Adaptation of an Open Source Semantic and Conceptual Retrieval Framework to the Astrobiological Domain

Lisa Miller and Rich Gazan
Department of Information and Computer Sciences
UH-NASA Astrobiology Institute
lisa.miller@hawaii.edu, gazan@hawaii.edu

April 27, 2010
Overview

- AIRFrame project rationale
- AIRFrame project goals
- Textpresso system
  - Ontology
  - Database
  - Adaptation to astrobiology
- Current and Future work
  - Database and Ontology building
  - Browsable visual interface
Keywords are inadequate for interdisciplinary work

Two searches on Thomson Reuters' ISI Web of Science:

1. **astrobio***
   - 791 results
   - astron* - 23,000+

2. **amino acid*** **Earth**
   - 940 results
   - Only 28 occur in both
   - No results from *Journal of Theoretical Biology* or *Origins of Life and Evolution of Biospheres* in **astrobio*** search.
   - Some articles with keyword **astrobiology** assigned by author appear in 2 but not returned with 1.
AIRFrame project goals

- Discover and relate diverse research concepts as a high level activity
- Eliminate the need to search for data using combinations of specific keywords
- Show users conceptual and functional relationships between diverse research documents
Textpresso-based system

- Open-source information retrieval and extraction system
  - Developed at CalTech for biological research
  - Currently deployed in 17 different, tightly focused literatures
- Two major components
  - Database of full-text scholarly documents
  - Ontology cataloging types of objects, abstract concepts, and relationships
Ontology

- Textpresso uses a shallow ontology to catalog terms

- **Concepts**
  - Nucleic Acid
    - Adenine
    - Cytosine
    - DNA
    - Guanine
    - ...

- **Descriptions**
  - Purpose
    - Fulfill
    - Make
    - Govern
    - Produce
    - ...

- **Relationships**
  - Comparison
    - Dissimilar
    - Equal sized
    - Like
    - Related
    - ...

- Comparison
We report the synthesis of glycine on interstellar ice-analog films composed of water, methylamine (MA), and carbon dioxide ...

Individual sentences

We report the synthesis of glycine on interstellar ice-analog films composed of water, methylamine (MA), and carbon dioxide ...

Ontology

XML Markup

We report the <biologicalprocess>synthesis</biologicalprocess> of <aminoacid>glycine</aminoacid> on interstellar ice-analog films <action>composed</action> of <volatile><water>water</water></volatile>, methylamine (MA), and <volatile><gases><chemicalelement>carbon</chemicalelement> dioxide</gases></volatile>
System query

Search by:
- keyword,
- keyword + synonyms,
- and/or whole categories

Results ranked by number of sentences with matching terms
Current Work

• “Proof-of-concept” AIRFrame/Textpresso is currently up and functioning
  - www.ifa.hawaii.edu/airframe/textpresso

• Workflow

**AIRFrame ontology**

- New articles
- Mark up documents & index
- Mine for new Terms
- Outside ontologies
Challenges and Fixes

- Previous implementations narrowly focused and based on existing ontologies
  - Fix: create a combined ontology from many sources
- System's ontology not in a standard form
  - Fix: convert to allow the use of SKOS formatted ontology
- Articles of interest to astrobiologists come from diverse sources
  - For now: manual gathering
  - Future: allow user submissions
Future work

- Create a browsable interface in addition to search using document clustering and classification techniques.
- Standardize the Textpresso output in order to provide visualization.
Thank you!

For updates visit our site at: www.ifa.hawaii.edu/airframe