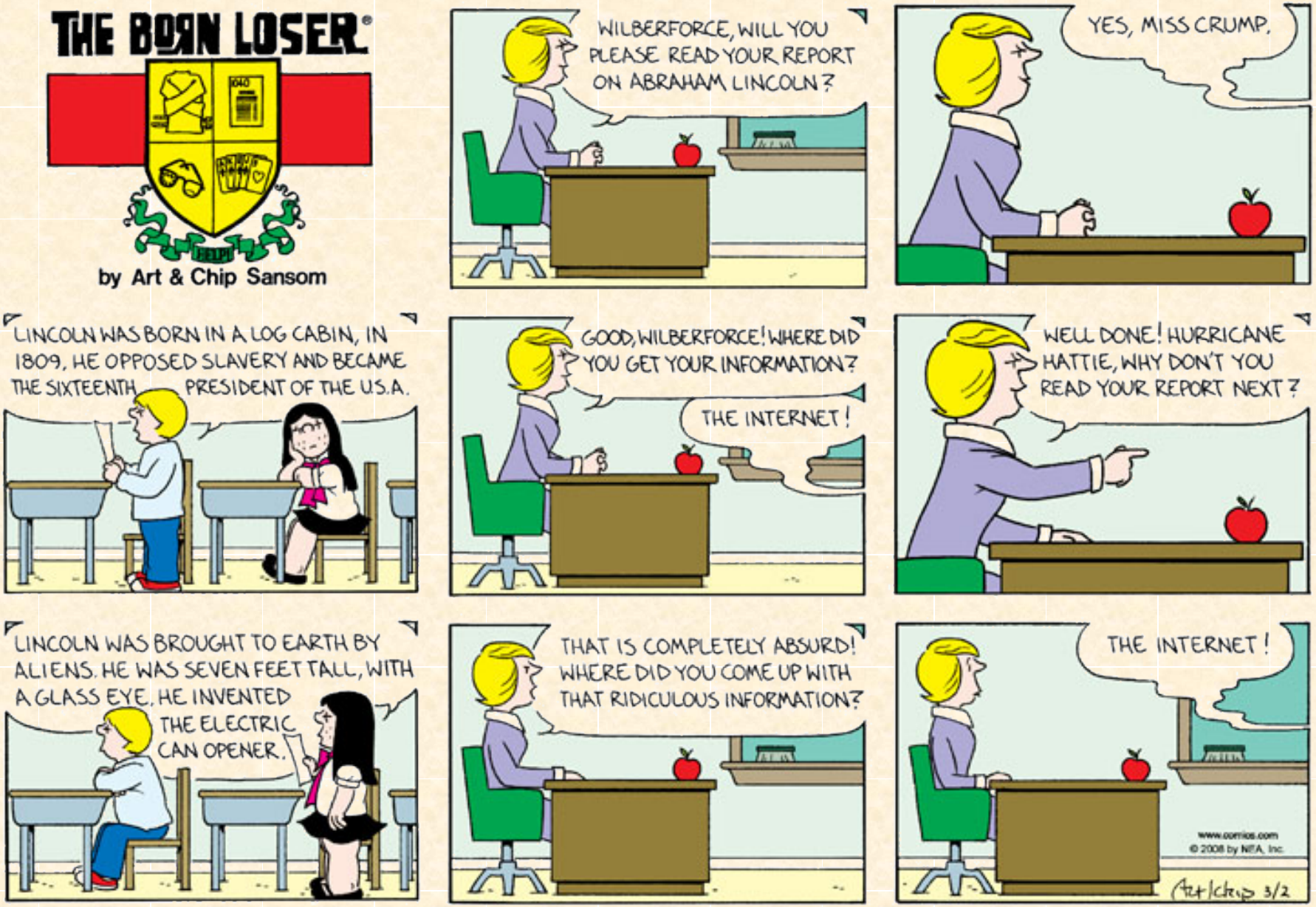
Persistent Identifiers & Language Resources

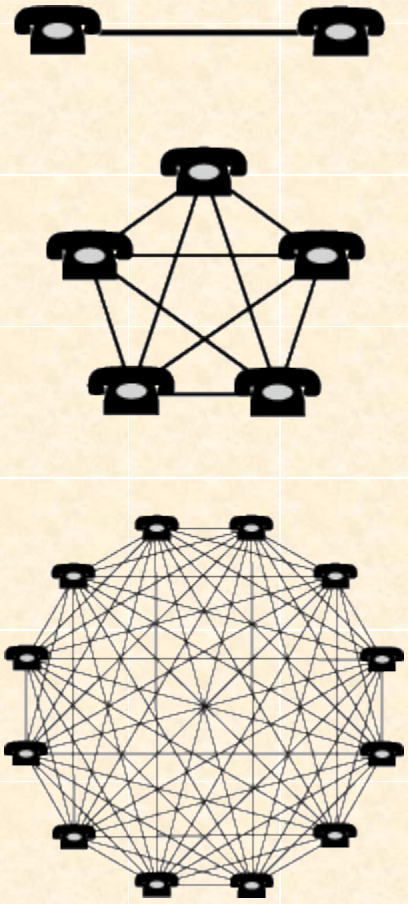
Potential for PIDs in TC 37 Resources

Sue Ellen Wright
Kent State University, Kent, Ohio
MPI Persistent Identifier eScience Seminar, Munich, March 2007



