Obtaining Precision when Integrating Information

July 2001

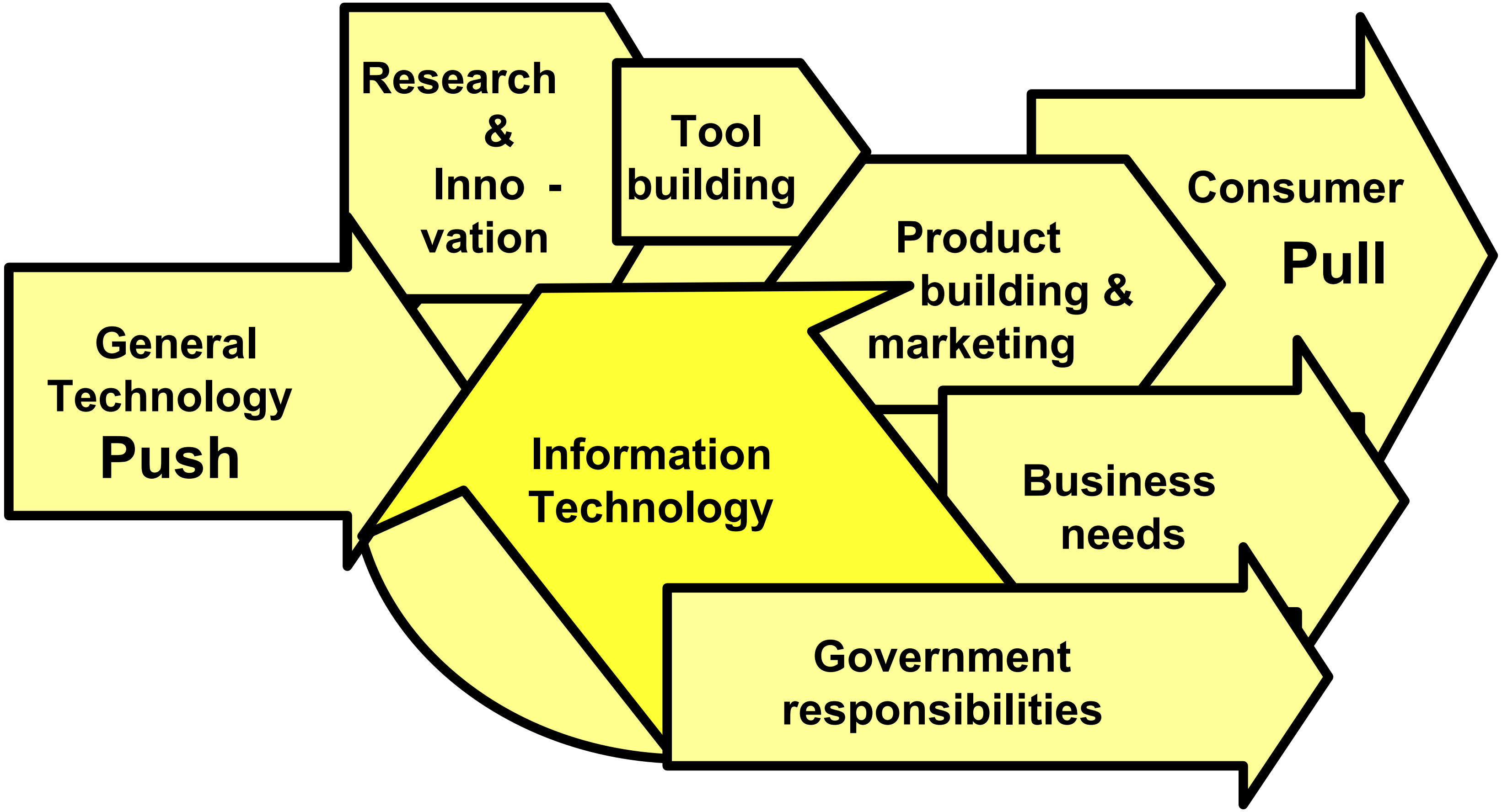
Gio Wiederhold

Stanford University

Outline

- **Setting** VG 3 - VG 6
- **Precision** VG 7 - VG 9
- **Lack of precision** VG 10 - VG 12
- **SKC solution** VG 13, VG 21- VG 27
- **Ontologies** VG 14 - VG 20
- **Early results** VG 28
- **Interoperation** VG 29 - VG 30
- **Tool & examples** VG 31 - VG 39
- **Composition and excution** VG 40 - VG 41
- **Summary – SKC to general** VG 42 - VG 45

Interactions

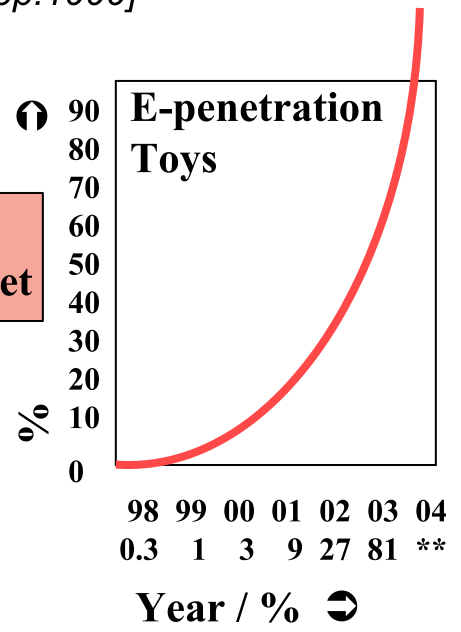


Trends 1998 : 1999

- Users of the Internet 40% ➲ 52% of U.S. population
- Growth of Net Sites (now 2.2M public sites with 288M pages)
- Expected growth in E-commerce by Internet users *[BW, 6 Sep.1999]*

segment	*1998*		*1999*
– books	7.2%	➲	16.0%
– music & video	6.3%	➲	16.4%
– toys	3.1%	➲	10.3%
– travel	2.6%	➲	4.0%
– tickets	1.4%	➲	4.2%
– Overall	8.0%	➲	33.0% = $9.5Billion

Centroid, in 1999 ~1% of total market



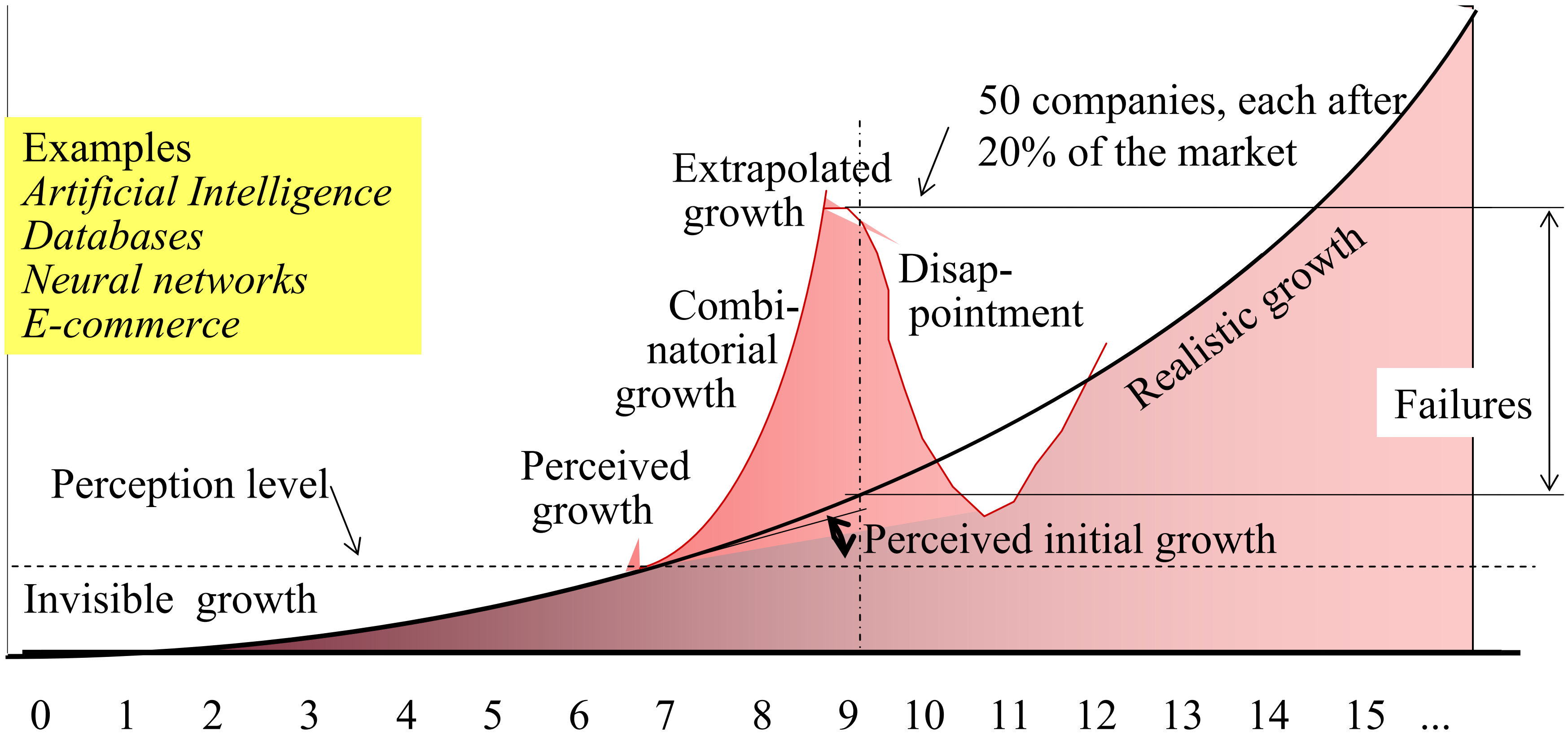
An unstainable trend cannot be sustained [Herbert Stein]

➲ *new services*

Growth and Perception

E-commerce

- **Gartner: 2000 prediction for 2004: 7.3 T$**
- **Revision:2001 prediction for 2004: 5.9 T$ *drastic loss?***



Our* Information Environment

*B2B, B2C, G2G, G2C, . . .

- **In the past: Scarcity**

  Customers needed more information to make better decisions

- **Today: Excess**

  The web provides more information than customers can digest

Effect: confusion in decision-making

Must I look at all possibly relevant information?

What is the penalty for missing something ?

