A Theory of Retrieval Using Structured Vocabularies

(SKOS: Preparation for Standardization)

Alistair Miles
CCLRC Rutherford Appleton Laboratory

NKOS Workshop, September 2006, Alicante

What Am I Presenting?

- A formal theory of retrieval using structured vocabularies.
- The main body of my masters dissertation, which is entitled “**Retrieval and the Semantic Web**”.
- N.B. This presentation is intended to give an overview, for the full text go to ...

purl.org/net/retrieval

Why?

- How do you **maximize the utility** and **minimize the cost** of **vocabulary control** ... ?
- Support standardization initiatives ...
  - SKOS to W3C Recommendation,
  - BS 8723 parts 3, 4 and 5.
- Check our working assumptions!

- See also “**SKOS: Requirements for Standardization**” to be presented at DC 2006.

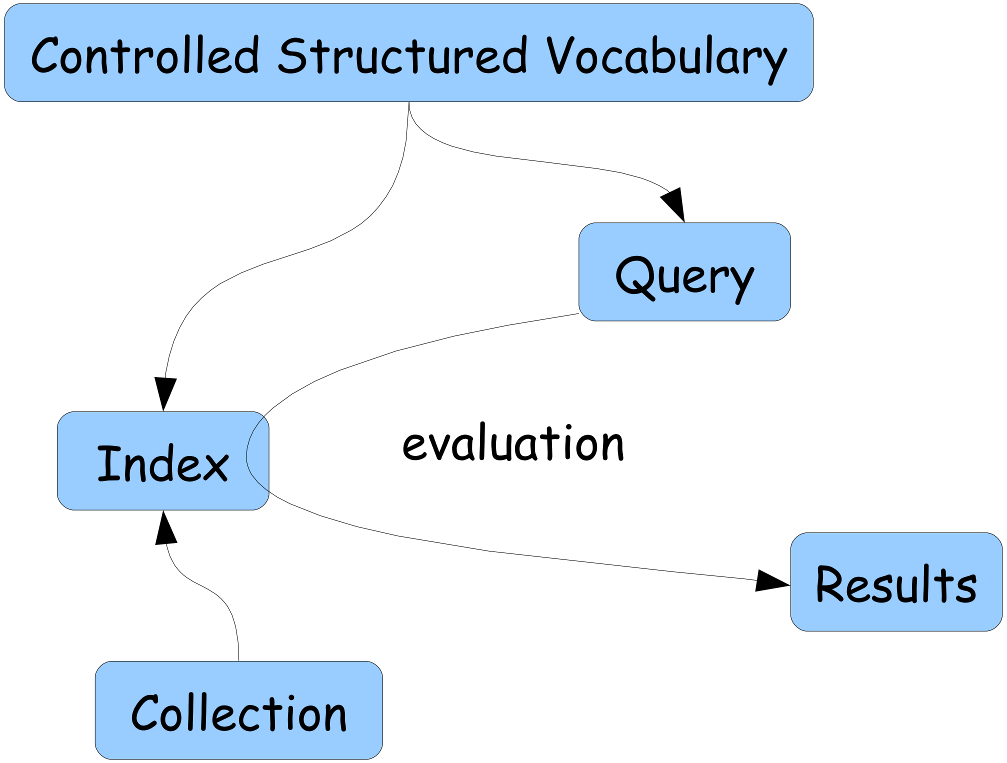
How?

- Use a **formal notation** (“Z”) to express underlying ideas with mathematical precision.
- Support formal specification with **explanatory prose**.
- N.B. This presentation is strictly **informal**!

Overview of the Theory

- Foundations (Chapter 3)
- Composite Queries (Chapter 4)
- Limited Cost Expansion (Chapter 5)
- Coordination (Chapter 6)
- Translation (Chapter 7)

General Scenario (1)



Controlled Structured Vocabulary

Query

Index

evaluation

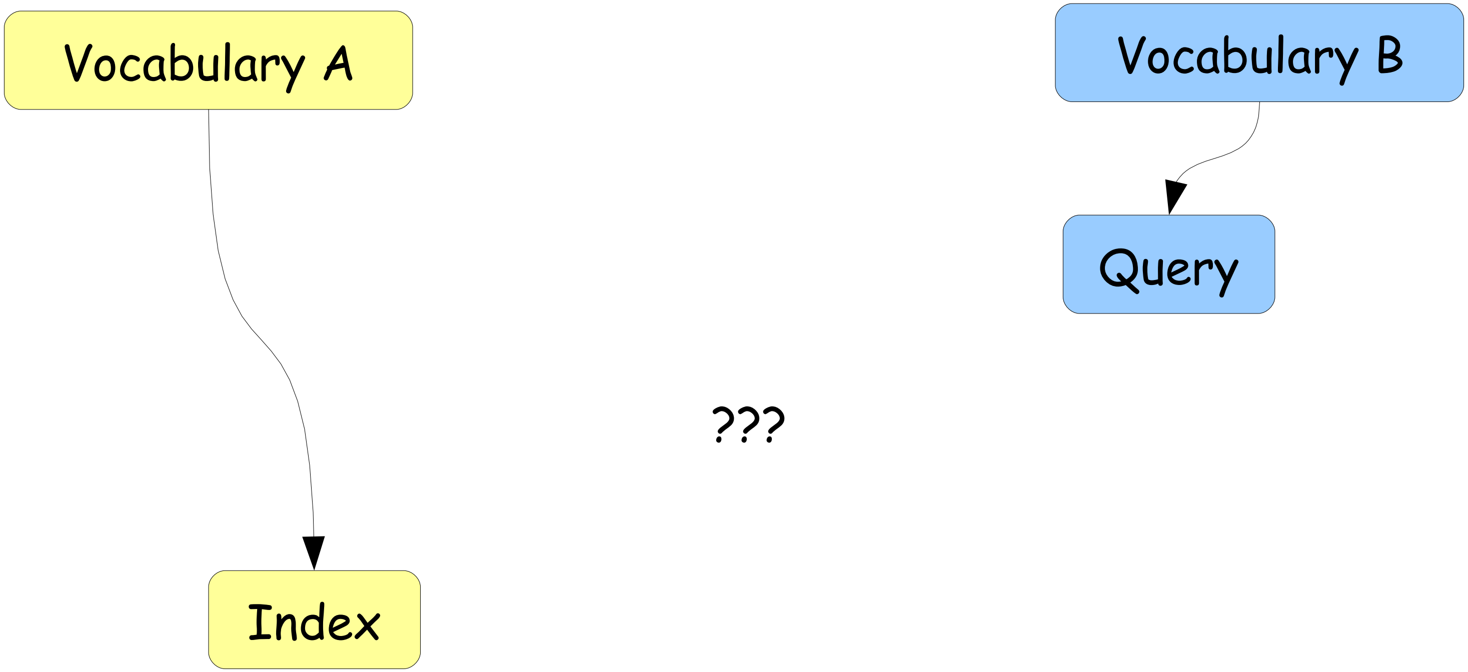
Results

Collection

Overview of the Theory

- Foundations (Chapter 3)
- Composite Queries (Chapter 4)
- Limited Cost Expansion (Chapter 5)
- Coordination (Chapter 6)
- Translation (Chapter 7)

General Scenario (2)



purl.org/net/retrieval

Lightning Tour (1) – Foundations

- Structured vocabulary.
- Index.
- Atomic query.
- Direct evaluation (of atomic queries).
- Naïve expansion (of an index).

