

---

Integration of distributed terminology resources to facilitate subject cross-browsing for library portal systems.

Libo Si, Ann O'Brien, Steve Probeta  
Department of Information Science,  
Loughborough University.

# Background: library portal products

---

- Library portal products incorporate federated search service
  - E.g. ddWiz, SirSi Rooms, MetaLib, WebFeat, Primo, etc.
- There is a lack of subject cross-browsing

# Current solution: terminology services

---

- A terminology service holds a number of controlled vocabularies and their mappings, and provides programmatic M2M interfaces which allow other services to manipulate their terminology data in M2M ways during the searching/browsing process. (e.g. HILT, STAR, OCLC).
- Terminology services include:
  - Controlled vocabularies which have been represented in particular encoding formats, and published on the Web
  - Mapping sets between different controlled vocabularies which have been represented in encoding formats, and published on the Web
  - Local vocabularies which are being used by a library portal system for local subject indexing and cataloguing.

# Issues with terminology services

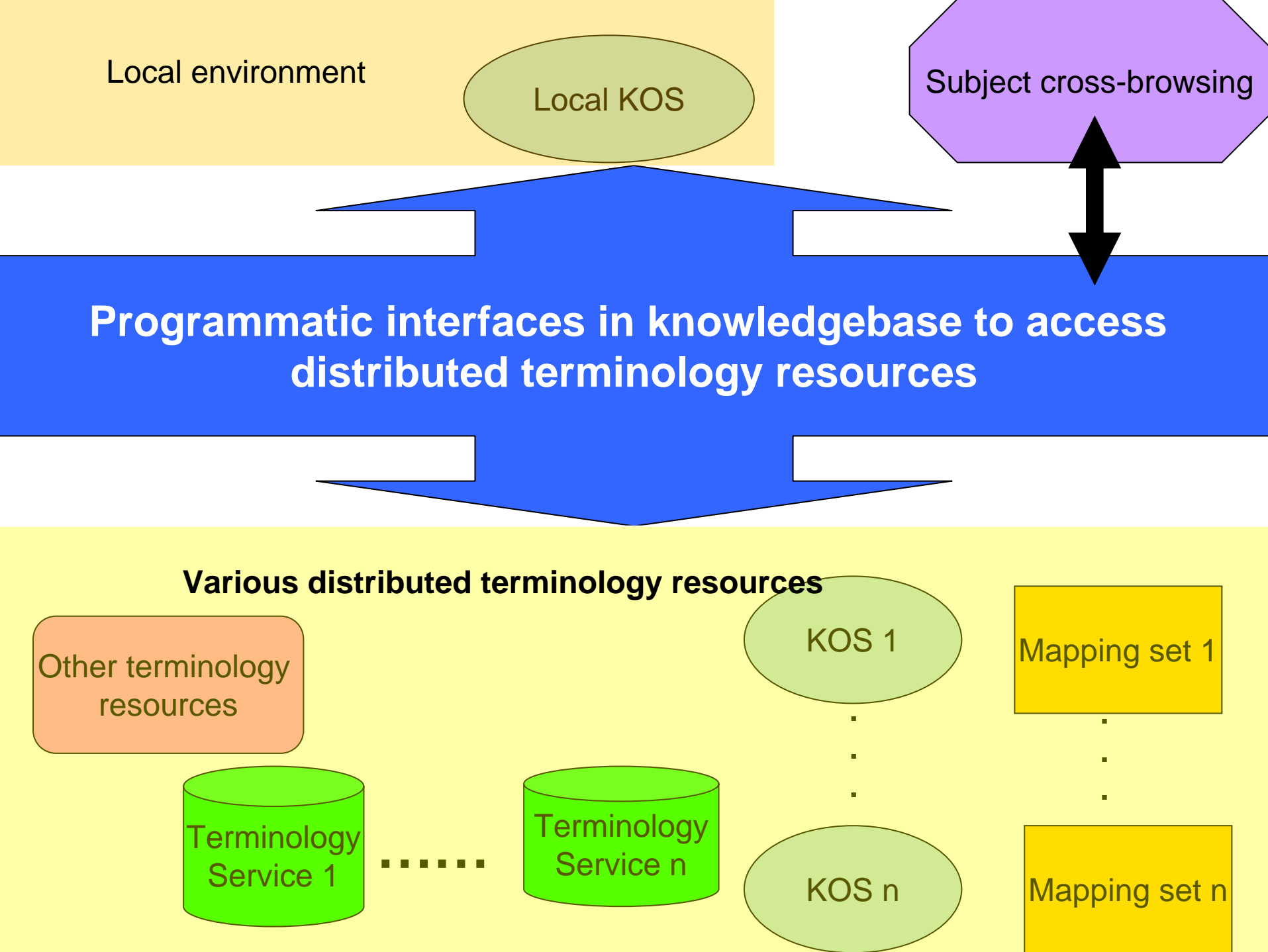
---

- It is impossible to have a terminology service which includes all the controlled vocabularies and mappings required in a library portal.
- It is important to consider methods that combine a variety of terminology services and local terminology resources to support subject cross-browsing within a local library portal.
- Semantic heterogeneity is desirable
- Technical heterogeneity is desirable

# Purpose of this research

---

- To develop a middleware platform to integrate technically and semantically, different terminology resources and also to provide subject cross-browsing services to different library portal systems.



# Methodology

---

- Nine expert interviews (based on the outcomes of a literature review) to collect in-depth ideas, and inform the development of a research framework
- Deep investigation of 50 widely-used KOS to gain a clear insight into the different characteristics of various KOS
- Development of a prototype system
- Expert evaluation

# Issues discussed in this presentation

---

- Semantic interoperability:
  - Structural model for establishing mappings
  - Indirection problems caused by the use of a central spine
  - Treatment of compound concepts
- Practical implementation of mappings
- Technical architecture:
  - Technical integration of different terminology resources
  - Technical integration of the middleware and a library portal system.