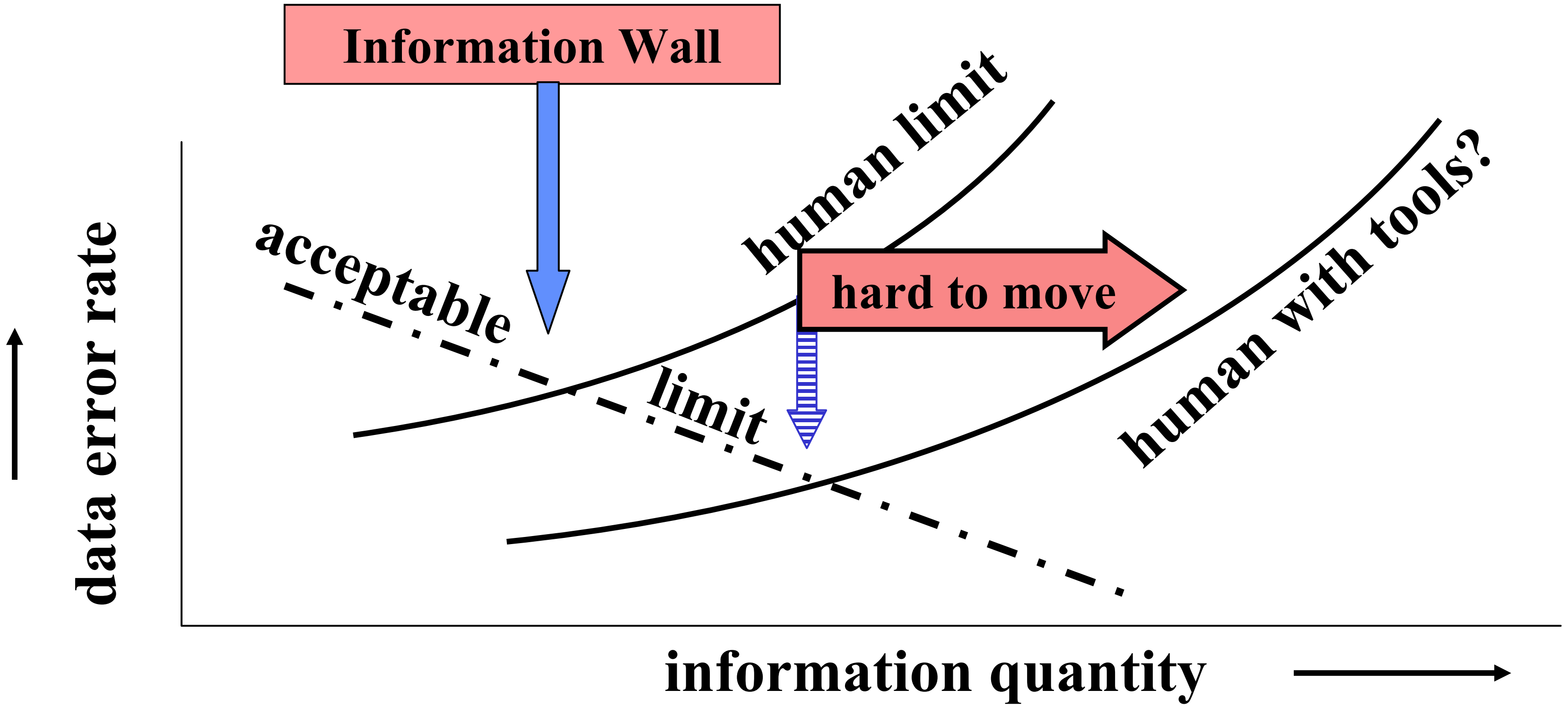
What is the cost of looking at everything ?

I am confused, best defer making any decision . . . .

Need for precision

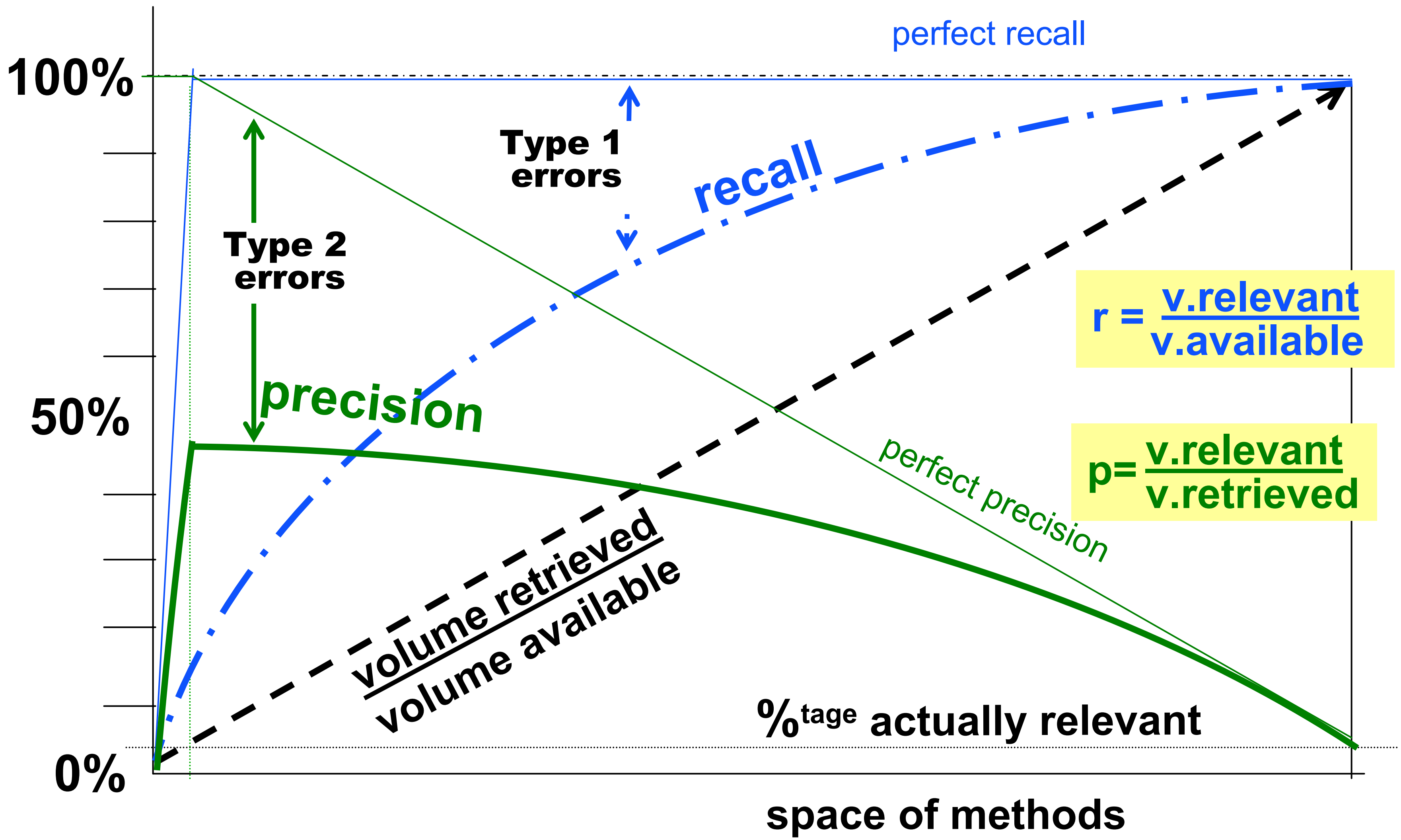
Precision: Few wrong or irrelevant results

More precision is needed as data volume increases
--- a small error rate still leads to too many errors



adapted from Warren Powell, Princeton Un.

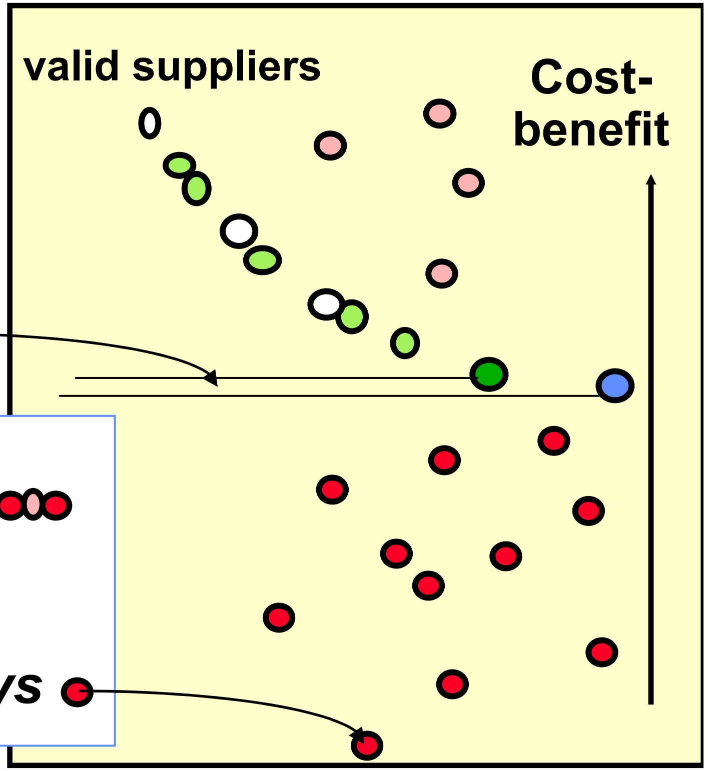
Relationships among parameters



Cost of Error types differs

1 **Missed Valid Information**
(False Negatives)
causes lost opportunities
cheapest shovel, . . .
suboptimal decision-making by *x*

2 **Excess Information**
(False Positives)
has to be investigated
attractive-looking supplier - makes toys



Space of results, ordered

Having many cases of excess information
costs more than some missing information

A Major Cause of Errors

Searches extend over many domains

- Domains have their own terminologies
  - Need autonomy to deal with knowledge growth
- The usage of terms in a domain is efficient
  - Appropriate granularity
    - Mechanic working on a truck vs. logistics manager
  - Shorthand notations
    - PSU vs. PSU
- Functions differ in scope
  - Payroll versus Personnel
    - getting paid vs. available (includes contract staff)

