A Semantic Classification Model for e-Catalogs

Dongkyu Kim, Ph.D.
Director of R&D,
Corelogix, Inc.,
Seoul, Korea

Sang-goo Lee, Ph.D.
School of Computer
Science and Engineering
Seoul National University
Seoul, Korea

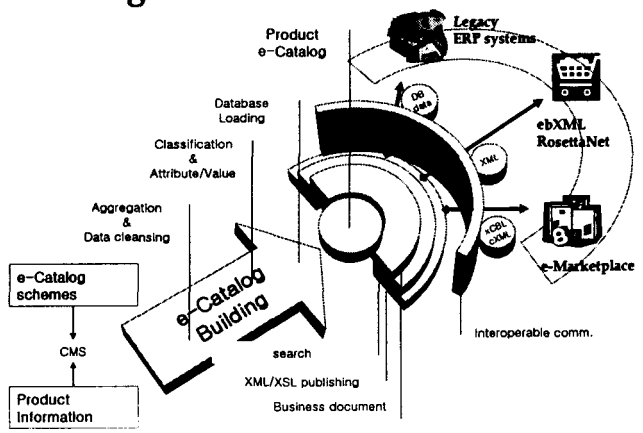
Junghoon Chun, Ph.D.
Department of Computer
Engineering
College of Engineering
Myongji University
Yongin, Kyunggi-Do, Korea

A table of contents

e-Catalog

- Catalogs
  - Information about products and services
  - Contents + Classification Schema + Presentation + Operational Issues
- What do we do with them?
  - [Schulten, et al, 2001]
    - Narrow down search for complete set of applicable products
    - Comprehend individual description to the precision needed
  - Support other applications that use product information
    - SCM, ERP, e-Procurement, etc.
- Catalog Management System
  - Design, storage, navigation & retrieval, transformation, communication, publication

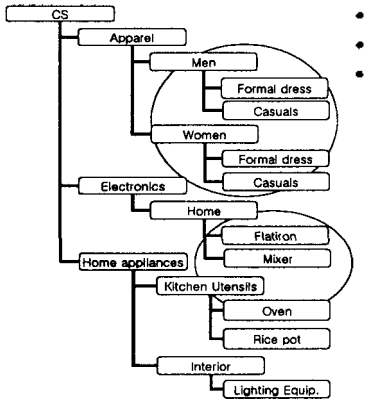
e-Catalog scenario



e-Catalog Issues

- e-Catalog Design
  - Contents schema
  - Classification schema
- e-Catalog Transformation
  - Information aggregation
  - Normalization
  - Classification of products
- e-Catalog Services
  - Authoring & Publishing
  - Search with classification & contents schema
- e-Catalog Globalization (Integration)
  - Classification schema mapping
  - Classification schema merging

