Tools To Improve Scholarly Metadata
Promising solutions for improving the accuracy of scholarly publishing metadata

Marjorie M.K. Hlava
Access Innovations, Inc
505-975-5578
Accessinn.com
Promising solutions for improving the accuracy of scholarly publishing metadata

• Standards  Check out the NISO Library on their web site
• Consortia for clean data
• Share, check, and enhance metadata
• Automate as much of manuscript submission & peer review as possible
• Clean up the Author Synonymy
• Enhance scholarly and learned publishing and the research communities it represents worldwide
• THE GOAL
• High integrity, accurate, consistent content
But How About Improving The **Content** Itself?

- Series of Metadata Filters and Enrichment
  - The varying names used in the content of the publication
  - Gene names – 19 or more synonyms per name
  - Medicinal Plant names – nearly 17 synonyms per name
  - Bad Cell Line references
  - Suspect Science topics / Fake news
- Semantic enrichment supports metadata and search
- Time savings for researchers both authors and readers
- It allows the disambiguated information in the formation of a platform for better science
- Being able to reference a widely available authoritative source is crucial to all world health
Every Walk of Life Uses
Constantly Changing Vernacular

- Homeless
- Unsheltered
- Unhoused
- Street people
- Hobos
- Vagrants
- Metastatic breast cancer
- Stage IV Breast Cancer
- Invasive Breast Cancer
- Covid-19
- Coronavirus
- SARS-CoV-2
- Omicron
- BA.4
- BA.5
- Taxonomy
- Ontology
- Thesaurus
- Knowledge Map
TaxoGene

- Automatically find all synonyms and insert the consensus approved name.
- Special characters and extensions
- Directing all readers to the preferred name in either search or publication allows
  - full retrieval recall of related material
  - insures precision in search
  - remove ambiguity in communication

- Synonymy: Average of 19 synonyms per gene name
- Sources:
  - Human Genome Project
Medicinal Plant Names Service (MPNS)

- How many kinds of “Ginger” are there??
  - At least 42
- Better communication between researchers worldwide – no misidentification
- Link to full plant name record at MPNS
- Includes all known scientific names, common names, homonyms, and more
- Global coverage – not just regional which is important an integrated world
- Constantly updated and linked to the
  - Kew International Plants Names Database.
  - International Plant Names Index (IPNI)

- Source Data: The Royal Botanical Gardens at Kew
- Synonymy: Average of over 16 names per plant are used.
- Includes all known scientific names, common names, homonyms, and more
- www.kew.org/mpns
Bad Cell Lines

- ICLAC There are about 437 cell lines which are documented as such and we mine for those as a way to highlight misuse....

- List of known contaminated cell lines (many of them invaded by HeLa cells)

- Don’t let your authors and researchers work with known bad data.
  - Over 32,000 papers that have worked on the wrong cells
  - Cited by *at least 500,000 more articles*,

Offering:
A rule base to quickly verify that the cell lines used are valid and not a contaminated line

Sources: 488 from ICLAC
[https://iclac.org/databases/cross-contaminations/](https://iclac.org/databases/cross-contaminations/)

757 from Swiss Institute of Bioinformatics (SIB)

Suspect Science Filter

• List of topics which require a closer look by acquisitions editors before sending out to potential peer reviewers
• Identifies questionable articles
• Autism and vaccination
  • Flag for assessment before sending to peer review
• Saves time in acquisitions review
• Auto Identify at time of submission using a rule base

Source: PLOS in conjunction with Access Innovations
Sci Gen

- The MIT Computer Science Department grad students bet that they could get papers which were programmatically generated accepted to peer reviewed publications.
  - https://pdos.csail.mit.edu/archive/scigen/
- This has been widely used as a mechanism to get the papers accepted.
- We have built a SciGen detector which identifies likely papers for closer review prior to acceptance.
- Also use to review your full collection to insure there are no bogus papers in the collections.
- GPT Bought by Microsoft to remove the threat.
Taxonomy Links in the PLOS Editorial Workflow

Submitted MS → Editorial Manager
- API Call to get indexing terms for editor/reviewer assignment

Editorial Manager → Peer Review
- API Call to get suggested final terms for editor selection

Peer Review → Finalized MS

Finalized MS → PLOS Subject Area Query DB
- Taxonomy ingested for platform display

PLOS Platform
- Display taxonomy
- Browse hierarchy
- Search Terms

PLOS Pre-Publication Filters
- SciGen Detector
- Bad Cell Lines
- Endangered Species
- Suspect Science
- Paper Fate Predictor
Increasing Audio Video Access

- The **new horizon** is indexing Audio Video content to make it accessible
  - Conference proceedings
  - Demonstrations, interviews online
  - News content
- All disappear without tagging of the content of the text
- Metadata without subject metadata does not give you access to the content
- Add taxonomy terms to the audio layer using transcription via auto tagging
- VATT – video to text and tagging from Data Harmony
Other Benefits beyond search and discovery

- Recommend articles to searches based on Indexing
- Gathering volunteers to communities of interest
- Recruitment based on topical areas
  - Trends analysis – where is the field headed?
- Auto sort conference tracks
- Peer review selection based on topical profile
- Discover new journal or conference topics
- Web site navigation
Fiscal Impacts: From Semantically Enriching your Content

- MUST use at both input and search
- 34% improvement in search
  - With just semantic enrichment
  - Ying-Hsang Liu, DC 2016, Copenhagen, Denmark

- 75% higher book sales with more complete metadata
  - NIELSEN BOOK US STUDY: THE IMPORTANCE OF METADATA FOR DISCOVERABILITY AND SALES,
  - David, Senior Director, Client Solutions, Nielsen Book’s Research and Commerce Solutions
    Published in the US December 31 2016

Thank you
Marjorie M.K. Hlava
Access Innovations, Inc
Accessinn.com
505-975-5578
Metadata is the Key!

Marjorie M.K. Hlava
Access Innovations, Inc
Accessinn.com
505-975-5578