



Building context into IA projects

A review of successful structures and processes

Jonathan Engel

WIAD 2024

Introduction – My background

- Reuters Manager for Multimedia News Production (19 years in multiple roles)
- Running own information management consultancy for last 22 years – InfoArk
- Designer of customised taxonomies and related metadata for classifying content in 30 major projects
- Specialist in linking classification schemes with automated tagging and search software, plus content filters and linked data



Introduction – Selected clients

- Dow Jones newswires online
- Times and Sunday Times online
- Institute of Chartered Accountants
- Clifford Chance law firm
- Which? (Consumer Association)
- Cambridge University
- Unilever
- Shop Direct Group (Littlewoods, Very)
- UK Care Quality Commission
- NHS Education for Scotland
- UK Department for International Development
- Oxfam International



Connections – puzzle needs context

Create four groups of four!



Nose

Head

Stiff

Wing

Bulb

Seal

Crayon

Rob

Ear

Engine

Hose

Candle

Cabin

Stalk

Honeycomb

Fleece

Connections – multiple links possible

Create four groups of four!

Nose ?

Head ?

Stiff

Wing

Bulb ?

Seal ?

Crayon

Rob

Ear ?

Engine ?

Hose ?

Candle ?

Cabin

Stalk

Honeycomb

Fleece

Connections – find common thread!

Create four groups of four!

RIP OFF

Fleece

Hose

Rob

Stiff

PARTS OF AN AIRPLANE

Cabin

Engine

Nose

Wing

UNITS OF VEGETABLES

Bulb

Ear

Head

Stalk

THINGS MADE OF WAX

Candle

Crayon

Honeycomb

Seal

Cycle of Context

Engage specialists – begin with advice on relevant vocabularies and documents

Build and test initial taxonomy around unifying topics

Extend taxonomy with synonyms

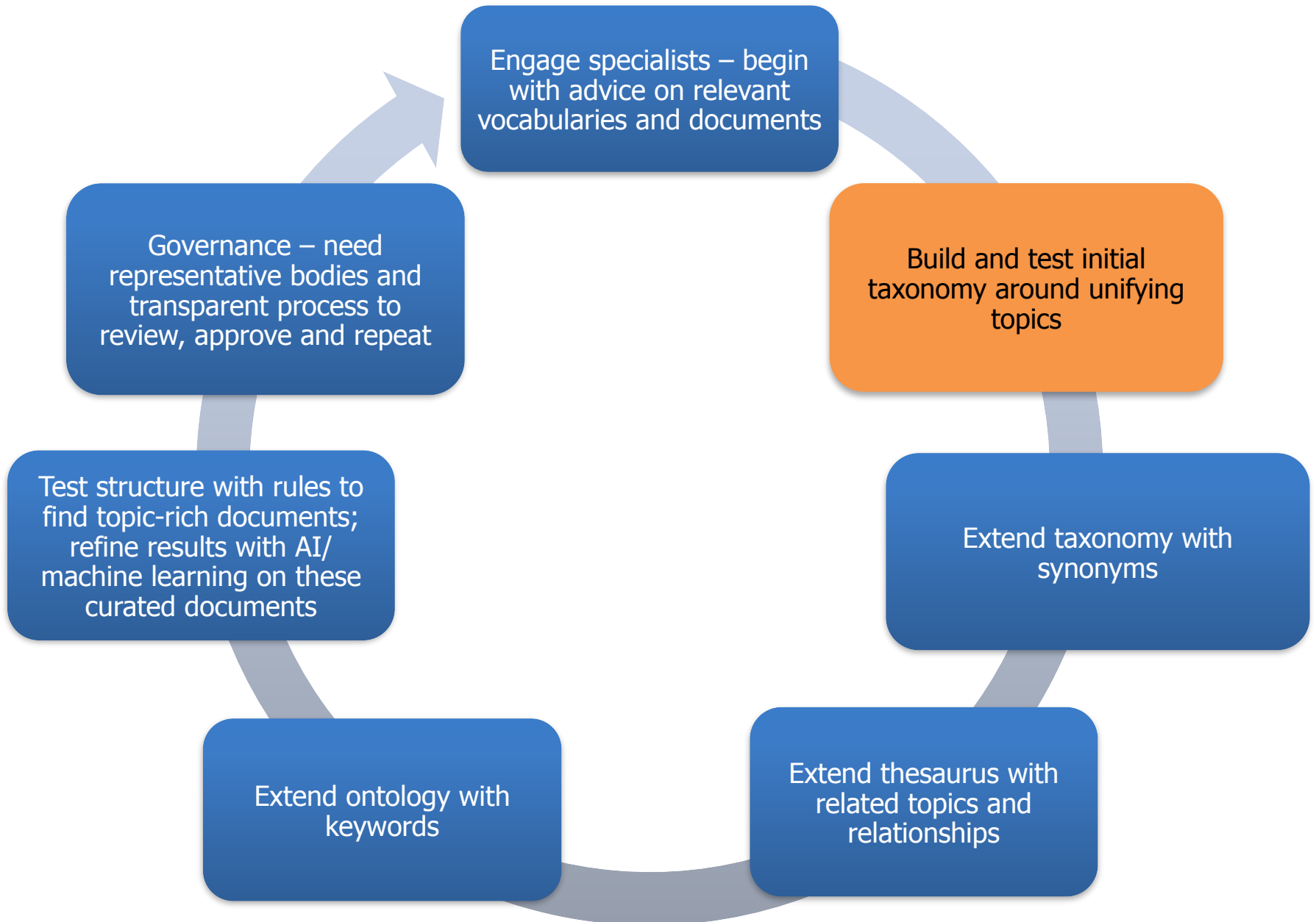
Extend thesaurus with related topics and relationships

