BHL/EOL Name Finding

3/11/09

Demo From Remsen

* + Search "Neoplasms" in MedLine to get all, then walk down the tree
  + Heirarchical searching/recovery
  + People Search with high level common names
  + BHL Literature is 'dry as dust', primarily used by scholarly users
  + Be able to filter across all books, by one book
  + Value in doing this across all
  + How much will be machine-based vs. human based
  + These mega searches will blow already lengthy results way out
  + Can choose which classification to use to view results
* Taxonomist Use Case
  + Find me everything, then filter results by nomenclatural acts
* No Single canonical hierarchy
  + What is classification?
  + How will we/EOL deal with alternate classifications?
    - Compiling an array of classifications in EOL (Patrick)
      * IP issues; Owners may not allow use
    - Currently aggregate results using standalone, siloed classifications
      * Vision to build consensus structure
      * Will need engagement of experts, many years
      * “This classification”, “That classification”, “Consensus/Aggregated”, “Curated”
  + Want to view BHL results for each silo
* EOL will develop a service
  + BHL need hold no synonymy or classification
  + We give them a name, they give us an expanded search
    - Browse hierarchy where only BHL content exists
* Is it acceptable to send these results to a user instead of accessing?
  + Taxonomist won’t necessarily
* Establish a ‘standing search’, then get alerts when new content is found
  + RSS feed
* Virtually no feedback at from EOL users about BHL
  + EOL is not much of an attractive environment
  + BHL is most expert part
  + Taxonomist would want to throw away anything that isn’t a nomenclatural act
    - Create the ‘gems’ and treat them as standalone objects
* How will ‘Passiflora’ classification from LifeDesk get into OEL
  + 1) export my classification into a repository and I can use that for my default browse
  + 2) Composite is created and any genus/species I have are algorithmically added to CoL
    - EOL is currently working on this
    - You get an alert “This name isn’t in Catalogue of Life but this name is in this family”
  + 3) Then I edit this structure as an expert; tidy up the environment in LifeDesk
* Need delivery dates for these
  + Put it into Jira
  + Don’t have enough specificity now
  + Need to create a ‘domain’ in Jira and subdivide
    - Responsibility for EOL to do
    - BHL says “We need this functionality by this date.”
      * Patrick rips out hair, figures out who & what is needed
* Action Item: BHL to describe high level description of today’s discussion and input into Jira; determine a date by which we need this
  + Will need specific technical requirements worked out with Patrick & BHL Dev Team
* Kickoff meeting with BHL-Europe, would be nice to

TF vs. TF2J

* Complementary name finding
* Do we run two tools against the literature, or make a best of breed for one?
  + Ubertool? Or 2 tools?
* Ahmed will have eval of TF2J by the end of the week
* Parallel developments aren’t bad
  + TAXAMATCH
  + TF
  + TF2J
* EOL has an agenda with GBIF
* Nomina Meetings
  + Meetings of developers
  + Want people to think about how to find names
  + Done before June 30?
  + Tony Rees, Patrick, GBIF
  + Could we do something around eBiosphere?
* Action Item: Organize meeting around NameFinding (with vernaculars) & NameReconcilliation before June 30th
  + Tom, Paddy, Chris
  + Determine invitees - Chris
  + Location? – Chris check on STL
  + Define budget – Tom & Paddy
  + 3 days
    - Day 1
      * show wares, hot air
    - Day 2-3
      * Enumerate visions, possibilities

Nomenclatural Acts

* Vision: assign a probability to a block of text being a nomenclatural act, such that it is highlighted in search results and maybe incorporated into LifeDesk/EOL objects
  + “nov. sp.”, “n. sp.”
  + 1) through algorithms/NLP
  + 2) through nomenclators
    - Tropicos knows this is the protologue for a name
  + 3) scholarly annotation
    - Pentcheff knows this is the article for his species
* GNI/GNA will include original descriptions when brought in from nomenclators like
* Zoology
  + Less standardization between ‘territories’
  + Style remains the same
    - Formal statement of what distinguishes this taxon from others
      * Could be 5 paragraphs, or 3 words
* Botany
  + Specific block of text defined by coordinate system
* Goal: Get to the Article!
  + Then, get to start of description
  + Extension: then, use highlighter to draw box around nomenclatural act
* No code for viruses
* Action Item: Review search interface to make it more article-friendly with existing metadata present in BHL
* Action Item: Tom to describe this issue in Jira

Taxonomic Acts

* More difficult to determine/find
  + Should be considered a next step after Nomenclatural Acts
* People make judgements on the validity of the species
* Taxonomic revisions
  + Of high value
  + But maybe hard to find
* “Forensic taxonomy”
  + equal weight for all taxonomic documents
    - 1830-now, all weighted equally
  + VS. looking for most recent revisionary documents
    - Most taxonomists will have this information at hand for their area of specialty
* Keywords for finding
  + “taxonomic revision”
  + others
* Implementation TBD, following Nomenclatural Acts

Automated Markup

* BHL needs to be able to identify the semantic & structural content of its corpus
  + Structural - Articles
  + Semantic – meaningful components of literature
* Don’t necessarily have to have article boundaries to get semantic
* OCR is a major barrier
* Wiki is a very real possibility
  + Might be slow uptake, but still involves community
  + Identify which pages have images
  + How do you get people to get engaged in this activity?
* Action Item: Chris to pick back up with reCAPTCHA
* Action Item: Tom to discuss OCR rekeying with Chinese Academy of Science
* Action Item: Chris & Mike L, with Patrick, to think about/look at leaving tags for names in OCR
* Action Item: Phil & Chris M to reevaluate wiki for rekeying/markup