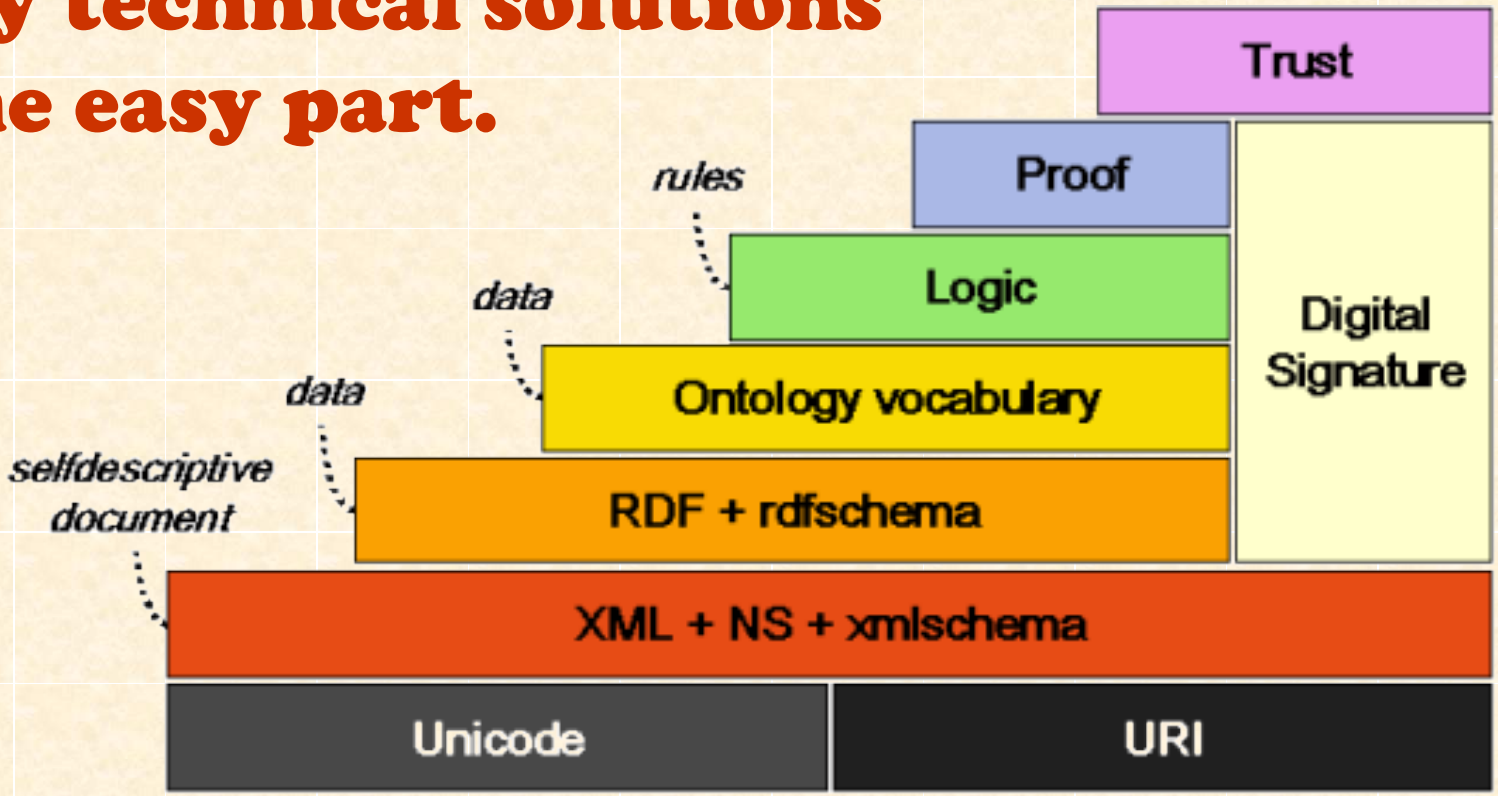
Web of Power & Impotence

- Metcalfe's law & various mathematical variants
- Indeterminacy: polysemy, homonymy, synonymy
- Issues of precision & recall
- Reliability, accuracy, relevance
- CLIR issues
- Unverified quality
- Inaccessibility of, unstable access to reliable resources



Securing the Trust Layer

Purely technical solutions are the easy part.

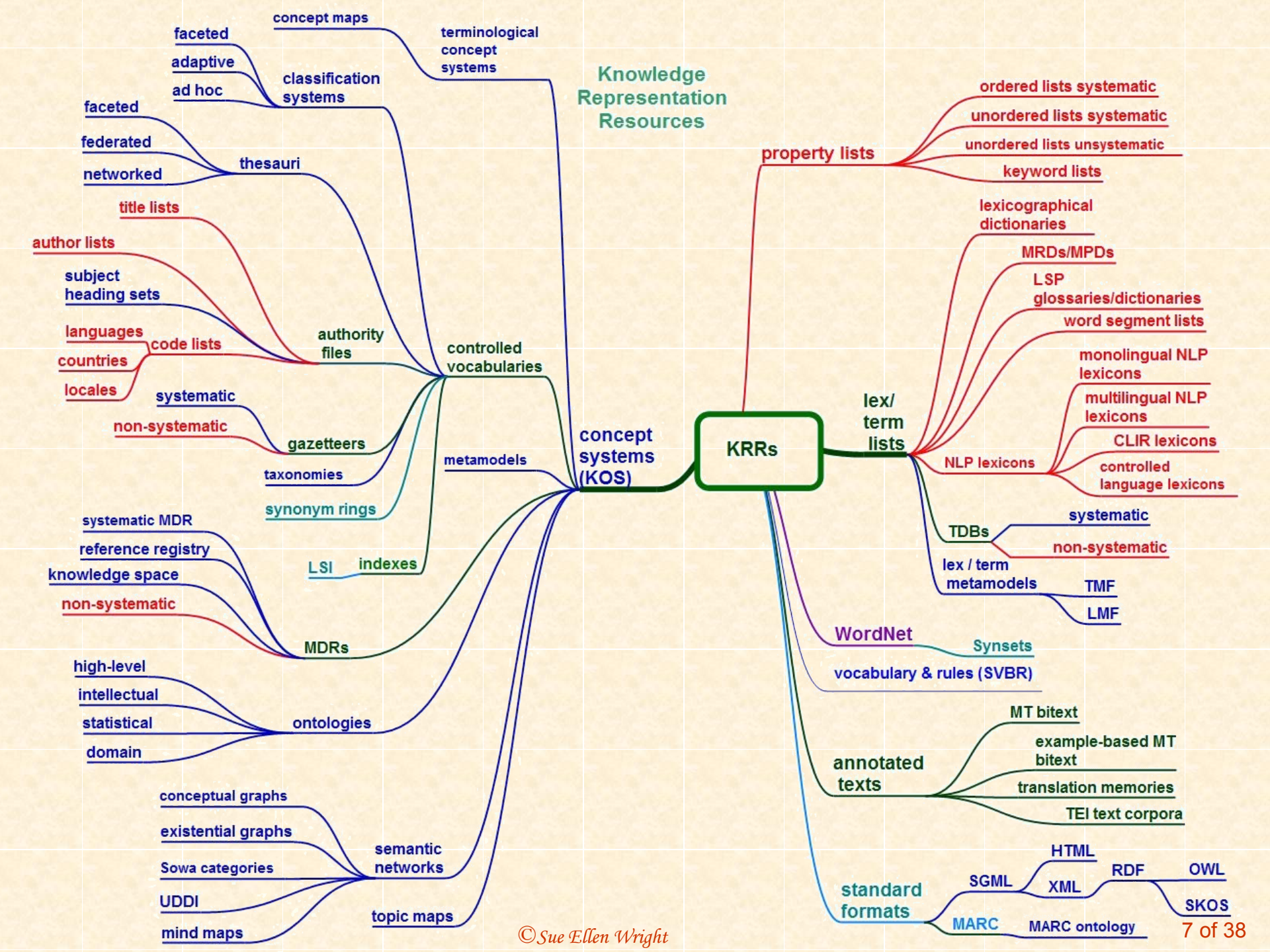


Web of (Dis)Trust

- The "*oh yeah*?" ("*says who*?") button
- The greatest mass is concentrated in un-substantiated link & tag space
  - Folksonomies, social networking
- Technical solutions (digital signatures, certificates, etc.)
- Tools for building data sets in authoritative repositories
- Subject experts create authoritative definitions in inaccessible frequently linguistic resources

Knowledge Representation Resources

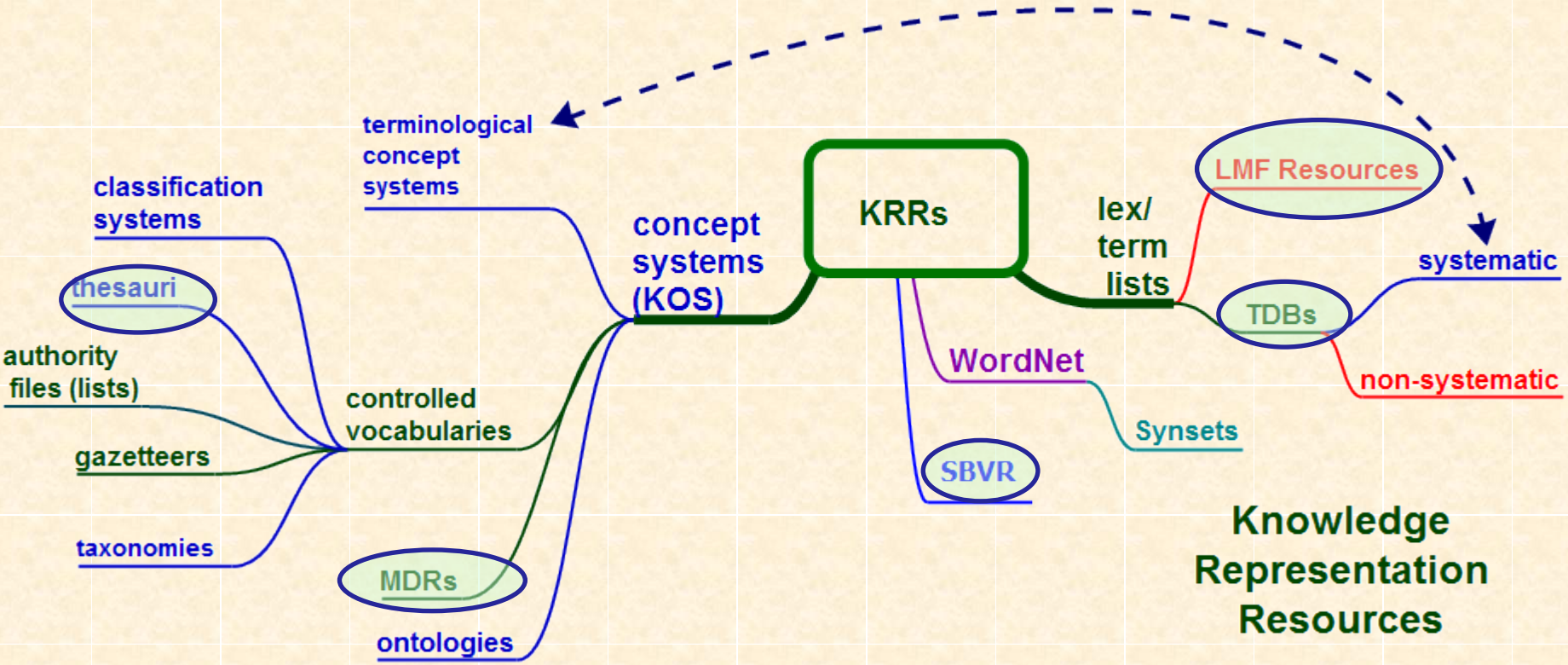
- **Resource that contains knowledge we can :**
  - **Manipulate**
  - **Mine or use to enrich other resources**
  - **Analyze and reuse (leverage)**
  - **Use to interact with various tools, either based on common environment planning (or not)**
- **Resources that are in themselves ontologically "organized" (systematic)**
- **Resources that contain latent organizational elements (non-systematic)**
- **Tagged or linked information, or not**