Lightning Tour (2) – Composite Queries

- Query expressions ...
  - “and”, “or”, “not”, “required-optional-prohibited”.
- Composition and decomposition of expressions.
- Direct evaluation (composite queries).
- Naïve expansion (of composite queries).
- Scoring and ranking of results.

Lightning Tour (3) – Limited Cost Expansion

- Beyond naïve expansion.
- Approximating numerical “relevance cost” of expansion.
- Limited cost expansion (of an index or query).
- Expansion weight and result scoring.

Lightning Tour (4) - Coordination

- Using vocabulary units in combination.
- Ordered and unordered coordination.
- Coordinated indexes and queries.
- Naïve expansion (of a coordinated index or query).
- Limited cost expansion (of a coordinated index or query).

Lightning Tour (5) – Translation

- Structural mapping.
- Query expression mapping.
- Naïve translation using a structural mapping.
- Naïve translation using a query expression mapping.
- Limited cost translation using a structural mapping.

Caveats

- Much of the prose was written in haste!
- I'm no mathematician or logician!
- My review of the literature is woefully incomplete!
- The chapter on RDF representations (chapter 8) is rather incomplete and at best only suggestive!
- Use cases need further development.

A Theory of Retrieval Using Structured Vocabularies

Foundations

Foundations – The Conceptual Basis of Controlled Vocabularies (1)

- The fundamental purpose of a controlled vocabulary is to **establish** a set of distinct meanings or “**concepts**” and to provide a means of **referring unambiguously** to those concepts.

Foundations – The Conceptual Basis of Controlled Vocabularies (2)

- I have modelled this means of reference as a set of “names”, which I have called “**concept names**”.
- A controlled vocabulary provides a set of “concept names” which constitutes an artificial language for use in constructing an “index”. (I.e. a **controlled indexing language**.)

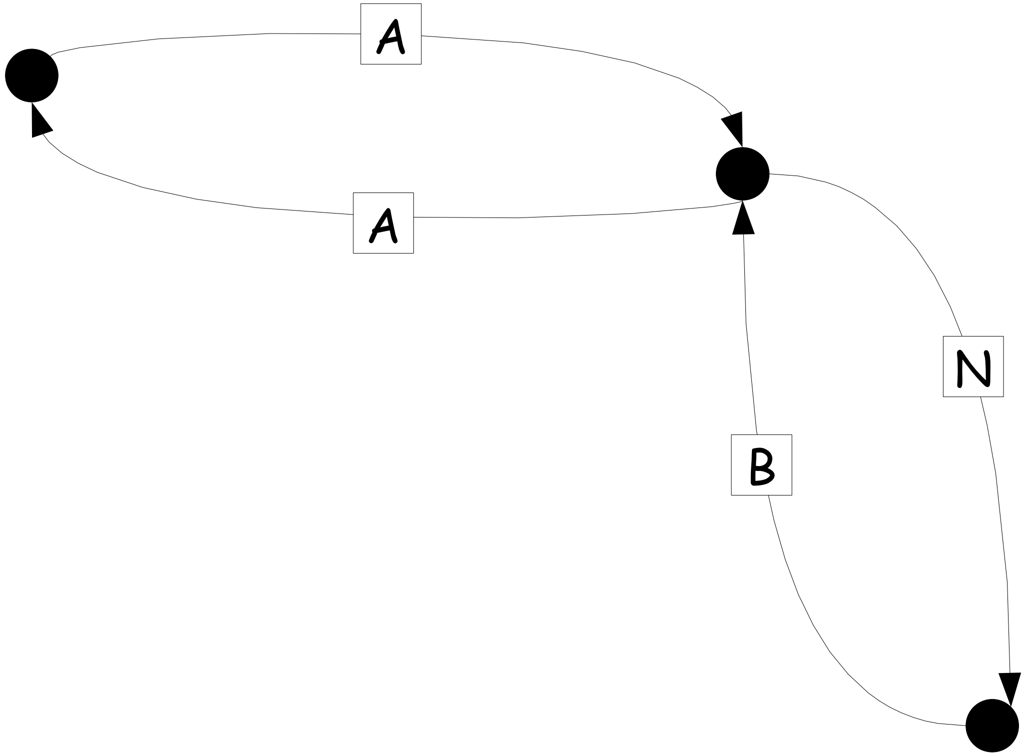
Foundations – Structure Relations (1)

- A controlled vocabulary may provide one or more binary relations on the set of concept names, which I refer to as “**structure relations**”.
- The structure relations of a controlled vocabulary together constitute the “**structure graph**”.

Foundations – Structure Relations (2)

- The theory considers only vocabularies that provide three structure relations, which I have called “**broader**”, “**narrower**” and “**associated**”.
- **N.B. No attempt is made to define** “broader”, “narrower” or “associated”!
- Their meaning is defined **entirely** in terms of **operational assumptions** that may be used to derive **retrieval operations**.

Foundations – A Structure Graph

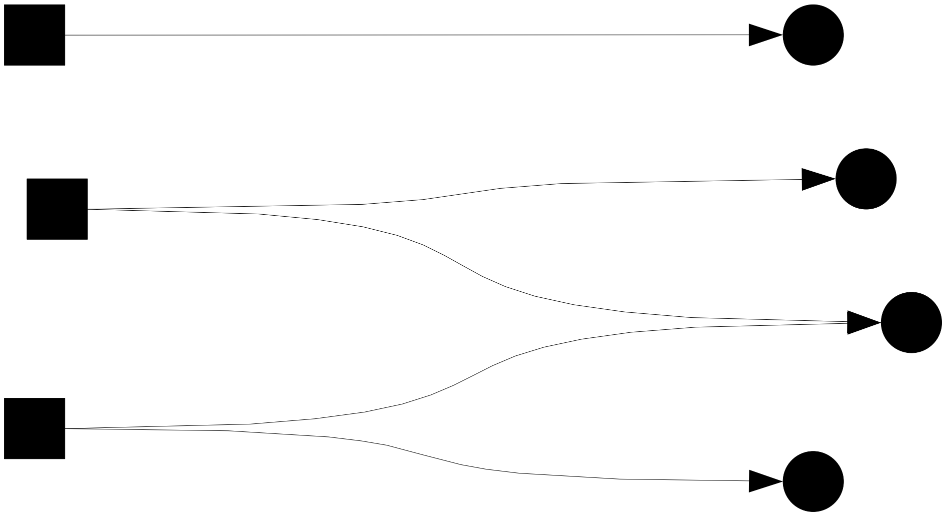


● CNAME

Foundations –
The Structure of an Index

- An “index” consists of one or more “fields”.
- A “field” is a binary relation between “document names” and “concept names”.
- (N.B. I use “document” to refer to any object we are interested in retrieving.)
- An index also provides a name for each field, so we can target particular fields in a query.

