Universität Leipzig

Institut für Informatik

Detecting Semantic Correspondences in Product Catalogs

WDI-Lab: Working Group 2 - Schema and Ontology Matching

Sabine Maßmann

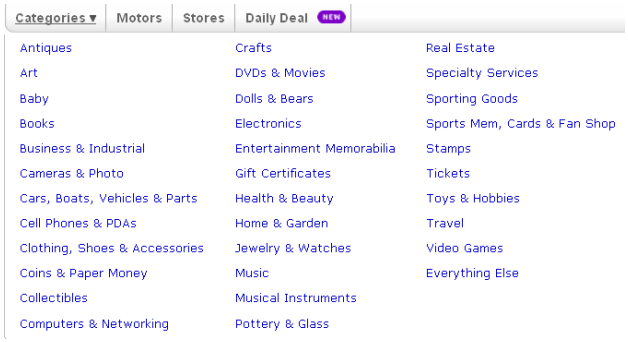
6th May 2010

Agenda

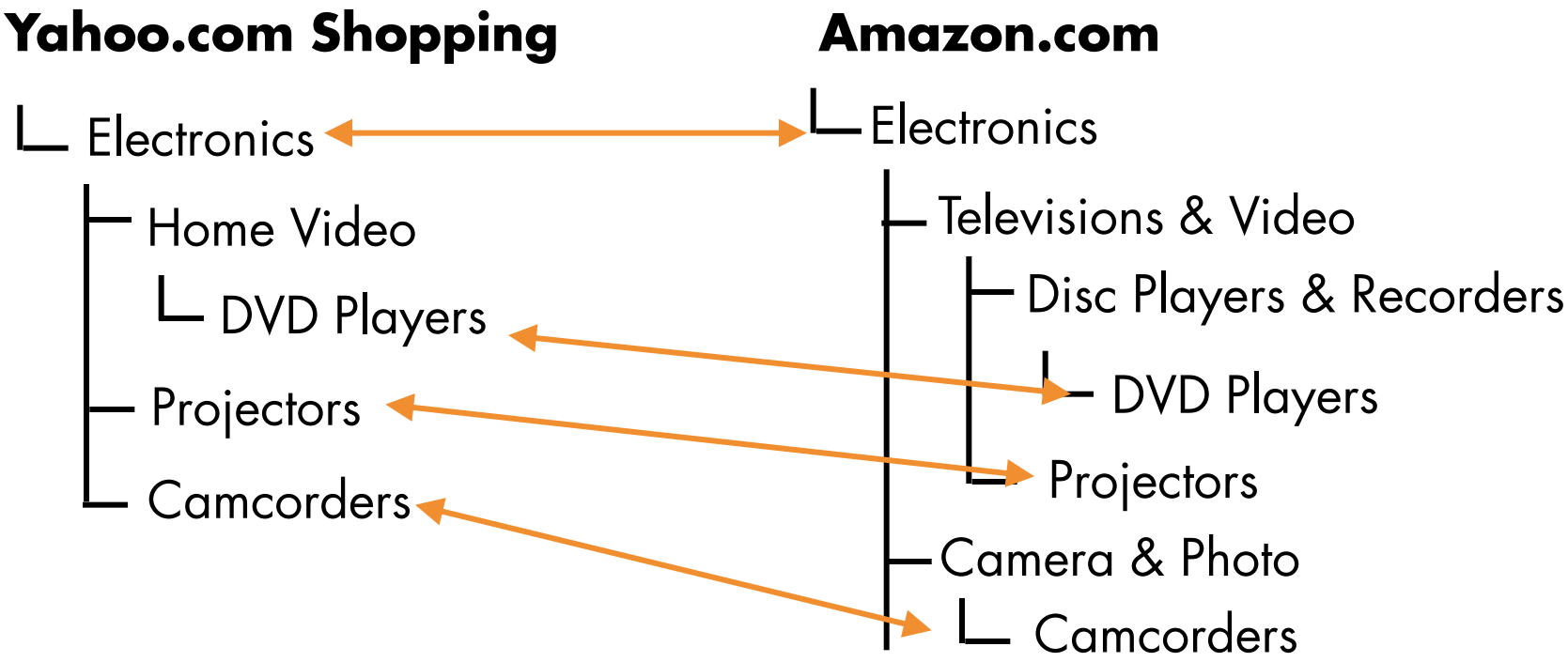
- Motivation & Background
- Challenges
- Our Approach: COMA++
- Applications