Colors

- Blue: systematic, represents shallow to deep semantic structures
- Red: non-systematic, primarily lists with random or conventional (e.g., alphabetical) ordering principles
- Green: hybrid superordinate nodes with both systematic and non-systematic children; texts of various kinds
- Purple: WordNet: internally hybrid system; shallow systematics, lexicographical approach

Resources of Interest for PIDs

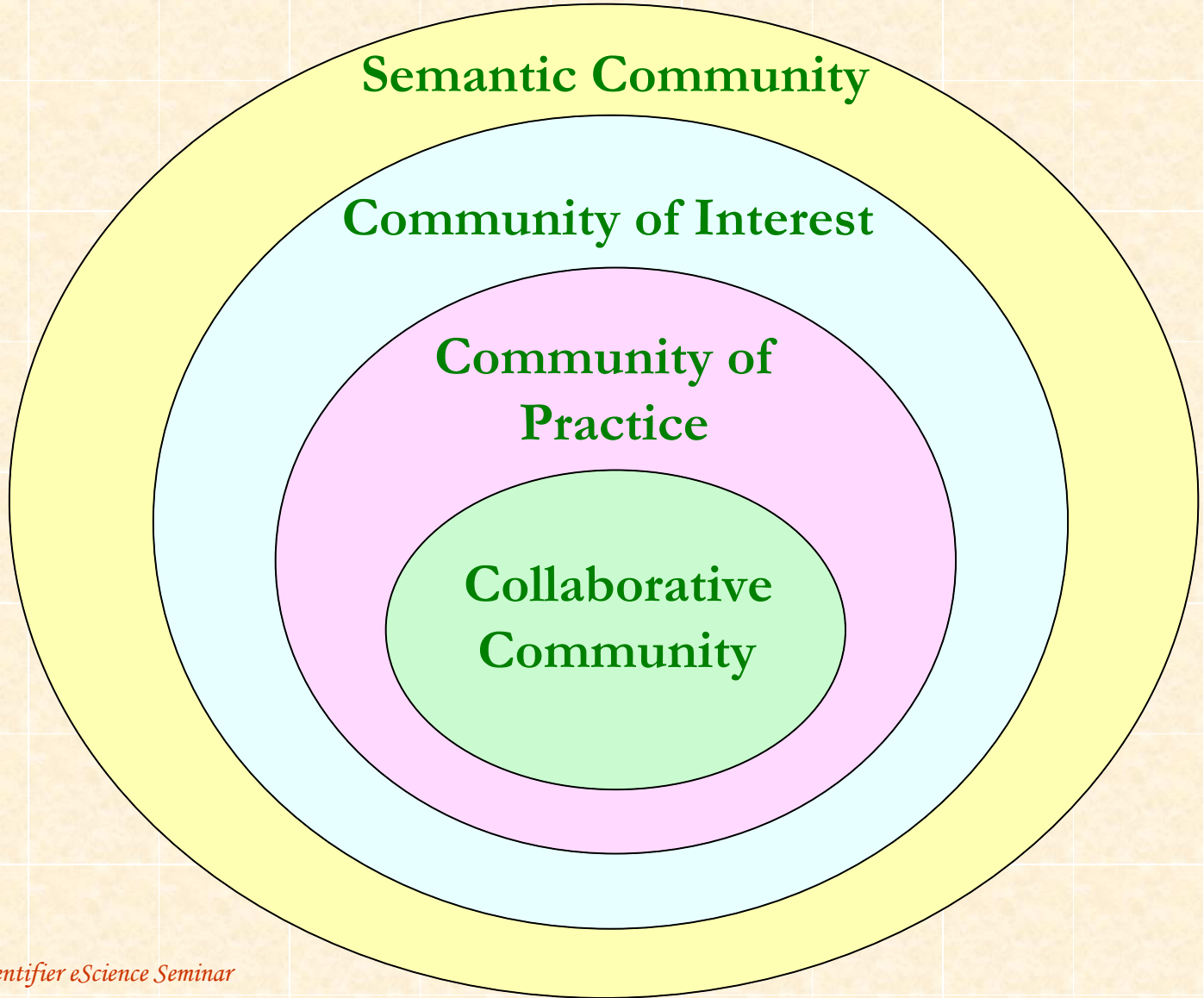


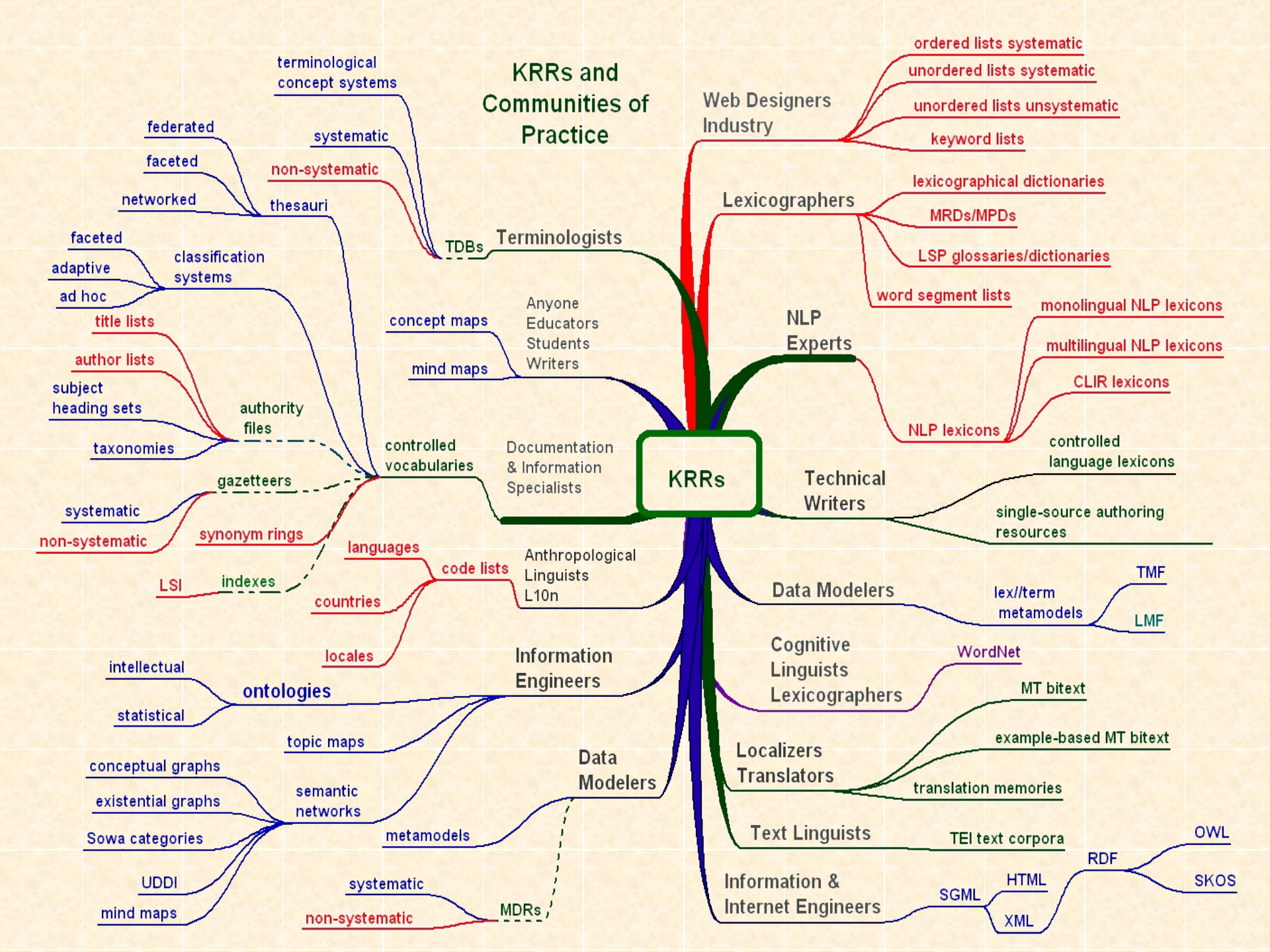
thesauri = subject language terminologies
TDB = discourse-oriented terminology
(terminology database)
MDR = metadata registry