Foundations – A Field



■ DNAME

● CNAME

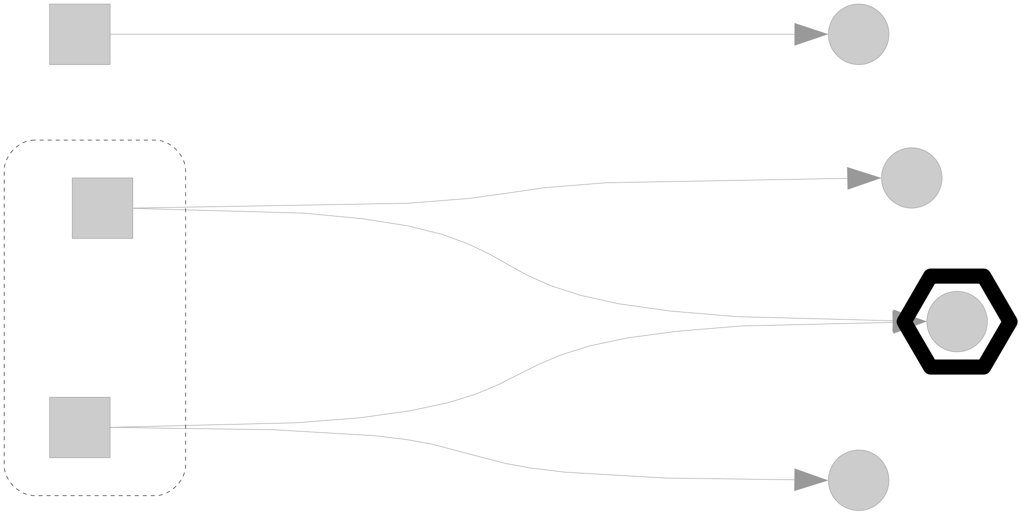
Foundations – Types of Index

- An index can have single or multiple fields.
- A field can be functional or relational.

Foundations – Atomic Queries

- An “atomic query expression” comprises a single field name and a single concept name.

Foundations – Direct Evaluation of Atomic Queries



⬡ QUERY

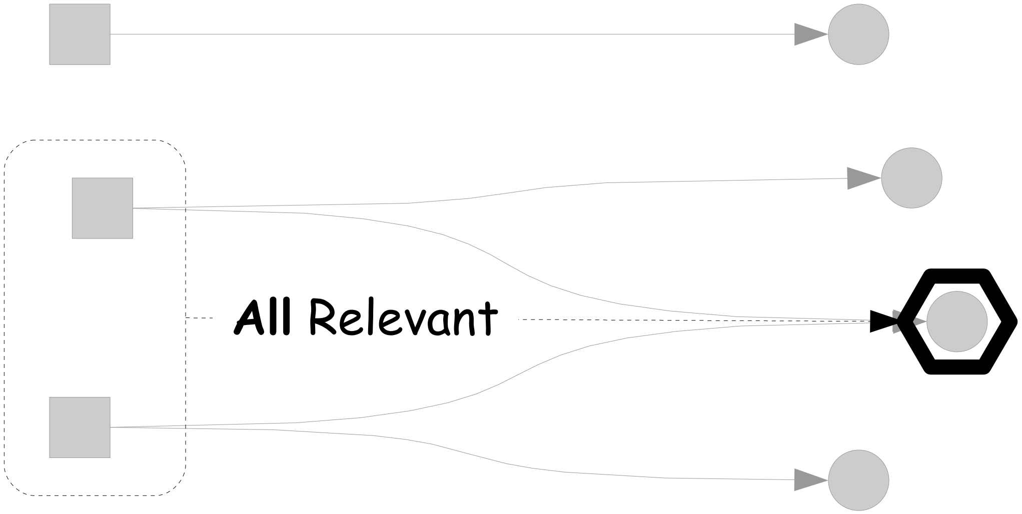
■ DNAME

● CNAME

Foundations –
Naïve Assumption of Ideal Indexing

- **All** documents indexed with a given concept name in a given field are **relevant** to an atomic query for that concept name in that field.

