

Tutorial – Vocabularies

Alistair Miles

<http://purl.org/net/aliman>

e-Science Centre

STFC Rutherford Appleton Laboratory

International Conference on Dublin Core and
Metadata Applications

27-31 August 2007, Singapore



Who am I?

- Research Associate
- STFC e-Science Centre
 - <http://www.e-science.stfc.ac.uk/>
 - *“Facilities such as synchrotrons, satellites, telescopes and lasers, collectively generate many terabytes of data every day. Their users require efficient access to geographically distributed leading edge data storage, computational and network resources in order to manage and analyse these data in a timely and cost effective way. e-Science builds the infrastructure which delivers this.”*



Information Management Group

- The group provides leading R&D with two major themes:
 - Distributed systems, especially web and grid middleware, trust and security modelling and deployment
 - ***Information Systems especially data and institutional repositories, controlled vocabulary management, data curation and the Semantic Web***



I fit in here!



Selected Contributions

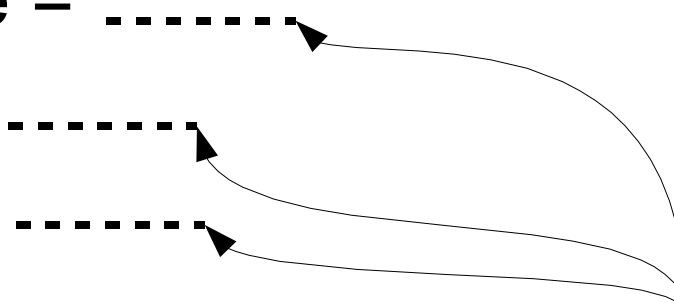
- SWAD-Europe
- W3C Semantic Web Best Practices WG
 - SKOS
 - Publishing RDF
- W3C Semantic Web Deployment WG
- CLADDIER
- STFC ePublications Archive (ePubs)
- Digital preservation / curation
 - CASPAR
 - DCC



What is a “vocabulary”?

- Loose definition ...
- “Something you can use on the RHS of the metadata equation.”

– Language =
– Format =
– Subject =



“values” from a “vocabulary” can go here



Scope of this tutorial?

- “Vocabularies” ... ?
- This tutorial: ***controlled structured vocabularies for retrieval***
- (retrieval = finding things)

- N.B. In this tutorial, when I say “vocabulary”, I mean “controlled structured vocabulary”



Brief Look Back

- DC-2004 Shanghai
 - Sutton & Tennis: ***Creating and Managing Controlled Vocabularies for Use in Metadata***
- DC-2005 Madrid
 - Ron Daniel Jr.: ***Controlled Vocabularies and the Dublin Core***
 - Alistair Miles: ***SKOS Core Tutorial***
- DC-2006 Manzanillo
 - Joseph Tennis: ***Vocabularies (In the Networked Environment)***



DC-2004 Tutorial

- ***Creating and Managing Controlled Vocabularies for Use in Metadata***
- Stuart Sutton & Joseph Tennis
- <http://dc2004.library.sh.cn/english/prog/ppt/tutorial4.ppt>
 - Vocabulary development
 - “Webized” vocabularies



DC-2005 Tutorial

- ***Controlled Vocabularies and the Dublin Core***
- Ron Daniel, Jr.
- http://dc2005.uc3m.es/program/tutorials/tutorial3_eng.ppt
 - Business context
 - Using vocabularies in DC elements
 - Factoring “Subject” into facets
 - Maintaining vocabularies



DC-2005 Tutorial

- ***SKOS Core Tutorial***
- Alistair Miles
- http://dc2005.uc3m.es/program/tutorials/tutorial4_eng.ppt
 - SKOS development and status
 - SKOS main features
 - Extending SKOS



DC-2006 Tutorial

- ***Vocabularies (in the Networked Environment)***
- Joseph Tennis
- <http://dc2006.ucof.mx/papers/JuevesJosephTennis.ppt>
 - Vocabulary development
 - Enabling reuse in a networked environment
 - (cf. “webize”)



This Tutorial

- Address some important themes from previous tutorials
 - ***Business context***
 - ***Networked environment***
- Say something new
 - ***Using vocabularies for retrieval***
- N.B. No time for lots of important themes
 - Vocabulary development, maintenance, change management...



Tutorial Outline

1. Informal vocabulary model
2. Business context
3. Networked environment
4. Using vocabularies for retrieval
5. Using vocabulary mappings for retrieval
(across heterogeneous metadata)



Key References

- ISO 2788
- ISO 5964
- ANSI/NISO Z39.19-2005
- BS 8723 (parts 1-4)



Caveat

- I have more experience of theory, less of practice...



Topic 1 – Vocabulary Model

Simple, informal model of controlled structured vocabularies...



Informal Model

- What is a controlled structured vocabulary and how does it work?
- Present an informal model, using diagrams
- Capture what is common to...
 - thesaurus, taxonomy, classification scheme, subject heading system ...



Informal Model – Caveats

- N.B. This is modern, DCMI/SemWeb view, representing best practice, constrained by requirements of distributed metadata systems...
- ... ideal – not all vocabularies (or standards) can be aligned with this model!



Diagram Conventions

- I am going to use ***my own*** diagram conventions to describe the ***informal*** model!
- Not related to any formal system or standard!



Basic Provisions – Identifiers and Labels

- A vocabulary must at least provide...
 - A set of ***globally unique identifiers***, for use in distributed computer systems (machine-to-machine communication)
 - A set of ***human-readable labels***, for use by people (person-to-person communication)



Diagram Conventions – Identifiers and Labels

URI-Y

(represents a globally unique *identifier* –
i.e. a URI)

“anthropology”@en

(represents a *label* – combination of
UNICODE character string and
language tag)



Diagram Conventions – Vocabulary

“acidification”@en

“archery”@en

“birthdays”@en

“personality”@en

“oxygen”@en

URI-A

URI-C

URI-D

URI-B

URI-E

URI-F

URI-G



Identifiers

- ***Identifiers are for computers!***
 - Used in databases, metadata ...
- Only people developing software should ever see them
- Identifiers must be globally unique, for use in distributed metadata
- Best practice – URIs (IRIs?)
- One URI, one meaning
 - No ambiguity
 - No redundancy



Labels

- ***Labels are for people!***
- Drawn from natural language (one or more)
- Accurately and unambiguously convey the intended meanings (concepts)