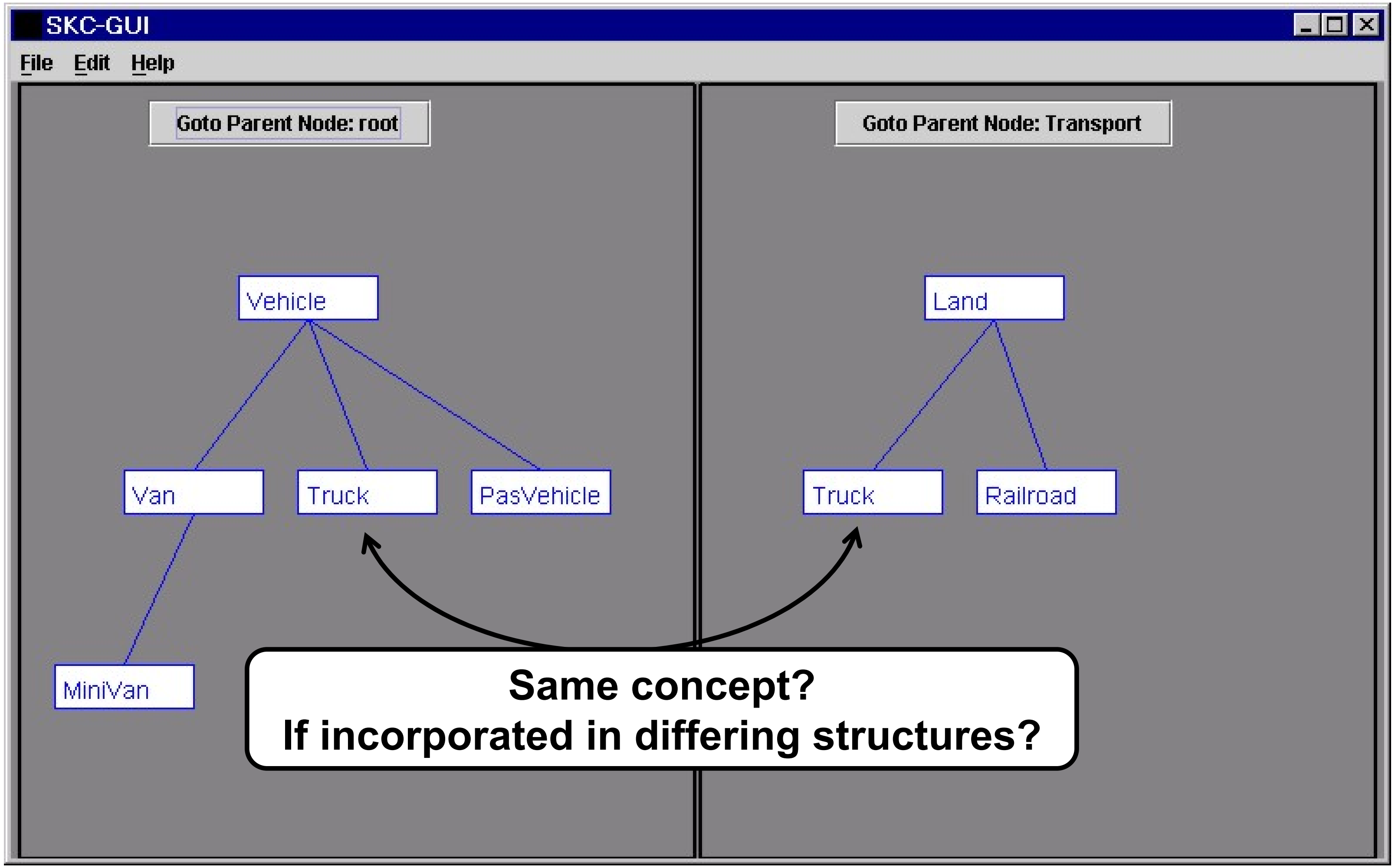
Semantic Mismatches

Information comes from many autonomous sources

- **Differing viewpoints** *(by source)*
  - **differing terms for similar items** *{ lorry, truck }*
  - **same terms for dissimilar items** *trunk(luggage, car)*
  - **differing coverage** *vehicles (DMV, AIA)*
  - **differing granularity** *trucks (shipper, manuf.)*
  - **different scope** *student (museum fee, Stanford)*
- **Hinders use of information from disjoint sources**
  - **missed linkages** *loss of information, opportunities*
  - **irrelevant linkages** *overload on user or application program*
- **Poor precision when merged**

Still ok for web browsing , poor for business & science

Structural Heterogeneity



Approach (SKC project)

Scalable Knowledge Composition – Stanford Univ. DB group

1. Define Terminology in a domain precisely
   - ➢ Schemas, XML DTDs ➔ Ontologies
2. Develop methods to permit interoperation among differing domains (not integration)
   - ➢ Articulation
   - ➢ Ontology Algebra
3. Develop tools to support the methods
   - ➢ Ontology matching

What are Ontologies?

Ontologies list the terms and their *relationships* that allow communication among partners in enterprises ***(in machine-readable form)***

Relationships determine meaning - **parent, school, company**

Databases use ontologies during design in their E-R diagrams ***(Implicitly)*** **and represent the leaf nodes in their schemas**

Knowledge-bases use ontologies ***(often implicitly)*** **add class definition** ***(to hold instances)*****, constraints, and, sometimes, operations among the terms**

Functions of Ontologies

- **Enable Precision in Understanding**

  People = designers, implementors, users, maintainers

  Systems = sources, mediators, applications

- **Share the Cost of Knowledge Acquisition & Maintenance**

  reuse encoded knowledge,
  remain up-to-date as domains change

- **Enable Information Interoperation** *

  Define the terms that link domains

Ancestors of Ontologies

- ↓ **Lexicons: collect terms used in information systems**
- ↓ **Taxonomies: categorize, abstract, classify terms**
- ↓ **Schemas of databases: attributes, ranges, constraints**
- ↓ **Data dictionaries: systems with multiple files, owners**
- ↓ **Object libraries: grouped attributes, inherit., methods**
- ↓ **Symbol tables: terms bound to implemented programs**
- ↓ **Domain object models: (XML DTD): interchange terms**

↓ **. . .** ***More Knowledge formalized***

Two Mismatch Solutions

1. A Single, Globally consistent Ontology *(Your Hope)*
   - wonderful for users and their programs
   - too many interacting sources
   - long time to achieve, *2 sources (UAL, LH), 3 (+ trucks), 4, … all ?*
   - costly maintenance, since all sources evolve
   - no world-wide authority to dictate conformance
2. Domain-specific ontologies (*XML DTD assumption*)
   - Small, focused, cooperating groups
   - high quality, some examples - *arthritis, Shakespeare plays*
   - allows sharable, formal tools
   - ongoing, local maintenance affecting users - *annual updates*
   - poor interoperation, users still face inter-domain mismatches

Global consistency: *Hope, but . .*

Common assumptions in assembling and integrating distributed information resources

- The language used by the resources is the same
- Sub languages used by the resources are subsets of a globally consistent language

These assumptions are provably false

Working towards the goal of globally consistency is

1. naïve -- the goal cannot be achieved
2. inefficient -- languages are efficient in local contexts
3. unmaintainable – terminology evolves with progress

Domain-specific Expertise

Knowledge needed is huge

- Partition into natural domains
- Determine domain responsibility and authority
- Empower domain owners
- Provide tools

Consider interaction

Domains and Consistency

- a domain will contain many objects
- the object configuration is consistent
- within a domain all *terms* are consistent &
- *relationships* among objects are consistent

Domain Ontology

No committee is needed to forge compromises * within a domain

- context is implicit

* Compromises hide valuable details

SKC *grounded* definition

- **Ontology:**
  a set of *terms* and their *relationships*
- **Term:**
  a reference to real-world and abstract objects
- **Relationship:**
  a named and typed set of links between objects
- **Reference:**
  a label that names objects
- **Abstract object:**
  a concept which refers to other objects
- **Real-world object:**
  an entity instance with a physical manifestation

(or its representation in a factual database)