Foundations – Naïve Assumption of Ideal Indexing

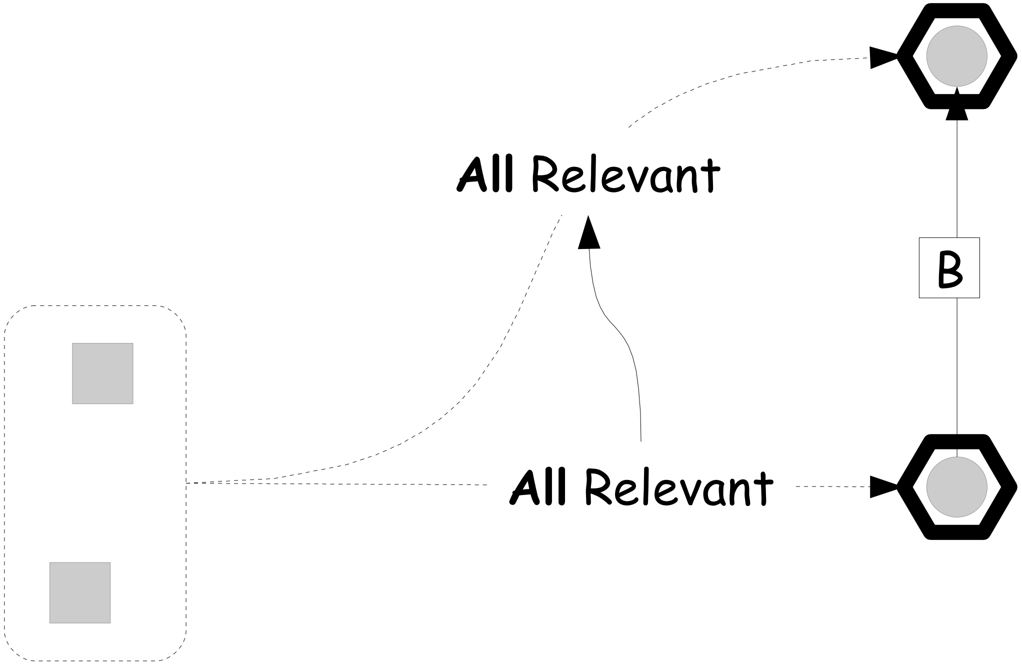


⬡ QUERY

■ DNAME

● CNAME

Foundations – Naïve Assumption of Broadening Relevance

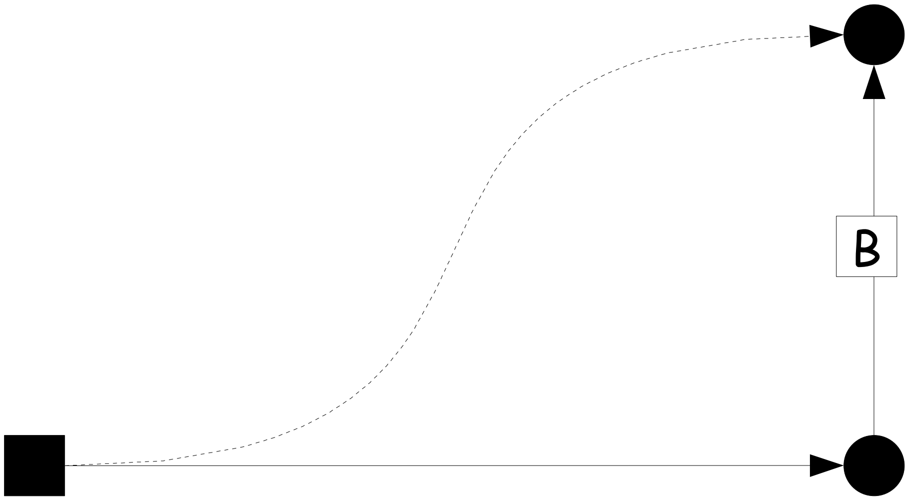


⬡ QUERY

■ DNAME

● CNAME

Foundations – Naïve Expansion of a Field



■ DNAME

● CNAME

purl.org/net/retrieval

Foundations – Naïve Expansion

- By including documents in a result set that are also relevant to the query, **recall is increased** at no cost to precision.

Foundations – Key Ideas

- Assumptions ...
  - Naïve assumption of ideal indexing.
  - Naïve assumption of broadening relevance.
- Operational definition for “broader/narrower”.
- Naïve expansion of an index to improve recall.

- N.B. This framework probably sufficient to cover the majority of applications!

A Theory of Retrieval Using Structured Vocabularies

Composite Queries

Composite Queries – Query Expressions (1)

- Composite query expression – has one or more “component” (or “child”) query expressions.
- Four types of composite query expression ...
  - *and*
  - *or*
  - *not*
  - *rop* (“required-optional-prohibited”)

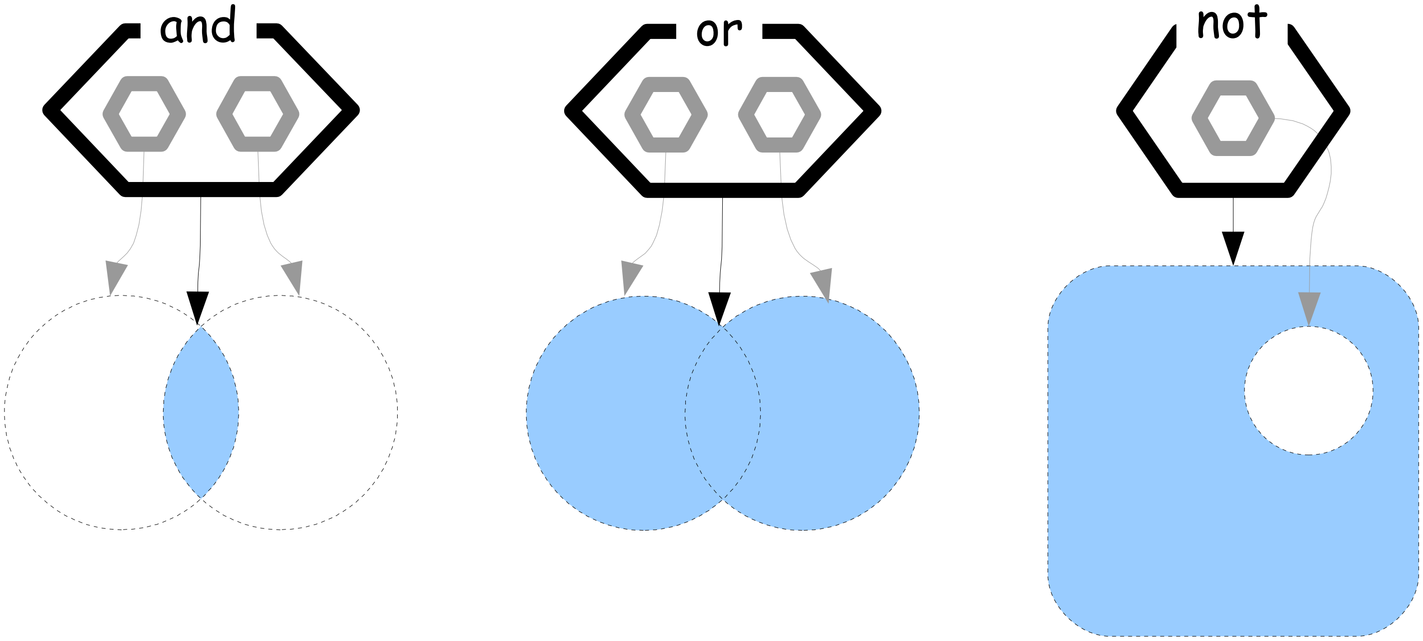
Composite Queries – Composition

- Child of a composite query expression can be an atomic expression or another composite expression.
- I.e. Expressions can be arbitrarily nested.

Composite Queries – Direct Evaluation

- Results of “**and**” expression ... set **intersection** of results of child expressions.
- Results of “**or**” expression ... set **union** of results of child expressions.
- Results of “**not**” expression ... set **complement** of results of child expression.
- Results of “**rop**” expression ... set **intersection** of results of “**required**” children minus set **union** of results of “**prohibited**” children ... N.B. “optional” children are truly optional.