Extend ontology with keywords

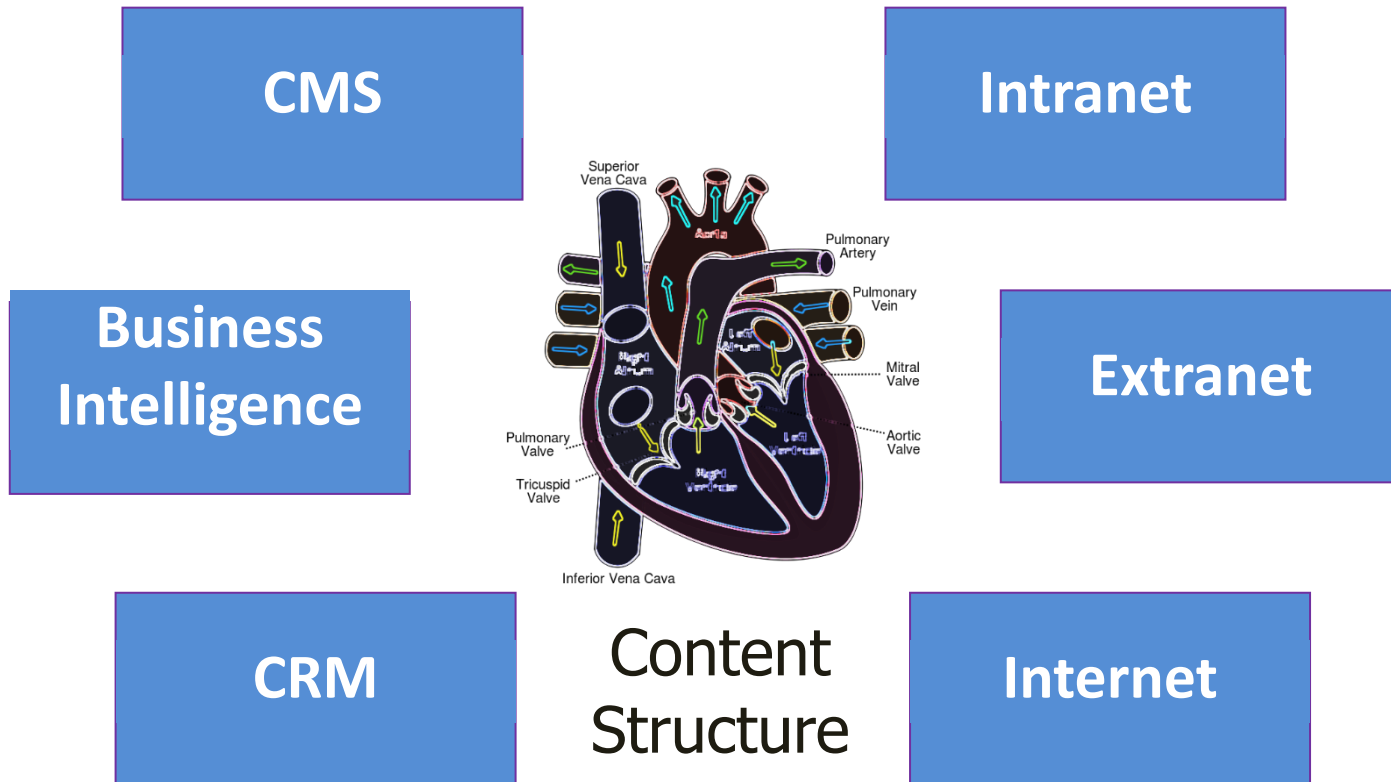
Test structure with rules to find topic-rich documents; refine results with AI/machine learning on these curated documents

Governance – need representative bodies and transparent process to review, approve and repeat

Cycle of Context



Taxonomy -- the heart of all IT systems



Multi-faceted taxonomy goes beyond “subjects”

Entities →
Who? Where? Who for?

- Geography for location, jurisdiction
- Organisation’s business units
- External organisations by type
- List of statutes, products, roles, etc.

Subject matter →
What? Why?

- Business activities and issues
- Business sectors, e.g. financial services

Focused filters →
How?

- Events, projects and initiatives
- Content types and level
- Language

Combine structures for “extended” taxonomy



Taxonomy

Hierarchical – logically nested topics for navigating



Thesaurus

Equivalence – adds synonyms for improved searching



Ontology

Associative -- with defined relationships for navigating, searching and filtering