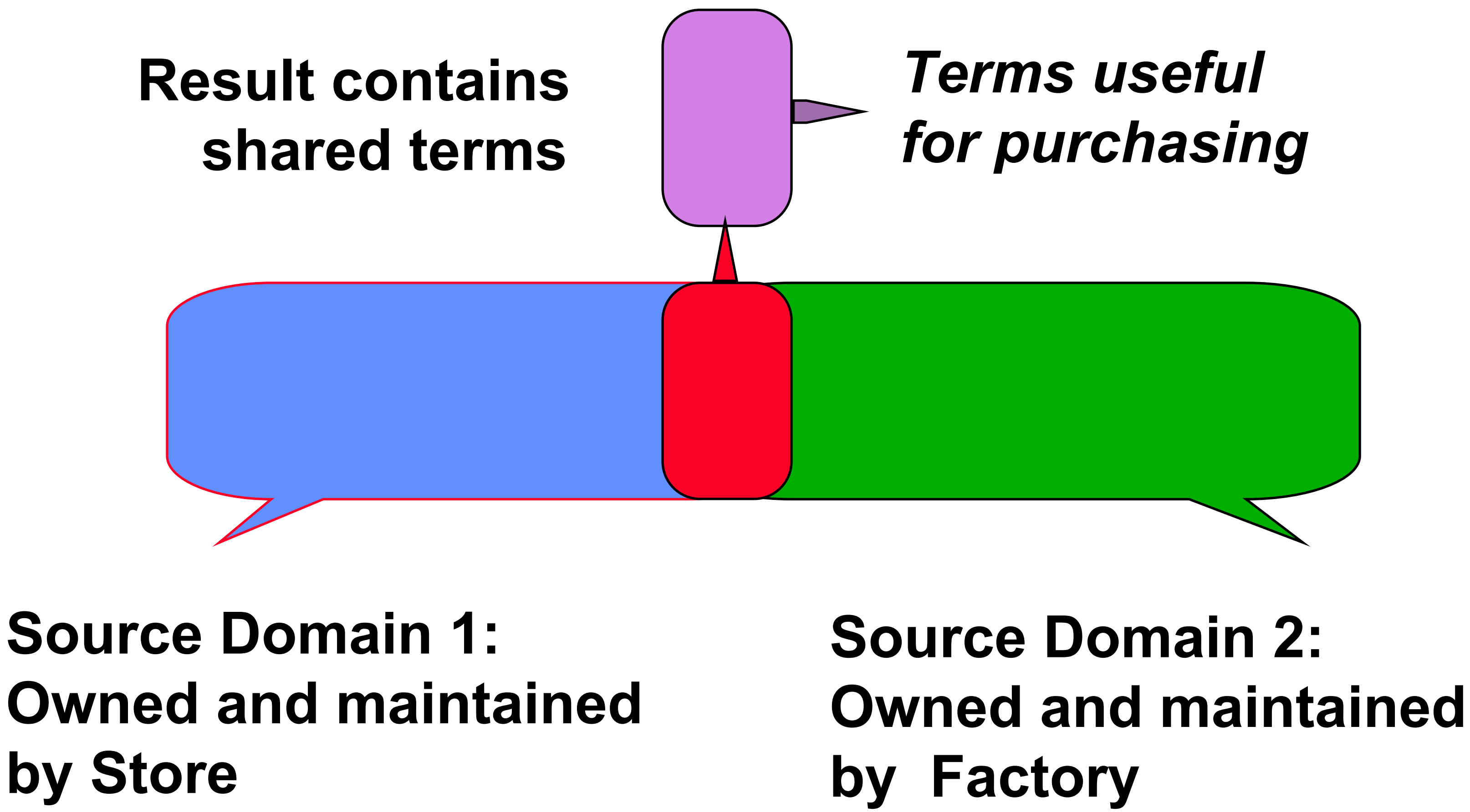
An Ontology Algebra

A knowledge-based algebra for ontologies

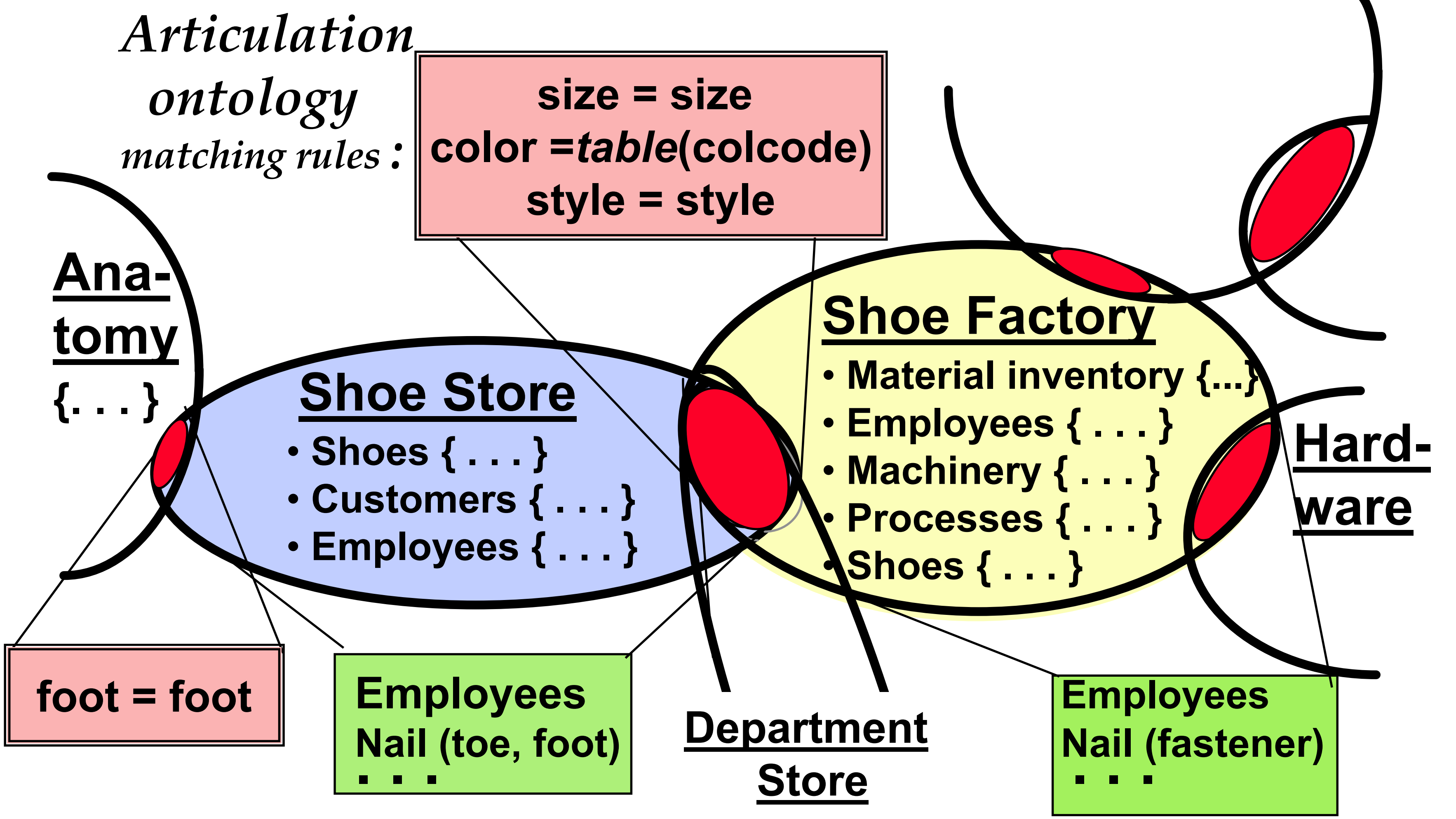
Intersection	$\cap$	create a subset ontology → keep sharable entries
Union	$\cup$	create a joint ontology → merge entries
Difference	$-$	create a distinct ontology → remove shared entries

The Articulation Ontology (AO) consists of matching rules that link domain ontologies

Sample Operation: INTERSECTION



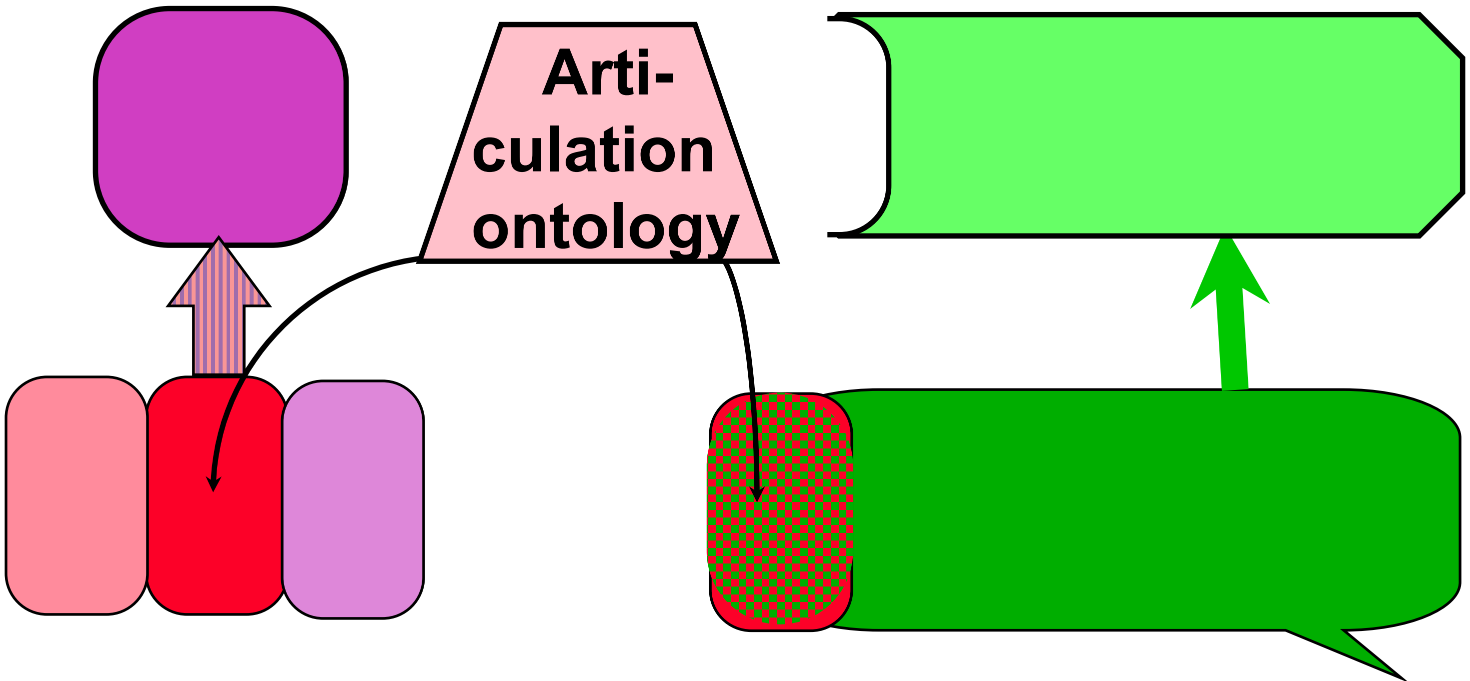
Sample Intersections



Other Basic Operations

UNION: ***merging entire ontologies***

DIFFERENCE: ***material fully under local control***



typically prior intersections

Features of an algebra

Operations can be composed

Operations can be rearranged

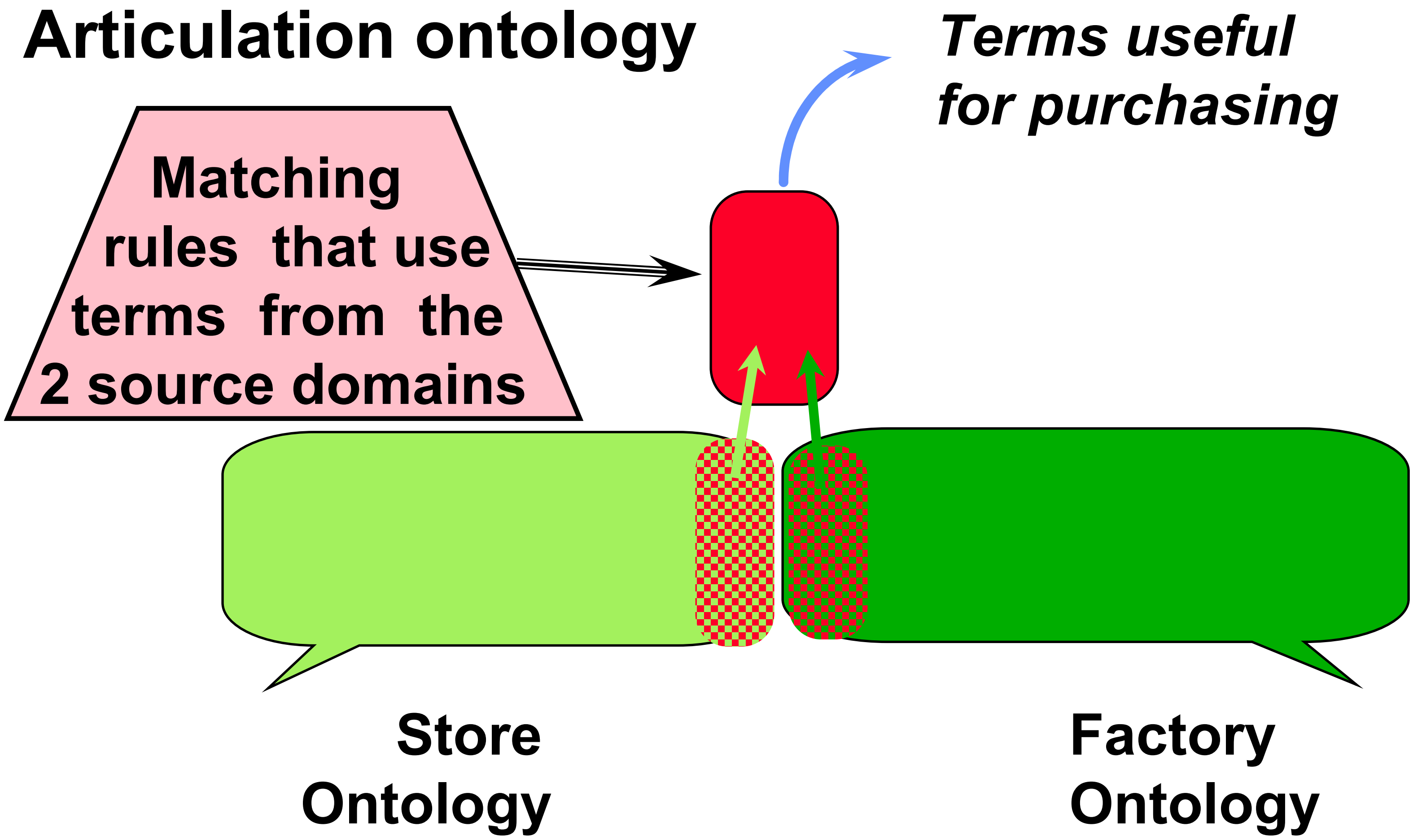
Alternate arrangements can be evaluated

Optimization is enabled

The record of past operations can be kept and reused

(experience: 3 months → 1 week for Webster's annual update,
→ 2 weeks for OED (6 x size **[Jannink:01]** ***)***