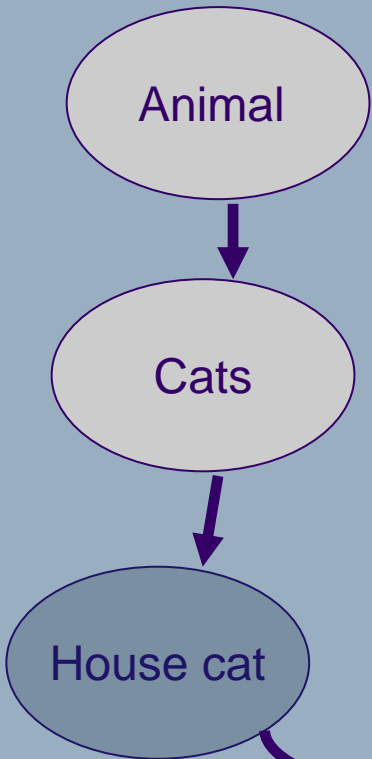
# Semantic integration: structural model for mappings

- Direct mappings VS. central spine;
- The selection of switch language.
- The use of Dewey Decimal Classification (DDC) as a switch language to exchange terminological data between different KOS still makes sense in the medium to long term.
  - DDC offers a limited capability of notational synthesis. For example, it is possible to combine the 026 (library) and 780 (music) into a compound concept 026.78 (music library)
  - DDC is not only a widely-used classification scheme, used by many academic libraries throughout the world, but also has been applied as a switch language by a number of terminology services such as the HILT terminology service, OCLC terminology service, Renardus, etc.
  - DDC has been encoded in MARC21 XML data format
  - Many metadata records have been indexed not only by DDC, but also by other vocabularies.