Motivation

- Product catalog: a systematic sorted collection of product information or service information
- Classification: regarding an identical attribute, e.g. manufacturer, use, feature, price, color
- Size: often thousands of categories, ten thousands products
- Examples



Example

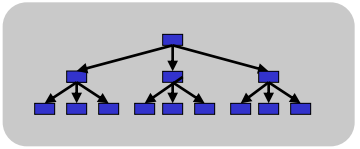

Yahoo.com Shopping
Amazon.com
Electronics
Home Video
DVD Players
Projectors
Camcorders
Electronics
Televisions & Video
Disc Players & Recorders
DVD Players
Projectors
Camera & Photo
Camcorders

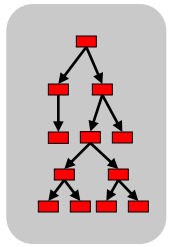
Background

- Ontology:
  - A formal representation of knowledge
  - Set of concepts within a domain and the relationship between them
- Ontology Matching
  - Process of identifying semantic correspondences between ontologies
  - Input: two ontologies O1 and O2
  - In addition maybe instances + auxiliary information
  - Output: Mapping (=set of correspondences) between O1 and O2

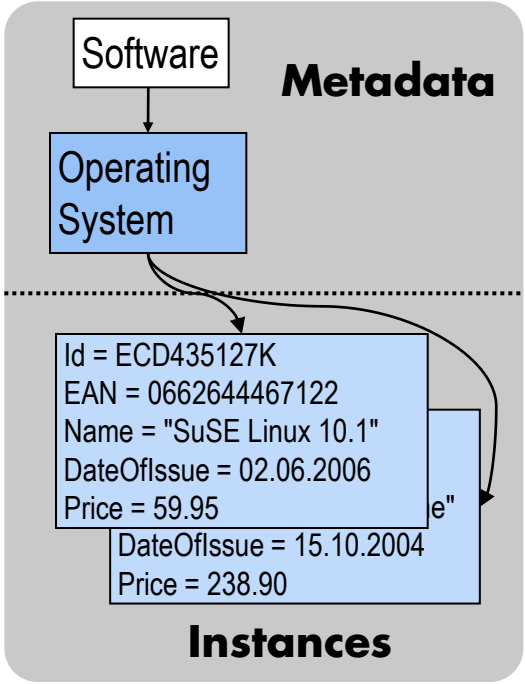
- Focus for this talk:
  content categorizations e.g. web directories and **product catalog**

Challenges





- Heterogeneity (general problem): terminological and conceptual
- Metadata:
  - Words used several times , e.g. “accessories”,…
- Instances:
  - Assigned to several categories (redundancy)
  - Some categories have only a few instances, other thousands



- Consequences
  - Matching difficult
    → one single algorithm is not enough
  - Correspondences not only 1:1 but n:m
  - Relationship between categories not just equal but overlap

Our Approach: COMA++

- Generic match system
- Supports matching of schemas and ontologies
- Different matchers using
  - Metadata + Auxiliary information
  - Structure
  - Instances
  - Reuse
- Combining match algorithms
- Strategies for matching large schemas and ontologies

Do, H.H., E. Rahm: *COMA - A System for Flexible Combination of Schema Matching Approaches*. VLDB 2002

Aumüller D., H.-H. Do, S. Massmann, E. Rahm: *Schema and Ontology Matching with COMA++*. Sigmod 2005