“acidification”@en

“archery”@en

“birthdays”@en

“personality”@en

“oxygen”@en

URI-A

URI-C

URI-D

URI-B

URI-E

URI-F

URI-G

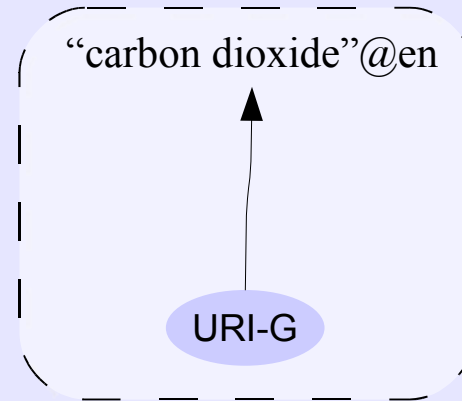
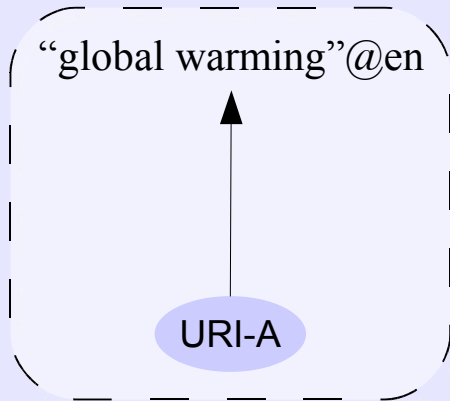


Conceptual Units

- Loose definition...
- A vocabulary can be divided into a set of “conceptual units”
- Each “conceptual unit” is intended to capture and convey a distinct meaning (concept).
- Cf. “node” (e.g. Daniel, 2005)



Diagram Conventions – Conceptual Units

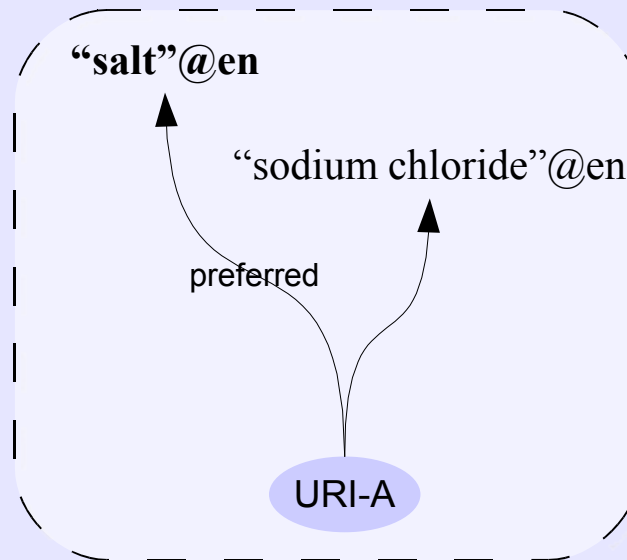


Preferred Labels

- Distinguished subset of labels which are “preferred”
 - One per conceptual unit (per language)
 - Provides non-redundant means of reference in person-to-person communication



Diagram Conventions – Preferred Labels

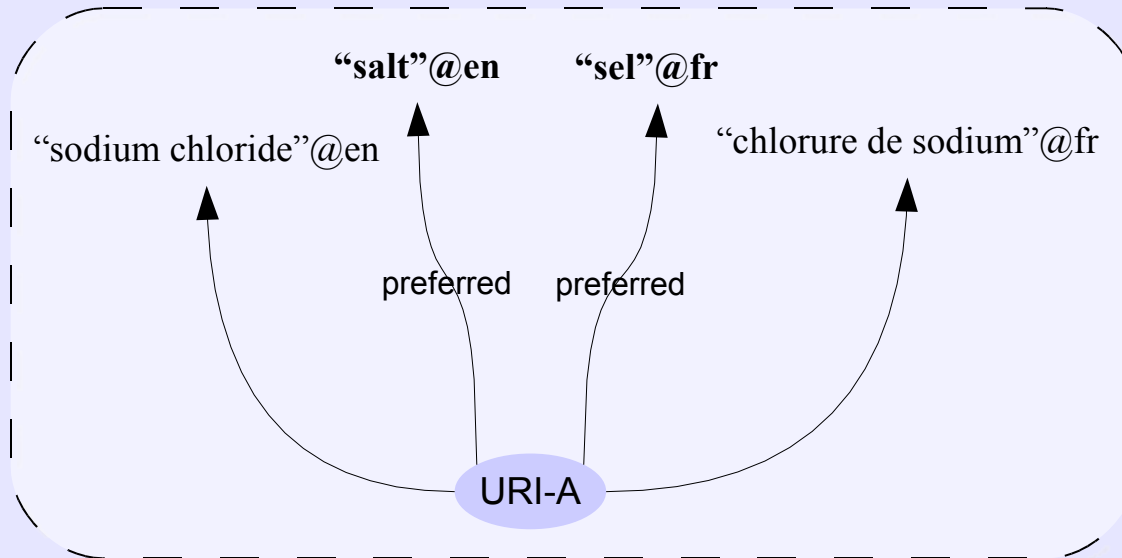


Multilingual vocabularies

- Provide labels in two or more languages
- One preferred label per conceptual unit per language
- Cf. BS 8723-4
 - “In a multilingual thesaurus, every concept is labelled by a descriptor in each of the languages.”



Diagram Conventions – Multilingual Labelling



Beware ISO 5964, BS 8723-4

- Standards for multilingual thesauri
- Process of constructing a multilingual CV
 - Resolve natural conceptual differences between language communities
 - Strategies for recognising and resolving different types of conceptual difference
- BUT goal is to end up with single, coherent set of conceptual units, across all languages provided.

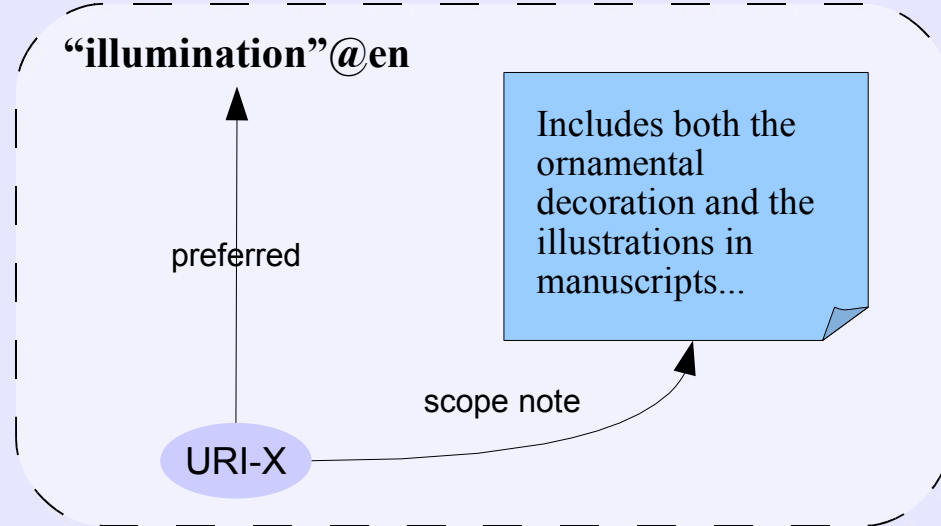


Notes

- Notes provide further useful information for human reader
- E.g. Clarify scope/meaning of conceptual unit
- E.g. Historical information