Components of an extended taxonomy

Preferred term

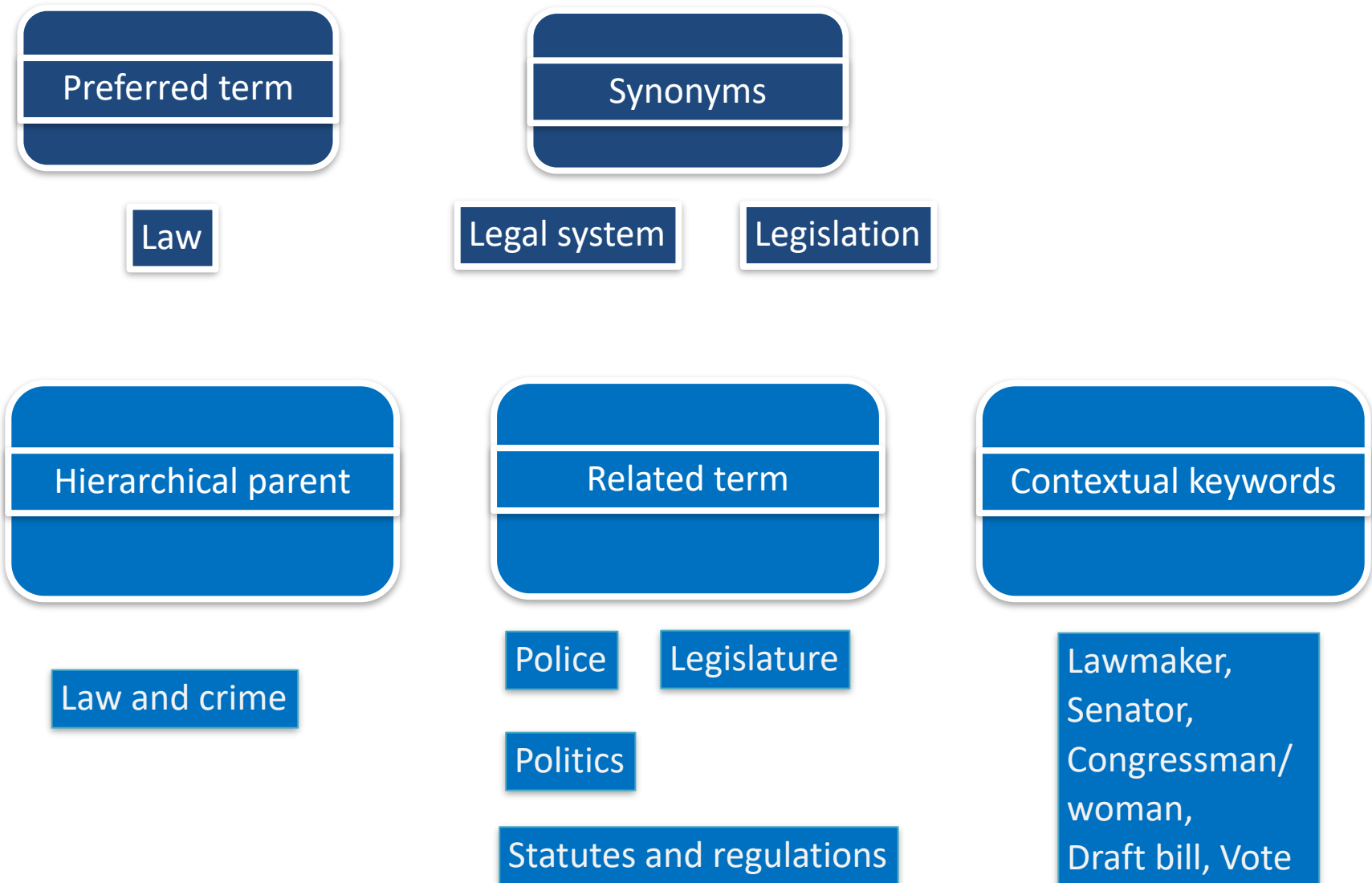
Synonyms

Hierarchical parent

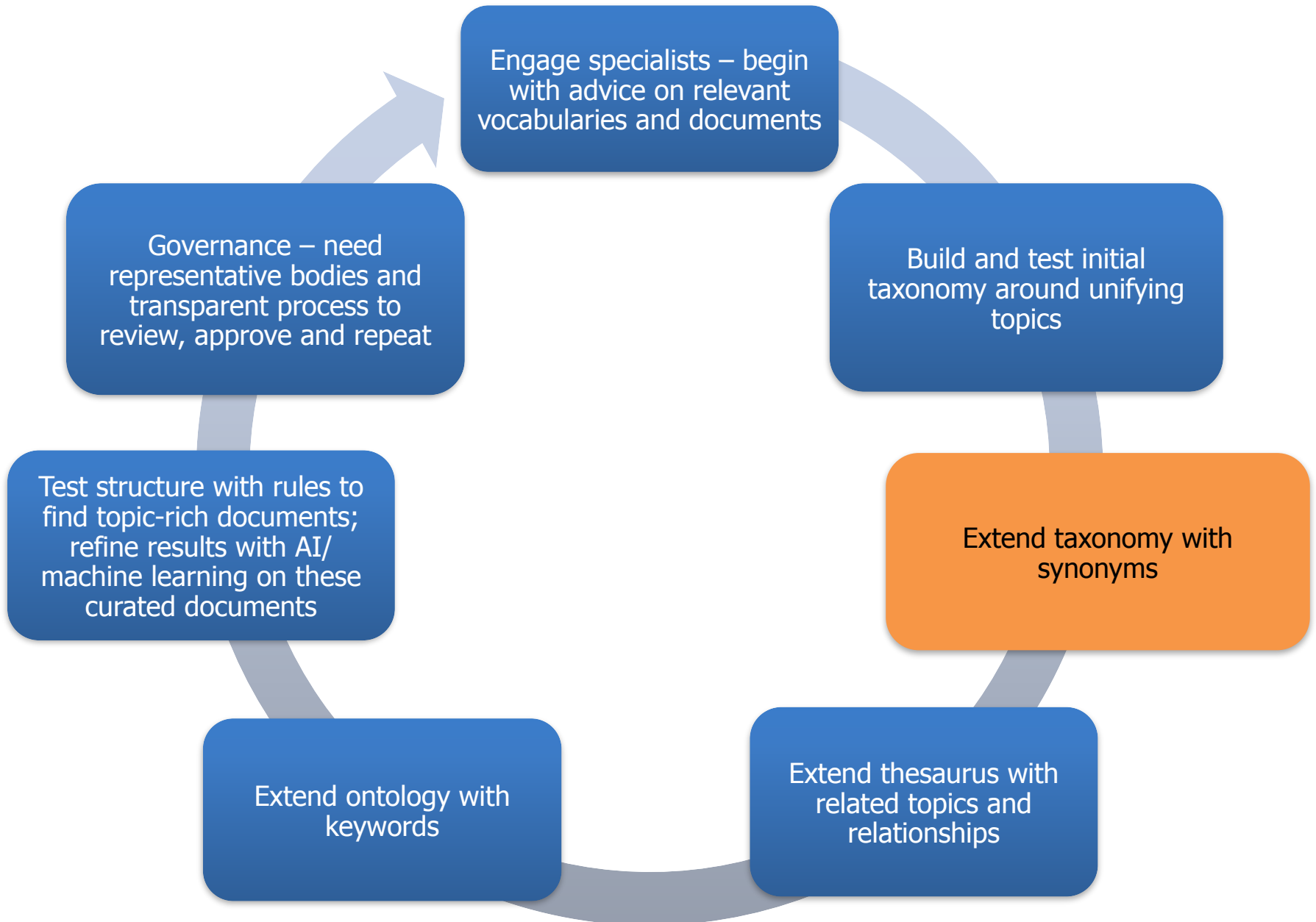
Related term

Contextual keywords

Extended taxonomy term – example

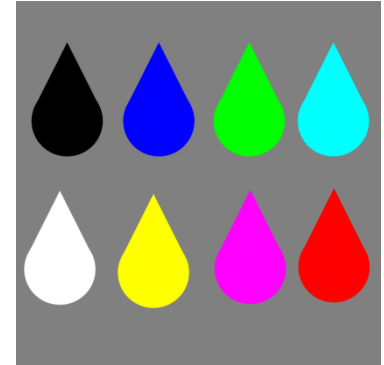


Cycle of Context

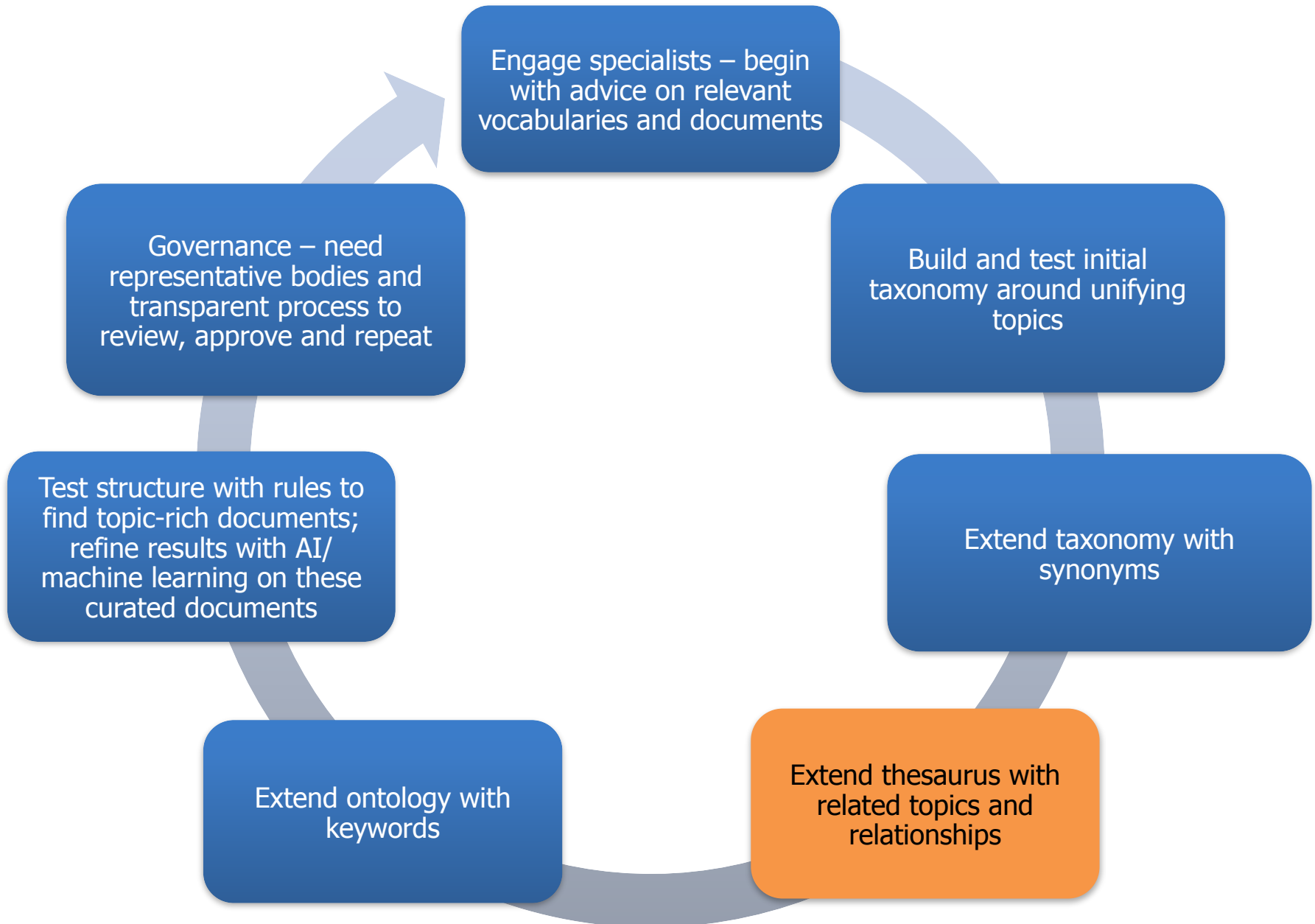


Synonyms

- Equivalent terms – exact or “near” match
- Example -- Cardiovascular disease
- Synonyms -- Heart disease, Atherosclerosis, Arterial disease, Cardiovascular condition, Cardiovascular illness
- Synonym rings – useful for recurring equivalencies, e.g. disease = illness = condition
- Can link rings to produce “semantic nets” to discover information, e.g. Danger + Southwest



