Constructing the InterPARES thesaurus: Challenges and opportunities

Jonathan Furner

Four terminological tools

- Register
- Dictionary
- Glossary
- Thesaurus

- four views of one database

Purpose

- to improve the effectiveness of IP-related communication and dissemination
- by helping authors, readers, and intermediaries to select and use terms **consistently**
- thus increasing the likelihood ...
  - that authors will use terms in the manner expected by their readers: effective **authorship**
  - that readers will find documents of the kind they want: effective **retrieval**
  - that readers will interpret documents in the manner intended by their authors: effective **understanding**

Some general challenges

- potential **conflict** of purpose: is partly descriptive, partly prescriptive
  - founded on current usage ...
  - ... but specifying preferred future usage
- **specificity** of (often multiple) senses of technical terms
- **domain**-specific variation
- **region**-specific variation
- facet analysis

Terms, qualifiers, definitions, concepts

- a qualified term is made up of ...
  - one **term**: e.g., "record"
  - three **qualifiers**:
    - a part of speech: e.g., "noun"
    - a facet: e.g., "object"
    - a context: e.g., "archival science"
- a qualified term + a **definition** = a **concept**
- a single term (homonym) may have multiple definitions
- different terms (synonyms) may have the same definition
- aim: disambiguate homonyms and collocate synonyms

Dictionary records: Example

Term	PoS	Facet	Context	Definition	Source
record	n.	object	arch.sci.	A document created or received and maintained by an agency, organization, or individual in pursuance of legal obligations or in the transaction of business.	IP
record	n.	object	arch.sci.	Document created by a physical or juridical person in the course of practical activity.	UBC
record	n.	object	comp.sci.	A grouping of interrelated data elements forming the basic unit of a file	SAA

Dictionary

- vocabulary:
  - IP- “accepted” terms
  - one or more definitions per qualified term
  - one record per definition
- arrangement:
  - alphabetical
- data
  - definitions + sources

Glossary records: Example

Term	PoS	Facet	Context	Definition
record	n.	object	arch. sci.	A document created or received and maintained by an agency, organization, or individual in pursuance of legal obligations or in the transaction of business.
record	n.	object	comp. sci.	A grouping of interrelated data elements forming the basic unit of a file

Glossary

- vocabulary:
  - IP- “preferred” terms
  - one definition per qualified term
  - one record per qualified term
- arrangement:
  - alphabetical
- data:
  - IP- “preferred” definition

Thesaurus records: Example

Qualified term	Link type	Linked term	Subfacet
record	BTG	document	
active record	BTG	record	level of activity
reliable record	BTG	record	level of reliability

Thesaurus

- vocabulary:
  - IP- “accepted” terms (cf. Dictionary)
- arrangement:
  - faceted and hierarchical
- data:
  - links among paradigmatically-related terms
    - hierarchical
      - genus-species
      - whole-part
      - class-instance
    - from IP-nonpreferred to IP-preferred

Current facets

- agents
  - "juridical person"; "creator"
- objects
  - "fonds"; "active record"
- properties
  - "authenticity"; "date of receipt"
- actions
  - "authentication"; "emulation"
- disciplines
  - "archival science"; "records management"

Register record: example

Term	record
PartOfSpeech	n
Facet	object
Context	archival science
RegisterEntry:CreatedDate	20030901
RegisterEntry:CreatedBy	jpm
RegisterEntry:LastModifiedDate	20030901
RegisterEntry:LastModifiedBy	jpm
Dictionary:CreatedDate	20031001
Dictionary:CreatedBy	ymc
Dictionary:LastModifiedDate	20031001
Dictionary:LastModifiedBy	ymc
Dictionary:DeletedDate	
Dictionary:DeletedBy	
GlossaryEntry:CreatedDate	20031001
GlossaryEntry:CreatedBy	ymc
GlossaryEntry:LastModifiedDate	20031001
GlossaryEntry:LastModifiedBy	ymc
GlossaryEntry:ApprovedDate	20031101
GlossaryEntry:ApprovedBy	ld
GlossaryEntry:DeletedDate	
GlossaryEntry:DeletedBy	
Thesaurus:CreatedDate	20040101
Thesaurus:CreatedBy	nr
Thesaurus:LastModifiedDate	20040101
Thesaurus:LastModifiedBy	nr
Thesaurus:ApprovedDate	20040201
Thesaurus:ApprovedBy	ld
Thesaurus:DeletedDate	
Thesaurus:DeletedBy	

Register

- **vocabulary**: all terms considered for acceptance into Dictionary
    - terms selected manually from automatically-created concordance of IP documents (case-study reports, data models, etc.)
    - "entailed" terms extracted from Glossary definitions and suggested by Thesaurus structure
    - terms selected manually from non-IP term lists
    - terms suggested by IP researchers
  - homonyms disambiguated by part-of-speech, facet, and contextual qualifiers
  - new terms added continuously as set of source documents grows
- **arrangement**: alphabetical
- **data**: dates and agents of creation, modification, approval, deletion of entries in Dictionary, Glossary, Thesaurus

Some specific challenges

- selection of additional vocabulary
  - which sources should be consulted?
  - what procedures and criteria should be used?
  - should composite terms be selected?
- establishing consistency among products of various IP groups
  - how?
- assignment of terms to facets
  - is the selection of facets useful and/or appropriate?
- categorization of terms within facets
  - within each facet, what set of categories would be useful and/or appropriate?

Some specific challenges, cont'd

- reconciliation of conflicting recommendations, in the various standards, as to form of terms
- maintenance of integrity of terminological database
- evaluation
  - how are tools to be evaluated w.r.t.
    - complying with international standards?
    - meeting users' requirements?