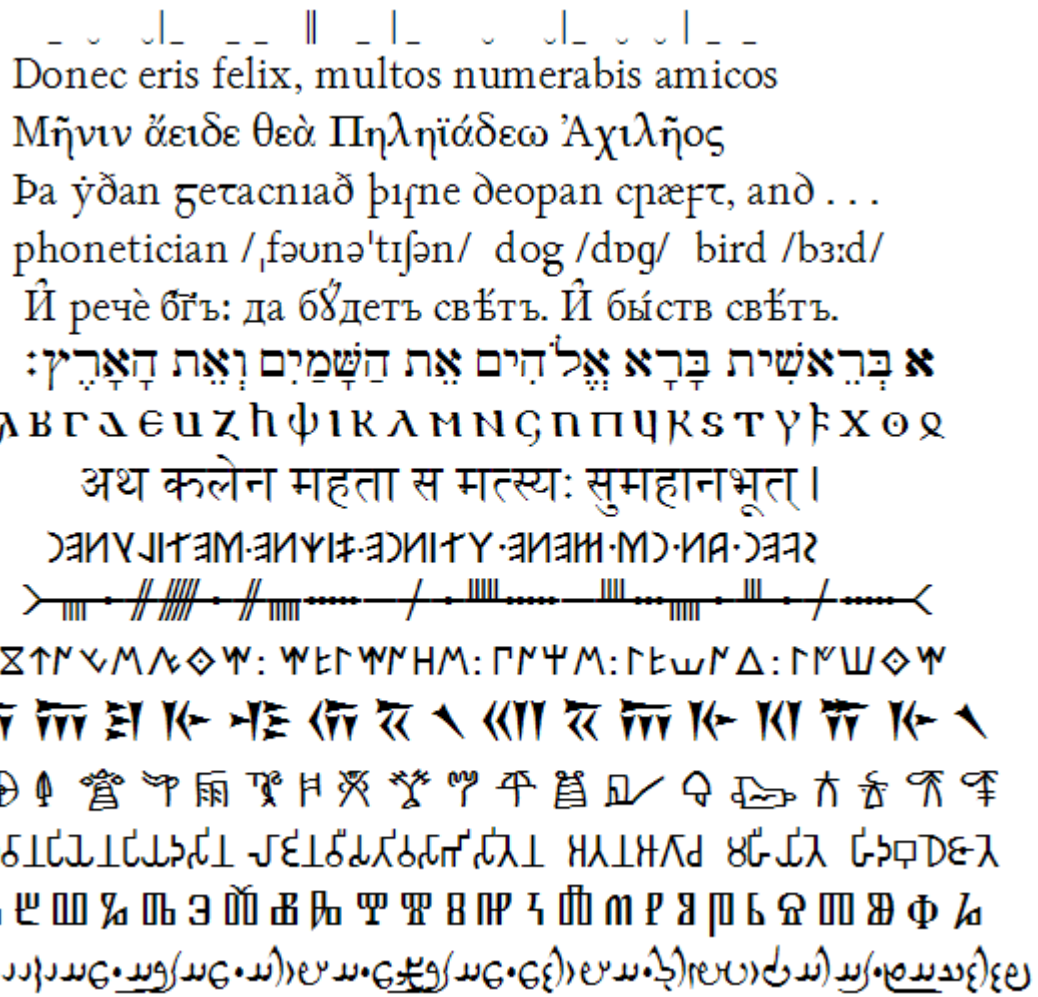
Encoding formats and consideration of requirements for terminology mapping

Libo Si, Department of Information Science, Loughborough University

Structure of this presentation

- Introduction to KOS mapping methods developed;
- Introduction to four encoding formats;
- Two frameworks to improve interoperability between different encoding formats;

Interoperability



Mapping to bridge the semantic gaps between different systems?

- “the process of associating elements of one set with elements of another set, or the set of associations that come out of such a process”. (www.semantic world.org)

Establishing semantic mapping between KOS

- [1] Zeng, Marcia Lei and Lois Mai Chan. 2004. Trends and issues in establishing interoperability among knowledge organization systems;
- [2] BS8723-Part 4;
- [3] Patel, Manjula, Koch, Traugott, Doerr, Martin and Tsinaraki, Chrisa (2005). Semantic Interoperability in Digital Library Systems.
- [4] Tudhope, D., Koch, T. and Heery, R. (2006). Terminology Services and Technology.