Cycle of Context

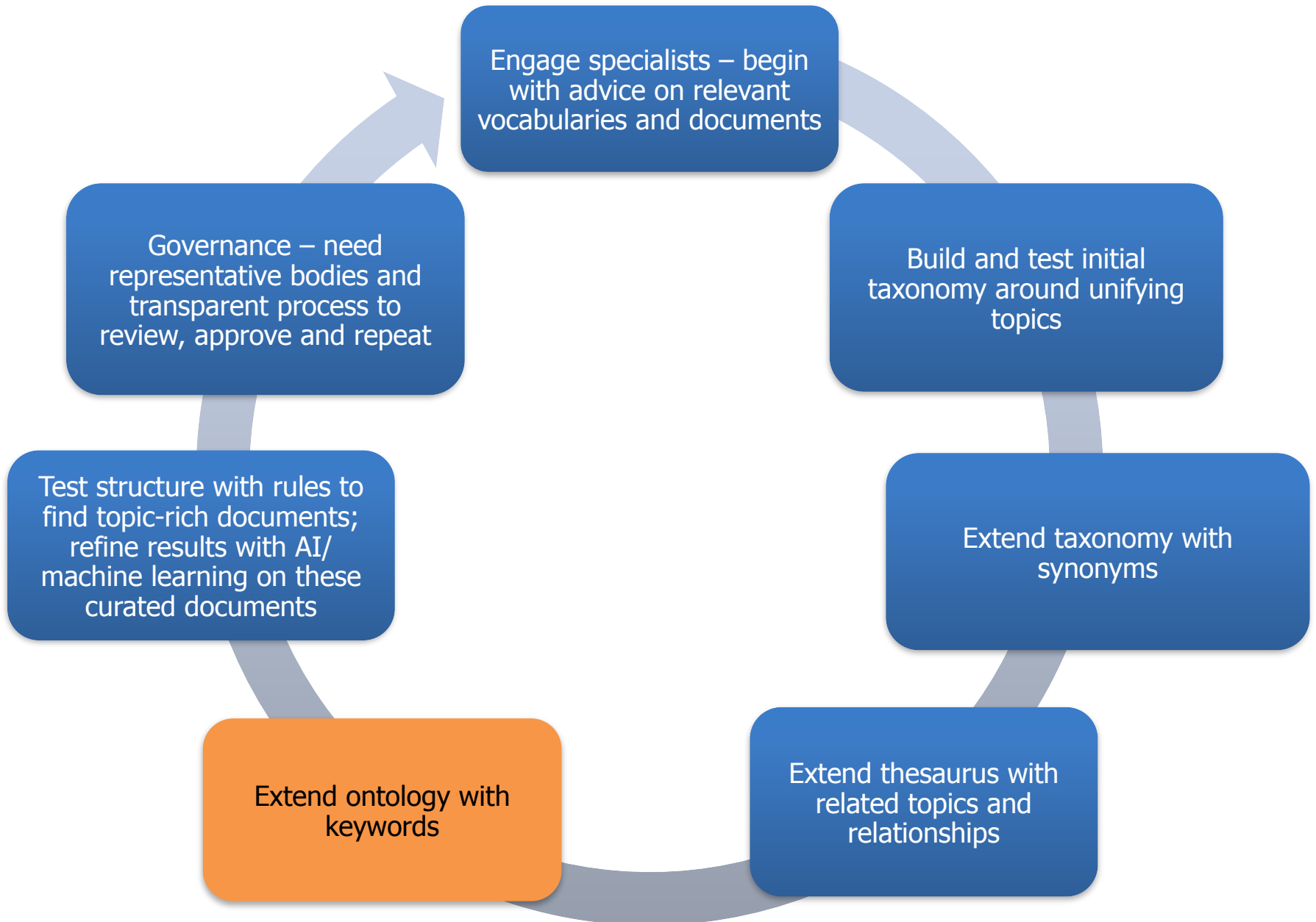


Related terms

- Already present in taxonomy
- Associated with the preferred term
- Useful to record strength of relationship for tagging, e.g. mandatory or discretionary
- The City of London police will always be linked to crime prevention, but crime prevention only sometimes will be linked to that specific police force
- Useful to capture type of relationship, e.g. organisation “comprises” specific members, while members are “part of” organisation