Diagram Conventions – Notes



Semantic Relationships (Links)

- Meaningful relationships between conceptual units
- a.k.a. “paradigmatic” relationships – inherent in meaning
- Give meaningful ***structure*** to a vocabulary
- Useful for navigation, browsing
- Can also be used “behind the scenes” to improve retrieval – see later in this tutorial



Semantic Relationships

- Two main types...
 - Hierarchical (broader/narrower)
 - Associative (related, see also)
- See ISO 2788, BS 8723-2, Z39.19 for guidance and explanation
- Later in this tutorial, I will focus on what these relationships mean from a retrieval point of view



Diagram Conventions – Hierarchical Relationships

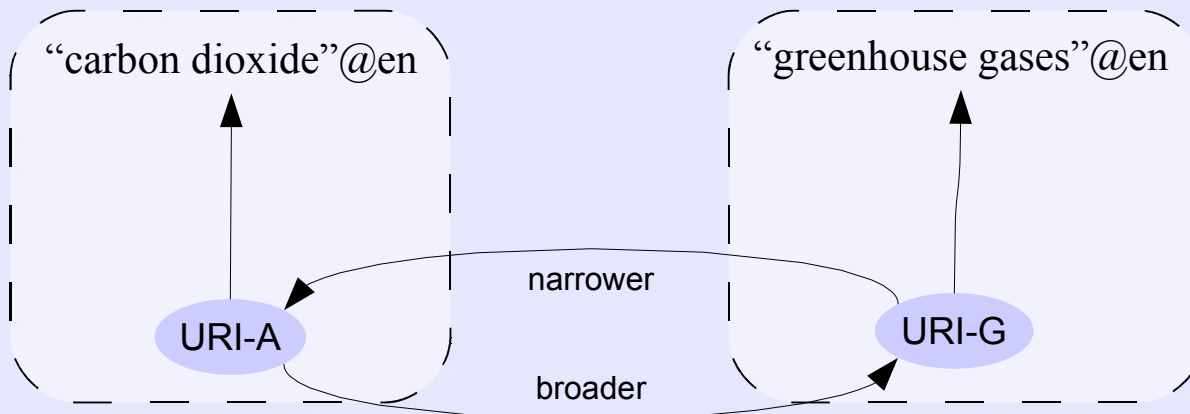
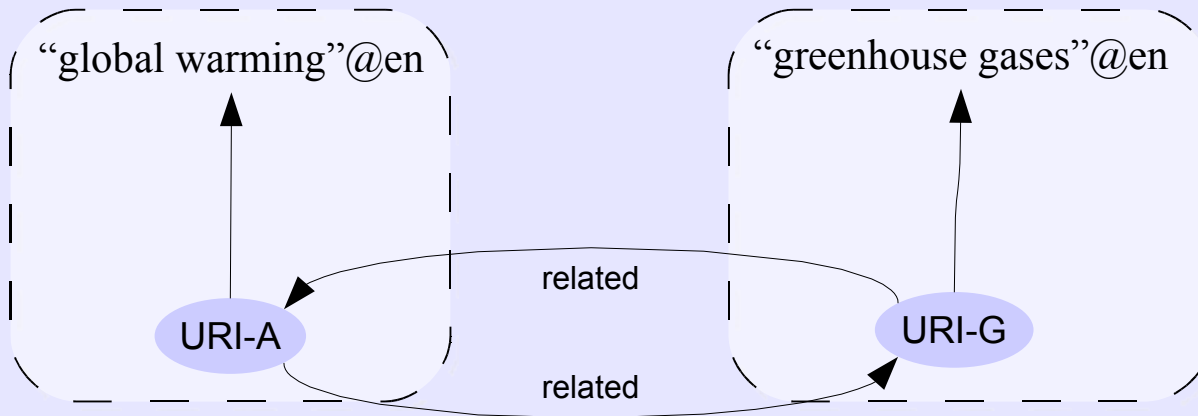


Diagram Conventions – Associative Relationships



Summary of Topic 1

- ***Informal vocabulary model***
 - ***Identifiers*** for computers, ***labels*** for people
 - Conceptual units
 - Preferred labels, multilingual labelling
 - Notes
 - Semantic relationships
 - Hierarchical and associative



Edge Cases

- The devil is in the detail!
- Lots of things don't fit into this simple model
 - Node labels; UF+; coordination ...
- But, we've got more than enough to start exploring design and implementation of retrieval systems...
- (And enough to raise lots of interesting issues :-)



Topic 2 – Business Context

The bigger picture ... How much will it cost?
Is it worth it?

ROI Problem

- Daniel 2005
- How are we going to use vocabularies, metadata, content in applications to obtain ***business benefits?***
 - What are the fundamental business problems we want to solve? I.e. business need?
 - How are we adding value?
 - Who are the “customers”?
 - How much value can we expect to add?
 - How much can we afford?



Alternatives

- If goal is improved functionality and performance of retrieval systems ...
- ... there are many ways to add value, other than by using a controlled vocabulary!
- A controlled vocabulary may simply not provide ROI
 - E.g. If no budget to do ongoing control and QA, a folksonomy/social tagging approach may be better (Daniel 2005)



Isolation and Integration

- Controlled vocabularies are hardly ever ***THE*** answer!
- More likely, part of ***integrated strategy***
 - Where content is textual – integrate with text retrieval
 - Where content is (hyper)linked – integrate with analysis of link topology
 - Where people are linked – integrate with social network systems



Strategy

- Strategy is important – how does vocabulary fit into integrated programme to solve business needs?
- N.B. Technical and business environment will define constraints ...



Tagging Problem

- Daniel 2005...
- How are we going to populate metadata elements with complete and consistent values?
 - What can we expect to get from automatic classifiers? Error detection, correction?
- One of many aspects of environment which ***influences where to focus effort!***



Early Returns?