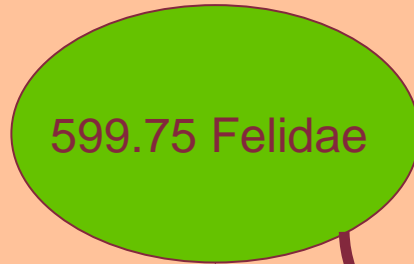
---

# Indirection problem caused by the use of the DDC

Local taxonomy

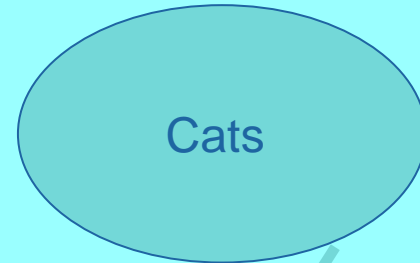


Switch language



Broad match

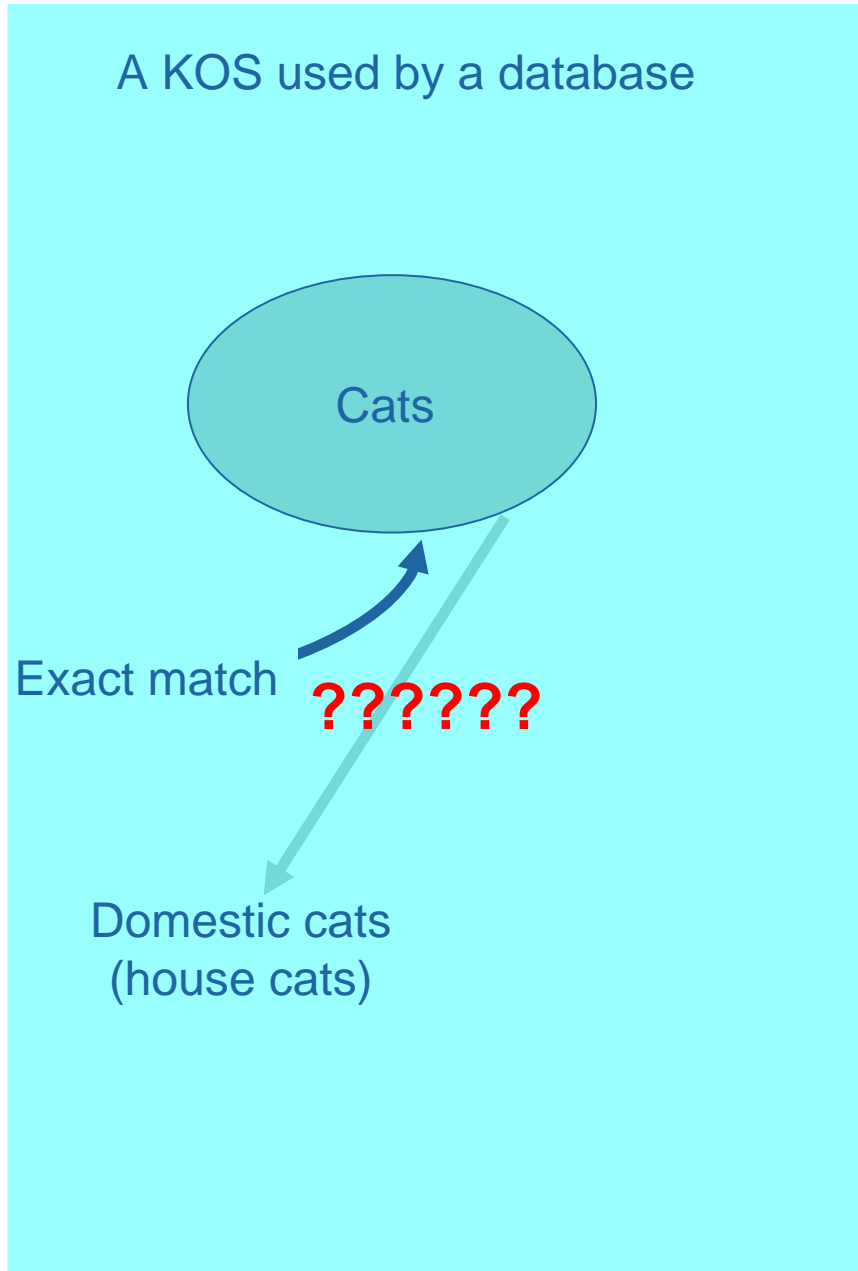
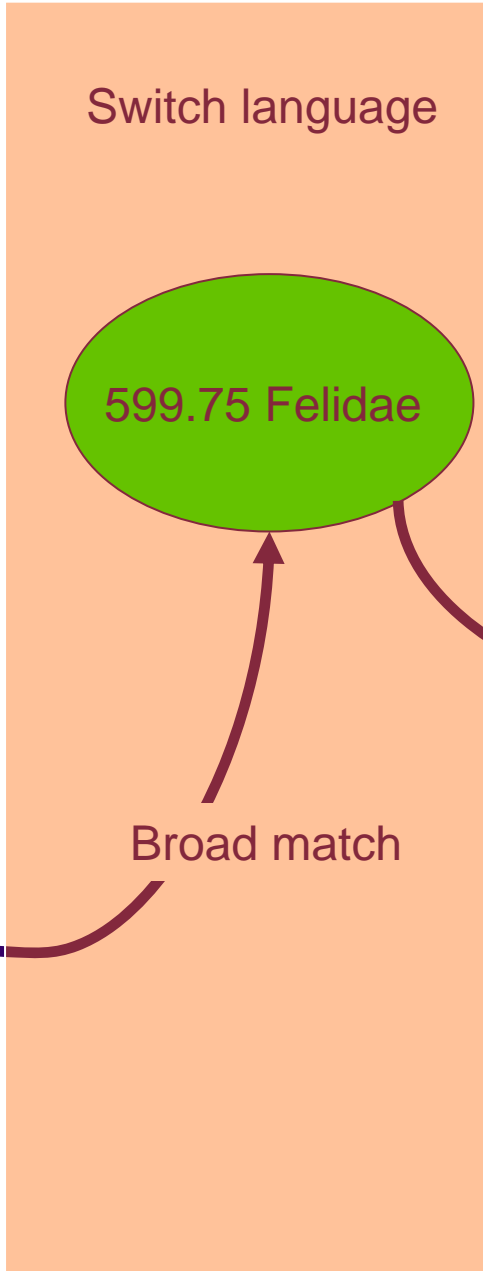
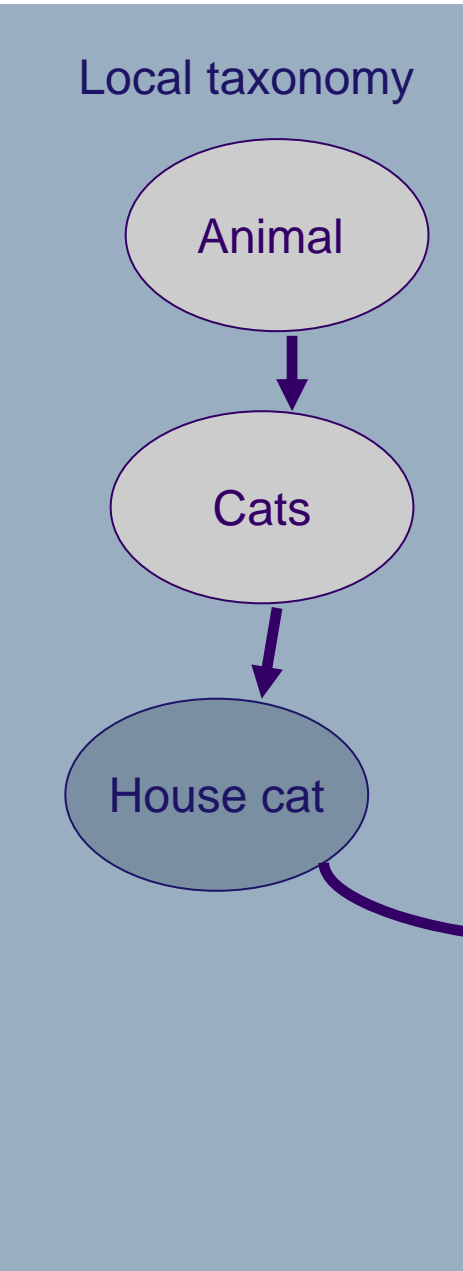
A KOS used by a database



Exact match

??????

Domestic cats  
(house cats)

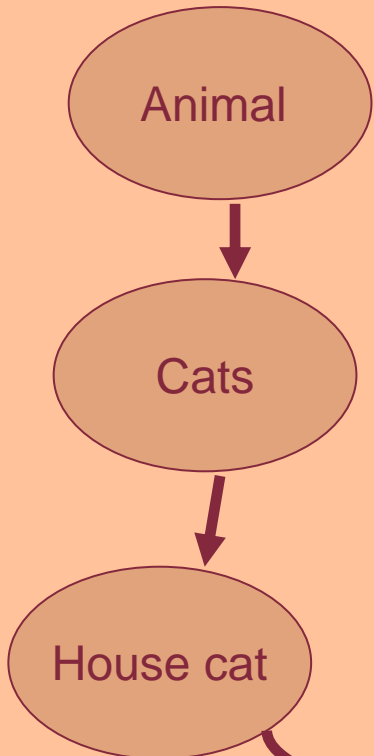


## Two possible solutions

---

- A combination of machine intelligence (query expansion algorithm) and mapping workers' intelligence;
- A combination of machine intelligence (query expansion algorithm) and users' intelligence.