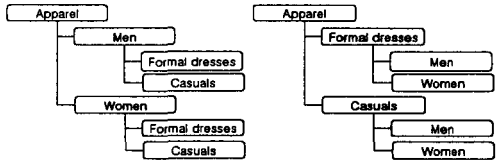
Issues: Classification Schema

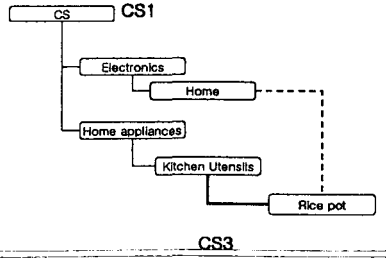


- Multiple classification
- Multiple Relationship
- Semantic Consistency

Issues: Search & Management

- View Change
- Schema Change
- Inference

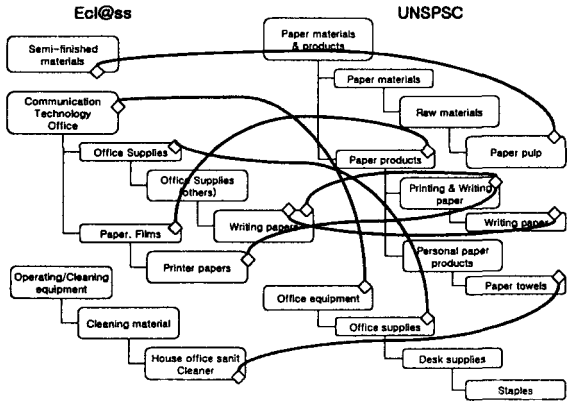




CS2

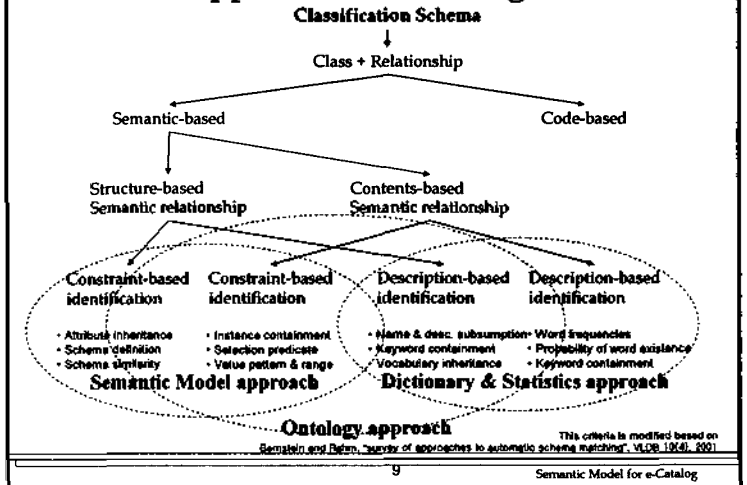
CS3

Issues: e-Catalog Integration



E. Schulten et al. "The E-Commerce Product Classification Challenge." *IEEE Intelligence Systems*, 16(4), 2001

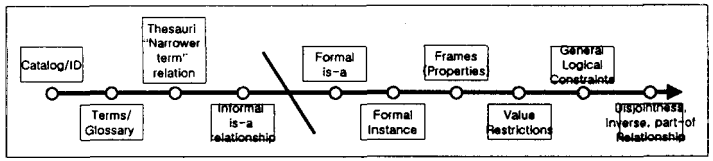
Semantic approach for e-catalog issues



Ontologies

- Critical role of Ontologies [Forrest Research]
  - Support of browsing and search for E-commerce
  - Support of interoperability for facilitation of knowledge management and configuration
- What is Ontologies…
  - A specification of a conceptualization [Gruber 1993]
  - Can be used to provide a concrete specification of term names and term meanings

An Ontology Spectrum



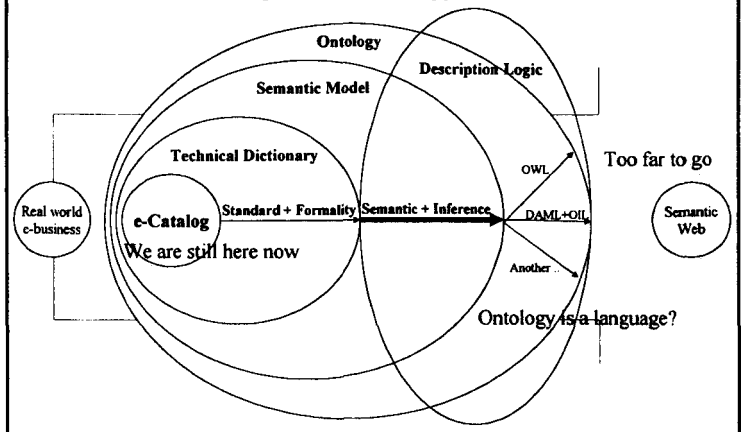
Deborah L. McGuinness,"Ontologies Come of Age",The Semantic Web: Why, What, and How,MIT Press,2001

- ***Catalog*** (controlled vocabulary-list of terms) Simplest notion of a possible ontology
- ***Glossary*** (a list of terms and meanings) Provide kind of semantics
- ***Thesaurus*** provide additional semantics in their relations between terms, but not explicit hierarchy

An Ontology Spectrum

- ***Informal Isa*** provide term hierarchy but not strict subclass hierarchy.
  Without true subclass(or true "isa") relationship, certain kind of deductive use of ontologies become problematic
- ***Strict subclass***
  Strict subclass hierarchies are necessary for exploitation of inheritance
- ***Formal Instance***
  Classification include not only class names but also ground individual content
- ***Frames***
  Classes include property information. (apparel-price, isMadeFrom)
- ***Value Restriction***
- ***Logical Statement***