- Full scale vocabulary project demands initial and ongoing investment, long term commitment
- Few organisations have the resources...
- ***Hard to argue the case!***
- Instead, can we add value in small steps?
 - Can we develop and exploit a fine-grained sequence of products, on the way to controlled structured vocabulary, ***getting a return early and often?***

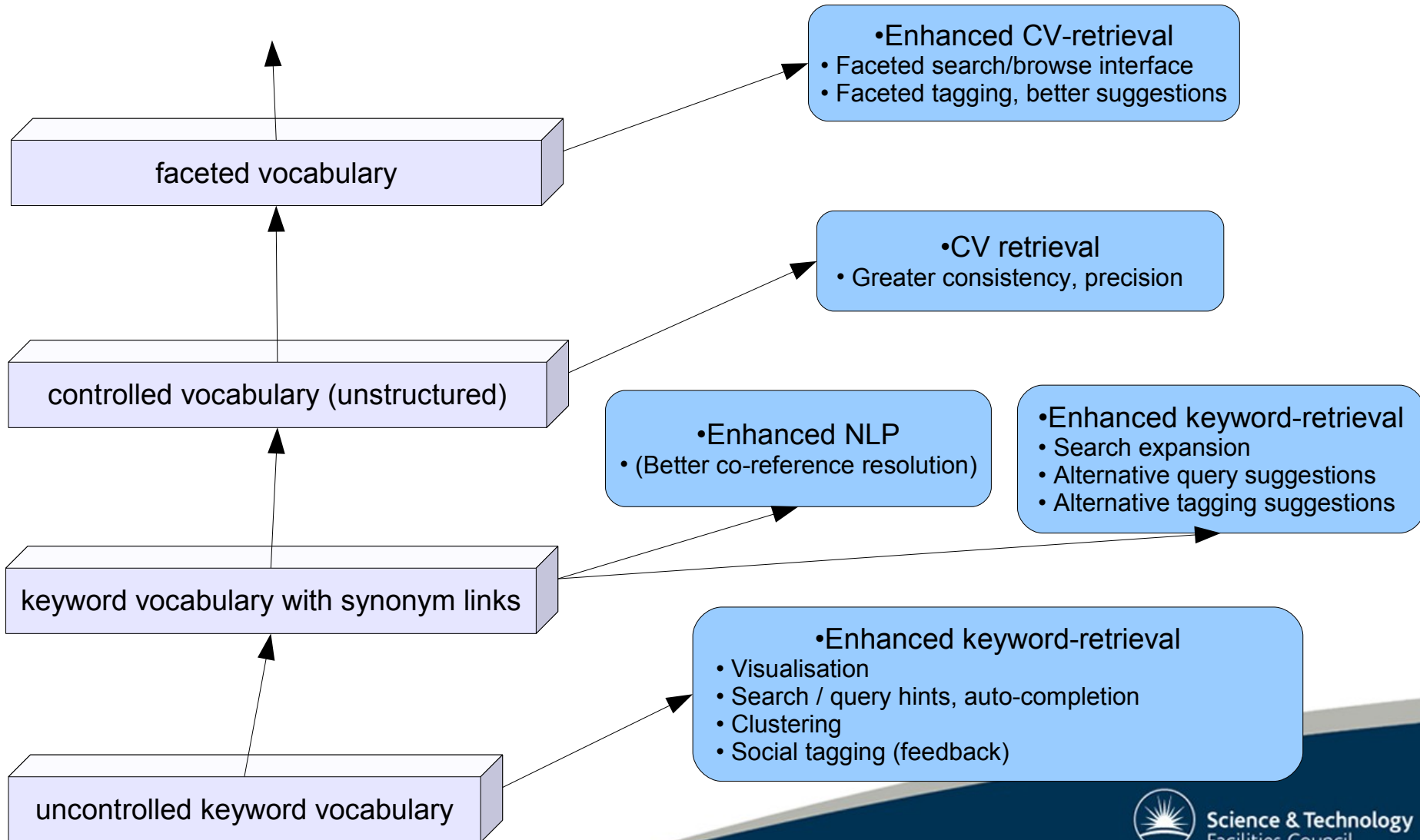


Value Grid

- One way of tackling this question is to construct a ***value grid***
- Consider step-wise development exploitation of ***intermediate products*** on the way to controlled vocabulary (and beyond?)
- ***Collaboration in the Value Grid for Semantic Technologies***, Miles 2007
 - <http://purl.org/net/value-grid>



Example Value Grid



Complications

- Different applications have different constraints ...
- ... may push intermediate products in different directions ...
- ... may be unsuitable as raw material for subsequent products



Vocabulary Problem

- How develop the vocabulary?
- Who will do it?
- Go faster for less?
- Maintenance?
- Roles, responsibilities, procedures?
- Better collaboration?
- (Daniel 2005)



Summary of Topic 2

- Business Context
 - ROI Problem
 - CVs as part of an integrated strategy
 - Tagging Problem
 - Early returns?
 - Value grid
 - Vocabulary development, maintenance and collaboration ...



Topic 3 – Networked Environment

Publishing vocabularies in a networked environment, for consumption by people and machines...



2 Basic Problems

- Identity & reference
 - All software systems depend on use of identifiers
 - In networked software systems
 - Identifier clash (one id, two referents)
 - Can be fatal
 - Aliases (two ids, one referent)
 - Less fatal, but can be costly)
- Dealing with partial information in an open-ended system



Identity & Reference

- How to avoid identifier clashes in an open-ended, global system?
- URI system
 - Shared “space” of identifiers
 - Mechanism for establishing ownership
 - Mechanism to “dereference” an identifier
 - N.B. For some URIs only!
- In practice, does not prevent clashes!
- Clashes can be subtle, e.g. dc elements...



URI

- Despite that, URIs are probably the best we have right now...
- Recommend use of URIs which can be “dereferenced”
 - E.g. http: URIs
- Why?
 - Provides a mechanism for resolving disagreements...
 - ... go GET a representation – the final answer



Vocabularies and URIs

- For controlled vocabularies, means allocating a URI for each conceptual unit
- (N.B. Allocation implies ownership!)



Partial Information

- Somebody gave me some metadata, it uses controlled vocabularies in some of the elements ... but all I've got is a big bag of URIs! That's not much good ...
- Classic example of partial information
- How to obtain further information from an authoritative source?



Partial Information

- One way of obtaining authoritative information is to rely on trust and social networks
 - I know organisation X is a clearing house for vocabularies, so I'll go ask them...
- Practical on a small-to-medium scale
- Breaks down at larger scales...
 - ... because it depends on prior knowledge, which may be patchy



Partial Information

- Alternative – ***unique dereference mechanism***
- Some unique mechanism for obtaining more information, which is guaranteed to come from the identifier owner...
 - E.g. The WWW & http URIs, ownership is delegated via DNS, which also provides decentralised system for locating appropriate network endpoints, and a protocol for requesting data (HTTP)



Publishing

- All well and good, but ***depends on vocabulary being out there!***
- Accessible from the right places...
- Publishing information on the Web, connecting up URIs to appropriate information resources...



Publishing Vocabularies

- Current best practice
 - Publish content for both people and machines
 - Publish HTML for people
 - Publish RDF for machines
 - Put this content on the Web somewhere
 - Set up vocabulary URIs for ***conditional redirects...***
 - Redirect to appropriate content (HTML or RDF), depending on the request header



How Much?