Cycle of Context



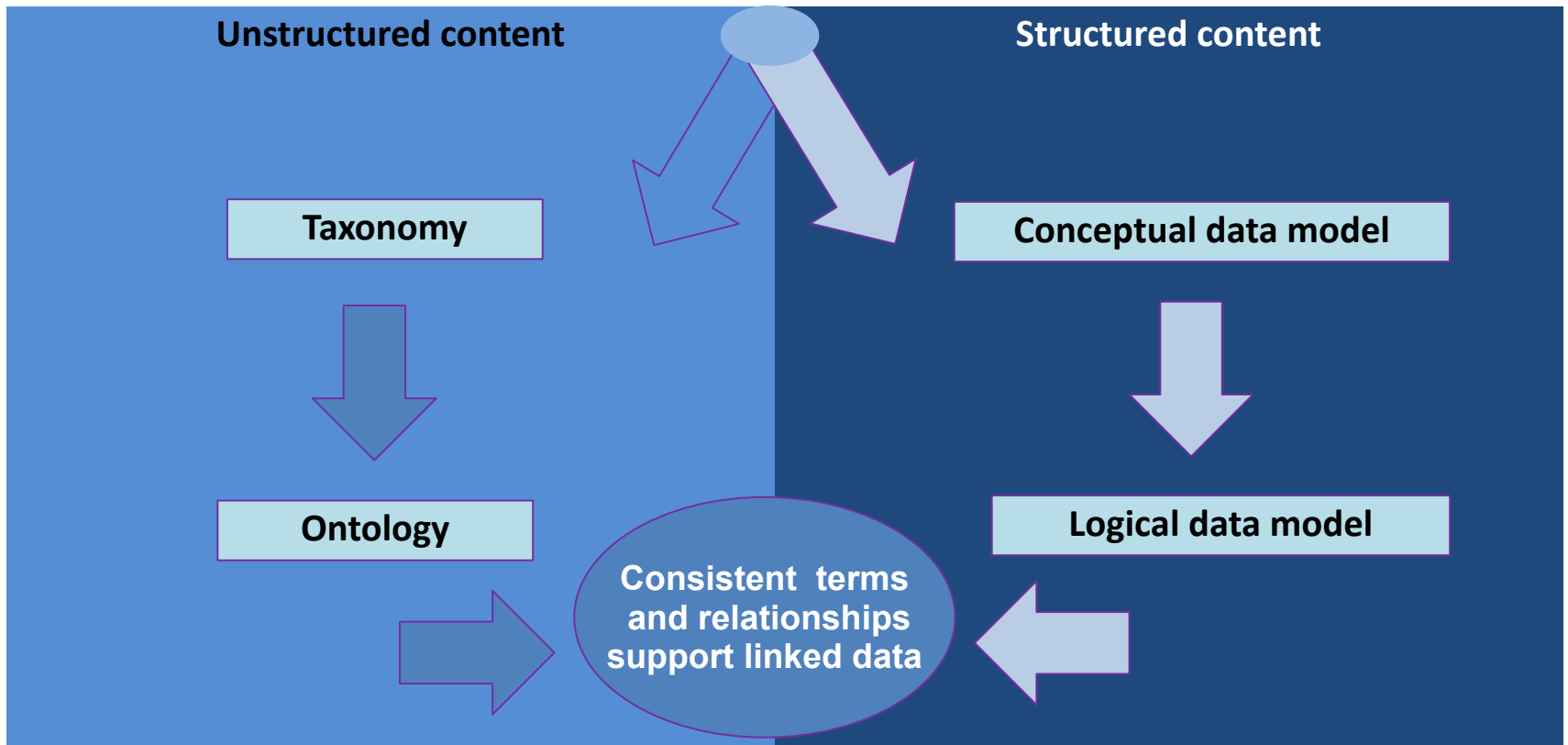
Sources for extended taxonomy

- “Runners up” to preferred term
- Acronyms
- Search queries
- Subject specialists
- Domain-specific documents
- Text-mining software
- Faceted-classification or search software (especially if employed when building taxonomy, not after)



Information strategy should unite realms of documents and data

Strategic controlled vocabulary



Linked Data connects internal and external

- Organisations often aim to collate and share internal and external data
- The Resource Description Framework (RDF) simplifies data structures into consistent “triples” or “triple-stores”
- It is similar to the way data bases contain the three elements of Entities, Attributes and Values
- Thus a Study is evidenced by a Content type that is a Report. This Report has an Author who is a named Person
- The entities or resources have a Uniform Resource Identifier (URI) that together reveal the entire linked chain of “triples”



Cycle of Context

Engage specialists – begin with advice on relevant vocabularies and documents

Build and test initial taxonomy around unifying topics

Extend taxonomy with synonyms

Extend thesaurus with related topics and relationships

Extend ontology with keywords

Test structure with rules to find topic-rich documents; refine results with AI/machine learning on these curated documents


Governance – need representative bodies and transparent process to review, approve and repeat

Case for assisted content classification

- High-volume tagging consistency requires automation – one person can tag fewer than 4,000 documents per year
- In same time, that staff member could define 2,400 tagging rules and templates – for 800 subjects, 100 focused filters (for content and event types) and 1,500 entities
- Using additional staff often undermines consistency -- Dow Jones' study found specialist editors' accuracy ranged from 40-100%, with nearly half of 500 sample stories failing to hit 80% accuracy target



Solution: Classification that leverages fully extended taxonomy structure



The image shows a large iceberg floating in the ocean. The visible tip of the iceberg is labeled "Synonyms", while the much larger, submerged part is labeled "Preferred term". This visualizes the concept of a taxonomy structure where the visible part represents a limited set of terms, and the submerged part represents a much larger, more detailed structure.

- Family hierarchy, plus related concepts, as "clues" to meaning
- Contextual keywords as additional "clues"
- Negative contextual examples to disambiguate, e.g. for "application"
- Related concepts as expansion tags

Use “combo” classification/search rule

- **Frequency test:**
Instances of Preferred term OR Synonyms in content

AND

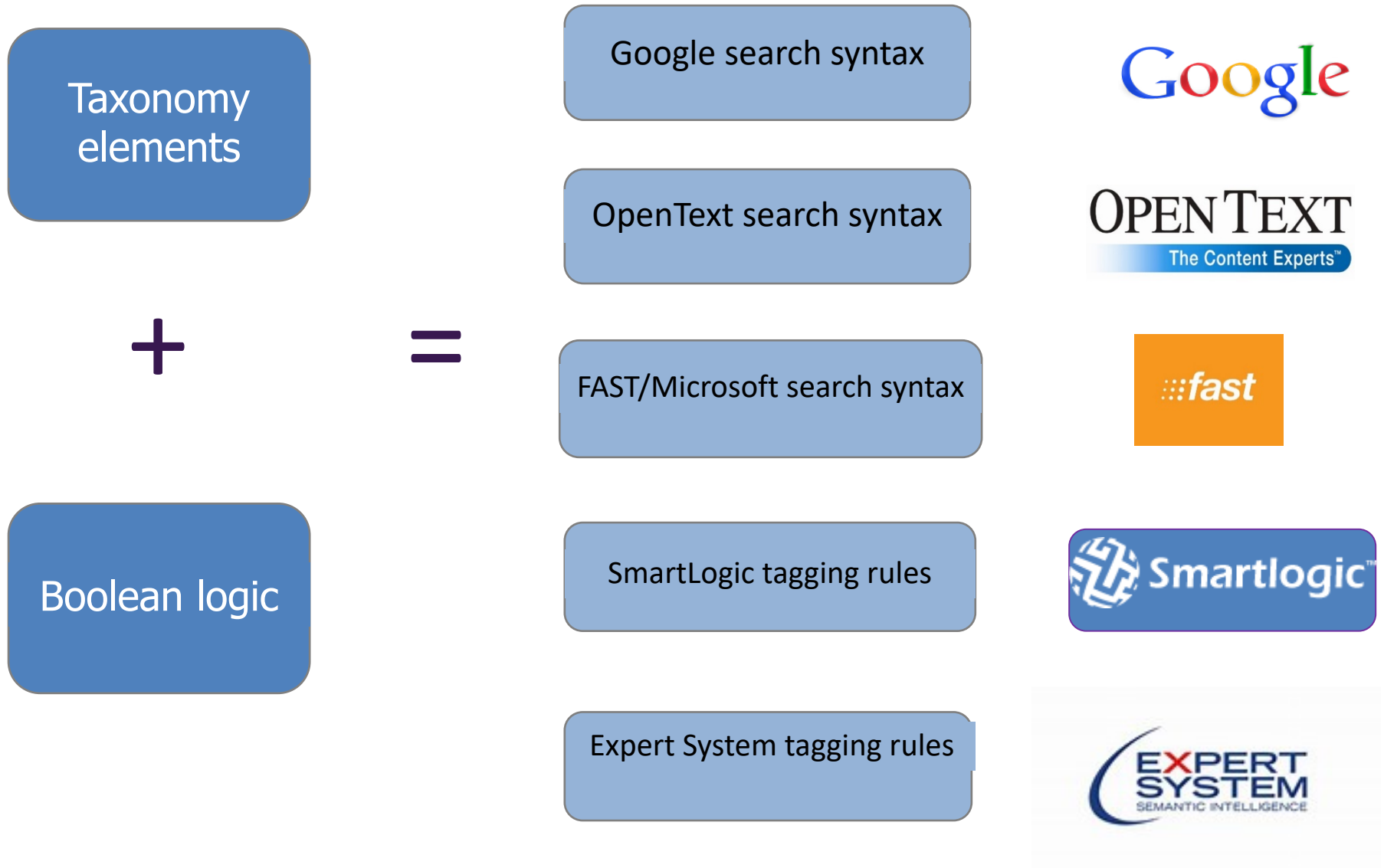
- **Prominent location test:**
Preferred term OR Synonyms in Title OR URL OR prominent Content element, e.g. Summary, Conclusion, etc.