From e-Catalog to Ontology

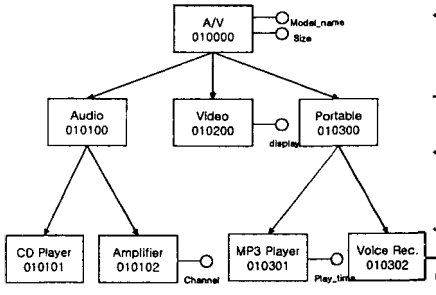


Technical Dictionaries

	eOTD	PLIB	RNTD	GDD
Formal Name	ECCMA Open Technical Dictionary	ISO 13584 Part Library	RosettaNet Technical Dictionary	Global Data Dictionary
Organization	ECCMA	ISO	RosettaNet	EAN / UCC
Sponsoring Country	US	France, Germany, UK, US, Japan, Swiss	Europe, Japan, Korea etc	Europe, US
Scope	Identification/ Description of Organization, Location, Goods, Service	Ontology Development except axiom rule	Common properties for defining products for RosettaNet PIP	Names, Definitions of all attributes used in Business include message and catalog
Expression	RDB Table Schema	EXPRESS	XML Excel File	RDB Table Schema
Ontology	△	△	X	X

Semantic Classification Model

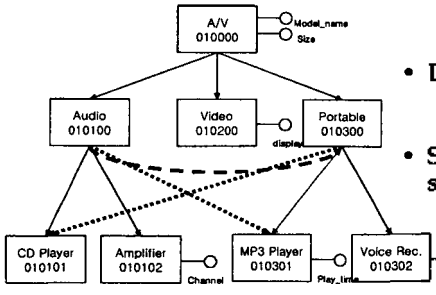
- Motivating Example



- Problem
  - Inference
    - Audio → MP3 player ?
    - Portable → CD player?
  - View change
    - Portable ←→ Audio?
  - Schema change
    - MP3 & Voice record Combo?
  - Semantic consistency
    - Is this classification scheme correct?

Semantic Classification Model

- Purpose



- Inference of Relationship
  - Inference
    - Audio → MP3 player ?
    - Portable → CD player?
- Decision for change
  - View change
    - Portable ←→ Audio?
- Schema change with semantic consistency
  - Schema change
    - MP3 & Voice record Combo?
  - Semantic consistency
    - Is this classification scheme correct?

Semantic Classification Model

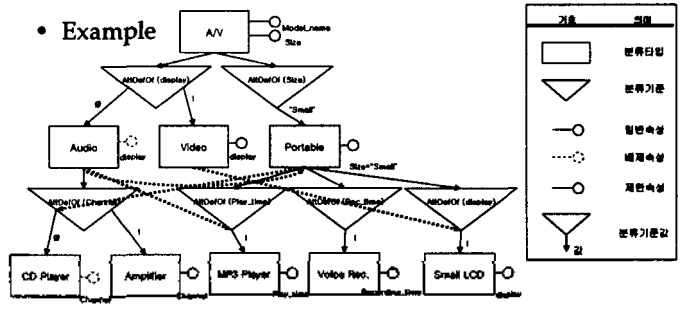
- Definition
  - Semantic Classification Schema of CS, $G_{CS} = \langle S, C, E \rangle$
    - S is a set of all the products
    - $C = \{C_0, C_1, C_2, \ldots, C_n\}$, a set of all the classes
    - A Class, $C_i \in C$
      $C_i = \langle D(C_i), R(C_i), M_i \rangle$
      - $D(C_i)$ = <class_code, name, description>
      - $R(C_i)$ is a set of common attributes for class C_i
      - **M_i is a Membership Predicate for Class C_i**
    - A Relationship, $E_j \in E$
      $E_j = \langle C_w, C_u, CP_{w,u} \rangle$ where $C_w, C_u \in C$
      - C_w is the parent class of C_u
      - **$CP_{w,u}$ is a Classification Predicate from C_w to C_u**

Semantic Classification Model

- Topological vs Semantic Relationships
  - **Topological Relationship, $\geq_T$**
    - $C_i \geq_T C_j$ iff C_i is an ancestor of C_j in G_{CL}
  - **Semantic Relationship, $\geq_S$**
    - $C_i \geq_S C_j$ iff M_j implies M_i
- A classification scheme is
  - *semantically sound* if each topological relationship implies semantic relationship
  - *semantically complete* if each semantic relationship implies topological relationship

