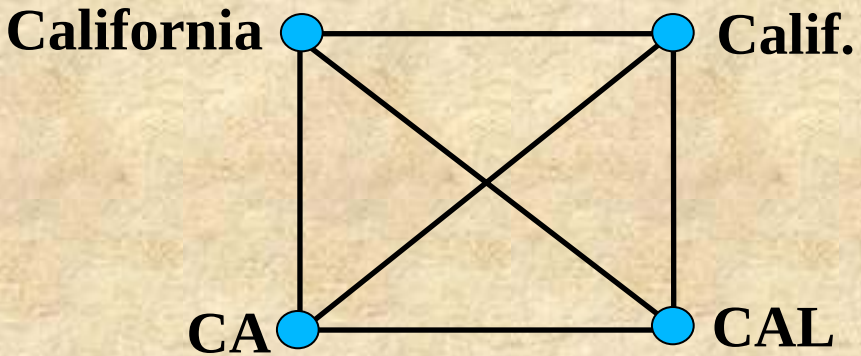
Example Bipartite Graph



Clique

- Clique = complete graph (or subgraph)
  - all possible edges are present
- Used to represent equivalence classes
- Typically, on undirected graphs

Example of Clique

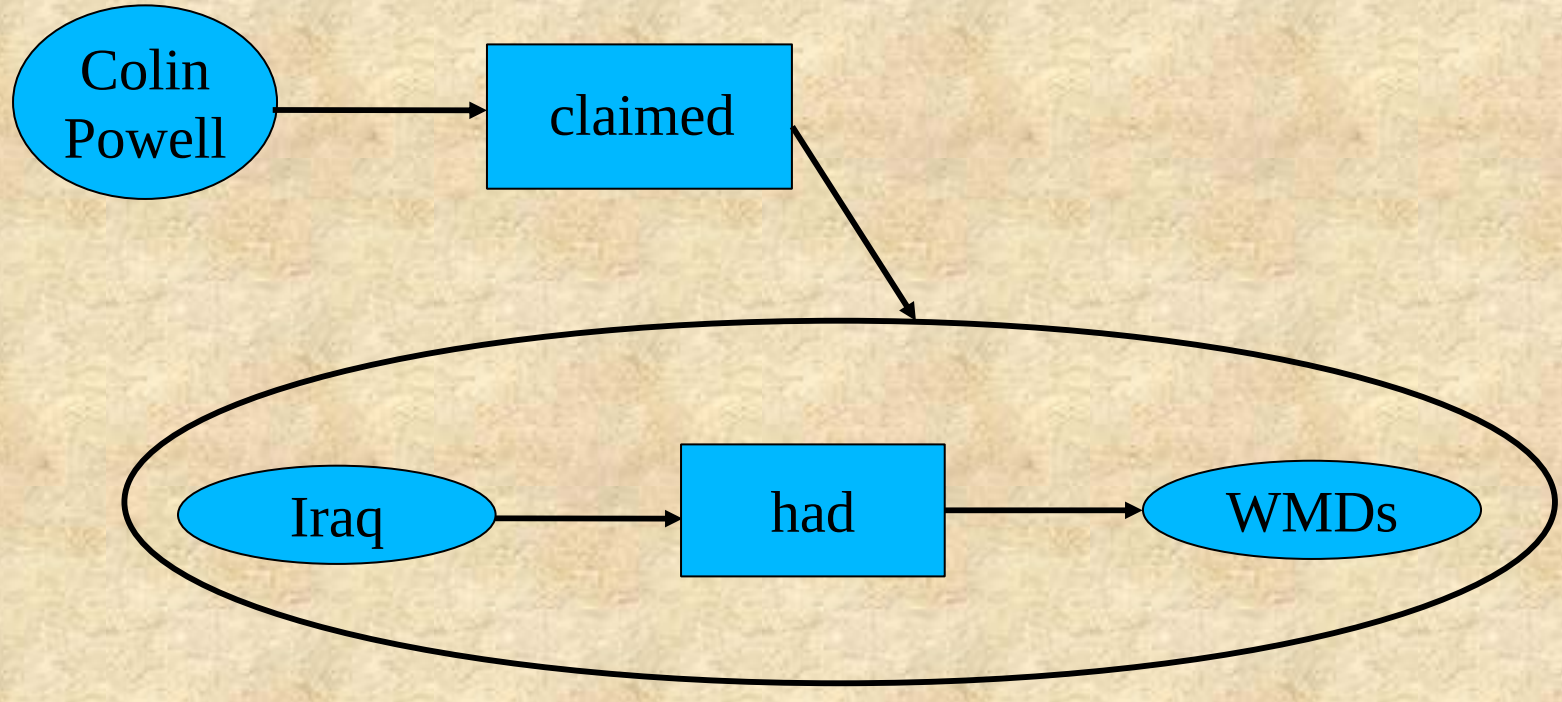


Here edges denote synonymy.

Compound Graphs

- Edges can point to/from subgraphs, not just simple nodes
- Used in conceptual graphs
- CG is isomorphic to First Order Logic
- Could be used to specify contexts for subgraphs

Example Compound Graph



Conclusions

- We can characterize metadata structure in terms of graph structures
- Partial Order Graphs are the most common structure:
  - used for taxonomies, partonomies
  - support multiple inheritance, faceted classification
  - implicit inclusion of inferred transitive closure edges

Challenges

- How to register & manage the various graph structures?
  - DBMS, File systems ….
- How to query the graph structures?
  - XQuery for XML
  - Poor to non-existent graph query languages
- How to get adequate performance, even in high performance computing environment
- User interface complexity
- How to manage semantic drift
- Versions
- How to interrelate graphs with other graphs and with data
- Granularity at which to register metadata (then point to greater detail elsewhere?)