Composite Queries – Direct Evaluation



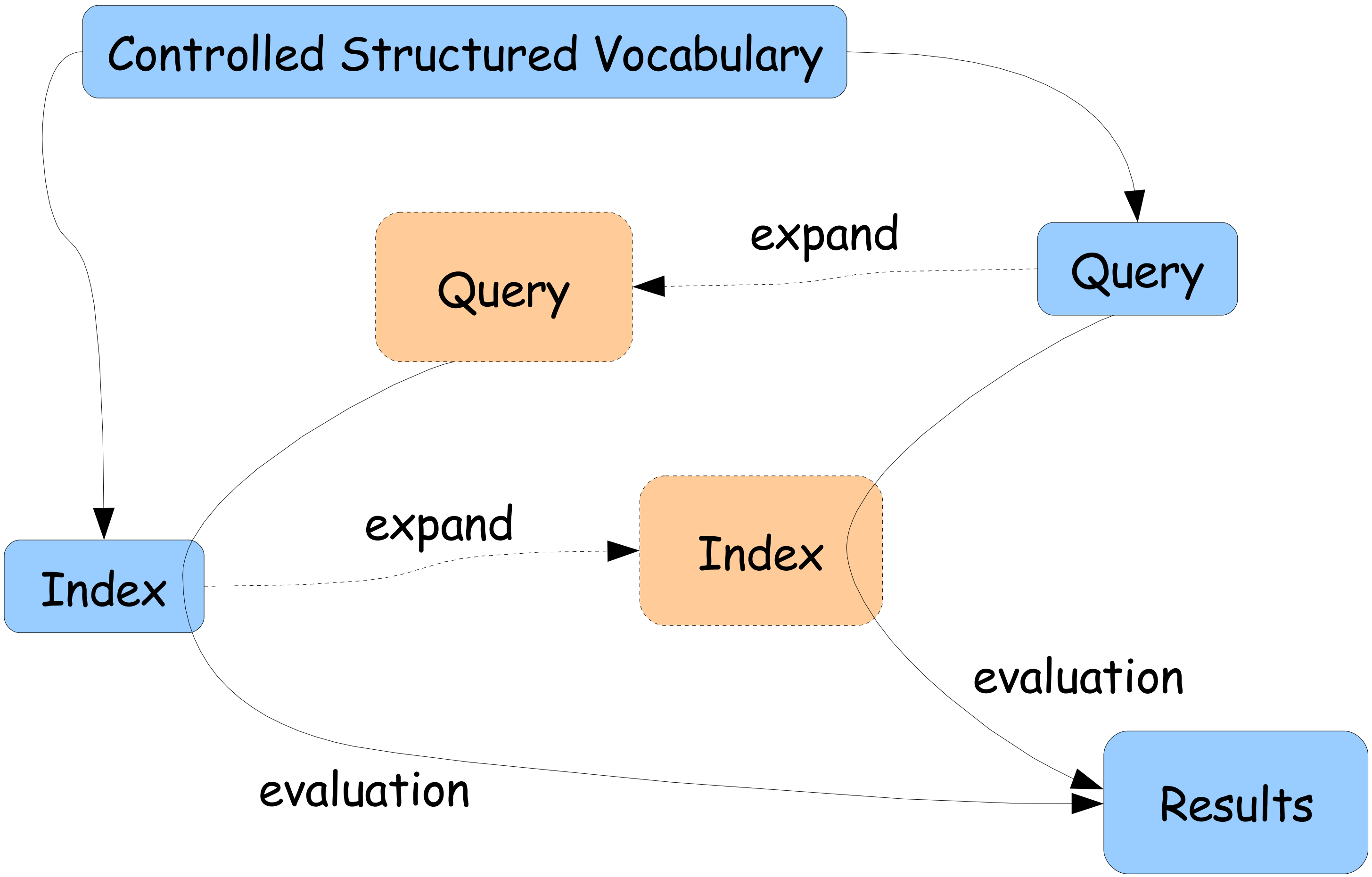
Composite Queries - Decomposition

- Decompose arbitrarily nested composite query into “**positive**” and “**negative**” atoms.

Composite Queries – Scoring Results

- Two metrics for scoring results of composite queries ...
  - **Unweighted** scoring (number of positive atoms matching the document).
  - **IDF weighted** scoring (take into account inverse document frequency of concept names in the index – greater weight to more "**discriminating**" atoms).
- Use scores to **rank** results (we assume in order of greatest relevance).

Composite Queries – Naïve Query/Index Expansion



Composite Queries – Naïve Query Expansion

- Expand arbitrarily nested query expressions.
- Mathematically equivalent to naïve index expansion (but not computationally equivalent).

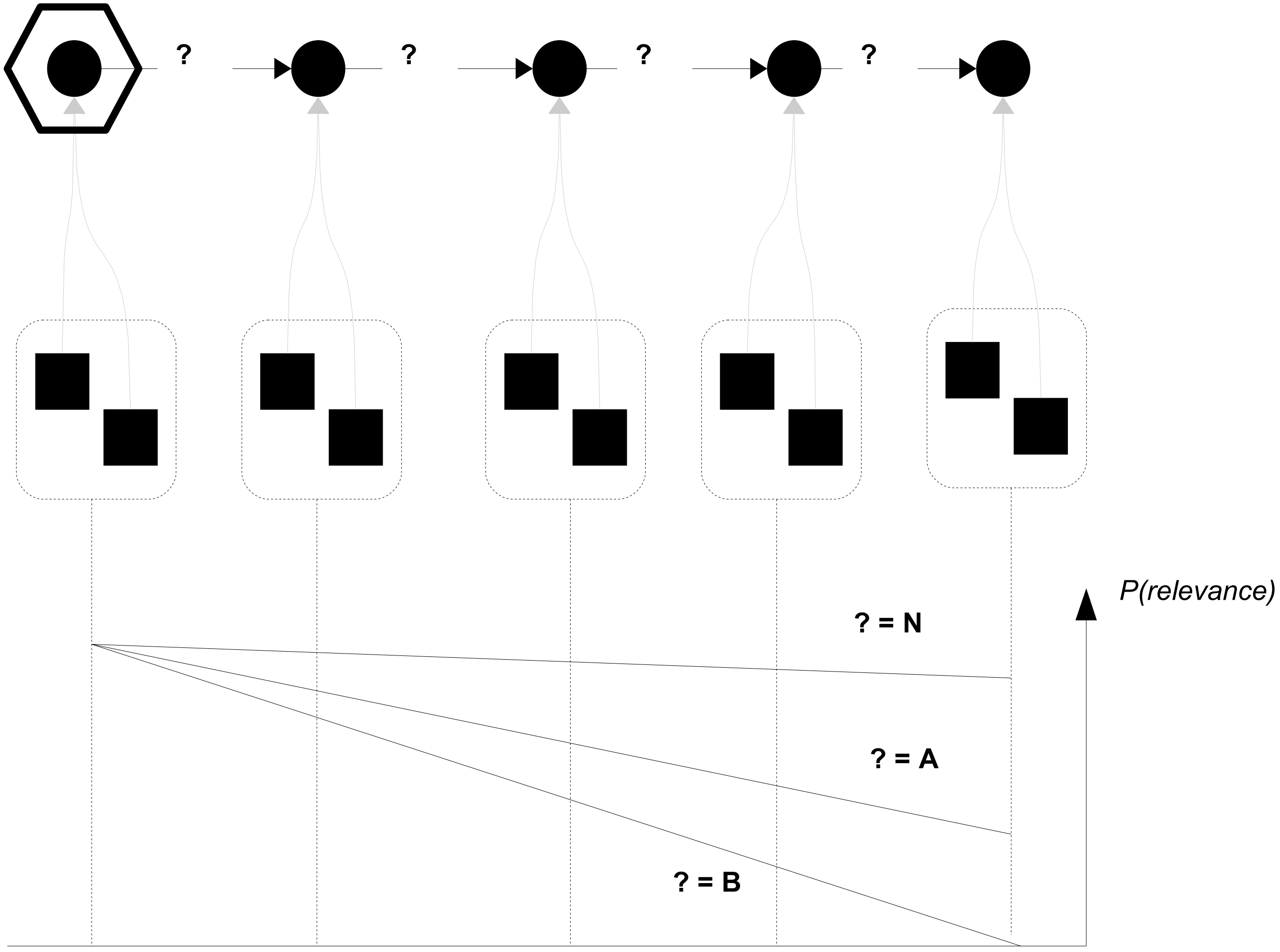
A Theory of Retrieval Using Structured Vocabularies