Language Resources

- Controlled vocabularies (SKOS space)
- LMF = Lexical Markup Framework
  - ISO 24613
- Terminology Database (TDB)
  - ISO 30042, TermBase eXchange (TBX)
- Metadata Registries
  - ISO/IEC 11179 Data Element standards
  - ISO 12620 Data categories, ISO TC 37 Data Category Registry (ISOcat)
- Semantics of Business Vocabulary & Business Rules
  - ISO DIS 24707

Semantic Communities




KRRs and Communities of Practice
KRRs
Terminologists
TDBs
terminological concept systems
systematic
non-systematic
thesauri
federated
faceted
networked
classification systems
faceted
adaptive
ad hoc
controlled vocabularies
authority files
title lists
author lists
subject heading sets
taxonomies
gazetteers
systematic
non-systematic
synonym rings
LSI
indexes
Anyone
Educators
Students
Writers
concept maps
mind maps
Documentation & Information Specialists
Anthropological Linguists L10n
code lists
languages
countries
locales
Information Engineers
ontologies
intellectual
statistical
topic maps
semantic networks
conceptual graphs
existential graphs
Sowa categories
UDDI
mind maps
Data Modelers
metamodels
MDRs
systematic
non-systematic
Web Designers Industry
ordered lists systematic
unordered lists systematic
unordered lists unsystematic
keyword lists
Lexicographers
lexicographical dictionaries
MRDs/MPDs
LSP glossaries/dictionaries
word segment lists
NLP Experts
NLP lexicons
monolingual NLP lexicons
multilingual NLP lexicons
CLIR lexicons
Technical Writers
controlled language lexicons
single-source authoring resources
Data Modelers
lex//term metamodels
TMF
LMF
Cognitive Linguists Lexicographers
WordNet
Localizers Translators
MT bitext
example-based MT bitext
translation memories
Text Linguists
TEI text corpora
Information & Internet Engineers
SGML
HTML
XML
RDF
OWL
SKOS

Communities of Practice

- Much more diverse than the entities shown in this slide
- Slide entities reflect prototypical driving forces behind specific applications
- Example of oversimplification: language and locale standards
  - Librarians (catalogers, documentationalists)
  - Terminologists
  - Anthropological linguists, Bible translators
  - Dialect specialists
  - Localization engineers & software translators
  - Business-oriented knowledge & information specialists

Latent Subtexts

- Different communities of practice (CoPs) use different terms for the same concepts and the same terms for different concepts.
- KOS are defined differently by different CoPs.
- Some terms are used indiscriminately.
- Result: indeterminacy in the form of hidden polysemy and synonymy even in the designation & definition of KRRs.

Semantic Community

- Definition: community whose unifying characteristic is a shared understanding (perception) of the things that they have to deal with
- Two kinds of (business) Semantic Communities:
  - Collaborative Community (CC)
    - e.g., a department, cross-function programme team, an internal service
    - e.g., potentially broader: an internal or external collaborative team
  - Community of Practice (CoP)
    - e.g., project managers, operational excellence champions, departmental budget managers
    - potentially broader: CC + other stakeholders, interested colleagues in other working groups?
    - Where does Community of Interest (CoI) come in?
- Three scopes for Semantic Communities:
  - internal to an organization
  - among parts of different organizations
  - across a discipline or set of related disciplines, potentially on an international level

Point of View

- ***Déformation professionnelle*** is a French phrase, meaning a tendency to look at things from the point of view of one's own profession and forget a broader perspective. It is a pun on the expression "formation professionnelle," meaning "professional training." The implication is that all (or most) professional training results to some extent in a distortion of the way the professional views the world.

Problematic: *term & terminology*

- Controlled vocabulary (subject-language terminology):
  - One or more words designating a concept
  - A descriptor in a controlled vocabulary used for information management & retrieval
  - Terms in a controlled vocabulary
- TDB (discourse-oriented terminology)
  - One or more words designating a concept
  - A single or multiword string used to designate a concept in human spoken or written, generally uncontrolled, discourse
  - May be structured (concept system) or not

Crosswalks and Mashups

- Controlled vocabulary developers tend to use *terminology* as a synonym *thesaurus*, whereas terminologists consider terminologies and even terminological concept systems to be different from thesauri.
- There is a need to identify crosswalk nodes between the systems in order to achieve any degree of interoperability.

References

- Berners-Lee, Tim. 1997.Cleaning up the User Interface, http://www.w3.org/DesignIssues/UI.html
- Haendler, James, and Golbeck, Jennifer. 2007. Metcalfe's Law, Web 2.0, and the Semantic Web. Gophttp://www.cs.umd.edu/~golbeck/downloads/Web20-SW-JWS-webVersion.pdf
- Hilse, Hans-Werner, and Kothe, Jochen. 2006. Implementing Persistent Identifiers. CERL 2006. http://www.knaw.nl/ecpa/publ/pdf/2732.pdf
- Lee, Jae Sung. Nov. 27 2007. Ontology Work Guide for Web Contents -- Principle and Methods (proposal for new work item in ISO TC37)
- Koivunen, Marja-Riitta and Miller, Eric. 2001. W3C Semantic Web Activity *Semantic Web Kick-off Seminar in Finland Nov 2, 2001*
- Svenonius, E. (2000). The Intellectual Foundation of Information Organization: Digital Libraries and Electronic Publishing. Cambridge, Mass. MIT Press.
- ISO TC37/SC4
- Wittenburg, Peter, and Wright, Sue Ellen (Ed.) . 2007. Infrastructure Note on Registry Databases. ISO TC 37/SC4 Memo

Knowledge Representation Resources

Principle of Systematicity

- Systematic resources include explicitation of relationships (parent-child, meronymy & metonymy, sequentiality, defined edges in logical triads (RDF), etc.)
- Non-systematic resources are not ordered or are conventionally ordered (alphabetical dictionaries, non-mnemonic numerical sequences, etc.)
- Some resources (terminologies, metadata registries [MDRs], etc.) are manifested in both variations
- Degrees of systematicity

Degrees of Systematicity

- "Lightweight ontologies"*
  - Thesauri (subject-language terminologies)
  - Classification systems
  - Taxonomies
  - Discourse-oriented terminologies & concept systems (TDBs)
- "Heavyweight ontologies"

…

- ⁂Thesauri & terminological concept systems are not equated here with ontologies because they do not generally contain explicit rules for automatic analysis and processing.