COMA++

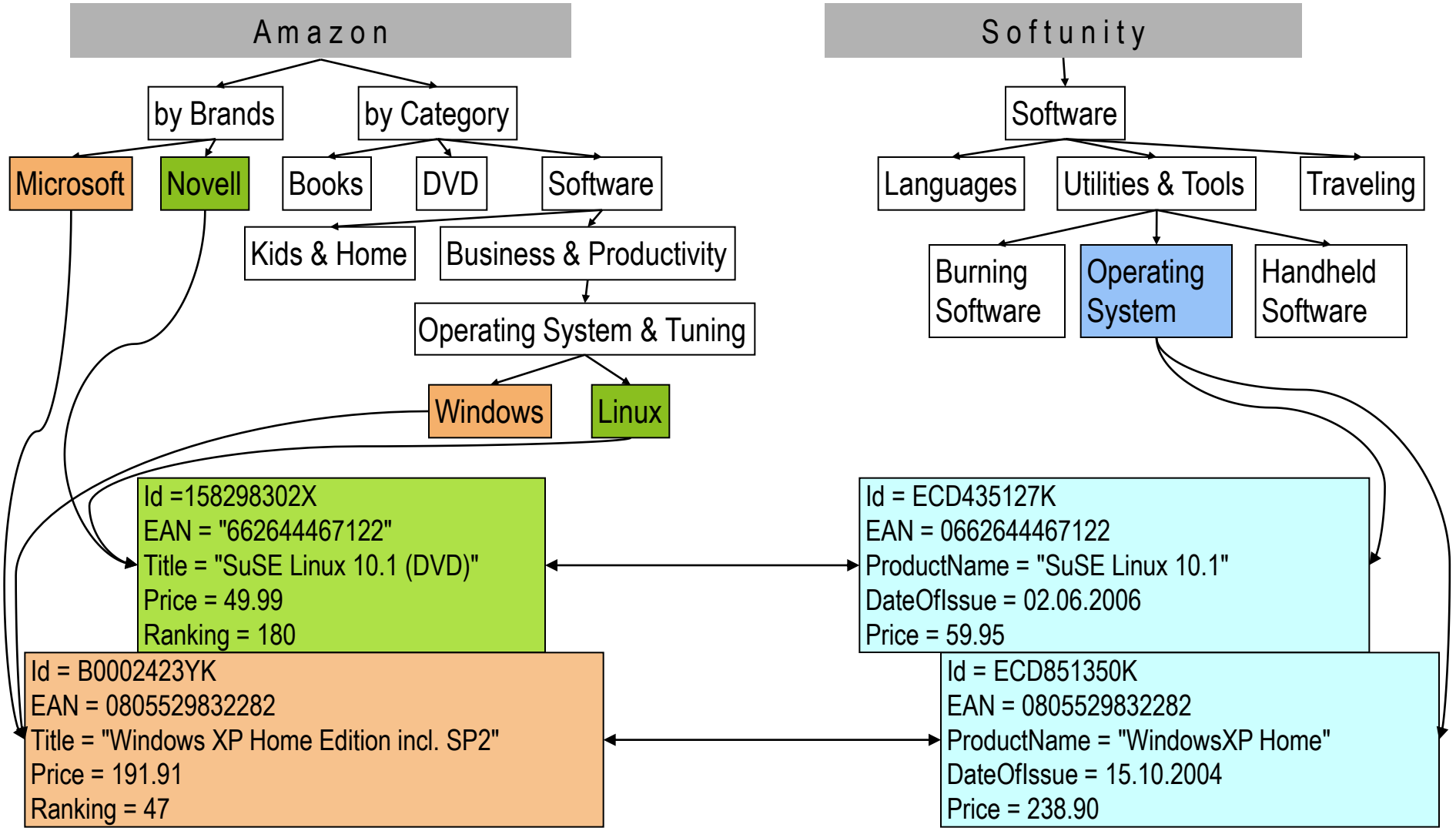
- Extended evaluation with e.g. PurchaseOrder, OAEI- Benchmark and web directories
- Hundred of research facilities and companies testet COMA++
- Used as reference system:
  - *"COMA++ is a generic, composite matcher with very effective match results." [Duchateau et al., OTM 2008]*
  - *"COMA++ is one of the best available schema matchers that enjoys from combining several available methods for schema matching" [Nezhad et al., WWW 2007]*
  - *"The best recall and the best F-measure were achieved by COMA++." [Kappel et al., BTW workshop 2007]*

Meta Data

- Category names → concept itself, path → for context
  - String similarity functions, e.g. trigram, edit distance, soundex
- Descriptions
  - Weighted document similarity (TFIDF)
- Id ?
  - Equal → ONLY if same source and thus same ids