INTERSECTION support



Sample Processing in HPKB

What is the most recent year an OPEC member nation was on the UN security council (SC)?

(An DARPA HPKB Challenge Problem)

- SKC resolves 3 Sources
  - » CIA Factbook '96 (nations)
  - » OPEC (members, dates)
  - » UN (SC members, years)
- SKC obtains the Correct Answer
  - » 1996 (Indonesia)
- Other groups obtained more, but factually wrong answers; they relied on one global source, the CIA factbook.

- Problems resolved by SKC
  - * Factbook – *a secondary source* -- has out of date OPEC & UN SC lists
    - • Indonesia not listed
    - • Gabon (left OPEC 1994)
  - * different country names
    - • Gambia => The Gambia
  - * historical country names
    - • Yugoslavia
  - » UN lists future security council members
    - • Gabon 1999

  needed ancillary data

Interoperation via Articulation

At application definition time

- Match relevant ontologies where needed
- Establish articulation rules among them.
- Record the process

At execution time

- Perform query rewriting to get to sources
- Optimize based on the ontology algebra.

For maintenance

- Regenerate rules using the stored formulation

Generation of the rules

Provide library of automatic match heuristics

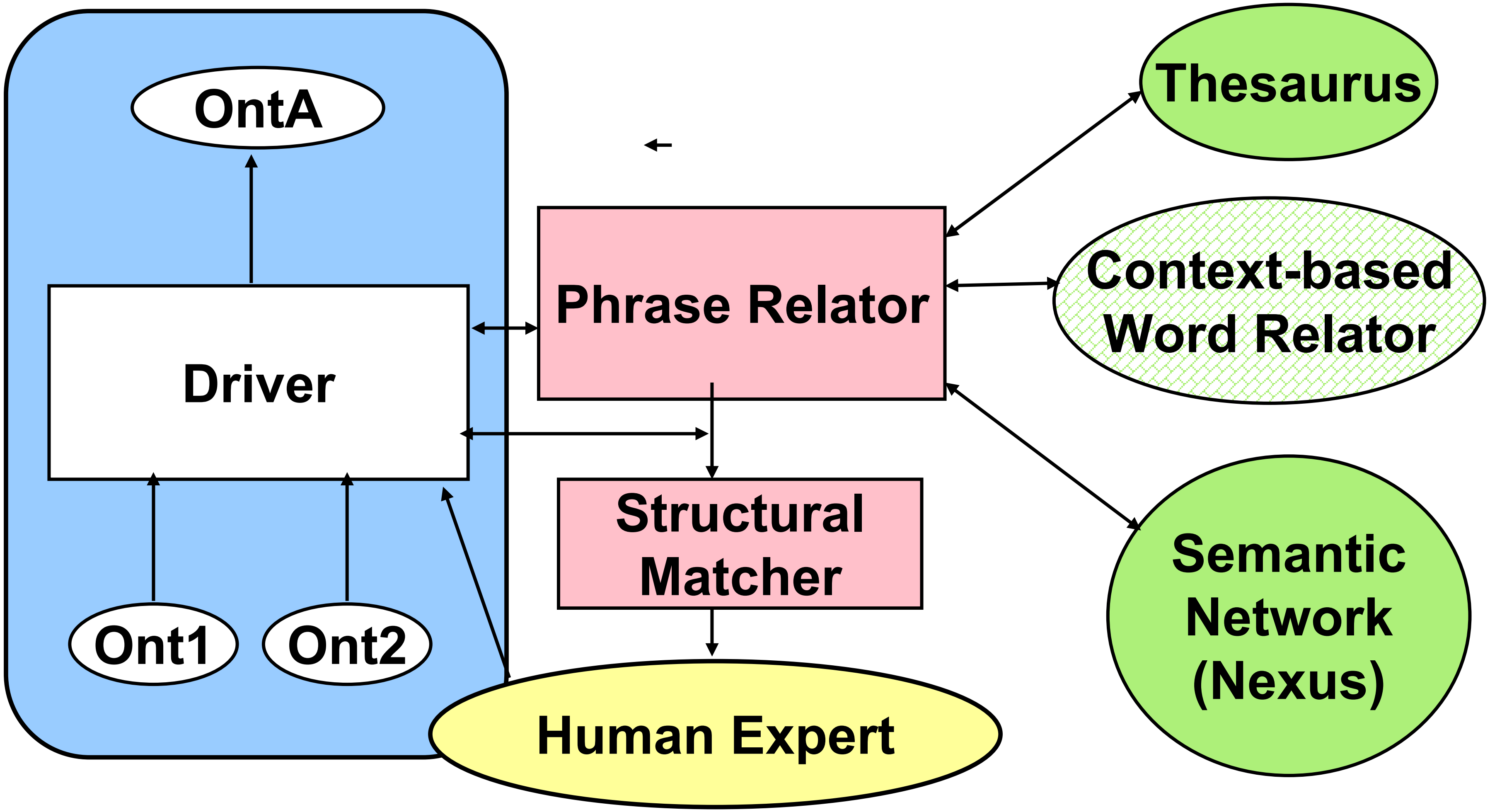
- Lexical Methods -- spelling
- Structural Methods -- relative graph position
- Reasoning-based Methods
- Nexus →
- Hybrid Methods
  - Iteratively, with an expert in control

GUI tool to

- display matches and
- verify generated matches using the human expert
- expert can also supply matching rules

Articulation Generator

Being built by Prasenjit Mitra



Lexical Methods

- Preprocessing rules.

  - Expert-generated seed rules.

    e.g., (Match O1.President O2.PrimeMinister)

  - Context-based preprocessing directives.

- Thesaurus - synonyms, generalizations

  yellow $\subset$ ochre, canary

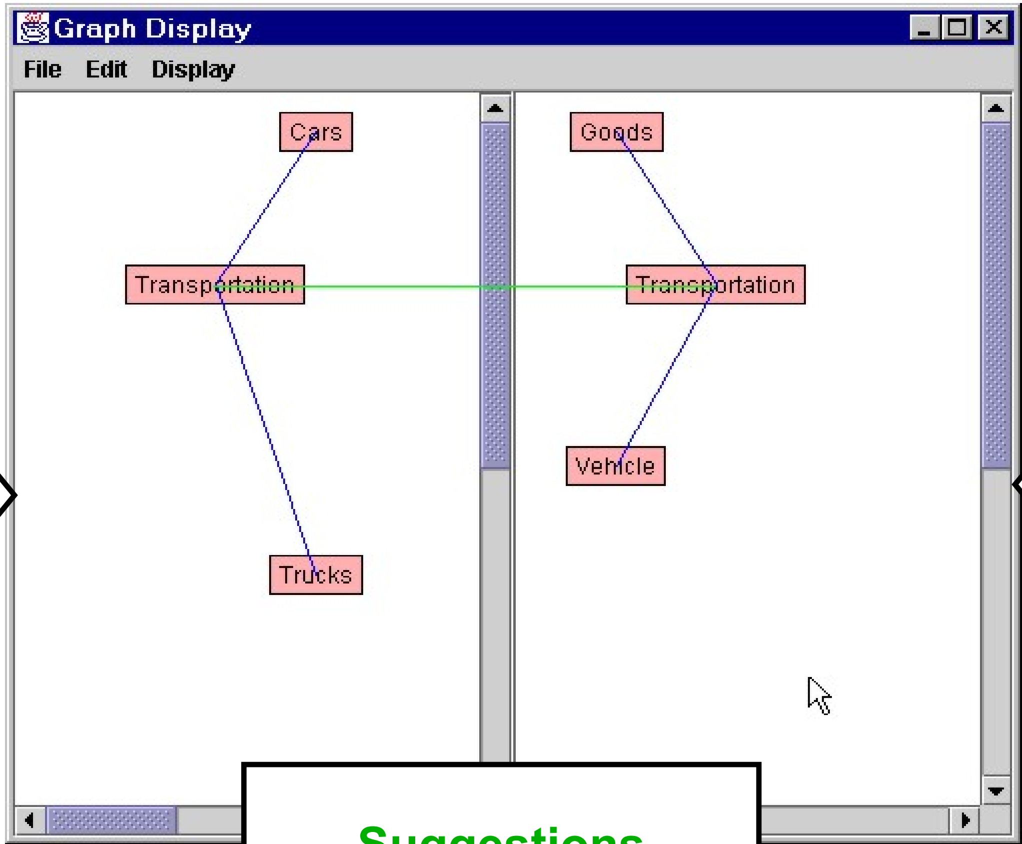
- Nexus – term relationship graph

  Owner = buyer

  – (Distance of words as measure of relatedness)