Mappings between KOS in the semantic level

- Derivation;
- Direct mapping;
- Switch language;
- Co-occurrence mapping;
- Satellite and leaf node linking;
- Merging;
- Linking through a temporary union list;
- Linking through a thesaurus server protocol.

Factors to challenge KOS interoperability in different levels

Levels of interoperability	Factors of interoperability
Scheme level	Different subject areas
	Different degree of pre-coordination/post-coordination
	Different granularity
	Different languages
Record level	Different encoding formats
	Different metadata schemes to describe KOS
System level	Different protocols to access KOS
	Different IR systems

Knowledge representation formats

- MARC21 for authority files;
- Zthes XML DTD/Schema;
- XML Topic Map for representing controlled vocabularies:
  - Techquila's Published Subject Identifiers for a thesaurus ontology;
  - Techquila's Published Subject Identifiers for a classification system ontology;
  - Techquila's Published Subject Identifiers for a faceted classification system;
  - Techquila's Published Subject Identifiers for modelling hierarchical relationships;
- SKOS: SKOS-Core, SKOS-Mapping, and SKOS-extension.

MARC 21 for authority file

```
<record>
<leader>…</leader>
<controlfield tag="001">GSAFD000002</controlfield>
<controlfield tag="003">IlchALCS</controlfield>
<controlfield tag="005">20000724203806.0</controlfield>
<datafield tag="040" ind1="" ind2="">  <subfield code="a">IlchaALCS</subfield>
  <subfield code="b">eng</subfield>  <subfield code="c">IEN</subfield>
  <subfield code="f">gsafd</subfield>
</datafield>
<datafield tag="155"> <subfield code="a">Adventure film</subfield> </datafield>
<datafield tag="455"> <subfield code="a">Swashbucklers</subfield></datafield>
<datafield tag="455"> <subfield code="a">Thrillers</subfield> </datafield>
<datafield tag="555"> <subfield code="w">h</subfield><subfield code="a">spy
      films</subfield></datafield>
<datafield tag="555"> <subfield code="w">h</subfield><subfield code="a">spy television
      programs</subfield></datafield>
<datafield tag="555"> <subfield code="w">h</subfield><subfield code="a">western
      films</subfield></datafield>
<datafield tag="555"> <subfield code="w">h</subfield><subfield code="a">western televsion
      programs</subfield></datafield>
<datafield tag="555"> <subfield code="a">sea film</subfield></datafield>
</record>
```

Preferred term

Nonpreferred term

Narrower term

Related term

Zthes XML Schema—term-based

```
<?xml version="1.0" encoding="utf-8" ?>
<Zthes>
 <term>
       <termId>1</termId>
       <termName>Brachiosauridae</termName> <termType>PT</termType>
       <termNote>Defined by Wilson and Sereno (1998) as the clade of all organisms more closely related to
       _Brachiosaurus_ than to _Saltasaurus_.</termNote>
       <postings>
        <sourceDb>z39.50s://example.zthes.z3950.org:3950/dino</sourceDb>
       <fieldName>title</fieldName>
       <hitCount>23</hitCount>
       </postings>
      <relation>
            <relationType>BT</relationType>
             <termId>2</termId>
             <termName>Titanosauriformes</termName>
             <termType>PT</termType>
       </relation>
     <relation>
             <relationType>NT</relationType>
              <termId>3</termId>
             <termName>Brachiosaurus</termName>
               <termType>PT</termType>
     </relation>
  </term>
</Zthes>
```

XTM for representing KOS

```
<topic id=”0001”>
<xtm:instanceOf>
  <xtm:subjectIndicatorRef
       xlink:href="http://www.techquila.com/psi/thesaurus/#concept" />
  </xtm:instanceOf>
  <subjectIdentity>
       <resourceRef
       xlink:href=http://www.zoologypark.org/animals.xtm#cats />
  </subjectIdentity>
  <baseName>
       <baseNameString>cats</baseNameString>
    <variant>
         <variantName>
           <resourceData>felines</resourceData>
         </variantName>
    </variant>
  </baseName>
</topic>
```

```
<topic id=”0012”>
<xtm:instanceOf>
  <xtm:subjectIndicatorRef
       xlink:href="http://www.techquila.com/psi/thesaurus/#concept" />
  </xtm:instanceOf>
  <subjectIdentity>
       <resourceRef
       xlink:href=http://www.zoologypark.org/animals.xtm#mammals />
  </subjectIdentity>
  <baseName>
      <baseNameString>mammals</baseNameString>
</baseName>
</topic>
```