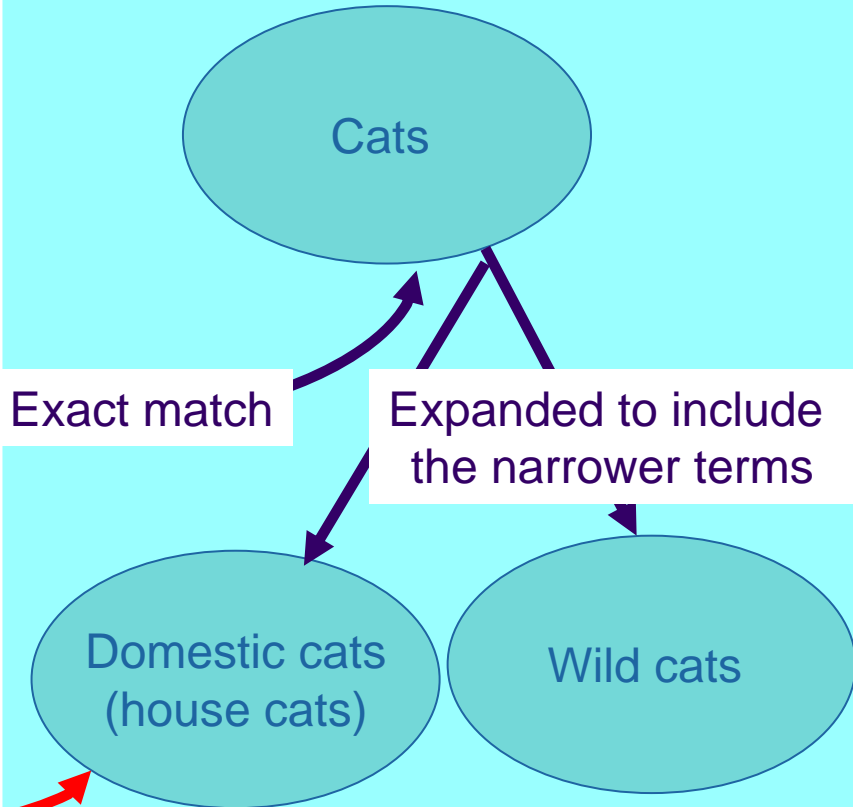
Local taxonomy



Switch language



A KOS used by a database



Exact match

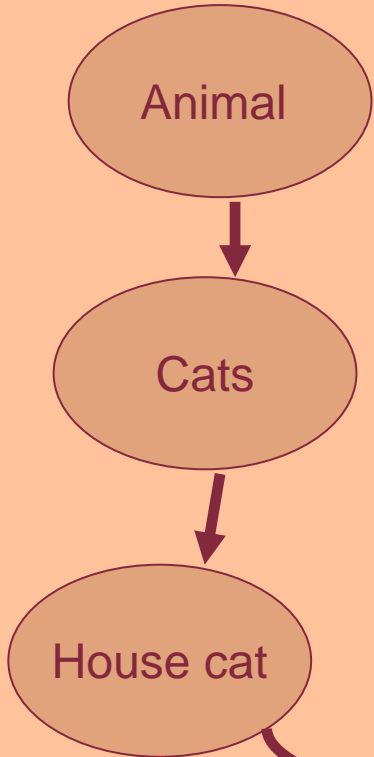
Broad match

Exact match

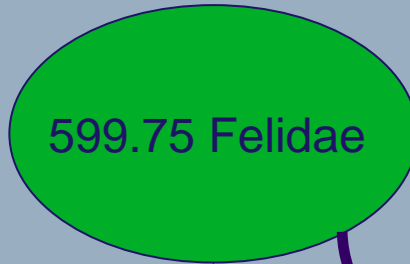
Expanded to include the narrower terms

Mapping workers can discover the direct mappings based on established mappings

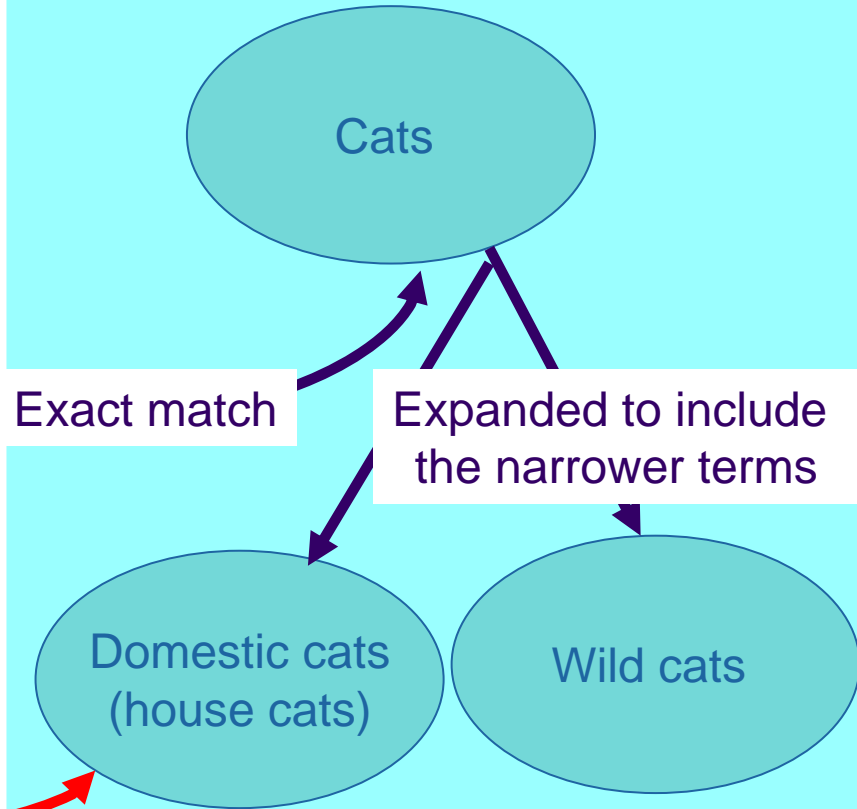
Local taxonomy



Switch language



A KOS used by a database



Exact match

Broad match

Exact match

Expanded to include the narrower terms

Domestic cats (house cats)

Wild cats

End-users can discover more appropriate subject terms for their search.