- *Gómez-Pérez & Fernández-López

Problematic Terms: concept system

- concept system
  - Set of concepts structured according to relations among them (ISO 1087)
  - Vs. *concept scheme* according to SKOS, where relations are theoretically optional
  - Nearly synonymous with KOS
  - Terminological concept system: traditionally, visual representation of semantic relations within a defined subject field

Language Resources

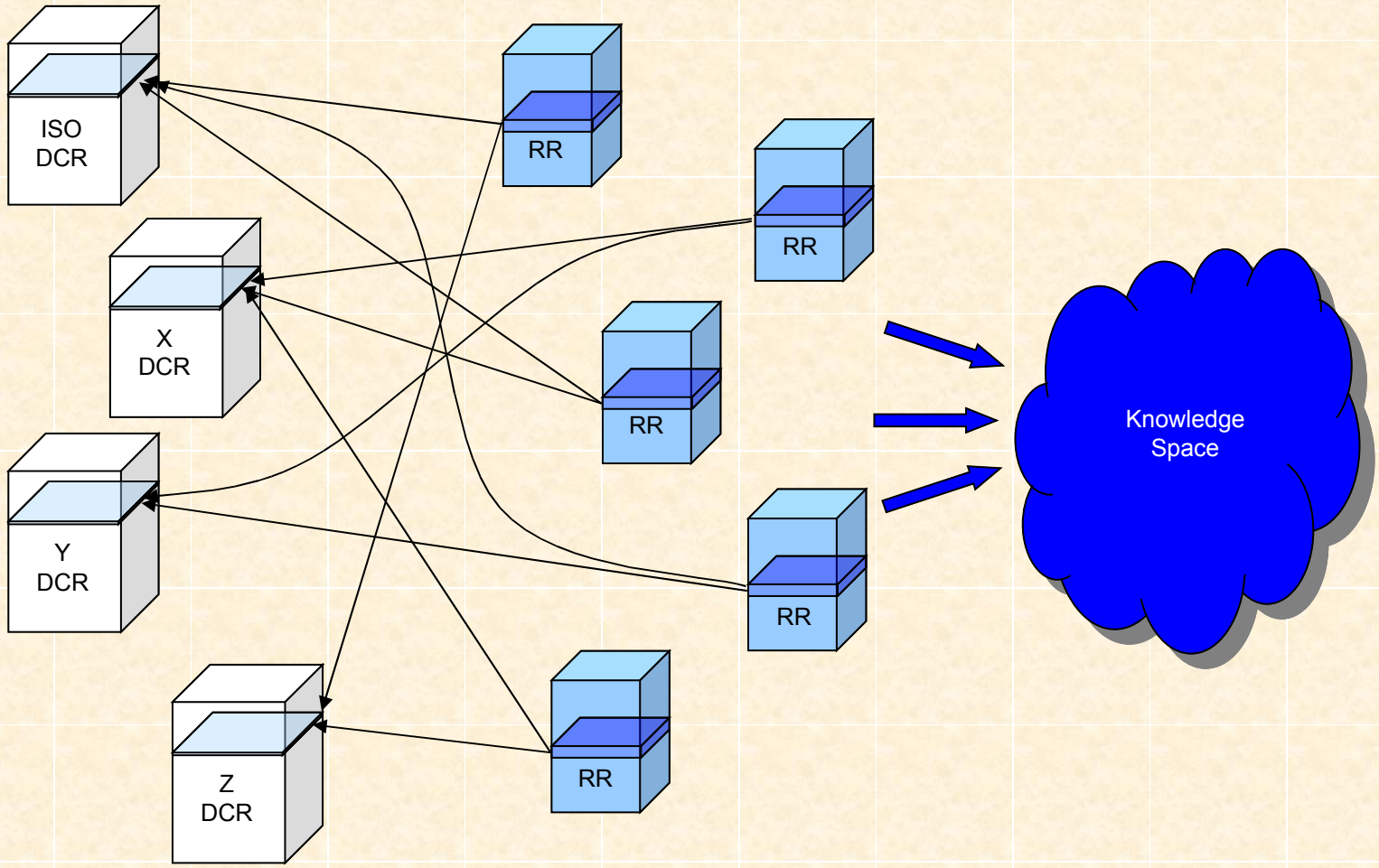
- What resources are we talking about?
- To what extent might they be useful in general web environments, particularly Web 2/Web 3 frameworks?
- What role might PIDs play in these resources?
- How might the availability of expert terminologies contribute to a trust layer in the evolving web environment?

Problematic Terms: concept system

- concept system
  - Set of concepts structured according to relations among them (ISO 1087)
  - Vs. *concept scheme* according to SKOS, where relations are theoretically optional
  - Nearly synonymous with KOS
  - Terminological concept system: traditionally, visual representation of semantic relations within a defined subject field

Problematic Terms

- ontology
  - Set of concepts structured according to relations among them (def. for *concept scheme/system*)
  - Prototypically systematic, sometimes with very deep structures (so-called high-level ontologies)
  - Some of whose nodes can be characterized by axioms or rules that can be used for machine processing of logical "intuitive" problems and queries



What does a 'Business Tool for Specifying Vocabulary & Policy/Rules' Contain?

Business Vocabulary + Rules

=

Business Glossary

(Noun Concepts, Definitions & Primary Terms)

(+)

Taxonomy

(General/Specific + Whole/Part Hierarchical Relationships)

+

Thesaurus

(Synonyms, Acronyms, Abbreviations, etc. + Multilingual)

(Instances of Concepts e.g. Business Events & Business Entities)

(Verb Concepts {Business Facts; Relations among Concepts})

+

Ontology

(Relations among Instances of Concepts)

(Definitional Rules)

(Definitions, Relationships & Rules specified in formal logic)