Purposes of XMDR Prototype for ISO/IEC 11179 Registry Standard

- Extend semantics management capabilities
- Explore uses of terminologies and ontologies
- Systematize representation of relationships
- Adapt & test emerging semantic technologies
- Help resolve registration & harmonization issues for different metadata standards
- Propose revisions to 11179 Parts 2 & 3 (3rd Ed.)
- Show how proposed revisions to metadata registry standards can be implemented
- Demonstrate Reference Implementation (RI)

How can Terminologies and Ontologies help Manage Metadata?

- At the level of metadata instances in a registry, connect metadata entities via shared terms
  - via automatic indexing of metadata words
  - via text values from specific metadata elements
- At the level of the 11179 (or other) metamodel, ontologies can help specify formal relationships
  - is-a and part-of hierarchies, etc.
  - Inheritance, aggregation, …
  - for automatic searching of sub-classes & inverses
  - to specify semantic pathways for indexing

Project Background

- Collaborative, Interagency Effort
  - DOD, EPA, LBNL, USGS, NCI, Mayo Clinic…Others?
- Draws on and Contributes to Interagency Cooperation on Ecoinformatics
- Involves International, National, State, Local Government Agencies, other Organizations
- Recognizes Great Potential of Semantics-based Computing, Management of Metadata
  - Improving Collection, Maintenance, Dissemination, Processing of Very Diverse Data Structures
- Collaboration Arises from *Need to Share Diverse Data* Across Multiple Organizations
- Project Duration Expected to be July 04 – Jun 05

Many Players, Many Interests…Shared Context

Major Tasks, Deliverables & Milestones

Task/Deliverable
Develop Project Plan
Identify, Select Technologies
Identify, Select Metadata Sources
Initial Architecture Design
Research and Development
System Test & Evaluation (Internal Participants)
Test Implementation (External Users)
Present Proposed 11179 Part 2 Revisions to SC32 WG2 mtg in DC
Prepare Draft Revision of 11179 Part 2 for SC32 mtg in Berlin

Gantt Chart Forthcoming

General Tasks/Intentions

Task/Intention
Limited User Interface - Initially
Prioritized Content Registration
Limited Help Functions
Limited Query Optimization
Documents will Recommend Choices
Brief IC Metadata Working Group
Brief DOD Metadata Working Group
Brief ISO/IEC L8

Will Seek to Promote Awareness

Potential Standards/Technologies

- DBMS
  - Object, XML, Relational, RDF/Graph, Logic, Text, Document, Multimedia
- Knowledge Representation
  - Web Ontology Language (OWL)
  - Common Logic (CL)
- Middleware/Messaging
  - Cocoon 2, Jini, CoABS, JMS, XMLBlaster, SOAP
- XML [Semantic] Web Services
  - Axis, JWSDP
- Agent Development
  - ABLE, JADE
- Engines/Servers
  - OMS (IBM), Federator/OMS (OWI)
  - Jess

Open Source and Risk Tolerant

Architecture Approach

- Fully modular approach
  - Exemplars:
    - Apache Web Server
    - Eclipse IDE
    - Protégé Ontology Editor
  - Benefits:
    - numerous modules are relatively easy to implement
    - clean separation of concerns and high reusability and portability
    - tooling support required is minimal

XMDR Prototype Architecture: Initial Implemented Modules

External Interface

Registry
Java

RegistryStore

WritableRegistryStore
Subversion

AuthenticationService

MetadataValidator
Jena, Xerces

RetrievalIndex

MappingEngine

FullTextIndex
Lucene

LogicBasedIndex
Jena, OWI KS
Racer

11179 OWL Ontology

Ontology Editor
Protege

● Generalization ◆ Composition (tight ownership) ◇ Aggregation (loose ownership)

XMDR Content Priority List

Phase 1

(V.A) National Drug File Reference Terminology (?)

DTIC Thesaurus (Defense Technology Info. Center Thesaurus)

NCI Thesaurus National Cancer Institute Thesaurus

NCI Data Elements (National Cancer Institute Data Standards Registry

UMLS (non-proprietary portions)

GEMET (General Multilingual Environmental Thesaurus)

EDR Data Elements (Environmental Data Registry)

ISO 3166 Country Codes – from EPA EDR

USGS Geographic Names Information System (GNIS)

XMDR Content Priority List

Phase 2

LOINC Logical Observation Identifiers Names and Codes

ITIS Integrated Taxonomic Information System

Getty Thesaurus of Geographic Names (TGN)

SIC (Standard Industrial Classification System)

NAICS (North American Industrial Classification System)

NAIC-SIC mappings

UNSPSC (United Nations Standard Products and Services Codes)

EPA Chemical Substance Registry System

EPA Terminology Reference System

ISO Language Identifiers ISO 639-3 Part 3

IETF Language Identifiers RFC 1766

Units Ontology

XMDR Content Priority List

Phase 3

HL7 Terminology

HL7 Data Elements

GO (Gene Ontology)

NBII Biocomplexity Thesaurus

EPA Web Registry Controlled Vocabulary

BioPAX Ontology

NASA SWEET Ontologies

NDRTF

Acknowledgements and References

- Frank Olken, LBNL
- Kevin Keck, LBNL
- John McCarthy, LBNL