# Using the User's Intelligence

---

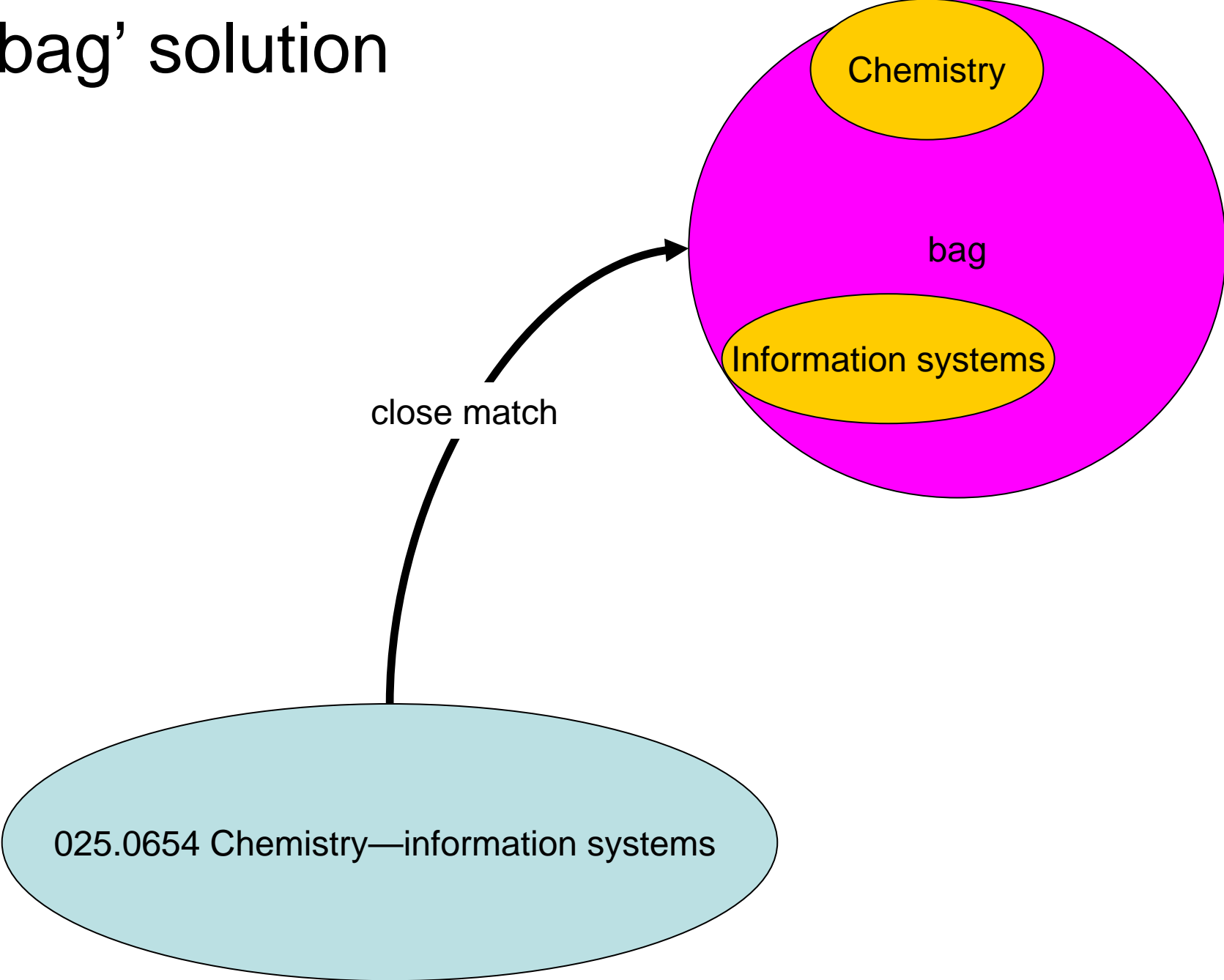
- Do you mean “Chemistry?”
- Query expansion:
  - Do you mean “Chemistry, Biochemistry, Biogeochemistry, Alchemy, Analytic chemistry, Clinical chemistry,”

# Treatment of compound concepts

---

- A DDC concept: 025.0654 Chemistry—information systems.
- Two concepts in UKAT: Chemistry, information system.
- Possible solutions:
  - Boolean operators;
  - each of UKAT concepts can be separately mapped against the DDC concept;
  - The use of a facet to combine these two UKAT terms;
  - Other solutions?

# A 'bag' solution



# Present a bag of concepts to users

---

- Do you mean “Chemistry, information system, Chemistry and information system, Chemistry or information system, Chemistry not information system, information system not chemistry?”