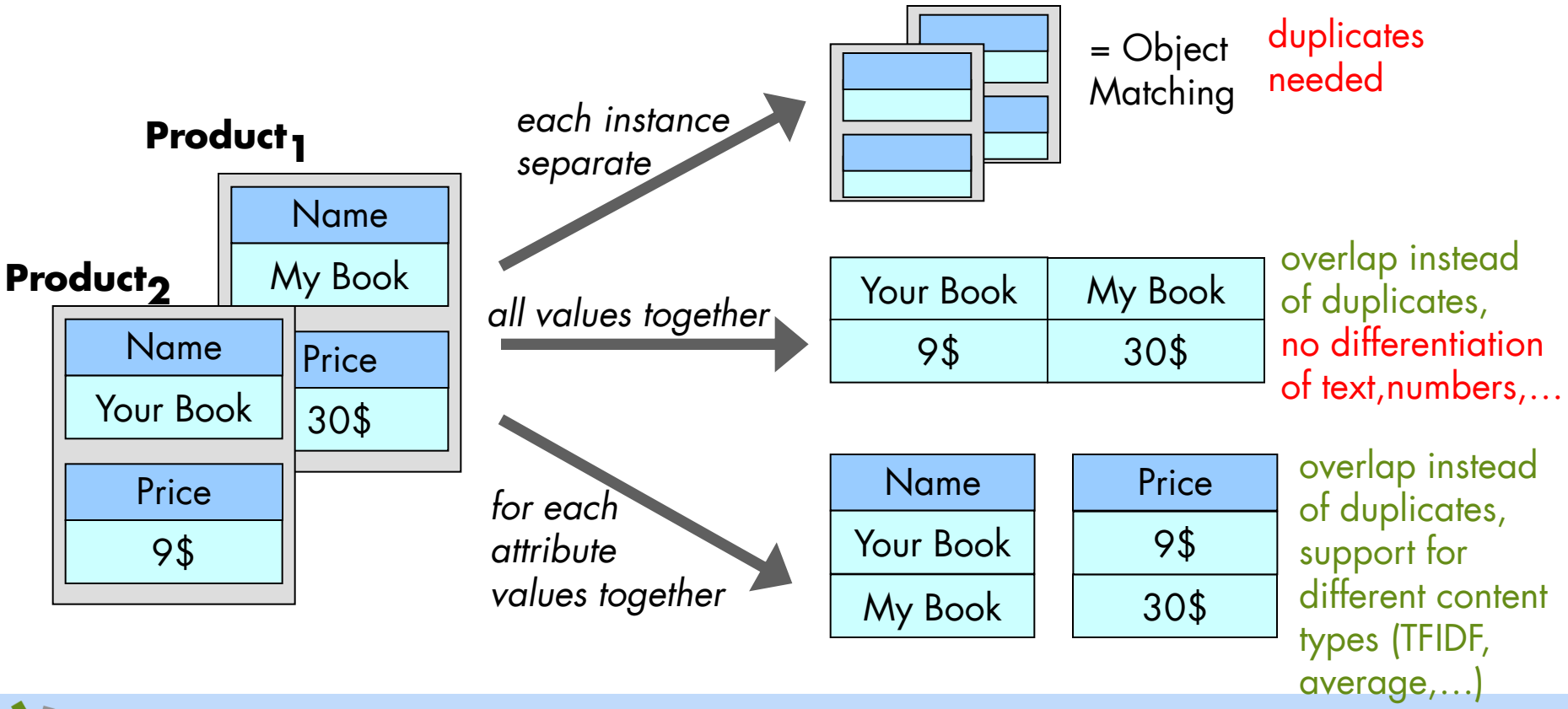
- Usage of auxiliary information
  - Synonyms, e.g. „beamer“ and „projector“
  - Abbreviations, e.g. „HP“ stands for „Hewlett-Packard“