Limited Cost Expansion

Limited Cost Expansion – Naïve Assumptions

- Likely to break down, especially for “deep” hierarchies (does not account for **specificity**).
- Does not take advantage of **associative** links.
- Expansion cannot be “tuned”, no possibility for dynamic functionality (“all or nothing”).
- Structure is not utilised for ranking of expanded result set.

Limited Cost Expansion – Quantitative Assumptions



⬡ QUERY

■ DNAME

● CNAME

Limited Cost Expansion – Relevance Cost

- Use a numerical function to model the accumulated “**relevance cost**” of expansion.
- Use a “**cost limit**” to provide a cut-off.
- Invert the minimum cost value to obtain an “**expansion weight**” between 0 and 1 (high weight suggests high probability of relevance).
- Factor expansion weight into result scoring and therefore **ranking**.

Limited Cost Expansion – Query/Index Expansion

- Limited cost expansion of either query or index.

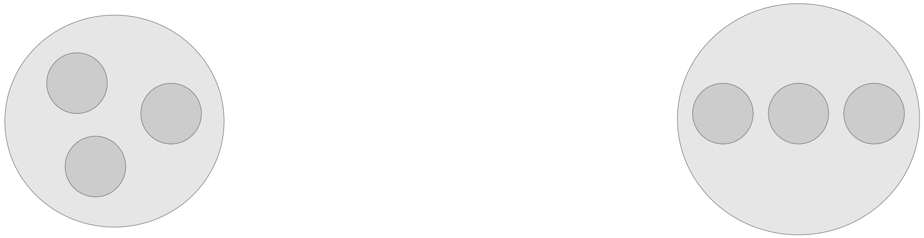
A Theory of Retrieval Using Structured Vocabularies

Coordination

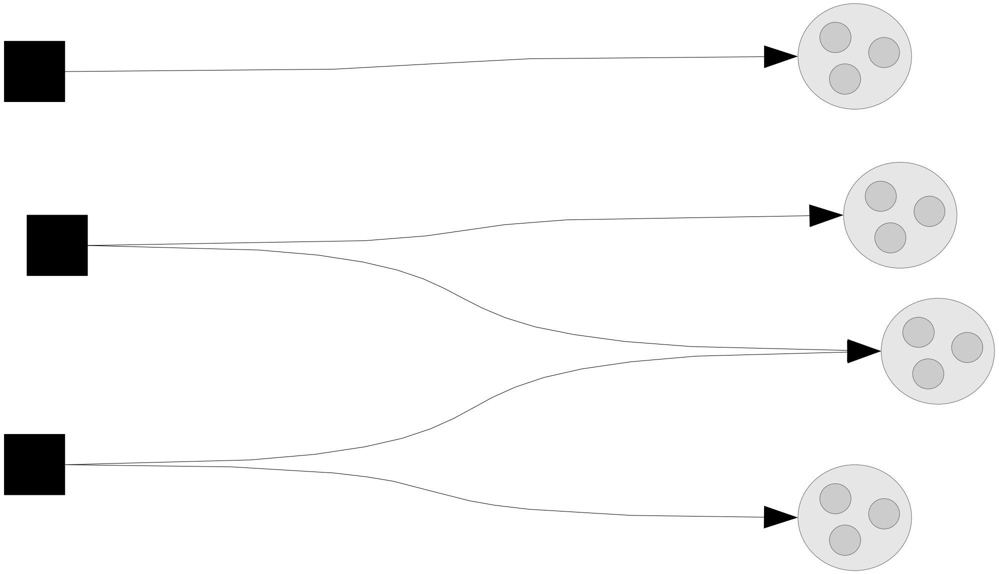
Coordination – Ordered and Unordered (1)

- Coordination is the act of combining concept names.
- **Ordered** – order of coordination **is** significant to meaning.
- **Unordered** – order of coordination **is not** significant to meaning.

Coordination – Ordered and Unordered (2)