Tools to create articulations

Graph matcher for Articulation-creating Expert



Vehicle ontology

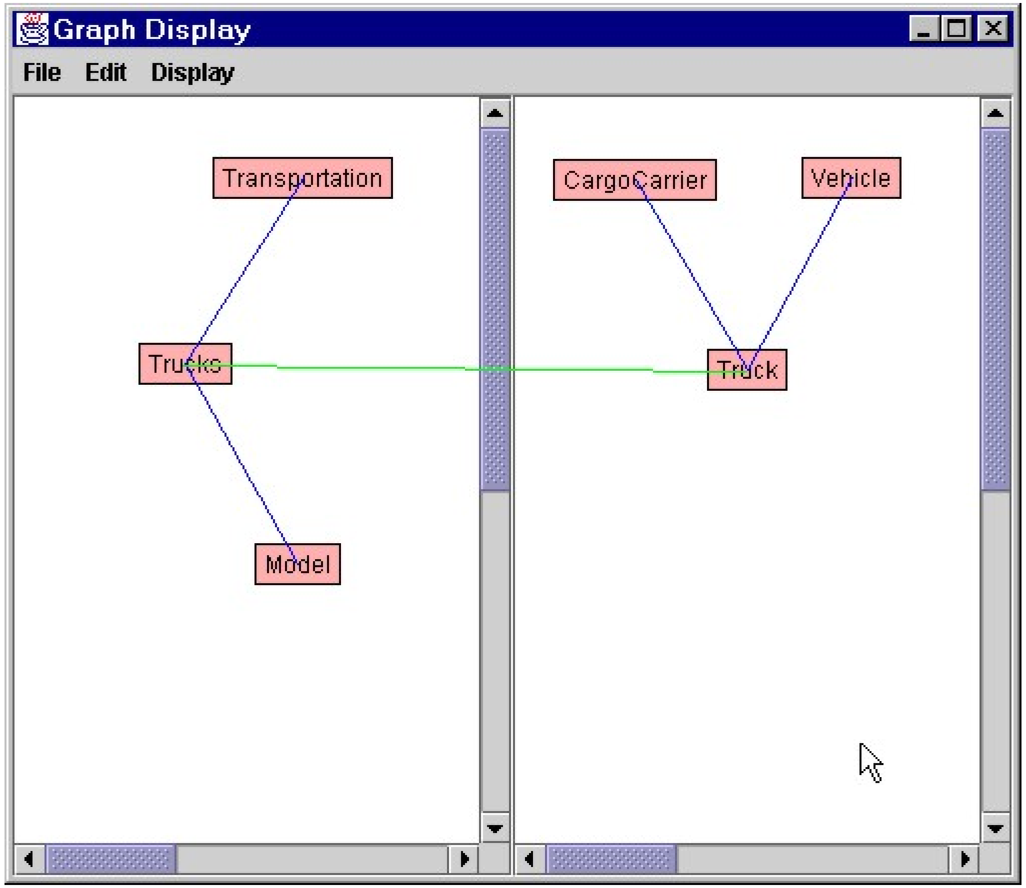
Transport ontology

Suggestions for articulations

continue from initial point

Also suggest similar terms for further articulation:

- **by spelling similarity,**
- **by graph position**
- **by term match repository**



Expert response:

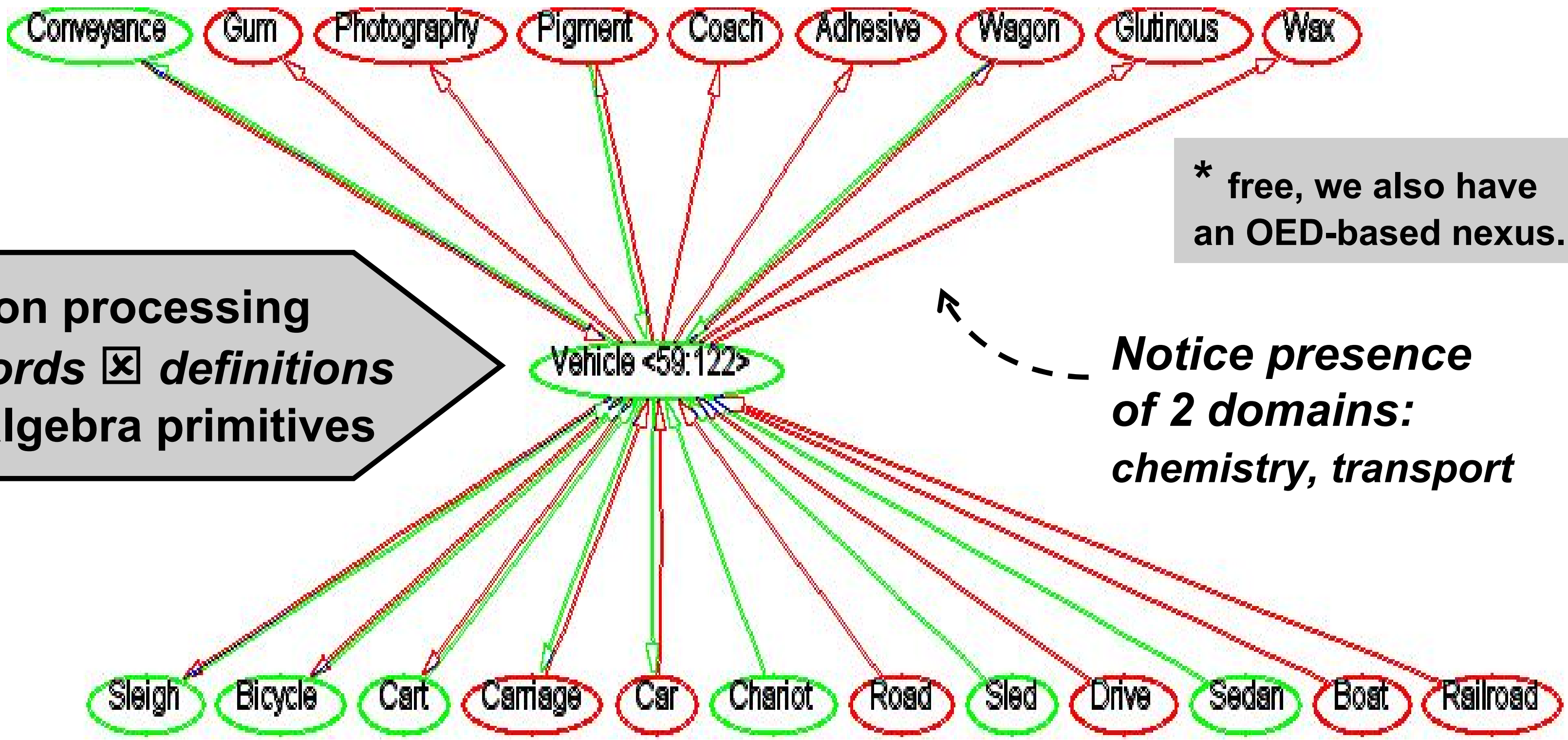
1. **Okay**
2. **False**
3. **Irrelevant to this articulation**