- Open questions...
 - URIs used in controlled vocabulary, how much data should you be able to GET?
 - Whole vocabulary? Part of vocabulary?
- See “Best Practice Recipes for Publishing RDF Vocabularies”
 - <http://www.w3.org/TR/swbp-vocab-pub/>
 - N.B. Recipe 6 most appropriate for medium to large vocabularies is still TODO!
 - N.B. Some technical issues also



Machine-Readable Representation

- SKOS – example of a “format” for machine-readable representation of a vocabulary
- Simple Knowledge Organisation System
- <http://www.w3.org/2004/02/skos>
- On W3C Recommendation Track
- Uses RDF
- Uses URIs
- Plugs (fairly) seamlessly into DCMI and Semantic Web frameworks



SKOSIFY!

- Creating a SKOS representation of a controlled vocabulary
 - If use RDBMS, can generate RDF/XML report, or can use RDB-RDF mapping e.g. D2RQ
 - If use XML, can use XSLT or XQuery to generate RDF/XML
 - If use spreadsheets, can output to CSV, XML or other intermediate form, then go from there



SPARQL Services

- One step on from simple publishing
- Expose vocabulary via SPARQL service
- Allows applications to selectively query and retrieve data
- Basis for further services



Summary of Topic 3

- Networked environment
 - Basic problems
 - Identity
 - Partial information
 - URIs for identity & reference
 - URIs + HTTP for publication and access
 - Publish for people *and* machines
 - SKOS example of machine format
 - SKOSIFICATION
 - SPARQL services



Topic 4 – Retrieval

Exploiting semantic relationships for better retrieval...



Vocabulary Control

- Fundamental problems of text retrieval...
 - **Synonymy** - two words have the same or very similar meaning
 - e.g. salinity/saltiness
 - Effect – **low recall** (i.e. some relevant items are missed)
 - **Ambiguity** - a word can have more than one meaning
 - e.g. Mercury (planet / metal / Roman god / ...)
 - Effect – **low precision** (i.e. lots of irrelevant results)



Vocabulary Control

- Primary goals...
 - Control synonyms – improve recall
 - Eliminate ambiguity – improve precision



Content Objects

- What do we want to retrieve?
 - What are we trying to find?
 - Could be a book, a web page...
 - ... a part of a document ...
 - ... a museum object ...
 - ... anything!
-
- For this tutorial, let's call them “content objects” (doesn't matter what they are)

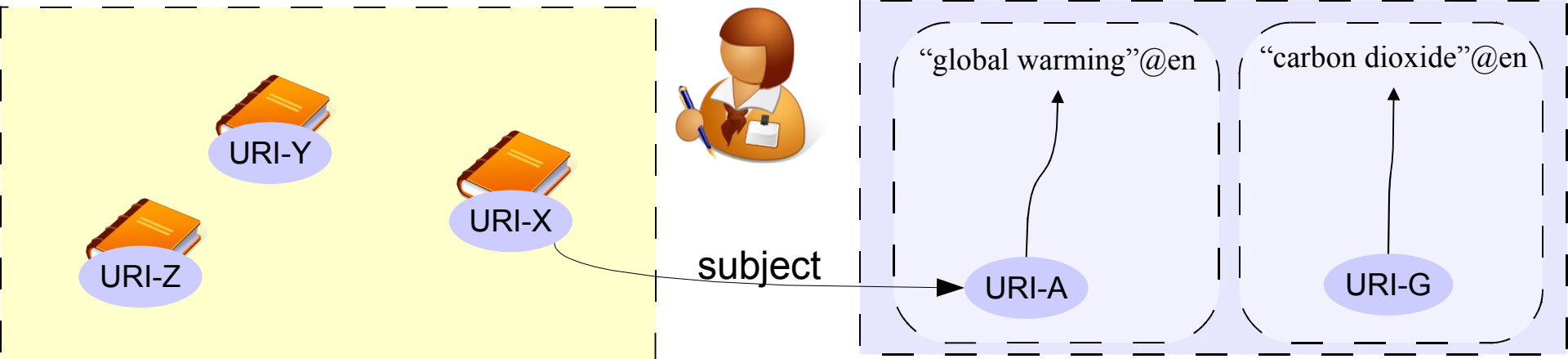


Collections

- A set of content objects identified by URIs



Tagging (Indexing)

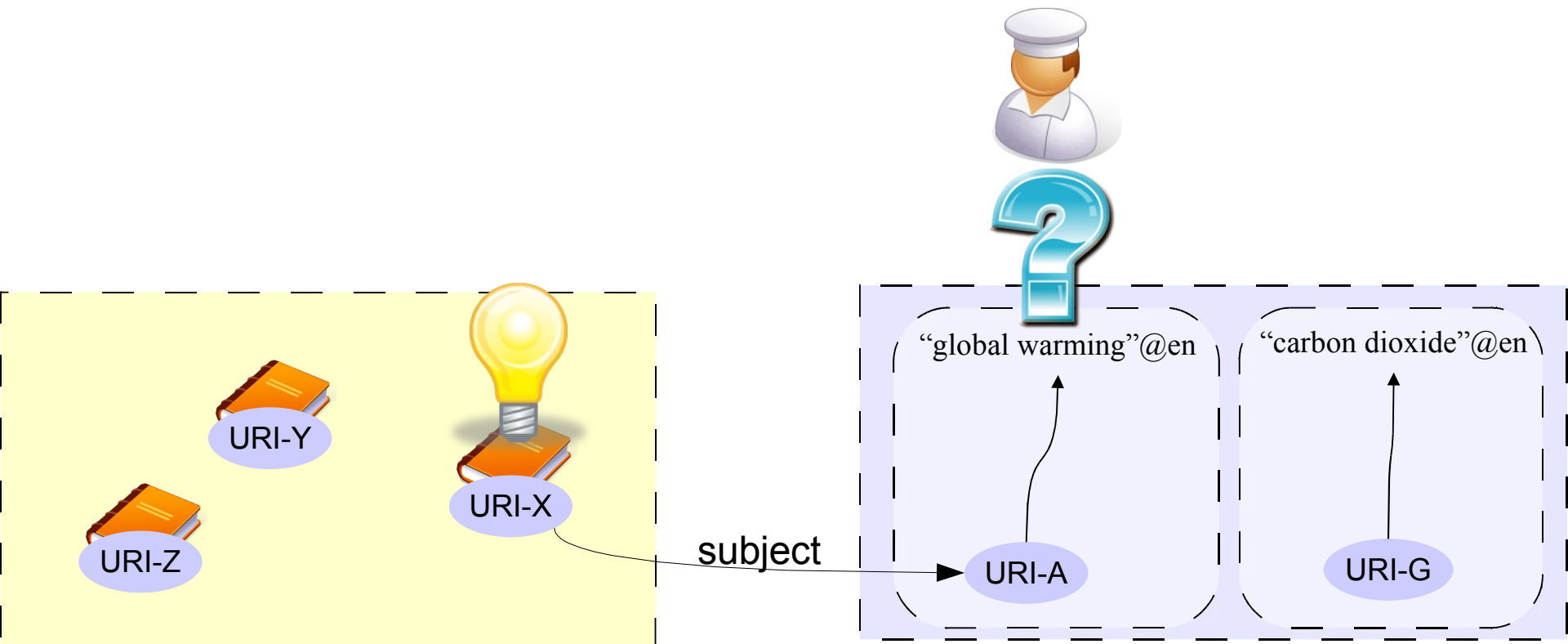


Subject ...

