Thesaurus use Cases and Maintenance (an ADS prospective)

Alberto Accomazzi
NASA Astrophysics Data System
Harvard-Smithsonian Center for Astrophysics

21 May 2012
IVOA Spring Interop, Urbana-Champaign

The ADS tower of Babel

- As an aggregator of bibliographic content, ADS deals with different keyword systems:
  - NASA/STI (1975-2000)
  - A&A keywords (1992-current)
  - PACS (mostly physics journals)
  - IAU Thesaurus (virtually never used)
  - Uncontrolled list (mostly user-submitted, arXiv)
- Only recent literature has keywords

Astronomy Journal Keywords

- Developed in 1992 by editors of major astronomy journals
- 374 items, of which 13 are top concepts
- Hierarchical categories, no definitions or synonyms
- Purpose is to tag and classify papers
- Well-known to scientists, editors, is updated ~ every 2 yrs
- Example:
  galaxies: active,
  quasars: emission lines,
  quasars: individual: PG 1416-129

Astronomy Journal Keywords

Topics	Nkwds
General	9
Physical Data and Processes	47
Astronomical Instrumentation, Methods and Techniques	32
Astronomical Databases	5
Astrometry and Celestial Mechanics	9
The Sun	31
Planetary Systems	30
Stars	69
Interstellar Medium (ISM), Nebulae	21
The Galaxy	19
Galaxies	46
Cosmology	14
Resolved and Unresolved Sources as a Function of Wavelength	42
Total	***374***

PACS for Astronomy/Astrophysics

- Subset of the PACS keyword system
- 589 items, of which 6 are top categories
- Hierarchical structure four levels deep
- Used mostly by some physics and planetary journals, updated ~ every 2 yrs
- Example:
  98.62.Ra - Intergalactic matter; quasar absorption and emission-line systems; Lyman forest

PACS for Astronomy/Astrophysics

Topics	Nkwds
04 - General Relativity & Gravitation	42
26 - Nuclear Astrophysics	20
95 - Fundamental astronomy and astrophysics; instrumentation, techniques, and astronomical observations	60
96 - Solar system; planetology	298
97 - Stars	61
98 - Stellar systems; interstellar medium; galactic and extragalactic objects and systems; the Universe	108
Total	***589***

IAU Thesaurus

- Published in 1993 by Librarians, sanctioned by the International Astronomical Union (IAU)
- 2,551 items, of which 274 are top concepts
- Rich set of relationships: BT, NT, RT, U, and UF (but no real definitions)
- Multilingual supplement published in 1995 (French, German, Italian and Spanish)
- Has not been updated since publication, rarely used, BUT...
- Used to "seed" the IVOAT

IAU Thesaurus

Example

quasar
- ALT
  - "QSO"
  - "quasi-stellar galaxy"
  - "quasi-stellar object"
  - "quasi-stellar radio_source"
- D
  - "quasar"
- BT
  - extragalactic_object
  - extragalactic_radio_source
- NT
  - X-ray_quasar
  - radio_quiet_quasar

RT
- BL_Lacertae_object
- Lyman_alpha_forest
- X-ray_source
- absorption_line_system
- active_galactic_nucleus
- blazar
- compact_galaxy
- double_quasar
- liner_galaxy
- luminous_arc
- quasar_galaxy_pair
- quasar_microlensing
- variable_source

IVOA Thesaurus

- On-going effort within the International Virtual Observatory Alliance (IVOA)
- Developed as SKOS document from the get-go
- Leverages on the IAU thesaurus, with significant updates, corrections (mostly by Rick Hessman)
- Current draft contains 2890 concepts, including 246 new additions and 120 modifications
- Hope is that community will pick up its maintenance as its usefulness is demonstrated in VO applications, services

How we deal with this

- Current use of keywords in ADS for classification:
  - Mapped keywords from STI, A&A and uncontrolled lists to a set of "normalized" keywords
  - Used PACS keywords as they are for physics content
  - We do not make use of hierarchy, relatedness
- Use of related terms:
  - We make use of synonyms (individual words) to enhance search when indexing title, abstract, keywords
  - No support for hypernyms or hyponyms

What to do with the UAT (1)

- Text mining
  - Search the full-text documents in ADS for terms in the thesaurus (preferred + alternate)
  - Assign terms from UAT to all documents in ADS, including historical literature
- Enhanced search and indexing
  - Incorporate current synonyms in thesaurus and use to expand search
  - Improve search recall and precision

What to do with the UAT (2)

- Document Classification
  - Map A&A keywords onto UAT terms
  - Create “Fingerprinting” of papers based on concepts
  - Support recommendations, notifications
- Faceted filtering & browsing of documents
  - Narrow search results by focusing on concepts
  - Expand search by selecting broader concepts

More in general

- UAT one component in larger astronomy Linked Data Ecosystem
- Link additional authority files to UAT concepts, e.g.
  - Object types
  - SIMBAD, NED object names
  - Facilities, Telescopes, Instruments
  - Funding agencies, grant numbers
  - Institutions, Collaborations

The opportunity

- Convergence of goals by several stakeholders
  - AIP, IOPP (AAS?)
  - ADS & IVOA
- Existing Thesauri being formalized / updated
  - IOPT, AIPT
  - IVOAT
- People ready to work on this
  - ADS in collaboration with CfA library
  - The usual suspects that brought you IVOAT

The Challenge

- Process
  - We want this to be a community effort, but...
  - We need clarity on editorial responsibilities, policies
- Unification and Maintenance
  - Data Harmony or Poolparty possible solutions (suggestions welcome)
  - Periodic updates to include terms from search logs, text mining
- IP and legalese: being worked out
  - IOP, AIP donating their work to ADS
  - ADS + IVOA merge with IVOAT, puts UAT in the public domain