AND/OR

- **Concurrent proximity test:**
Preferred term and synonyms within 10 words of Hierarchical parent, Child term, Related terms and Contextual keywords (or within same paragraph or same Content section, or within same five rows of text)

Use **OR** for more Recall; **AND** for more Precision

Same taxonomy can drive multiple rules



Effective use of "mail merge"

el Taxonomy Data

Word Template

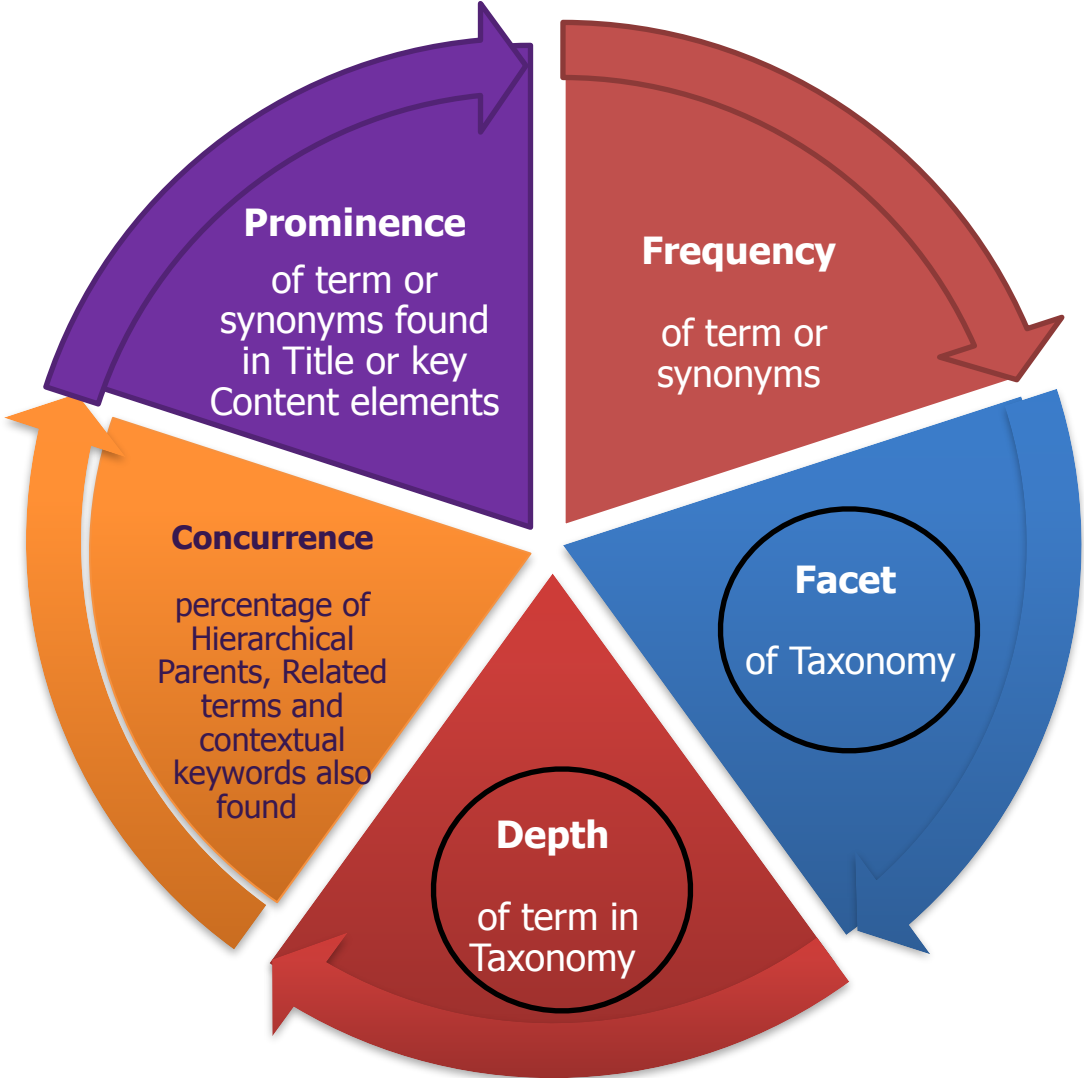
Classification rule or Search query

Short Description	Hierarchical Parent	Synonym1	Synonym2
Diabetes mellitus	Glucose metabolism disorders	Diabetes	High blood sugar
Physico-chemical characteristics	04. Substances	Physical characteristics	Chemical characteristics
Food Standards Agency	Key external organisations	FSA	UK Food Standards department
Danish Health and Medicines Authority	Key external organisations	DHMA	Danish Health Authority
Health Products Regulatory Authority	Key external organisations	HPRA	Irish Medicines Board