http://www.techquila.com/psi/

XTM for representing KOS

```
<association>
  <instanceOf>
    <subjectIndicatorRef
      xlink:href="http://www.techquila.com/psi/thesaurus/thesaurus.xtm#broader-narrower"/>
  </instanceOf>
  <member>
    <roleSpec>
      <subjectIndicatorRef
        xlink:href=" http://www.techquila.com/psi/thesaurus/thesaurus.xtm#broader"/>
    </roleSpec>
    <topicRef xlink:href="#0012"/>
  </member>
  <member>
    <roleSpec>
      <subjectIndicatorRef
        xlink:href=" http://www.techquila.com/psi/thesaurus/thesaurus.xtm#narrower "/>
    </roleSpec>
    <topicRef xlink:href="#0001"/>
  </member>
</association>
```

SKOS

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<skos:Concept rdf:about= "http://www.socialsciencepark.org/thesaurus/concept/a092">
      <skos:prefLabel>freedom</skos:prefLabel>
 <skos:altLabel>liberty </skos:altLabel>
<skos:scopeNote>the rights to control one's own right</skos:scopeNote>
<skos:broader rdf:resource="http://www.socialsciencepark.org/thesaurus/concept/a045"/>
      <skos:narrower rdf:resource="http://www.socialsciencepark.org/thesaurus/concept/a0945"/>
      <skos:narrower rdf:resource=
      "http://www.socialsciencepark.org/thesaurus/concept/a0946"/> <skos:narrower
      rdf:resource= "http://www.socialsciencepark.org/thesaurus/concept/a097"/> <skos:related
      rdf:resource=
"http://www.socialsciencepark.org/thesaurus/concept/b056"/>
<skos:inScheme rdf:resource=