All results are recorded

***Okay*’s are converted into articulation rules**

Candidate Match Nexus

Term linkages automatically extracted from 1912 Webster's dictionary *

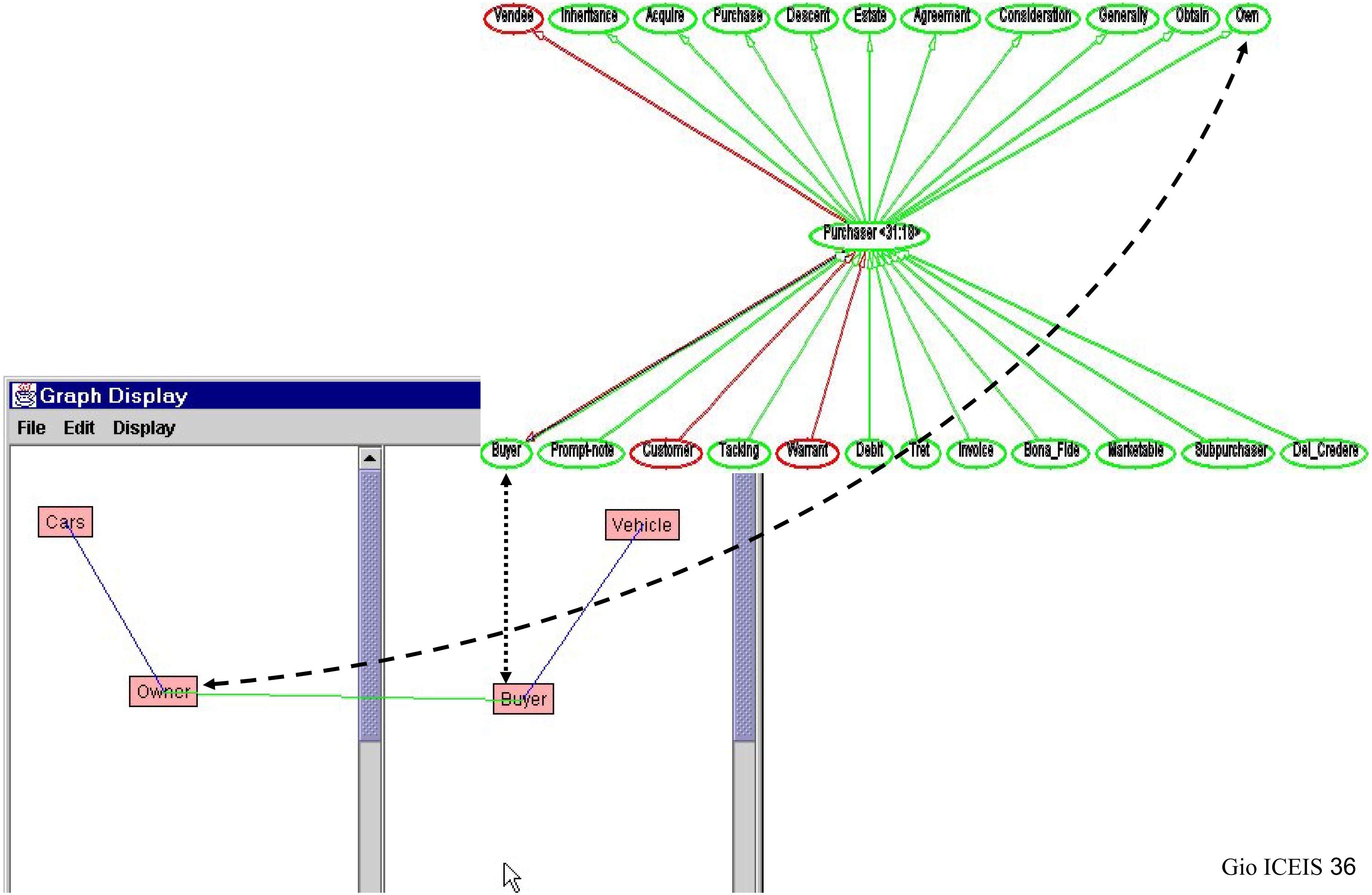


*** free, we also have an OED-based nexus.**

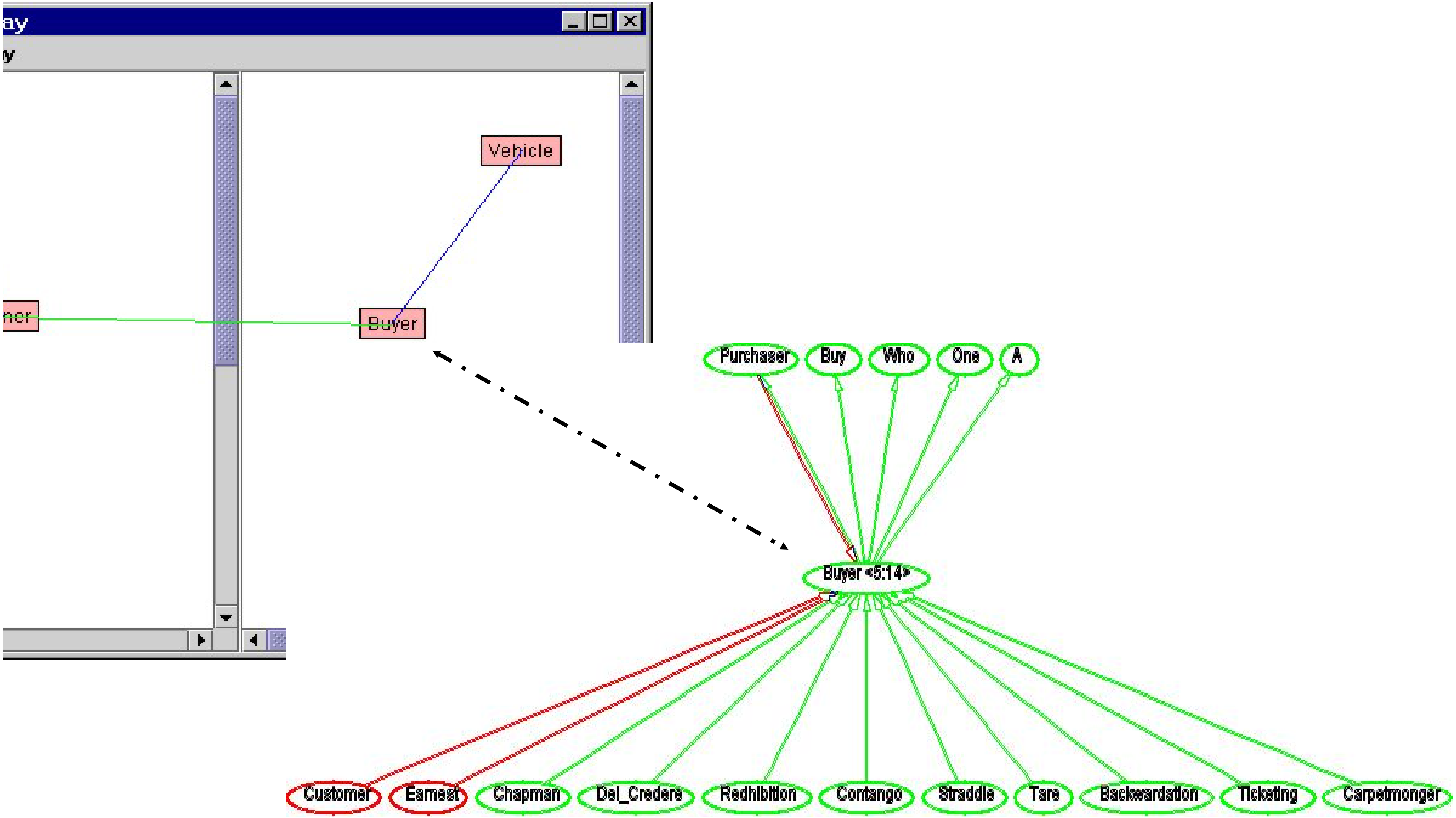
Based on processing *headwords* ⊠ *definitions* using algebra primitives

Notice presence of 2 domains: chemistry, transport

Using the nexus

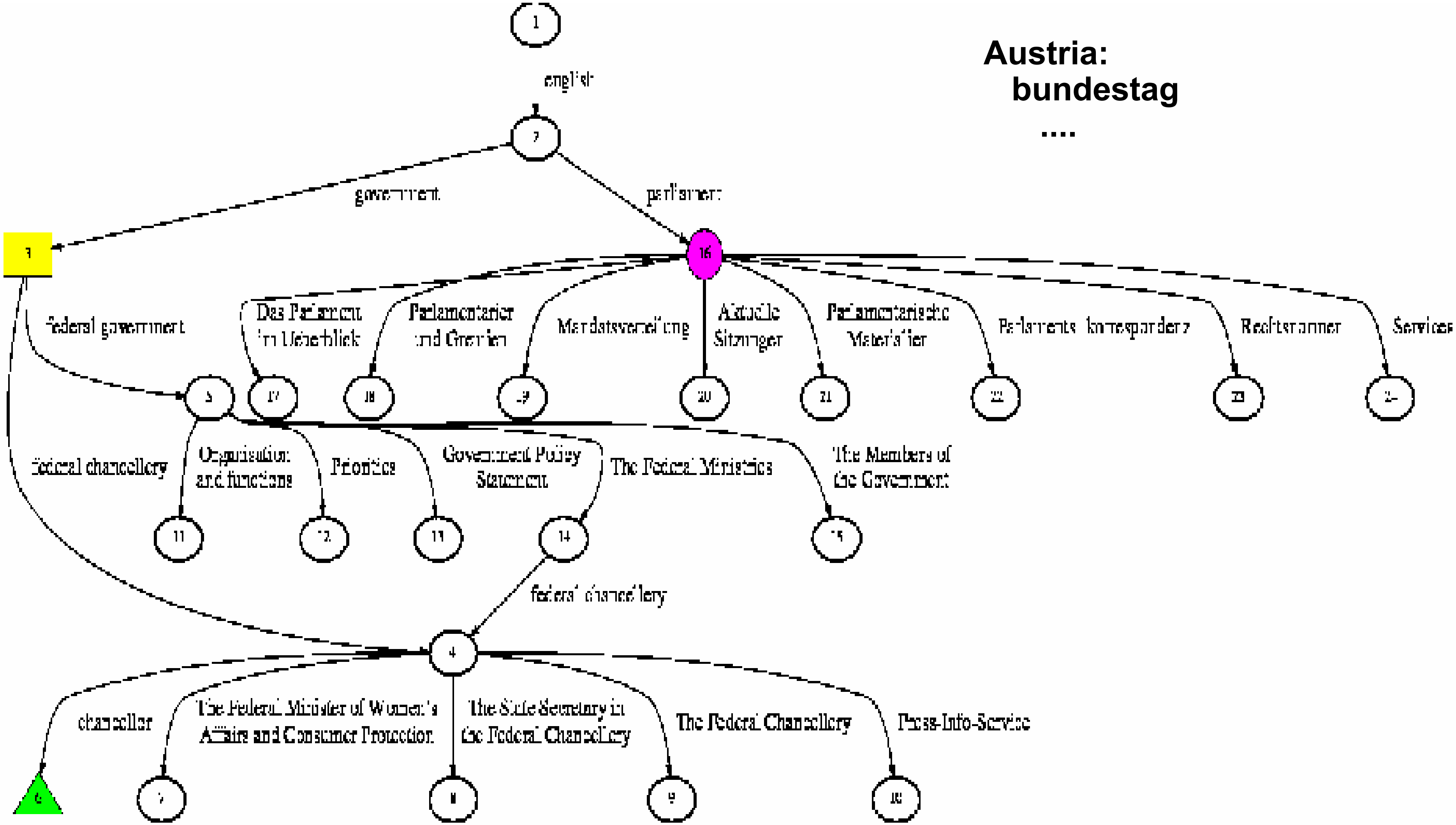


Navigating the match repository



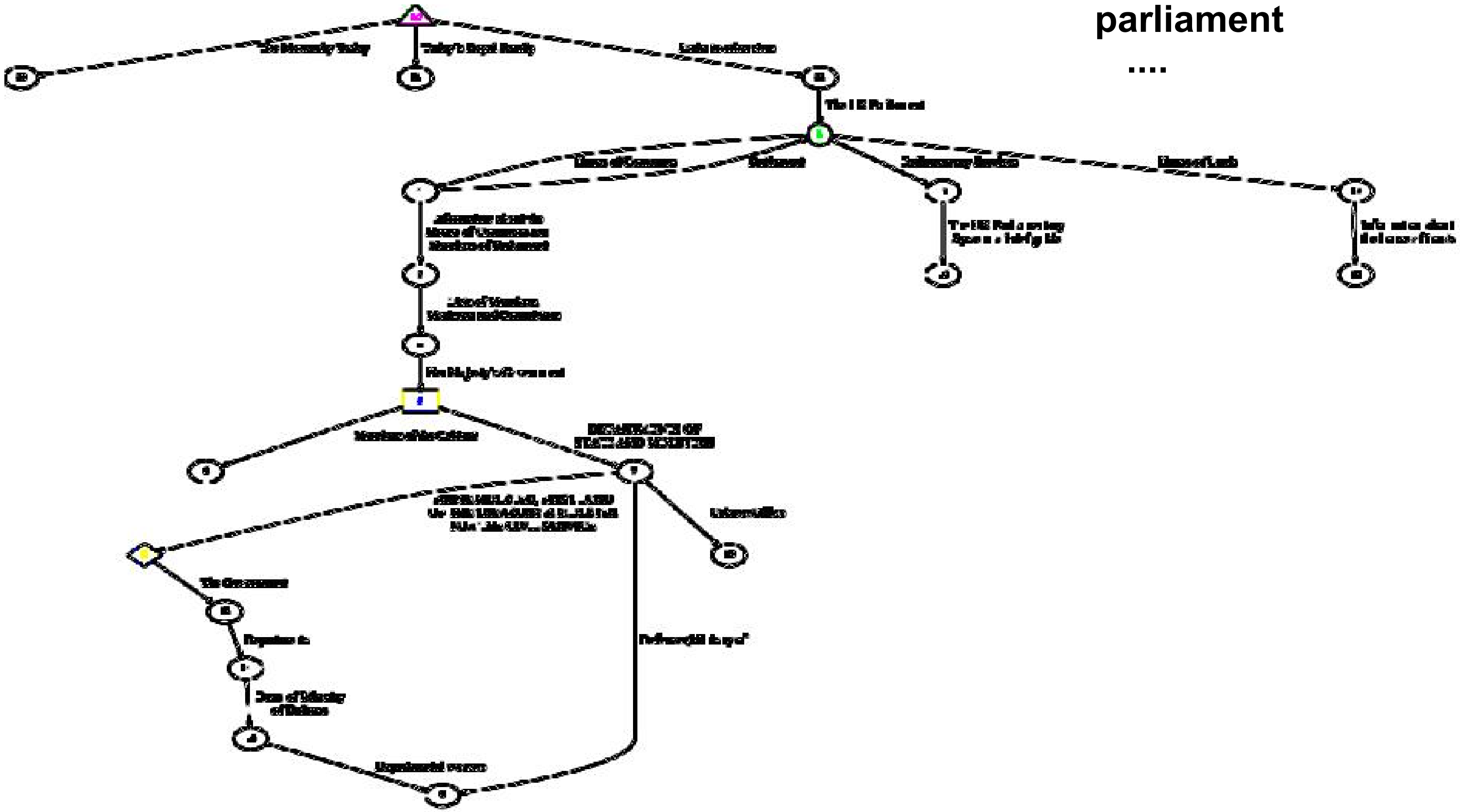
Example: NATO Country Graphs

Austria:
bundestag
....