- N.B. “Subject” is not the only tagging field where you can use vocabularies ...
- See Daniel 2005, especially notes on breaking subject into *facets*
- Today, will consider single tagging field for simplicity
- ***N.B. Everything I say here can be extended to faceted retrieval***



Simple Retrieval



Retrieval Basics

- Relevance – does result satisfy need?
- Precision – proportion of search results which are actually relevant?
 - Low precision = lots of irrelevant search results
- Recall – proportion of all relevant objects returned in search results?
 - Low recall = lots of relevant objects not in search results



Perfect Indexing?

- Assumption of perfect (ideal) indexing...
 - All indexing is correct
 - Nothing missed
- ... predicts 100% precision and recall for simple retrieval
- ... can test prediction!



Imperfect Indexing!

- In practice, perfect indexing is virtually impossible
- Can we use semantic relationships to help?



Semantic Relationships

- ISO 2788, BS 8723-2, Z39.19 provide guidelines on use and meaning of hierarchical and associative relationships...
- ... from philosophical point of view
 - e.g. Hierarchical = IsA (generic), instance of (instantive) or part of (partitive)
 - e.g. Associative = anything else (e.g. cause/effect ...)
- ***But what do semantic relationships mean for retrieval?***



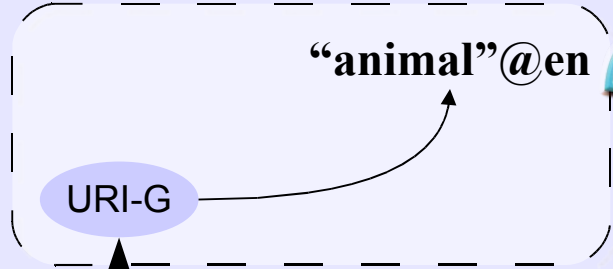
Hierarchies & Relevance

Is this relevant?

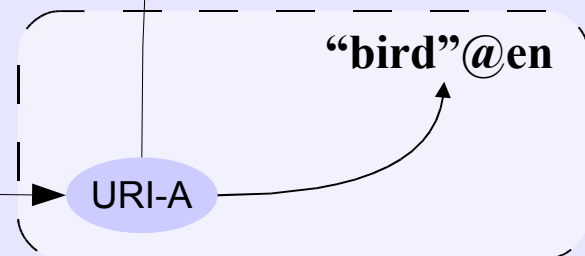


URI-X

subject



broader



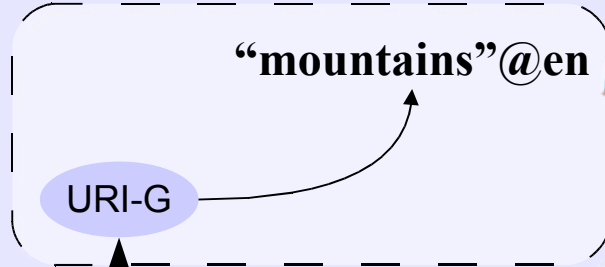
Hierarchies & Relevance

Is this relevant?

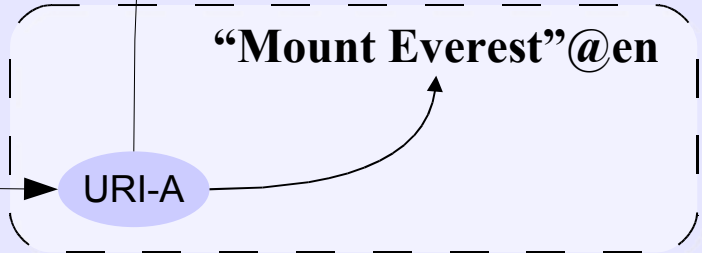


URI-X

subject



broader



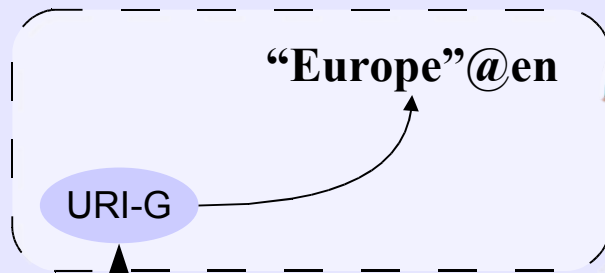
Hierarchies & Relevance

Is this relevant?

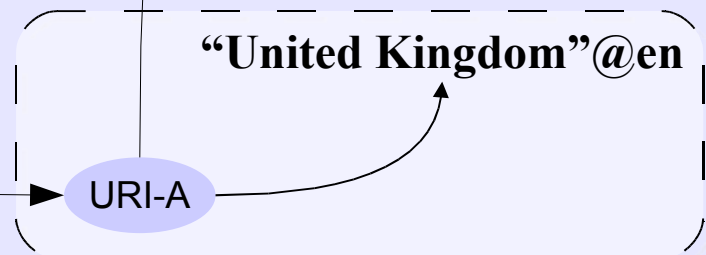


URI-X

subject



broader



Naïve Assumption – Broader

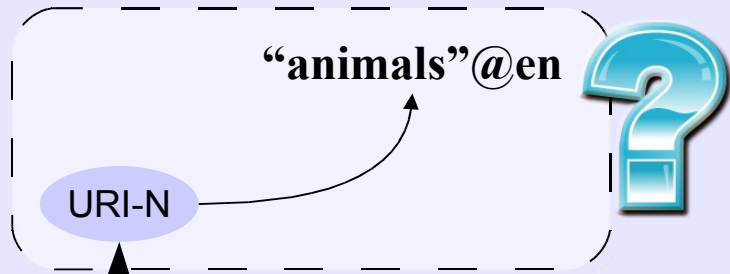
- Can make a “naive” assumption ...
- ... if X -broader- \rightarrow Y and A is relevant to X , then A is also relevant to Y
- E.g. If an object is tagged with *birds of paradise*, then it is also relevant to a query for *birds*, and to a query for *animals* ...