"http://www.socialsciencepark.org/thesaurus"/>
</skos:Concept>
</rdf:RDF>
```

	MARC21 for AF	Zthes XML Schema	XTM	SKOS
Specificity	Cannot represent some complex relationships, e.g. part-whole, etc.	No support on faceted classifications	Can represent various complicated KOS	Can represent various complicated KOS, but lack of power of validating the RDF data
Ontological extensibility	Cannot be extended to an ontology	Cannot be extended to an ontology	Can be extended to a topic map ontology.	Can be extended to an OWL ontology
Term-based or concept-based	Concept-based	Term-based	Both concept-based and term-based	Concept-based
Tools, protocols or APIs to access	XSLT-related technologies, MARC systems.	XSLT-based technologies	XTM APIs, such as, TMQL,	RDF-APIs, SKOS-APIs, and SPARQL protocol
Capability of supporting mapping	Cannot encode very specific mapping relationships	No mapping capability	Can be extended to support mapping	SKOS-mapping

Issues (1)

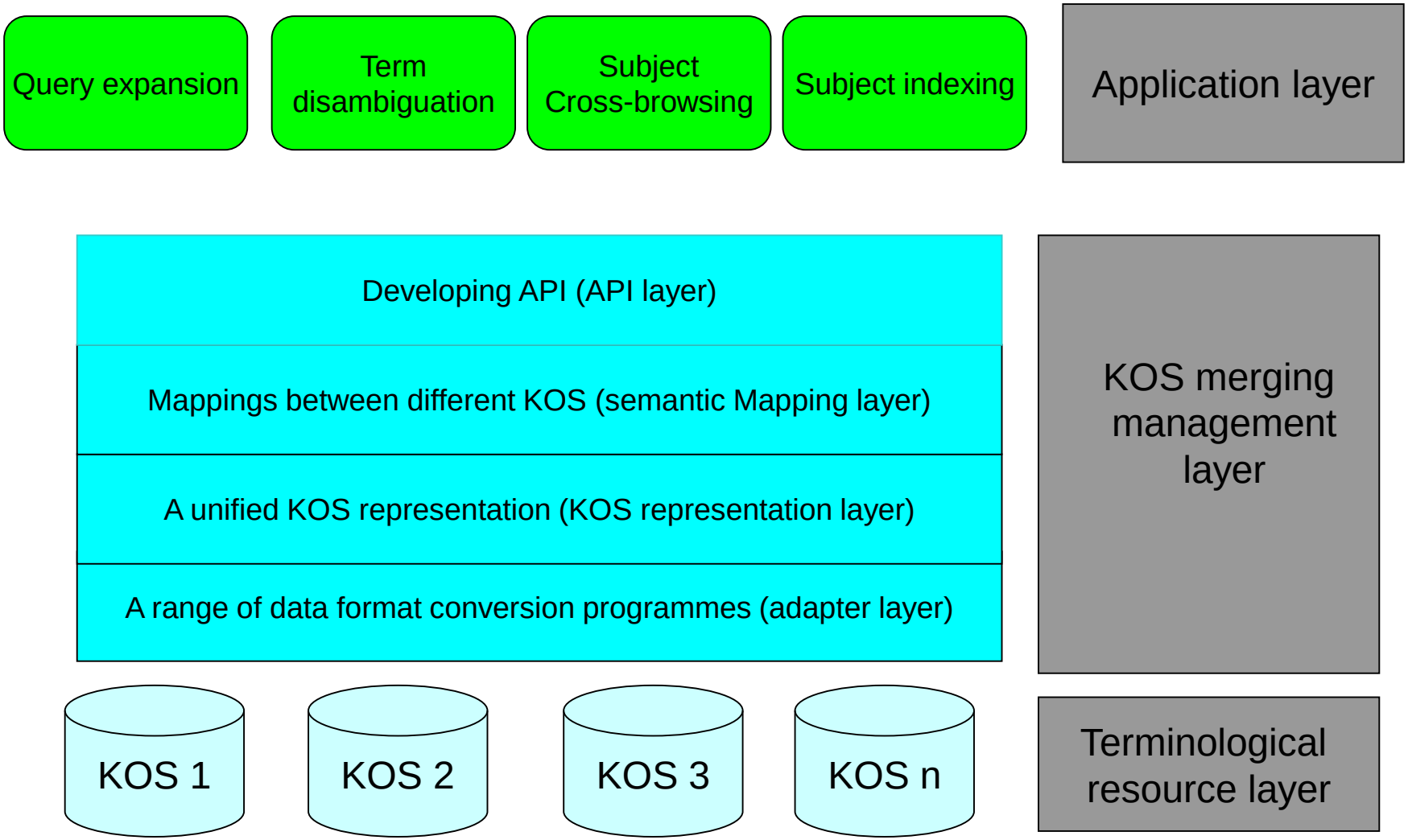
1. XML-based formats are limited and cannot represent some of the more complex thesauri or ontologies and the mappings between them, and therefore RDF-based or XTM-based formats are more appropriate to be extended to encode ontological vocabularies;
2. It is impractical to use only one representation format to encode all the controlled vocabularies, because each has its own structures and syntax. More importantly, different representation formats can be converted into each other depending on the specific requirements.