# Who should create the mappings?

---

- KOS owners
- The reuse of established mappings provided by terminology services
- Local librarians
- Library portal providers

# The inconsistency in establishing mapping sets

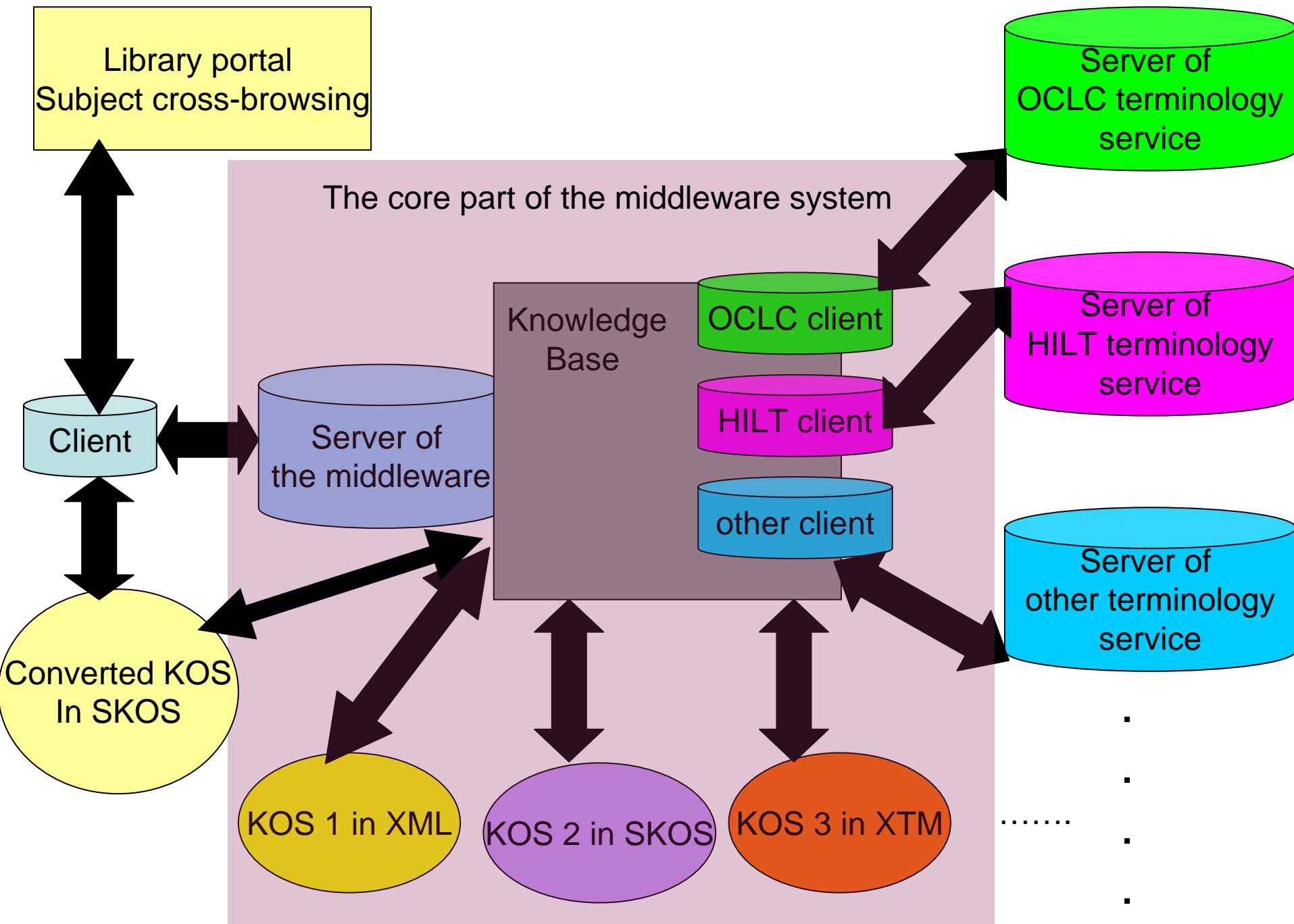
---

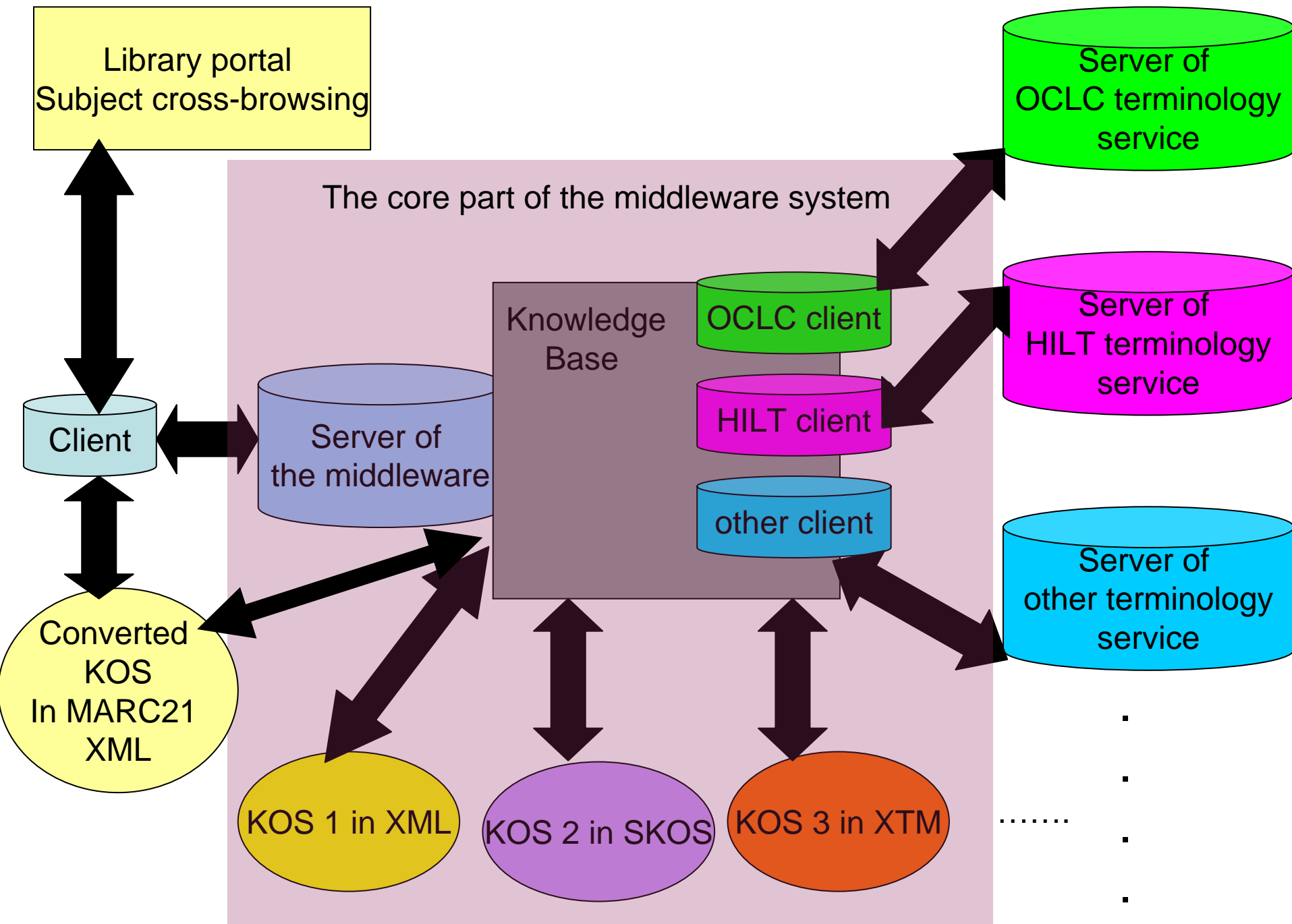
- Different terminology services may use different mapping strategies, such as provenance (source), methods (intellectual, co-occurrence, other automatic, etc), concept indicators, and so on...
- **Ideally**, one central team with sufficient expertise should be formed to assess these distributed mapping data sets and should have the responsibility for improving the consistency and quality of distributed mapping resources.
- **However**, It may be possible to create a terminology mapping registry to record these different characteristics of distributed mapping sets;