Semantic Classification Model

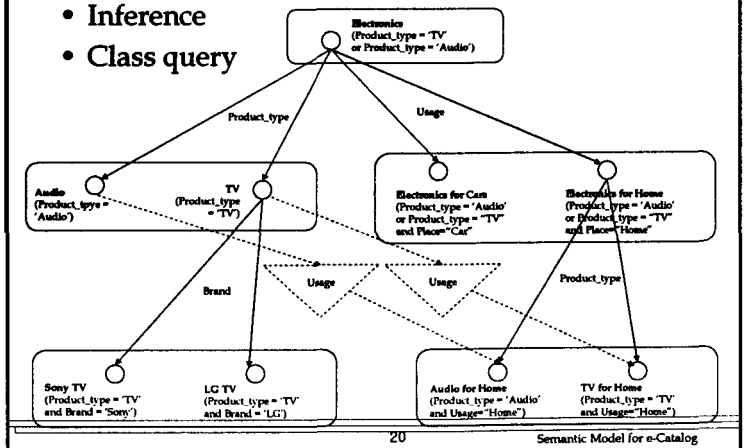
- Example



$M(C_{\text{오디오 플레이어}}) = \{Model\_name<1>, Size<1>, display<*>, Channel<*>, Play\_time<*>, Recording\_time<*>\}$
$M(C_{\text{오디오}}) = \{Model\_name<1>, Size<1>, display<∞>, Channel<*>\}$
$M(C_{\text{비디오}}) = \{Model\_name<1>, Size<1>, display<1>, Channel<*>\}$
$M(C_{\text{휴대용 기기}}) = \{Model\_name<1>, Size<'Small'>, Play\_time<*>, Recording\_time<*>\}$
$M(C_{\text{CD 플레이어}}) = \{Model\_name<1>, Size, Channel<∞>\}$
$M(C_{\text{앰프}}) = \{Model\_name<1>, Size<1>, Channel<1>\}$
$M(C_{\text{MP3 플레이어}}) = \{Model\_name<1>, Size<'Small'>, Play\_time<1>\}$
$M(C_{\text{음성 녹음기}}) = \{Model\_name<1>, Size<'Small'>, Recording\_time<1>\}$
$M(C_{\text{소형 LCD}}) = \{Model\_name<1>, Size<'Small'>, display<1>\}$

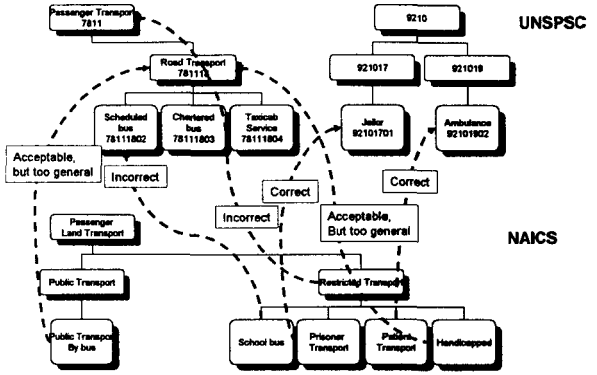
Semantic Classification Model

- Inference
- Class query



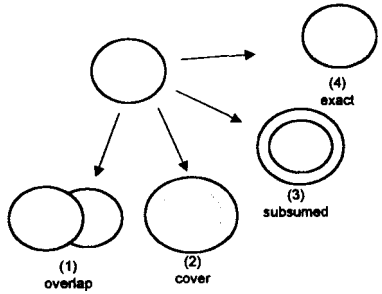
Classification Schema Mapping

- Crosswalk Table Mapping



Classification Schema Mapping

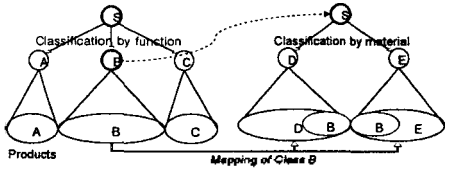
- Cases of Mapping Semantics

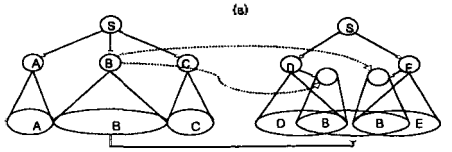


- We want to achieve cases (3) & (4)
- System should be aware of the type of each mapping entry

Semantic Class Mapping

- Semantic Class Mapping





Mapping of Class B

(Class D and $(M(B) \cap M(D))$) U (Class E and $(M(B) \cap M(E))$)

(b)

Classification schema Integration

- Merge two different classification schemas G_1 and G_2 into one integrated classification schema G
  - cross industry standards, mergers of companies with different but overlapping domains
- preserve soundness

  For all $C_i \geq_T C_j \in G$, it always holds $C_i \geq_S C_j$
- preserve all classes and relationships in the original schemas