Issues (2)

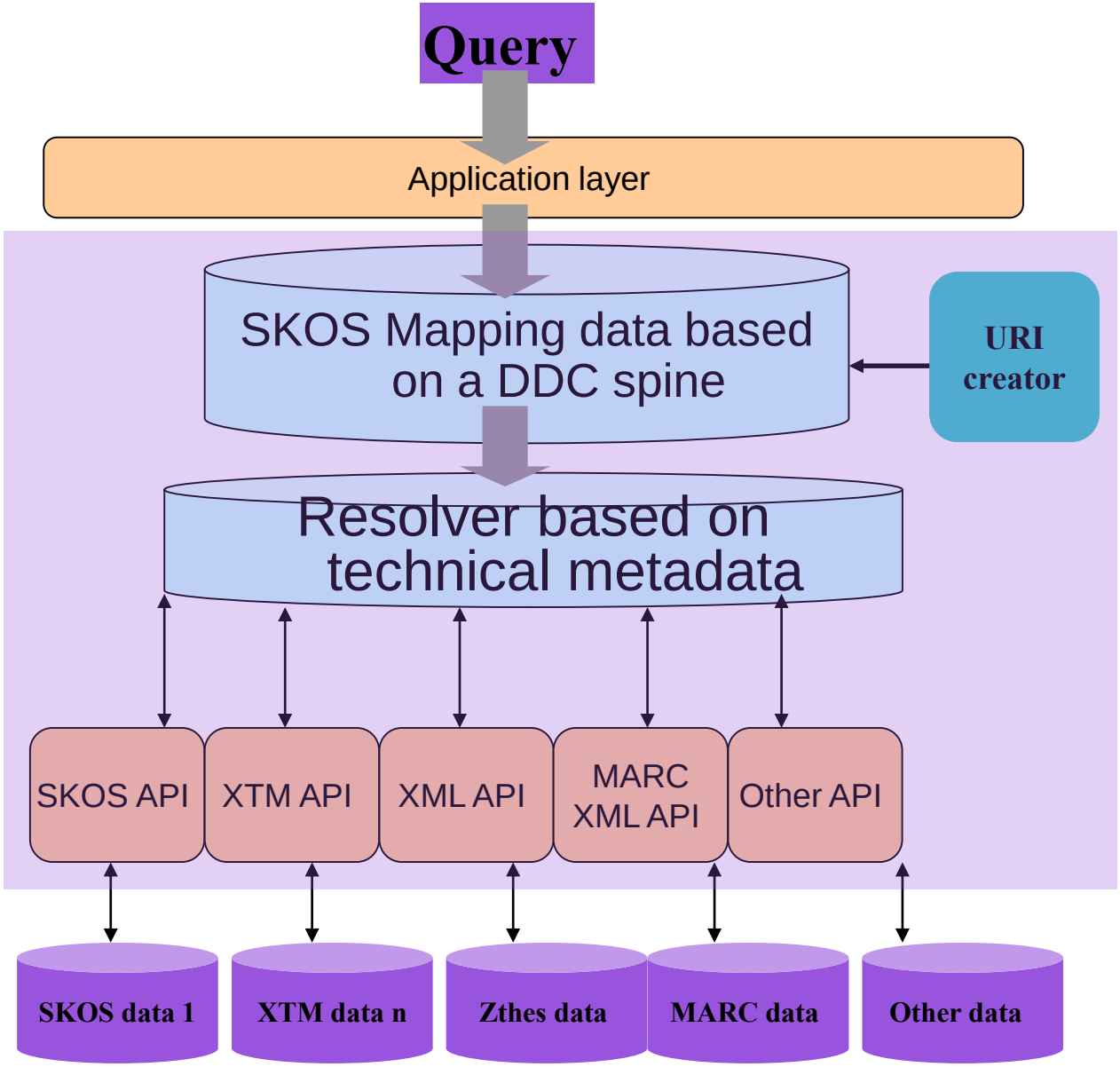
3. In the KOS community, there is continuing argument about whether to apply term-based or concept-based representation formats to encode the KOS. Most term-based encoding formats are designated to represent thesauri where the basic description element is based on terms. However, end-users may prefer to use different KOS as knowledge navigators, which emphasises the need to group relevant terms into a concept and represent a tree of the concepts to the users. Thus, it is important to develop a variety of algorithms and applications to encode KOS in both term-based and concept-based forms. An in-depth usability study on the use of subject access services based on KOS is required.
4. Different representation formats will co-exist for a long time, and there are a number of protocols and applications available to support access to encoded data in different formats.

Thus, when developing a terminology mapping service, it is hoped that different formats and protocols can be applied together to improve interoperability between different KOS in different formats.

Data conversion model



- A terminology mapping system is proposed to support multiple KOS format and protocols which is based on a knowledgebase.

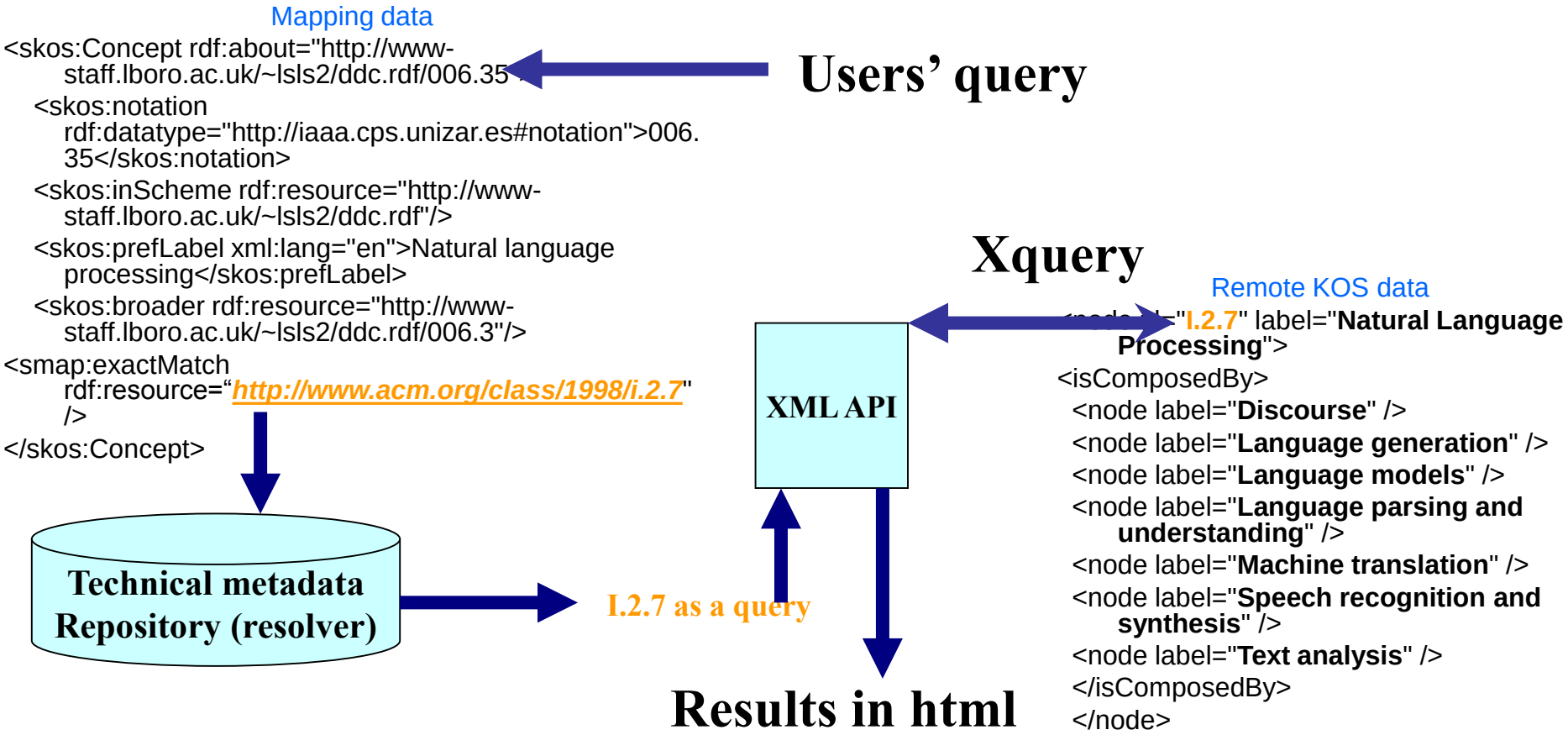


Application layer