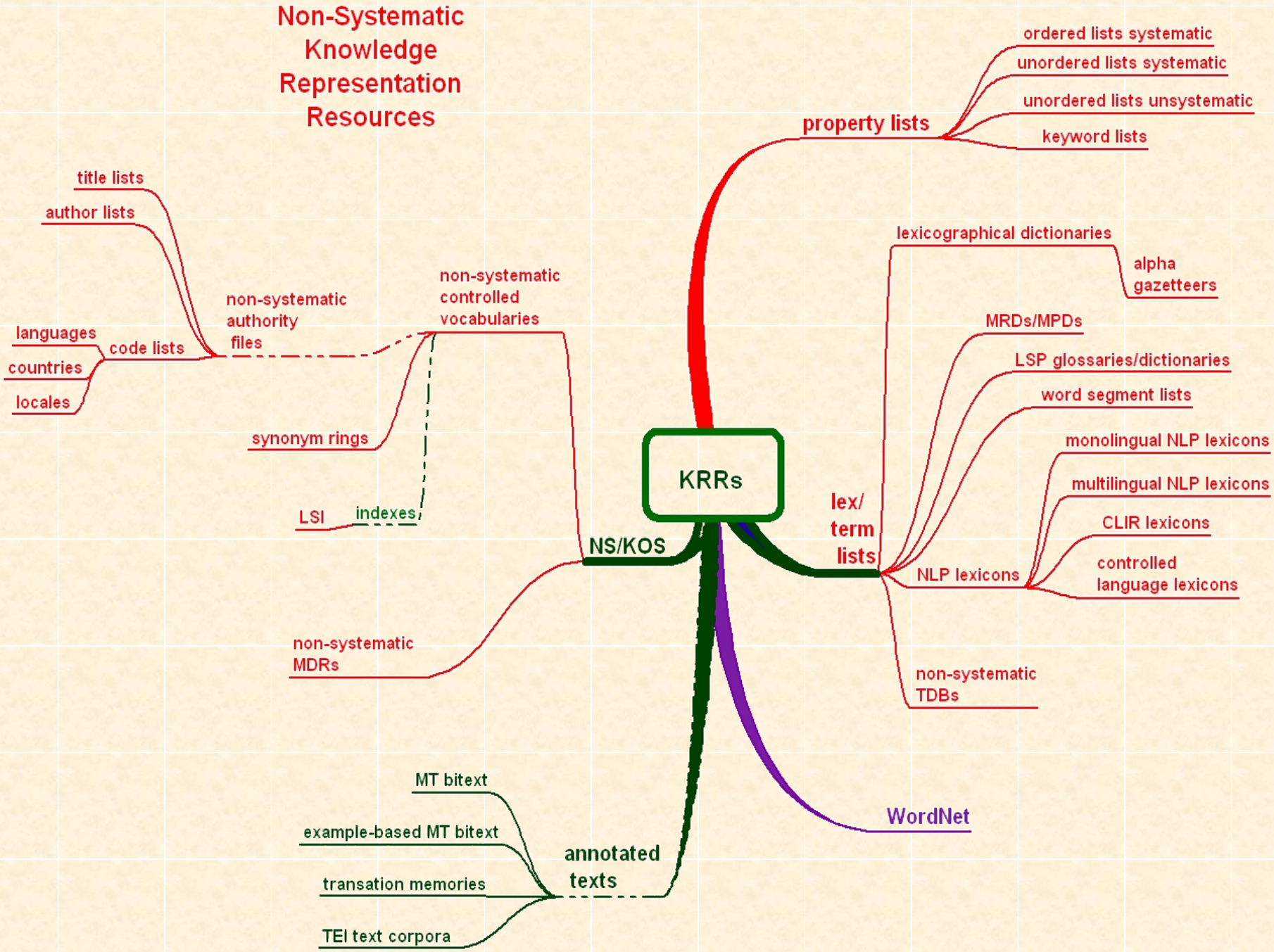
+

Business Rules

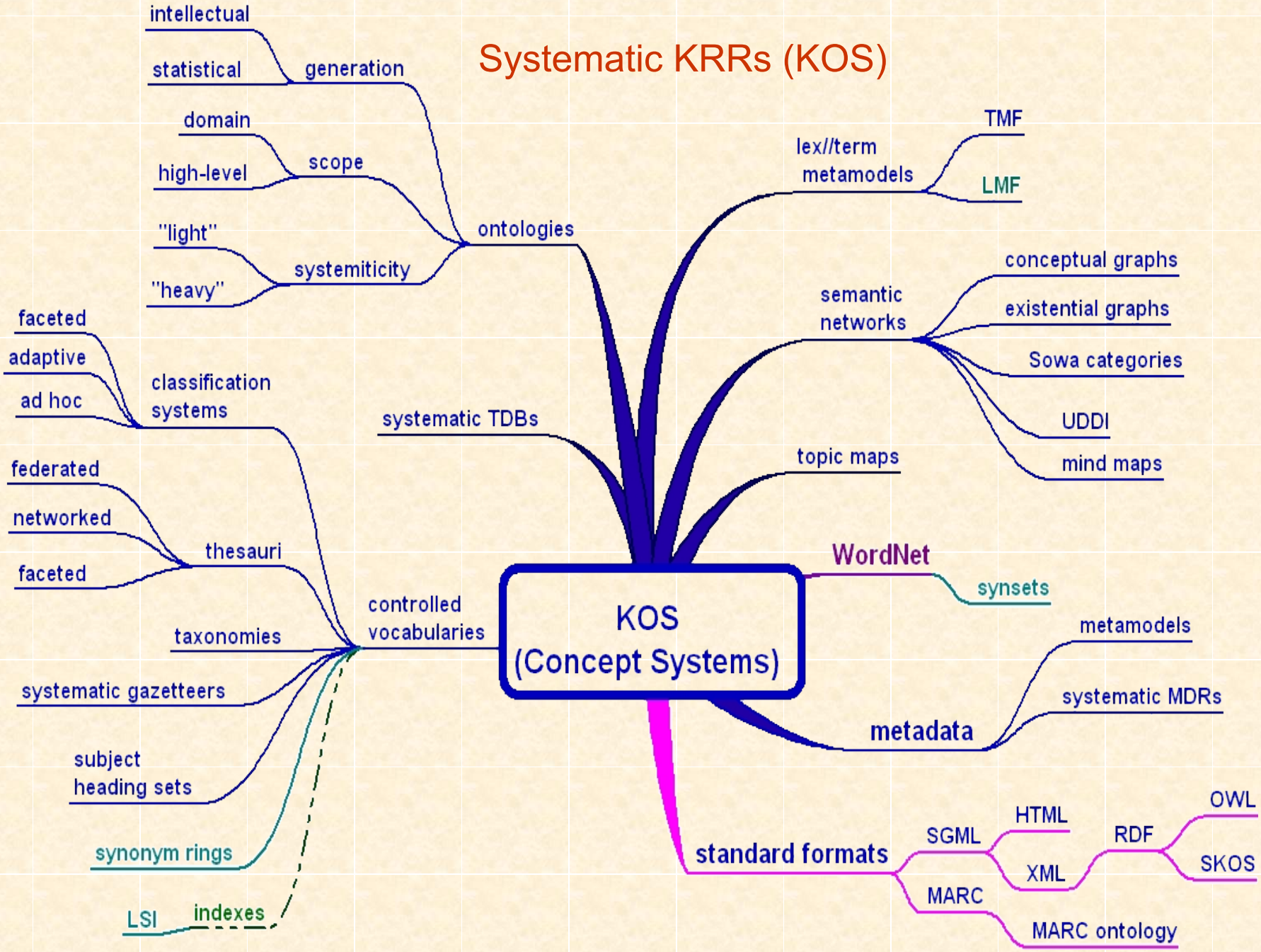
(Rules Governing Business Actions)

Potential for Faceting

- Communities of Practice
- Systematic resources
- Non-systematic resources
- Technology orientation (Semphuber)
- Degrees of indeterminacy
- Language & knowledge-oriented standards
- Standards bodies