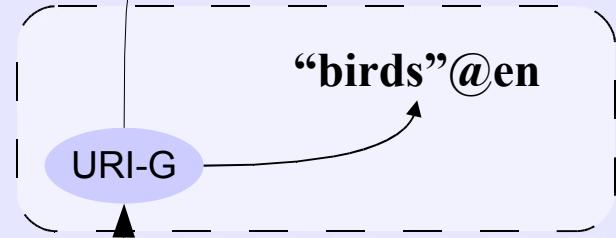
Naïve Assumption – Broader



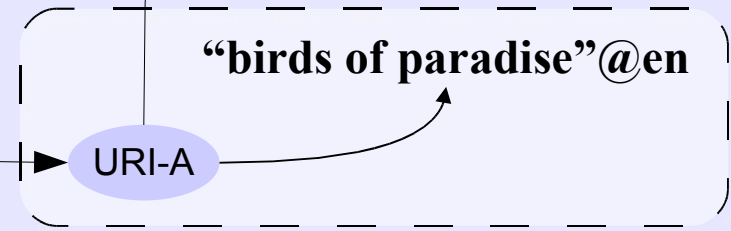
subject



broader



broader



Simple Search Expansion

- Use naïve assumption to expand search
- Predicts increase in recall
- N.B. Two ways to implement (for any type of expansion)
 - Expand query
 - Expand index (tagging data – metadata)



Expand Query

- *Birds of paradise* ->
- *Birds of paradise* OR *birds* OR *animals*

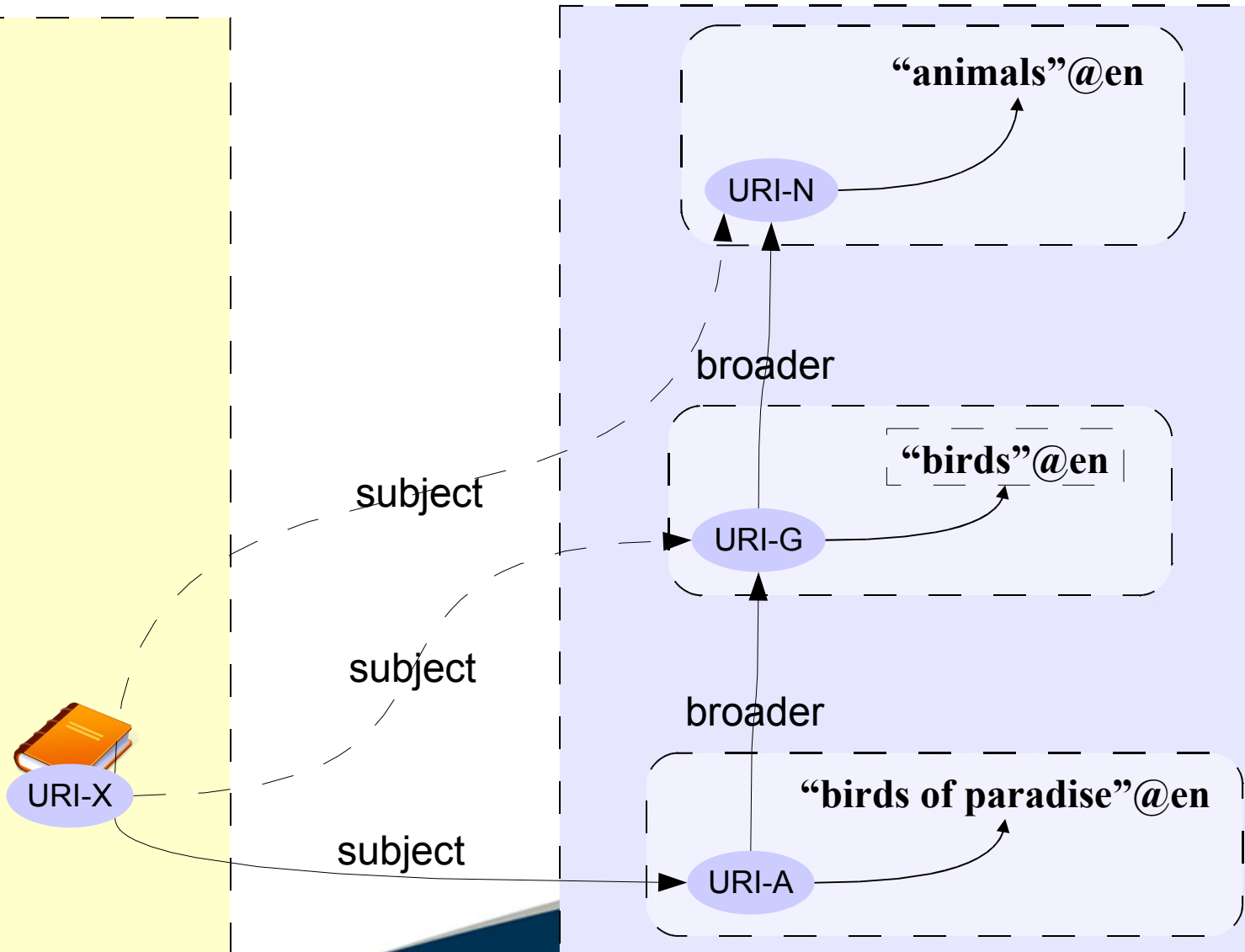
- Expand queries “on the fly” then evaluate against original index



Expand Index

- Expand index once
- Evaluate queries against expanded index

Expand Index



Simple Expansion - Limitations

- What about narrower?
 - E.g. If an object is tagged with *birds*, is it relevant to a query for *birds of paradise*?
 - ... sometimes?
- What about associative relationships?
 - E.g. If an object is tagged with *birds*, is it relevant to a query for *ornithology*?
 - ... sometimes?



Simple Expansion – Limitations

- What about levels of specificity/generalality?
 - E.g. If one object is tagged with *animals*, another is tagged with *birds of paradise*, which is ***more relevant*** to a query for *animals*?