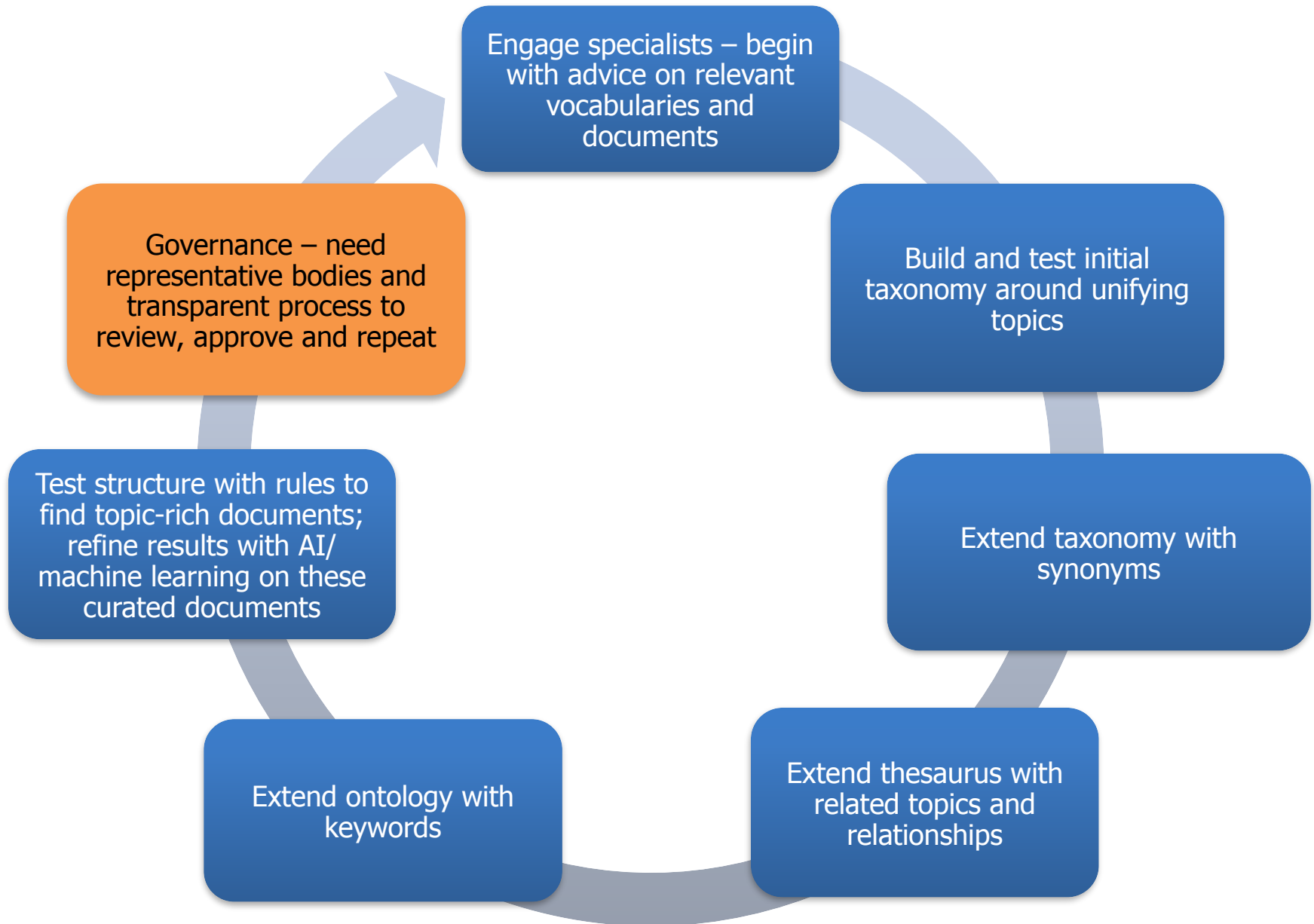
```
near(and(or(title:or("«ShortDescription»",
"«Synonym1»", "«Synonym2»",
"«Synonym3»", "«Synonym4»",
"«Synonym5»", "«Synonym6»",
"«Synonym7»"),
(or("«ShortDescription»",
"«Synonym1»", "«Synonym2»",
"«Synonym3»", "«Synonym4»",
"«Synonym5»", "«Synonym6»",
"«Synonym7»"))),
or("«CollectiveRelatedTerm»",
"«MandatoryRelatedTerm2»",
"«MandatoryRelatedTerm3»",
"«DiscretionaryRelatedTerm1»",
"«DiscretionaryRelatedTerm2»",
"«DiscretionaryRelatedTerm3»",
"«HighEvTerm»",
"«LowEvTerm»")),n=10)
```

```
near(and(or(title:or("Diabetes mellitus", "Diabetes", "High blood sugar", "Type 1 diabetes", "Type 2 diabetes", "High blood glucose", "Hyperglycaemia"), (or("Diabetes mellitus", "Diabetes", "High blood sugar", "Type 1 diabetes", "Type 2 diabetes", "High blood glucose", "Hyperglycaemia"))), or("Glucose metabolism disorders", "cardiovascular system", "obesity", "Insulin")),n=10)
```

Taxonomy structure can also contribute to relevance weighting



Cycle of Context



Time for questions





Jonathan Engel
Consultant Information Architect

W: www.infoark.co.uk

E: j.engel@infoark.co.uk

M: +44 (0) 7966 754614