Example

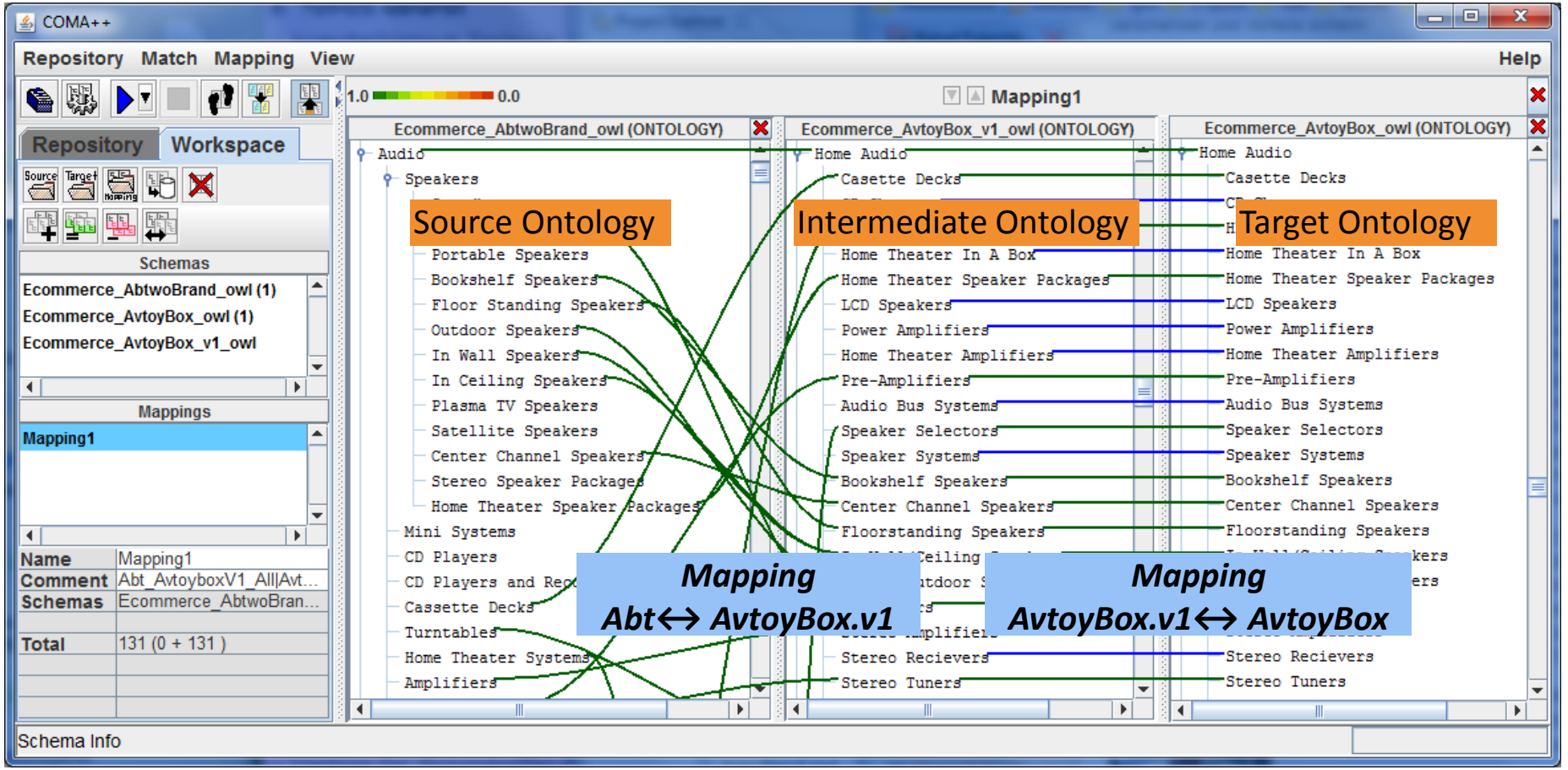

Amazon
by Brands
by Category
Microsoft
Novell
Books
DVD
Software
Kids & Home
Business & Productivity
Operating System & Tuning
Windows
Linux
Softunity
Software
Languages
Utilities & Tools
Traveling
Burning Software
Operating System
Handheld Software
Id =158298302X
EAN = "662644467122"
Title = "SuSE Linux 10.1 (DVD)"
Price = 49.99
Ranking = 180
Id = ECD435127K
EAN = 0662644467122
ProductName = "SuSE Linux 10.1"
DateOfIssue = 02.06.2006
Price = 59.95
Id = B0002423YK
EAN = 0805529832282
Title = "Windows XP Home Edition incl. SP2"
Price = 191.91
Ranking = 47
Id = ECD851350K
EAN = 0805529832282
ProductName = "WindowsXP Home"
DateOfIssue = 15.10.2004
Price = 238.90