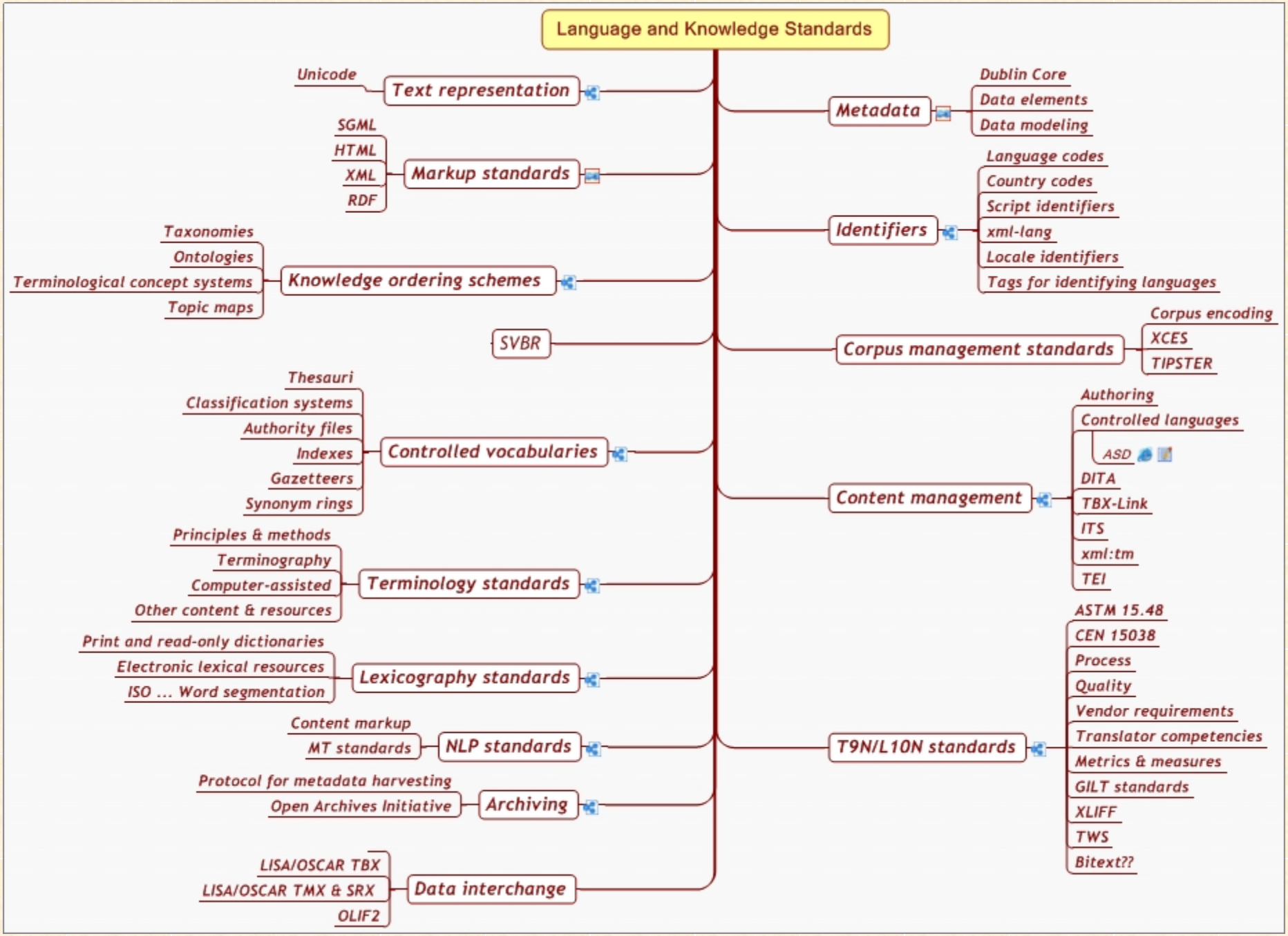

Non-Systematic
Knowledge
Representation
Resources
KRRs
property lists
ordered lists systematic
unordered lists systematic
unordered lists unsystematic
keyword lists
lexicographical dictionaries
alpha
gazetteers
MRDs/MPDs
LSP glossaries/dictionaries
word segment lists
monolingual NLP lexicons
multilingual NLP lexicons
CLIR lexicons
controlled
language lexicons
lex/
term
lists
NLP lexicons
non-systematic
TDBs
WordNet
NS/KOS
non-systematic
controlled
vocabularies
non-systematic
authority
files
title lists
author lists
code lists
languages
countries
locales
synonym rings
LSI
indexes
non-systematic
MDRs
annotated
texts
MT bitext
example-based MT bitext
transation memories
TEI text corpora

Systematic KRRs (KOS)



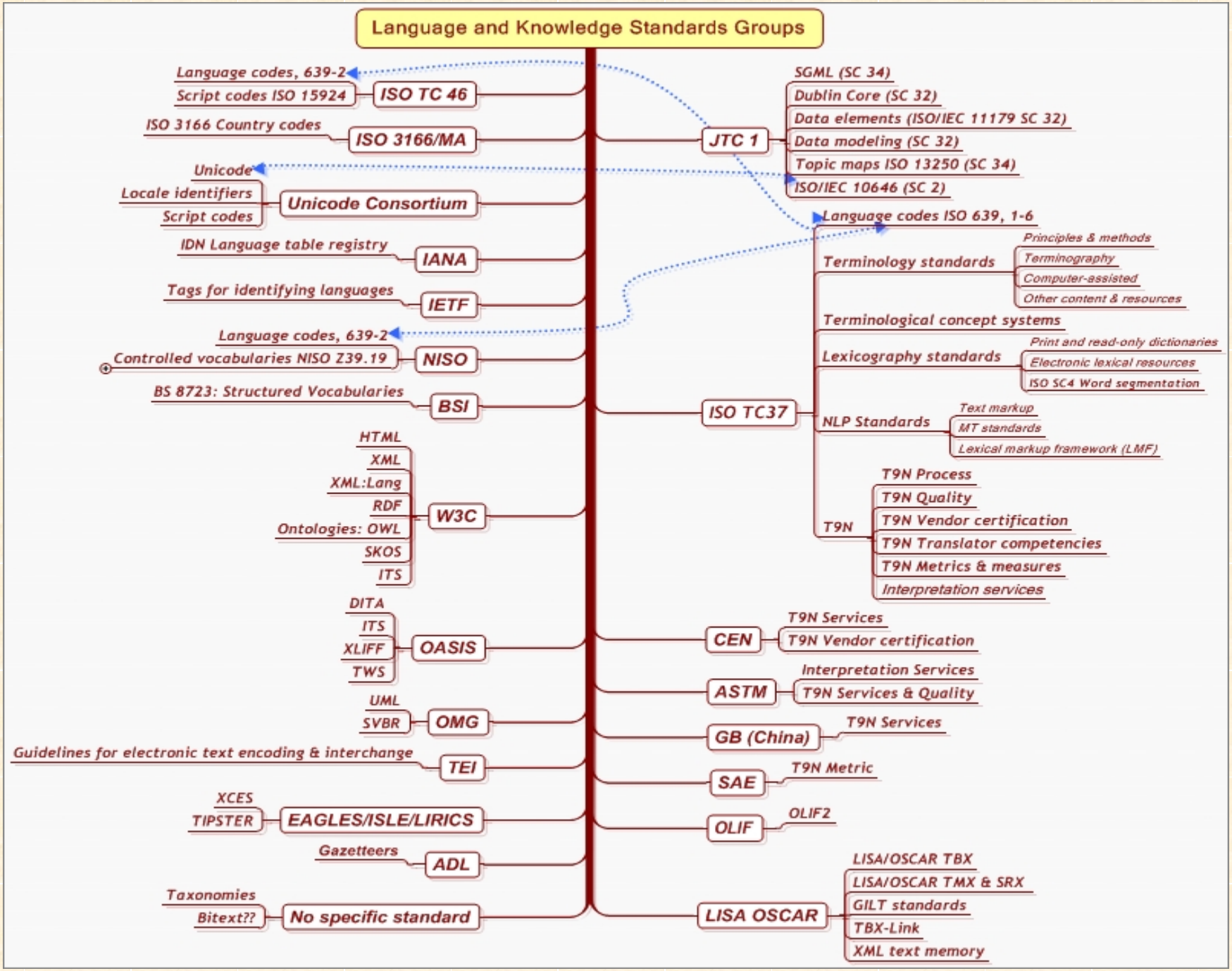
Standards for "Language Industry"

- Not as detailed on the KOS side of the equation because these slides cast an even broader net
- Focus on the so-called "language industry," much of which is potentially interesting for information science perspectives
- Major missing elements that need to be added on the KOS side


Language and Knowledge Standards
Text representation
Unicode
Markup standards
SGML
HTML
XML
RDF
Knowledge ordering schemes
Taxonomies
Ontologies
Terminological concept systems
Topic maps
SVBR
Controlled vocabularies
Thesauri
Classification systems
Authority files
Indexes
Gazetteers
Synonym rings
Terminology standards
Principles & methods
Terminography
Computer-assisted
Other content & resources
Lexicography standards
Print and read-only dictionaries
Electronic lexical resources
ISO ... Word segmentation
NLP standards
Content markup
MT standards
Archiving
Protocol for metadata harvesting
Open Archives Initiative
Data interchange
LISA/OSCAR TBX
LISA/OSCAR TMX & SRX
OLIF2
Metadata
Dublin Core
Data elements
Data modeling
Identifiers
Language codes
Country codes
Script identifiers
xml-lang
Locale identifiers
Tags for identifying languages
Corpus management standards
Corpus encoding
XCES
TIPSTER
Content management
Authoring
Controlled languages
ASD
DITA
TBX-Link
ITS
xml:tm
TEI
T9N/L10N standards
ASTM 15.48
CEN 15038
Process
Quality
Vendor requirements
Translator competencies
Metrics & measures
GILT standards
XLIFF
TWS
Bitext??