Beyond naïve Assumption – Relevance Cost

- Can use any type of semantic relationship to expand search ...
- ... but every time you expand across a single relationship, you incur a **cost** in terms of **decreased probability of relevance**
- Different relationship types have different costs
 - broader < related < narrower (index expansion)

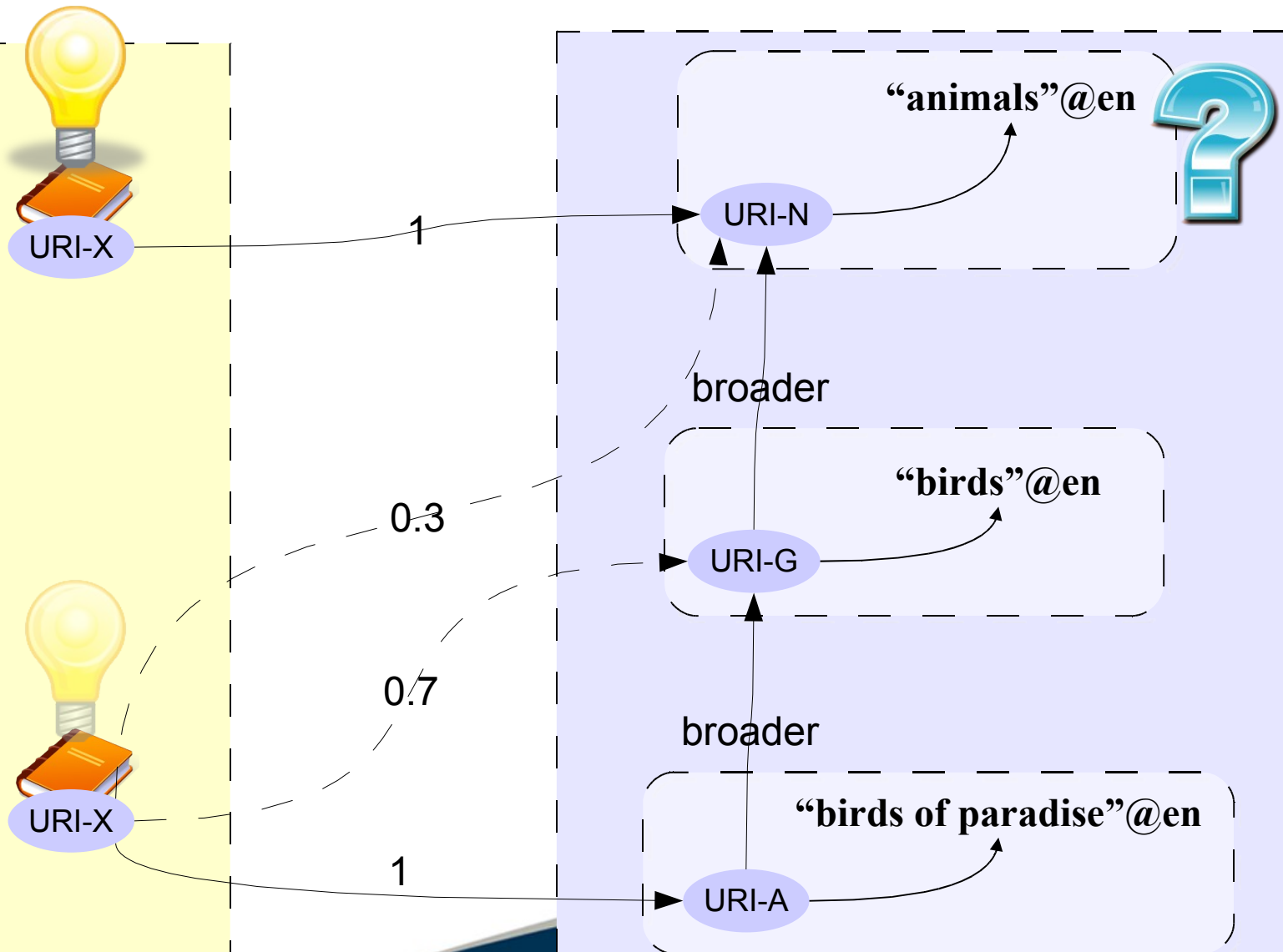


Limited Cost Expansion

- Use numeric function to model relevance cost ... use this to rank results
- Expand until cost reaches a predefined limit
- A.k.a. Weighted expansion



Limited Cost Expansion



Expansion Comparison

- Limited cost expansion vs. simple expansion
 - Limited cost is better, because...
 - Uses all relationship types
 - Takes account of specificity in ranking
 - Can be fine tuned



Issues & Complications

- Expansion and negation in boolean queries?
- Tuning, choice of cost function?
- For more, see
 - <http://purl.org/net/retrieval>



Further Applications

- Handling vocabulary mapping!



Summary of Topic 4

- Using vocabularies for retrieval
 - Tagging (indexing)
 - Retrieval principles
 - Using semantic relationships to improve retrieval
 - Assumptions about relevance
 - Expansion Strategies
 - Expand query or index?
 - Simple Expansion
 - Limited Cost Expansion



Topic 5 – Mapping

Mapping (linking) between vocabularies...



Why Map?

- How get uniform retrieval across heterogeneous metadata?
 - i.e. metadata where two or more different vocabularies used
 - i.e. use single vocabulary of choice to query across all resources
- Requires mapping (linking) ***between vocabularies***
- In standards, often referred to as “interoperability”



Guidance?

- Beware ISO 5964
 - Not really a guide to mapping between vocabularies!
- BS 8723-3 “Interoperability”
 - Although beware, subtle but important distinction between ...
 - resolving problems in construction of multilingual thesauri
 - Mapping between two different vocabularies
- Z39.19 “Interoperability”
 - Some discussion



Simple Mapping Model

- Most agree on basic mapping relationships...
 - Exact
 - Broader
 - Narrower
 - Inexact / Related?
- N.B. Mapping between conceptual units (a.k.a. Semantic mapping)



One-To-Many

- Some uncertainty about one-to-many mappings
- E.g. BS 8723-3 has special notation
- Others use “AND”, “OR” and “NOT”, often loosely without stating what they mean, how they should be interpreted for retrieval



Mapping & Retrieval

- This tutorial's question:
 - Given set of mappings between two vocabularies...
 - ... how implement seamless retrieval?



Translation & Expansion

- Goal is to ***translate*** either query or metadata (index)
- Can treat ***translation*** in very similar way to ***expansion***
 - I.e. using mapping relationships to translate similar to using semantic relationships to expand
 - Both depend on assumptions about how mapping / semantic relationships impact on relevance



Relevance Assumptions

- Everything depends on the assumptions you make about what mapping relationships mean for relevance
- Can make naïve assumption
 - Use simple translation
- Can also make relevance cost assumption
 - Use limited cost (weighted) translation
- N.B. Also depends on goals
 - Preserve recall or precision?



Summary of Topic 5

- Mapping between vocabularies
- Retrieval across heterogeneous metadata
- Simple mapping model
- Expansion and translation
- Relevance assumptions
- Retrieval Goals



Tutorial Summary

- Previous tutorials!
- Topic 1 – informal vocabulary model
- Topic 2 – business context
- Topic 3 – networked environment
- Topic 4 – vocabularies & retrieval
- Topic 5 – vocabulary mapping & retrieval



Contact Me (& work on SKOS!)

- a.j.miles@rl.ac.uk
- <http://purl.org/net/aliman>
- (See Tom Baker to join SWDWG and help make SKOS a standard ... and help solve all those niggling vocabulary management problems :-)

- Thanks for listening!
- Questions?