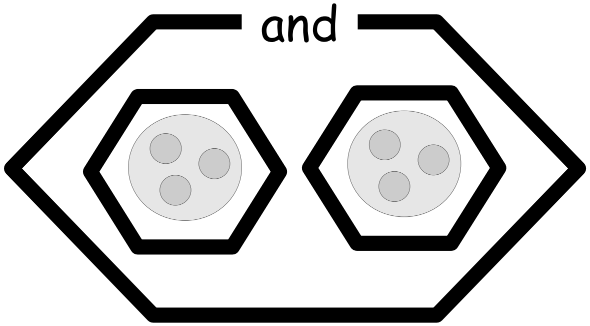
Coordination – A Coordinated Field



■ DNAME

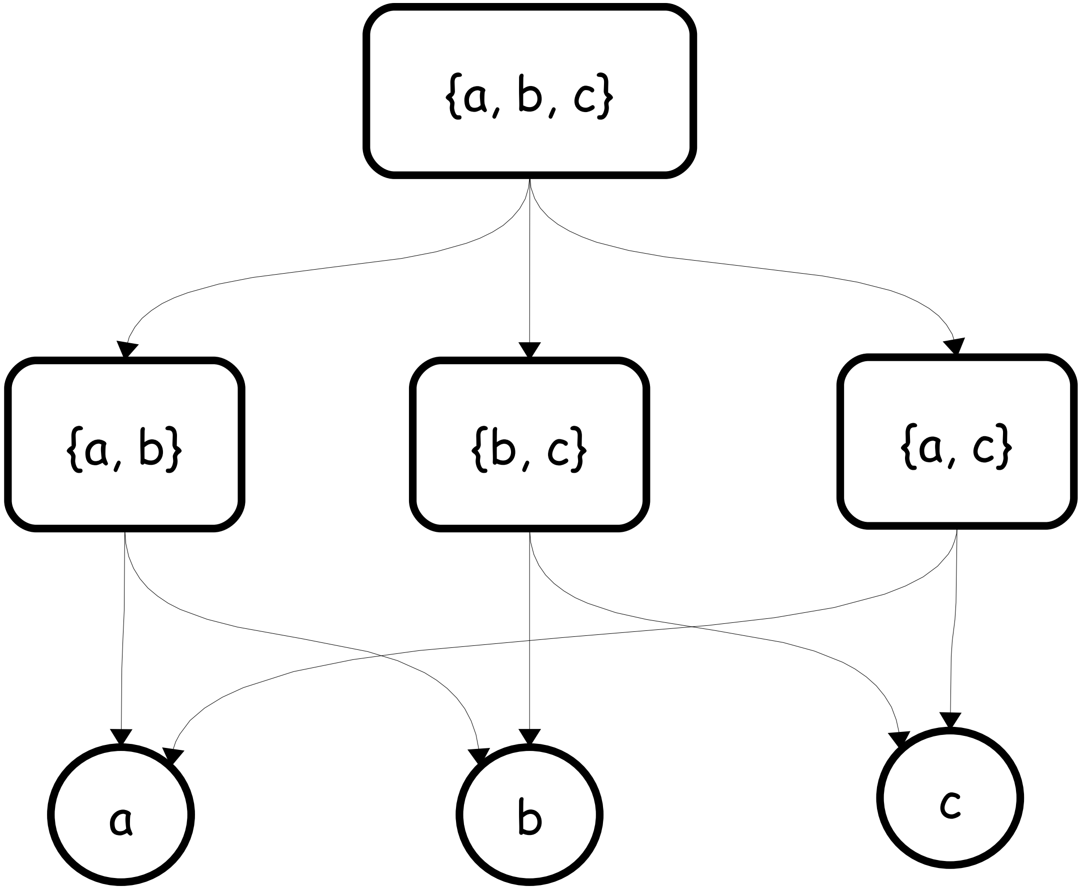
● CNAME

Coordination – A Coordinated Query

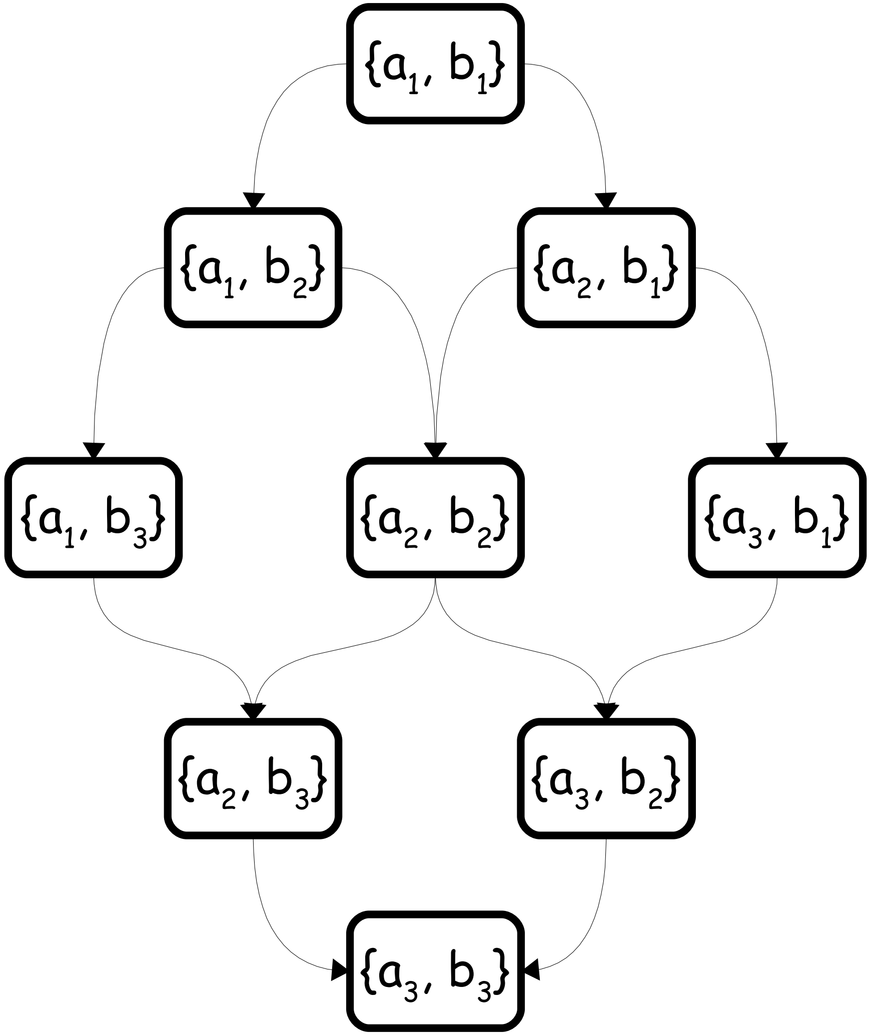


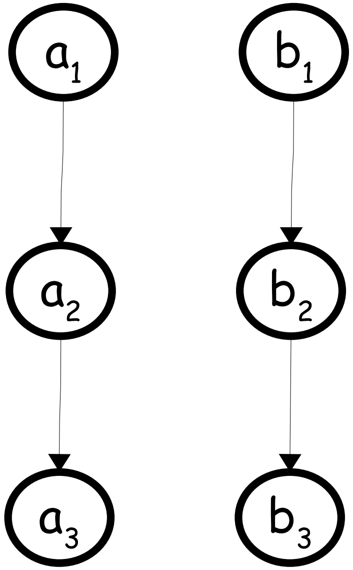
purl.org/net/retrieval

Coordination - Decomposition



Coordination – Structure Relations





purl.org/net/retrieval

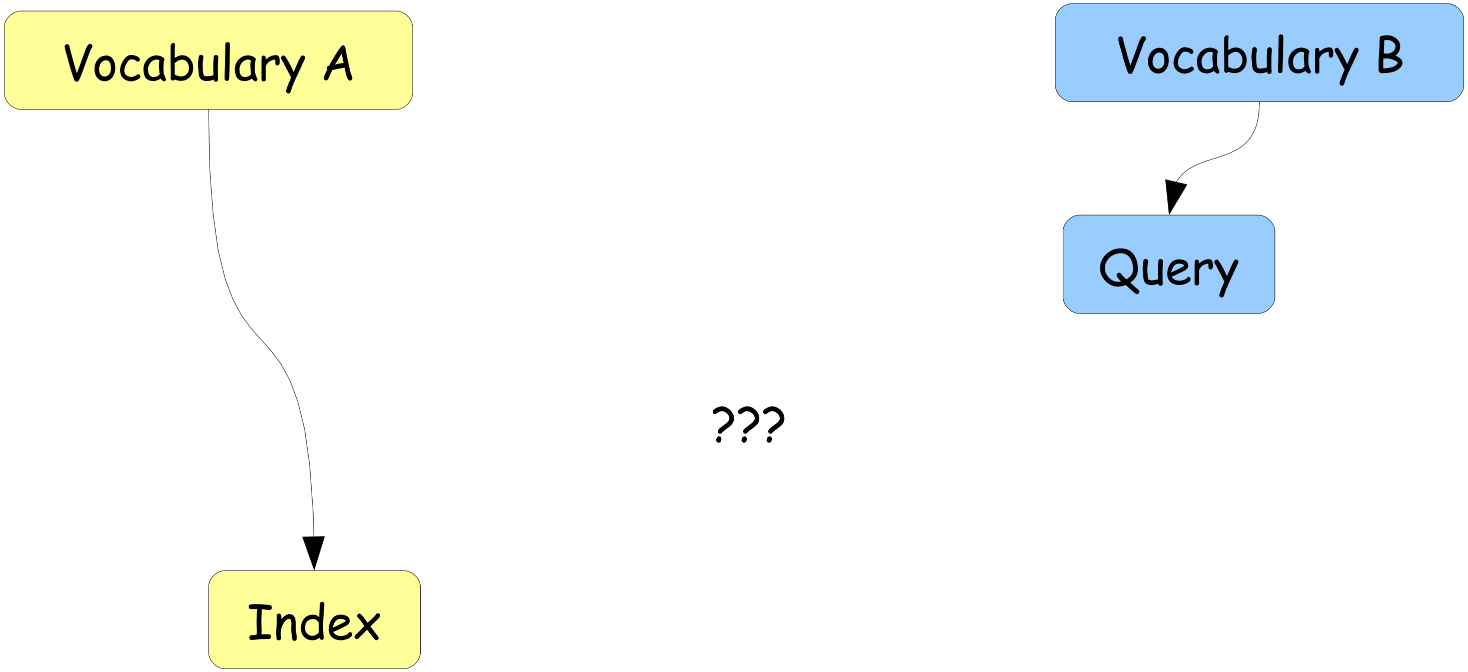
Coordination - Expansion

- Naïve expansion of coordinated queries or indexes.
- Limited cost expansion of coordinated queries or indexes.

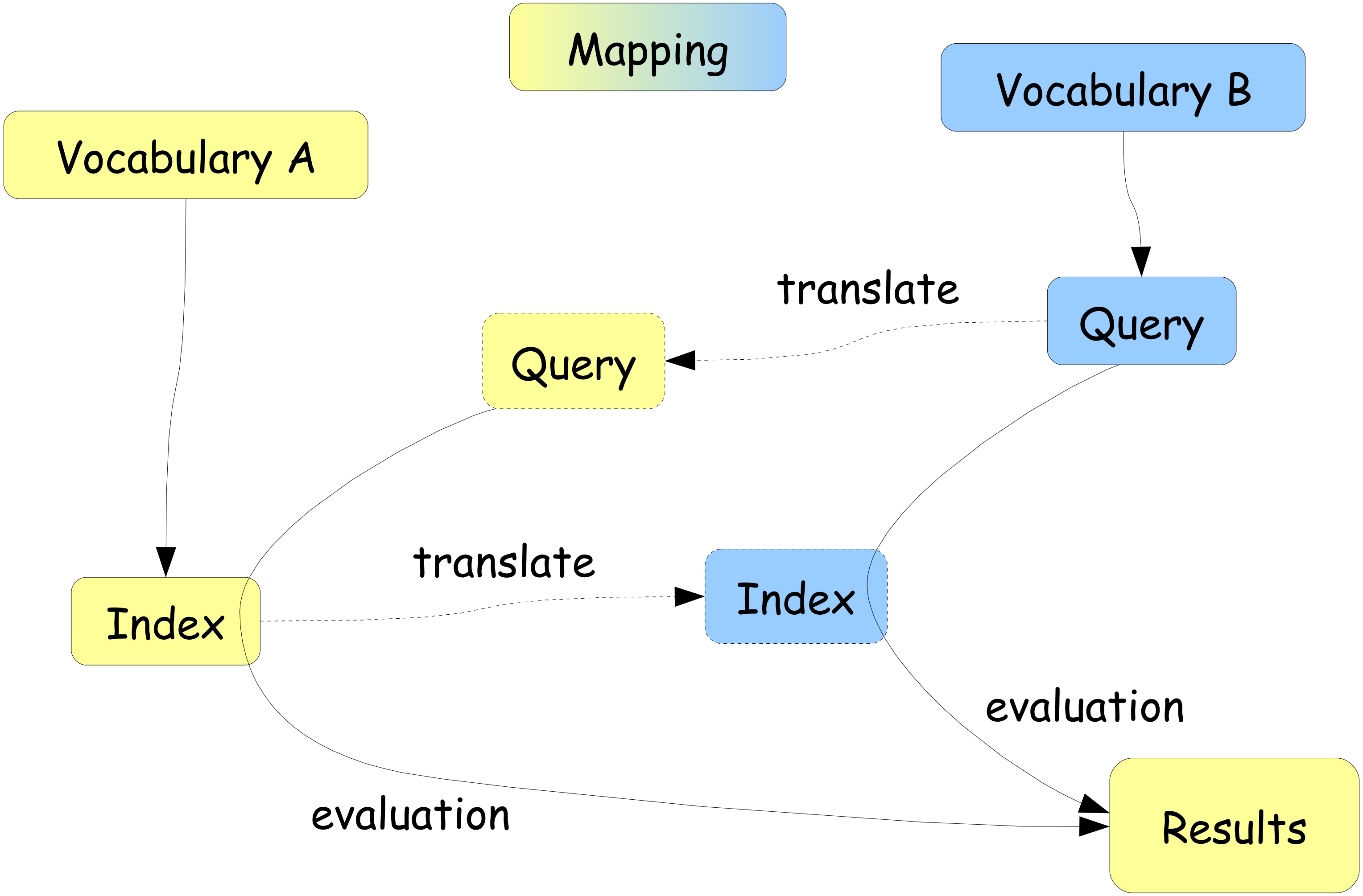
A Theory of Retrieval Using Structured Vocabularies

Translation

General Scenario (2)



Translation



Translation - Goals

- Automated translation.
- Understand consequences for precision and recall.
- Minimise loss of precision and recall.

Translation – Mapping

- Structural mapping ...
  - Use “broader”, “narrower”, “associated” and “equivalent” mapping relations.
- Query expression mapping ...
  - Use composite query expression as the target of the mapping.

Translation – Methods

- Naïve translation.
- Limited cost translation ...
  - Translation weight.

- N.B. Limited cost translation is much less demanding on the completeness of the mapping!

A Theory of Retrieval Using Structured Vocabularies

Next Steps ...

Adaptation and Change

- Use mappings to express change in vocabularies.
- Use translations to adapt indexes and/or queries.

- N.B. Requires vocabulary management tools that capture change information at the point of change!

Summary

- Pragmatic, operational approach to describing the use of structured vocabularies for retrieval.
- Formalise the underlying assumptions.
- Support standardization, especially of representations for index, vocabulary and mapping data.