Standards Bodies

- ISO family of standards bodies
- National bodies (NISO, DIN, BSI, etc.)
- Web-oriented standards bodies (W3C, IETF, Unicode, etc.)
- Industry standards (OMG, OASIS, LISA, etc.)
- Professional organizations (ATA, FIT, etc.)
- Research groups and grant teams
- Others?


Language and Knowledge Standards Groups
Language codes, 639-2
Script codes ISO 15924
ISO TC 46
ISO 3166 Country codes
ISO 3166/MA
Unicode
Locale identifiers
Script codes
Unicode Consortium
IDN Language table registry
IANA
Tags for identifying languages
IETF
Language codes, 639-2
Controlled vocabularies NISO Z39.19
NISO
BS 8723: Structured Vocabularies
BSI
HTML
XML
XML:Lang
RDF
Ontologies: OWL
SKOS
ITS
W3C
DITA
ITS
XLIFF
TWS
OASIS
UML
SVBR
OMG
Guidelines for electronic text encoding & interchange
TEI
XCES
TIPSTER
EAGLES/ISLE/LIRICS
Gazetteers
ADL
Taxonomies
Bitext??
No specific standard
SGML (SC 34)
Dublin Core (SC 32)
Data elements (ISO/IEC 11179 SC 32)
Data modeling (SC 32)
Topic maps ISO 13250 (SC 34)
ISO/IEC 10646 (SC 2)
JTC 1
Language codes ISO 639, 1-6
Terminology standards
Principles & methods
Terminography
Computer-assisted
Other content & resources
Terminological concept systems
Lexicography standards
Print and read-only dictionaries
Electronic lexical resources
ISO SC4 Word segmentation
NLP Standards
Text markup
MT standards
Lexical markup framework (LMF)
T9N
T9N Process
T9N Quality
T9N Vendor certification
T9N Translator competencies
T9N Metrics & measures
Interpretation services
ISO TC37
CEN
T9N Services
T9N Vendor certification
ASTM
Interpretation Services
T9N Services & Quality
GB (China)
T9N Services
SAE
T9N Metric
OLIF
OLIF2
LISA OSCAR
LISA/OSCAR TBX
LISA/OSCAR TMX & SRX
GILT standards
TBX-Link
XML text memory

Outlook

- Incorporating all types of knowledge resources into a single system is a daunting task.
- The inclusion or exclusion of certain resources is very much subject to personal opinion.
- The exclusion of resources may mean loss of interoperability in the future.
- Opportunities for faceting are numerous and highly interesting.

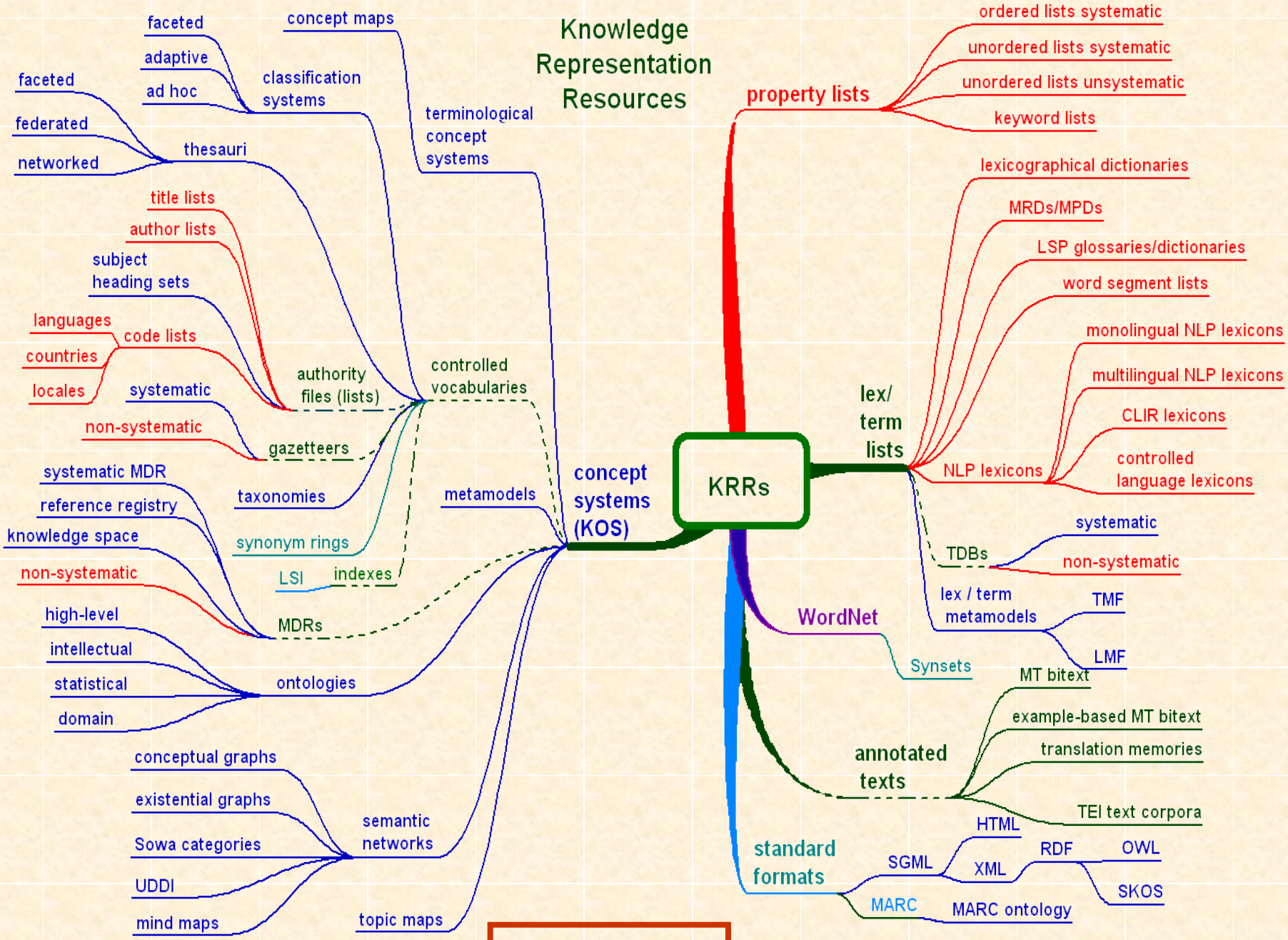
Do *YOU* have any answers?

- What do we need to add?
  - **Resource types?**
  - **Communities of practice?**
  - **Standards?**
  - **Standards bodies?**
  - **What kinds of crosswalks ...**
    - **Are the most important?**
    - **Are the most challenging?**

For More Information

- Sue Ellen Wright
  Institute for Applied Linguistics
  Kent State University
  109 Satterfield Hall
  Kent, Ohio 44242, USA

  sellenwright@gmail.com


Knowledge Representation Resources
KRRs
concept systems (KOS)
terminological concept systems
concept maps
classification systems
faceted
adaptive
ad hoc
thesauri
faceted
federated
networked
controlled vocabularies
authority files (lists)
title lists
author lists
subject heading sets
code lists
languages
countries
locales
gazetteers
systematic
non-systematic
taxonomies
synonym rings
indexes
LSI
metamodels
MDRs
systematic MDR
reference registry
knowledge space
non-systematic
ontologies
high-level
intellectual
statistical
domain
semantic networks
conceptual graphs
existential graphs
Sowa categories
UDDI
mind maps
topic maps
property lists
ordered lists systematic
unordered lists systematic
unordered lists unsystematic
keyword lists
lex/ term lists
lexicographical dictionaries
MRDs/MPDs
LSP glossaries/dictionaries
word segment lists
NLP lexicons
monolingual NLP lexicons
multilingual NLP lexicons
CLIR lexicons
controlled language lexicons
TDBs
systematic
non-systematic
lex / term metamodels
TMF
LMF
WordNet
Synsets
annotated texts
MT bitext
example-based MT bitext
translation memories
TEI text corpora
standard formats
HTML
SGML
XML
RDF
OWL
SKOS
MARC
MARC ontology
2007-11-10