# Technical integration

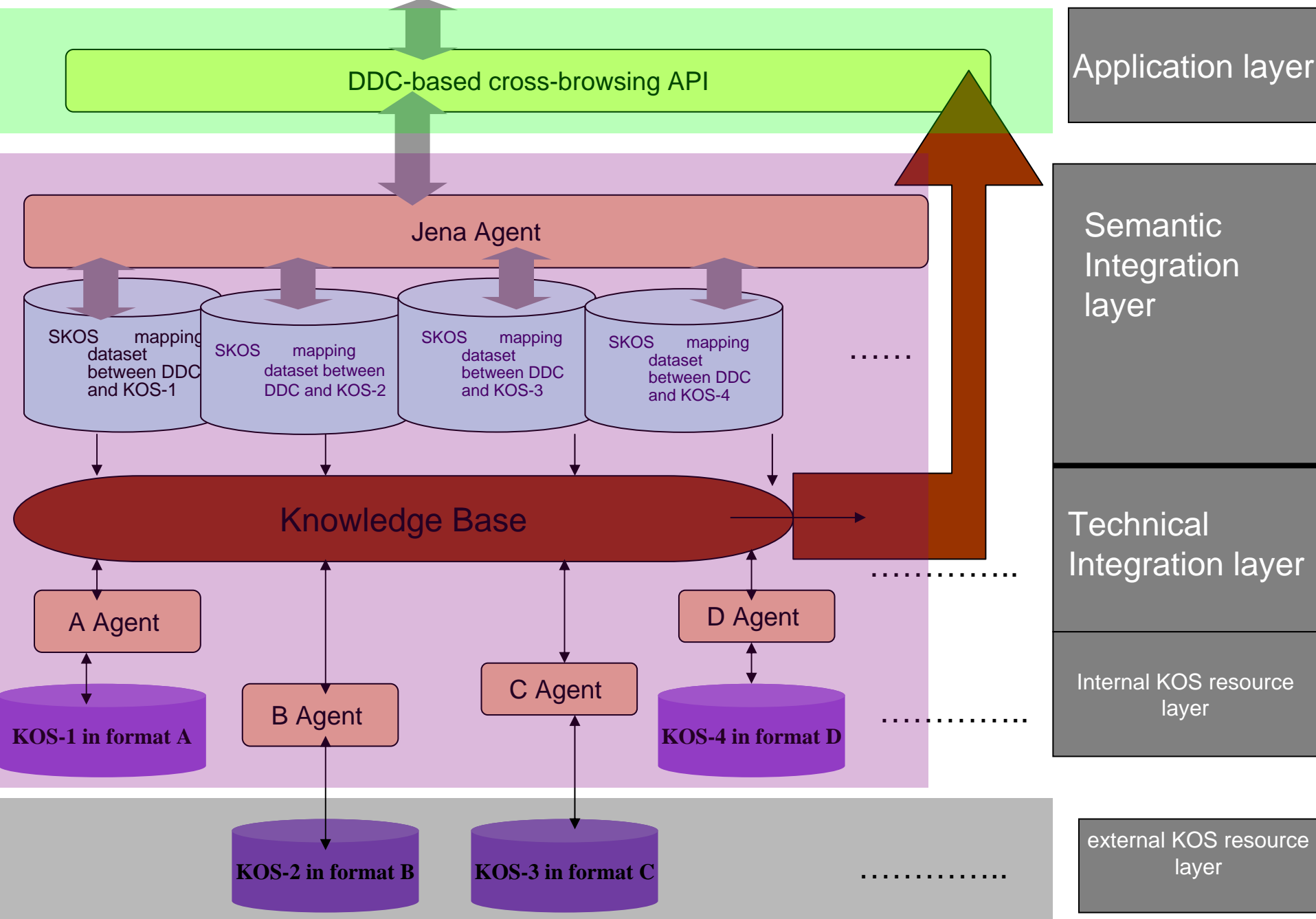
---

- A knowledge base was developed to store connectivity details of different terminology resources.
- Two core components in the knowledge base:
  - Query transmission: translate the users' queries into appropriately structured queries which the different terminology resources could understand.
  - Format conversion: convert the returned terminological records into a consistent format.



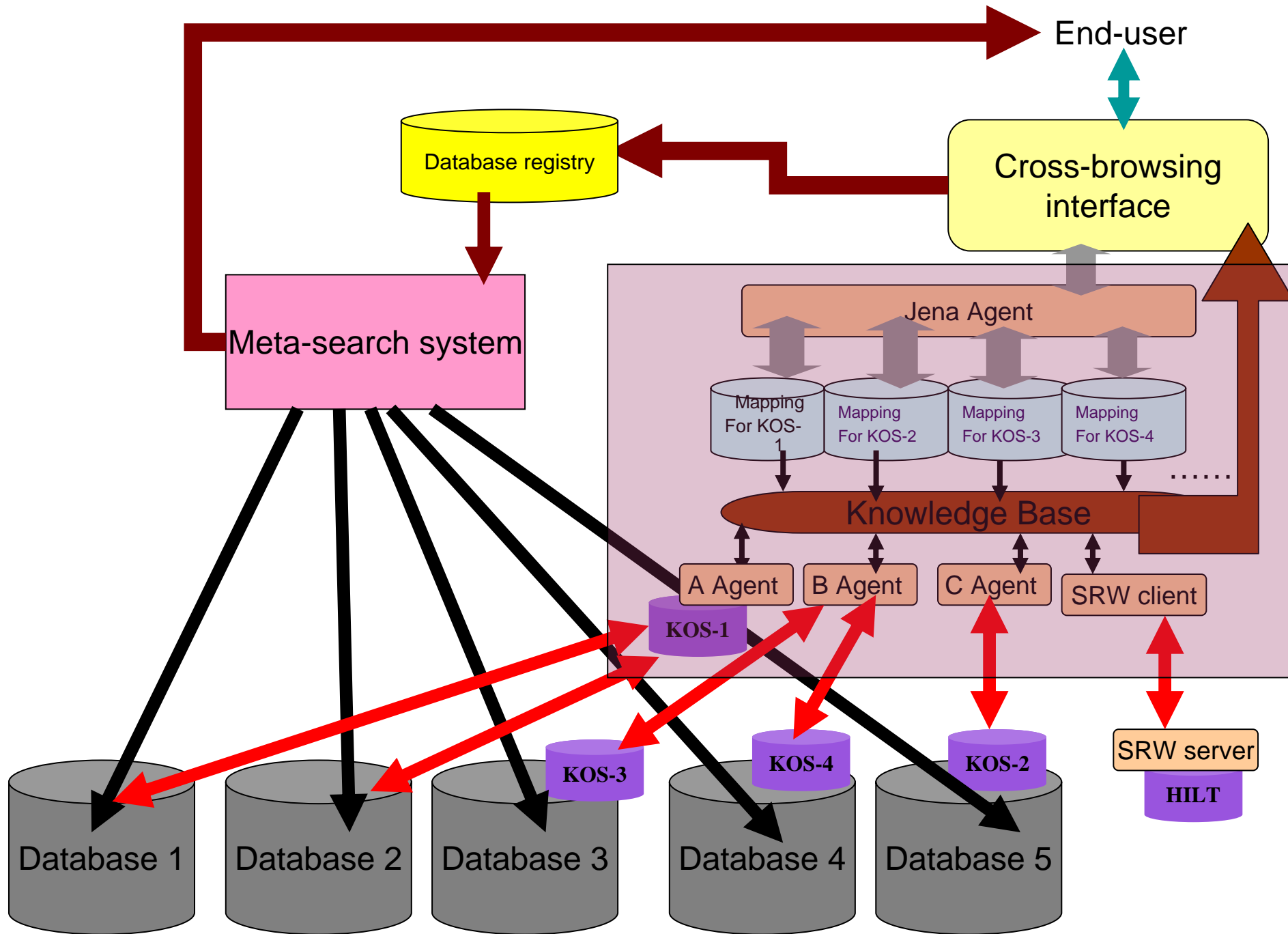


# End-user



---

# **The integration of terminology resources and federated search**



# Recommendations: semantic integration

---

- To use DDC as a switch language, but in consider UDC, BLISS as a future alternative
- To develop a collaborative mapping registry where different mapping work can be stored, maintained, and reused
- To use query expansion algorithms to expand mapped terms, and enable users or mapping workers to discover direct mappings
- To use a bag to combine these individual concepts, map this bag against the compound concept, and then use a Google-styled “Do you mean...?” plus all possible Boolean combinations of mapped terms

# Recommendations: technical integration

---

- To develop a knowledge base, which can record the functionality provided by the different terminology services, translate the users' query into different forms of the queries that different terminology services can accept, and convert into a consistent format.
- To integrate the middleware platform between different terminology resources with the federated search service provided by a library portal product, and then the enable users to gain relevant metadata records through subject cross-browsing

---

**Thank you!**  
Any questions?