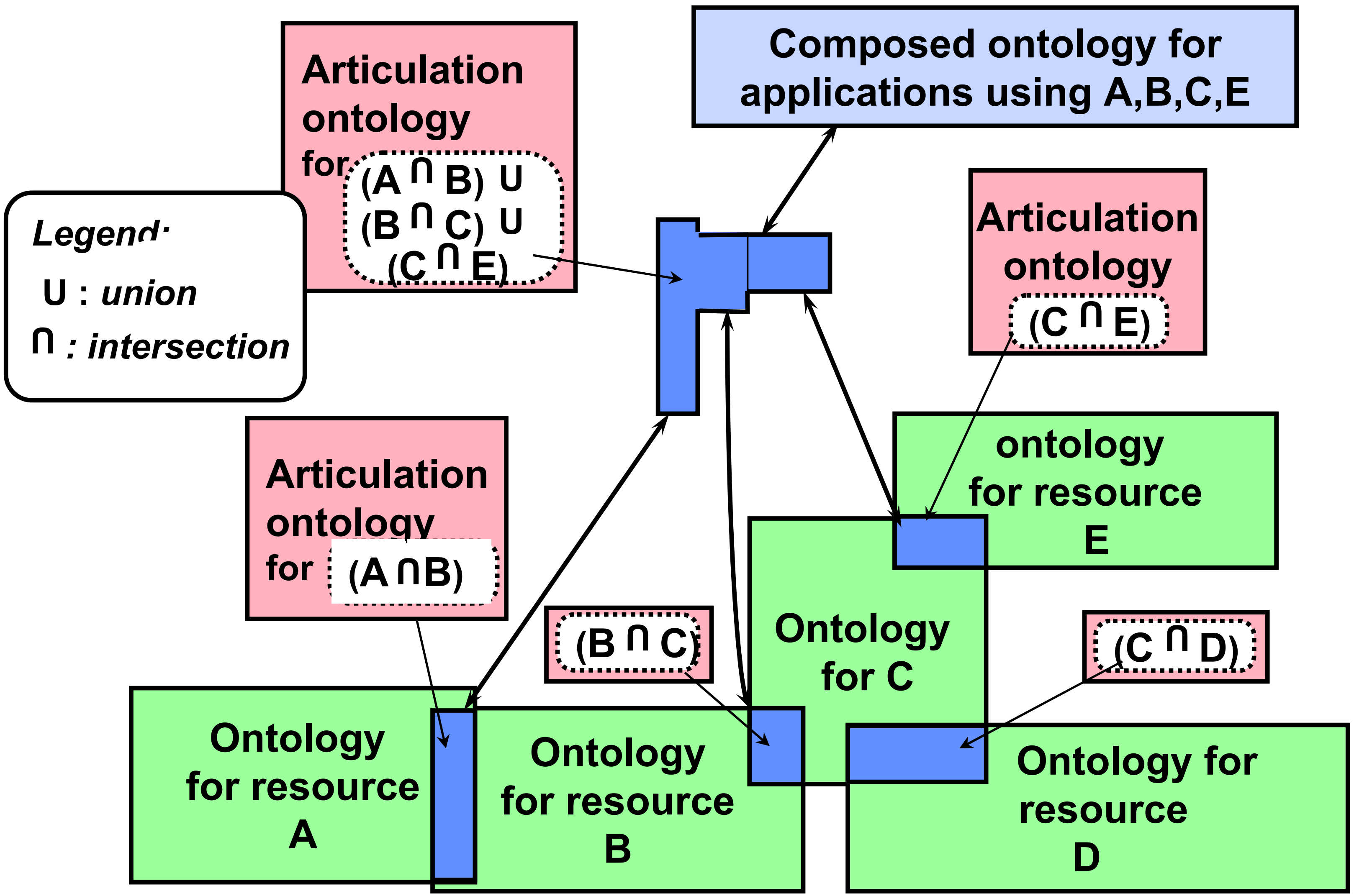


To be matched to

Great Britain
parliament
....



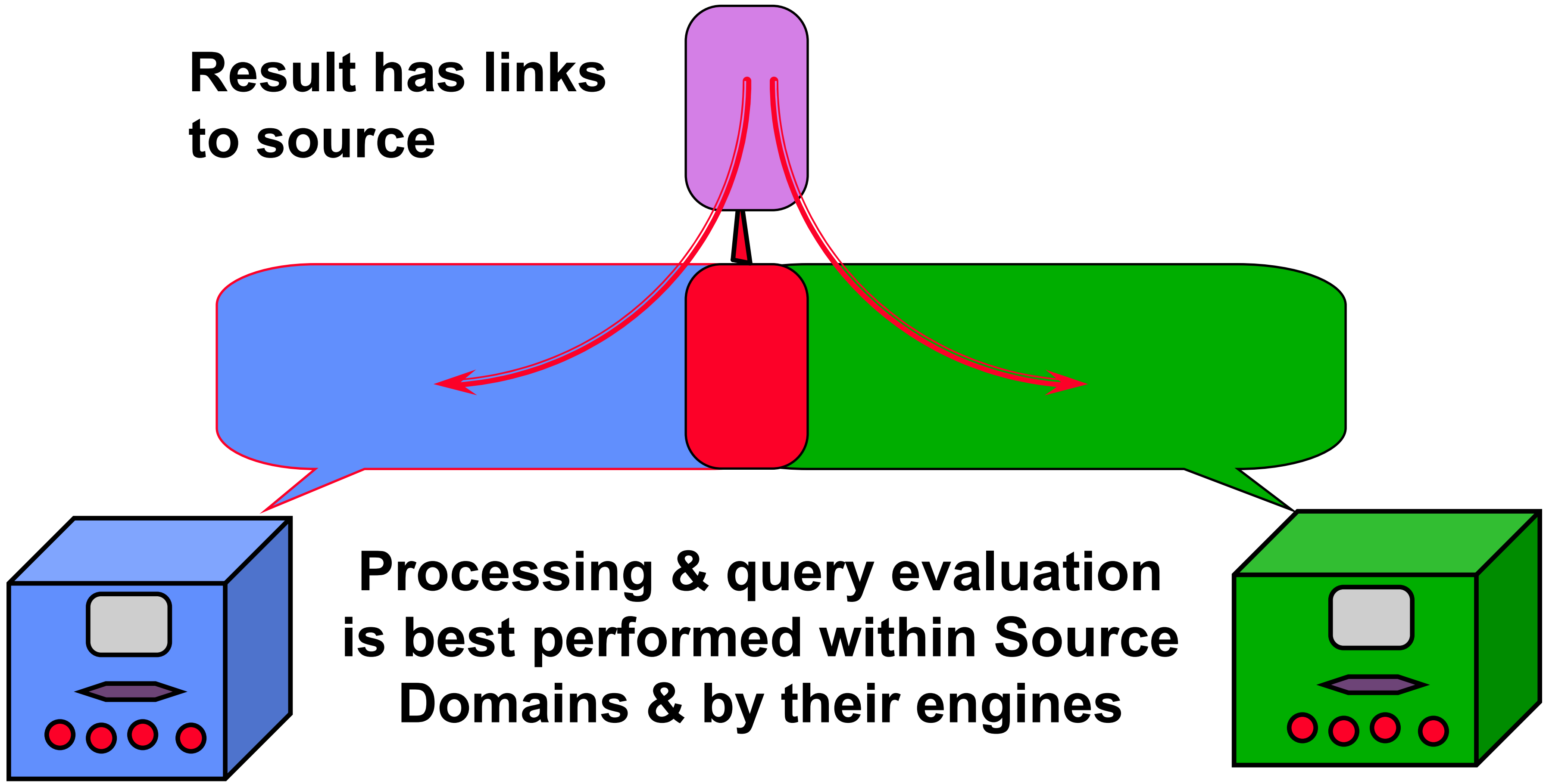
Broader Applications Compose



Exploiting the result

Future work

Result has links to source



Processing & query evaluation is best performed within Source Domains & by their engines

SKC Synopsis

- Research Objective:
  – Precise answers from heterogeneous, imperfect, scalably many data sources
- Sources for Ontologies:
  – General: CIA World Factbook '96, UN-www, OPEC-www Webster's Dictionary, Thesaurus, Oxford English Dictionary
  – Topical: NATO, BattleSpace Sensors, Logistics Servers
- Theory:
  – Rule-based algebra over ontologies
  – Translation & Composition primitives
- Sponsor and collaboration
  – AFOSR; DARPA DAML program; W3C; Stanford KSL and SMI; Univ. of Karlsruhe, Germany; others.

Domain Specialization

- **Knowledge Acquisition** ***(20% effort)*** **&**
- **Knowledge Maintenance** ***(80% effort *)***

to be performed by

- **Domain specialists**
- **Professional organizations**
- **Field teams of modest size**

autonomously maintainable

Empowerment

*** based on experience with software**

Innovation in SKC

- No need to harmonize full ontologies
- Focus on what is critical for interoperation
- Rules specific for articulation
- Tools for creation and maintenance
  - Maintenance is distributed
    - » to *n* sources
    - » to *m* articulation agents
- Potentially many sets of articulation rules

is $m < n^2$, depends on semantic architecture density

a research question: density

Conclusion

- **High precision is important for enterprise applications**
  - **cost of overload versus opportunity loss**
- **Semantic differences cause problems**
  - **Today solved by human intermediate experts**
  - **Will need automation support**
  - **Tools so that expert knowledge is captured**
- **Scalability requires a thorough foundation**
  - **Algebra provides composition, formal basis, delegation**
  - **Formal composition supports maintenance**
  - **Delegation of responsibility and authority enhances quality**
- **Many research tasks left**