KOS merging management layer

Terminological resource layer

The process of URI resolver



Advantages of knowledge base model

- Do not need to create a lot of XSL files to convert the data, so avoid the terminological data loss;
- Different APIs are applied to maximise the use of different KOS;
- The KOS owners do not need to put their KOS into a centralised database.

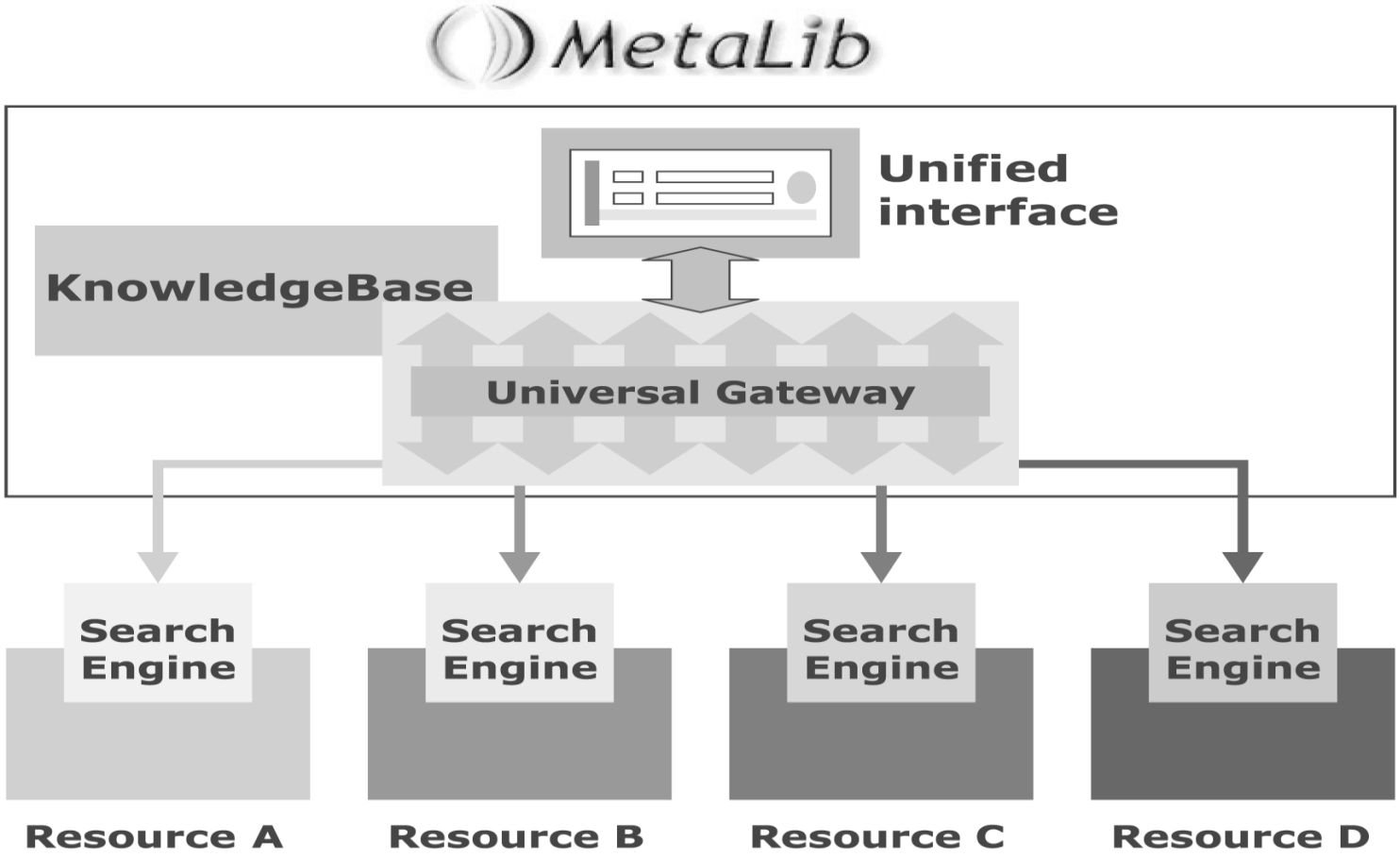
Questions?

- Thank you very much!
- l.si@lboro.ac.uk

Methods of establishing mappings

Methods of mapping KOS from [1]	Methods of Mapping metadata from [2] and [3]
Derivation/modelling	Derivation
Satellite and leaf node linking	Application profile
Direct mapping	Crosswalk
Co-occurrence mapping through metadata records	Co-occurrence mapping through subject terms in KOS
Merging	Metadata framework
Switch language	Switch-across
Fairly thinking about extending the methods to develop the KOS mapping service in the level of record and repository? **For example, JISC is conducting some research project on the development of KOS registry.**	Metadata registry
	Conversion of metadata records
	Data reuse and integration
	A metadata repository based on OAI-PMH
	A metadata repository supporting multiple formats without conversion
	Aggregation
	Value-based mapping based for cross-searching
	Element-based and value-based crosswalking services

A case study: MetaLib's Knowledge Base



- The MARC 21 Authority Format is applied to code common controlled vocabulary elements, such as preferred and non-preferred terms, term relationships, term mappings, the source of the content and the origin of changes.

Access steps

1. The users input some queries to the applications;
2. The users' query will access relevant DDC concepts in the SKOS mapping data, and then a range of concept URIs for other KOS are found;
3. Theses URIs will be resolved by the resolver, and then the resolver will convert the URIs to become appropriate queries for relevant APIs;
4. Different APIs will use converted queries to access different KOS in different formats, and get the results.
5. The final results from different KOS will be converted in a consistent format to present to the users.

The structure of the knowledge base

1. SKOS mapping data:
    1. Different KOS are mapped (manually or automatically) to a DDC spine;
    2. Use SKOS-Mapping to represent the mapping work;
    3. Give all the concepts from different KOS a URI as the identifiers of the concept, although in some less-well developed KOS, they may not use URI as identifiers.
2. A resolver to convert the URIs to appropriate queries for different KOS:
    1. The type of protocols that the remote KOS support;
    2. The encoding formats that the remote KOS use;
    3. The formats of results that are retrieved;
3. Different APIs are employed to manipulate different KOS in different formats.