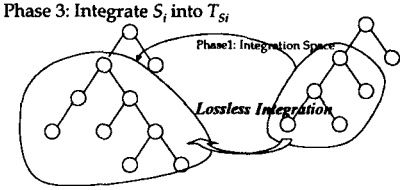
  For all $C_i \in G_1$, $C_i \in G$, and for all $C_m \in G_2$, $C_m \in G$,

  and

  if $C_i \geq_T C_j$ in G_1, then $C_i \geq_T C_j$ in G, and likewise for G_2

Semantic Classification Schema Integration

- Algorithm
  - $G_S = <S_S, C_S, E_S>$, $G_T = <S_T, C_T, E_T>$
  - Phase 1: Make a decision for integration space
    - Find out the set of minimal subgraph pairs of $<G_S, G_T>$
    - Source root class S_0 should subsume target root class T_0
  - For each $S_i \in C_S$,
    - Phase 2: Find the most suitable class for integration, $T_{Si} \in C_T$
    - Phase 3: Integrate S_i into T_{Si}

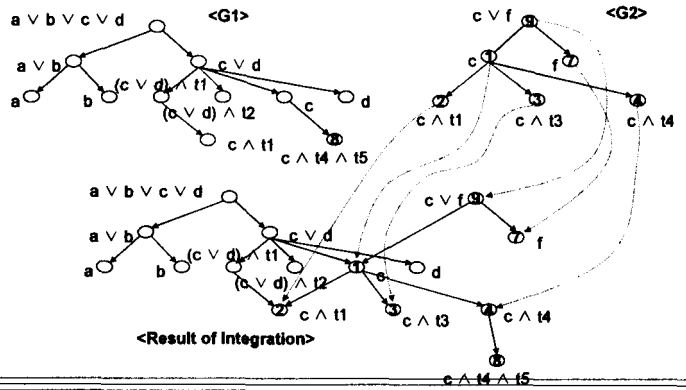


Phase2:Find the most suitable class for integration
Phase3:Integrate two classes

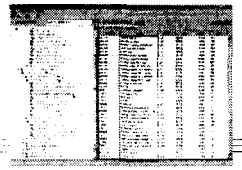
Target Classification schema, G_T Source Classification schema, G_S

Semantic Classification Schema Integration

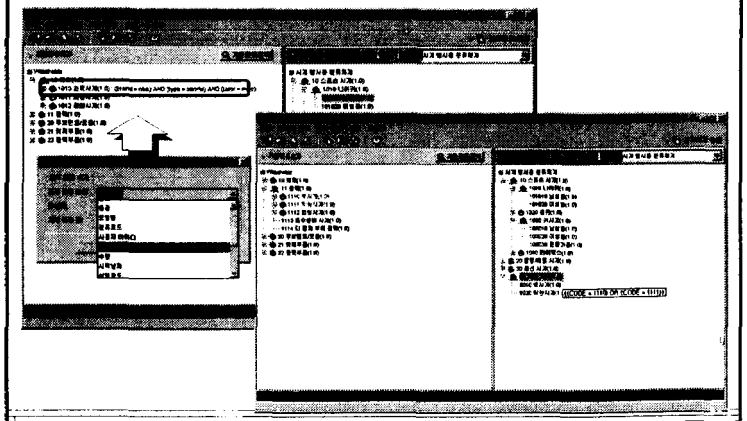
- Example



Applications



Applications



Comparison

	Code-based	*Structure, Description-based*	*Contents, Description-based*	*Hybrid, Constraint-based*
Example	UNSPSC	TD-based systems	Statistical classifiers	Semantic CS systems
Selection of instances	Code (Selection by class code)	Code (no additional view definition for each class)	Code (no additional view definition for each class)	View definition by Selection predicate
Multiple classification view	No	No	Yes, but need long computation	Yes (by identifying semantic relationships)
Inference of semantic relationships	No	Yes, but too general results Low precision	Yes, but no guarantee of accuracy and consistency	Yes (by identifying subsumption relationships)
Classification schema evolution	No	Yes, but not support class move	Yes, but its result may cause wrong modifications	Yes (by maintaining explicit & semantic relationships)
Guarantee of semantic consistency	No	Yes	No, it depends on distribution of its instances	Yes (guarantee of semantic soundness)
Classification schema mapping	Manual mapping No guarantee of correct mapping	Semi-automatic, No support for overlapped classes	Yes, but it is just statistic decision	Yes (by comparing class definition & addition predicates)
Classification schema integration	Manual integration No guarantee of lossless integration	Semi-automatic, Not deterministic	No, No guarantee of sound integration	Yes (recursive approach of class mapping)

Conclusion