Instance-based Matching

- Idea:
  - Instances describe content of category better than just the name
  - Overlapping instance values indicate similar categorie

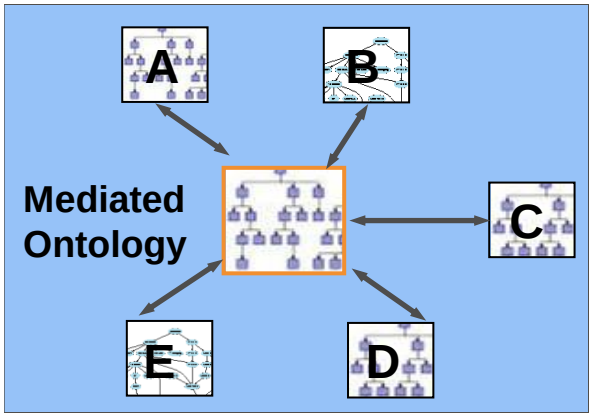


Reuse of Mappings



Mediated Ontology

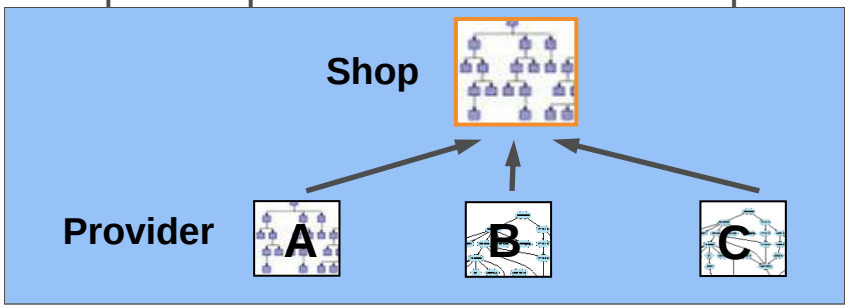
- Mediated Product Catalog
  - Contains every category that appears in one or more catalogs
  - Crucial: coverage, granularity



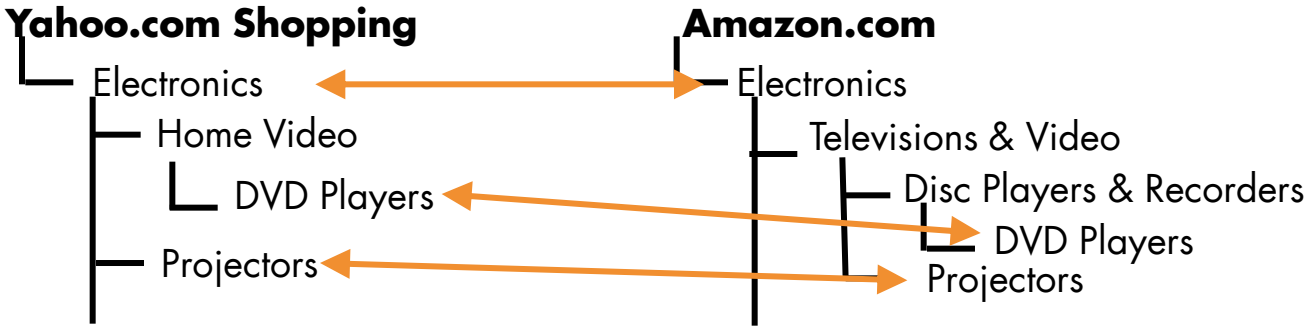
- Adding new product catalog:
  - One Match Task: new product catalog to mediated catalog
  - Calculating other mappings via reuse/compose

Applications

- Creating an integrated (master) product catalog
  - an ecommerce shop sells products from different providers



- Find additional/faster products for a query
  - one query but products from many product catalogs



- Automatic adaptations due to ontology evolution