AIRFrame:
Astrobiology Integrative Research Framework

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

August 5, 2010
Overview

- AIRFrame project rationale
- AIRFrame project goals
- Textpresso system
  - Ontology
  - Database
  - Adaptation to astrobiology
- Current and Future work
  - Database and Ontology building
  - Clustering/Classification
  - Browsable visual interface
Nature of Astrobiology

● System-level science
  – Concerned with complex, multidisciplinary, multi-phenomena behaviors of large physical and biological systems
  – Information technology needed to consolidate and represent knowledge and data across many disciplines

● New field
  – No centralized repositories of knowledge
  – No established, standardized vocabulary
Interdisciplinary Collaboration

- Proven to be difficult
  - Different disciplines have different:
    - Vocabularies/terminologies
    - Methods and formats for sharing and presenting research
    - Assumed levels of precision
  - Institutional and cultural boundaries

- Concern about duplication of research
  - Lack of access to existing knowledge
  - Lack of discipline-specific knowledge to efficiently access what is available
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
Keyword Search

Inadequate without a well-defined vocabulary

- Keyword searches on Elsevier's ScienceDirect using some astrobiologically relevant synonyms:

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Number of results</th>
<th>An article found only with this search</th>
</tr>
</thead>
<tbody>
<tr>
<td>“prebiotic synthesis”</td>
<td>566</td>
<td>The origin of life – a review of facts and speculations</td>
</tr>
<tr>
<td>“prebiotic chemistry”</td>
<td>611</td>
<td>A possible circular RNA at the origin of life</td>
</tr>
<tr>
<td>“abiotic synthesis”</td>
<td>312</td>
<td>Recent advances in the chemical evolution and the origins of life</td>
</tr>
<tr>
<td>“organic synthesis”</td>
<td>53,434</td>
<td>Prebiotic organic synthesis in early Earth and Mars atmospheres</td>
</tr>
</tbody>
</table>
Keyword Search
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
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
Onotology

- 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
    - ...

- Description:
  - Adenine
  - Cytosine
  - DNA
  - Guanine
  - ...

- Comparison:
  - Dissimilar
  - Equal sized
  - Like
  - Related
  - ...

- Purpose:
  - Fulfill
  - Make
  - Govern
  - Produce
  - ...

- Relationships:
  - Nucleic Acid
  - Descriptions
  - Purpose
  - Comparison
  - Produce
  - ...

- Ontology:
  - Textpresso
  - Shallow ontology
  - Term cataloging
We report the synthesis of glycine on interstellar ice-analog films composed of water, methylamine (MA), and carbon dioxide...
System query

Allows search by *meaning* rather than specific keyword

Search by:
- keyword,
- keyword + synonyms,
- and/or whole categories
Results ranked by number of sentences with matching terms

**Category 2**
- **Keyword**: water, H2O
- **Synonym**: carbon dioxide

**Category 1**
- **Keyword**: water, synthesize

**Category 2**
- **Synonym**: processing

Additional results:
- Title: Experimental constraints on the chemical evolution of large icy satellites
  - **Authors**: Scott HP Williams, Q Ryerson FJ
  - **Journal**: EARTH AND PLANETARY SCIENCE LETTERS
  - **Year**: 2002
  - **Abstract**: Both experimental and calculated results show that primordial organics are likely to have been largely oxidized to carbonates through hydrothermal processing early in Ganymede's history, potentially sterilizing Ganymede's H2O layer. [Field: abstract, subscore: 3.00]
Adaptation challenges

- Previous implementations of Textpresso have been based on pre-existing ontologies such as the Gene Ontology.
- Previous implementations have been much more narrowly focused such as genetics for a single organism.
- Most biological journals have open full-text access through a single source, PubMed.
Current Work: Ontology Standards Adoption

- To allow more breadth to the ontology
  - Build a new ontology using SKOS standard developed by Vocabulary Explorer at IVOA
  - Adapt Textpresso to directly read SKOS ontology

- Standards based ontology can
  - Ease porting existing ontologies to ours
  - Allow easy sharing of our data with other systems
Why SKOS?

- Allows fast addition of terms from existing sources such as International Astronomical Union Thesaurus and IVOA.
- http://explicator.dcs.gla.ac.uk/WebVocabularyExplorer/
- Created to allow conversion from other formats
- SKOS allows the kind of relationships we want to leverage with AIRFrame

```xml
<rdf:Description rdf:about="http://iaaa.cps.unizar.es/thesaurus/SPACE SCIENCES">
  <skos:related rdf:resource="http://iaaa.cps.unizar.es/thesaurus/CELESTIAL MECHANICS"/>
  <skos:prefLabel xml:lang="fr">SCIENCES DE L'ESPACE</skos:prefLabel>
  <skos:broader rdf:resource="http://iaaa.cps.unizar.es/thesaurus/MT_2.30 SPACE SCIENCES"/>
  <skos:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:prefLabel xml:lang="es">CIENCIAS DEL ESPACIO</skos:prefLabel>
  <skos:related rdf:resource="http://iaaa.cps.unizar.es/thesaurus/COSMOLOGY"/>
  <skos:prefLabel xml:lang="en">SPACE SCIENCES</skos:prefLabel>
</rdf:Description>
```
Current Work: Database Building

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

Workflow:
- New articles
- AIRFrame ontology
- Mark up documents & index
- Mine for new terms
- Outside ontologies
- AIRFrame searchable database
Current Work: Initial Steps in Document Classification

- Want to use machine learning methods to discover connections in the data
  - Possibly use several to give users many views.
- Some work done on phrase-based clustering
- My research area is in information theoretic clustering
- Open to other ideas
Information Bottleneck Method

• Developed by Tishby, Pereira, Bialek in 1999

• Based on Shannon Information
  - Measures the reduction in uncertainty or the distortion between an original signal and its compressed representation

$$I[x, y] = H[x] - H[x|y]$$

- Where uncertainty is the entropy

$$H[x] = \sum_{x}^{N} p(x) \log \left(1 / p(x)\right)$$
IB Method 2

- Data is compressed so that information about a quantity of interest (in this case words) is kept maximally.

- \( \max_{p(c|x)} [I(y,c) - TI(x,c)] \)

- Relative entropy emerges as the distortion function (Kullback-Leibler divergence)

- With the optimal assignment rule

- \( p(c|x) = \frac{p(c)}{Z(x,\beta)} \exp\left(\frac{-1}{T} D_{KL}[p(y|x)|p(y|c)]\right) \)
IB Method 3

- Has been shown to be one of the most accurate unsupervised clustering algorithms
- Downside: slow & computationally expensive
- I am working to establish the optimal number of clusters automatically within the same process using annealing.
Future Plans

- Visualizations!
Future plans

- Create a cleaner, browsable interface which displays results and links in an easily understandable way

Current Interface
- Ugly!
- Not intuitive
- Results spread across pages - Hard to see overall picture
Future Plans

• Allow users to input a document and get back information such as
  - Related articles
  - Connections to other researchers
  - Relevant NASA/NAI goals
Some Ideas

- Use document classification and clustering assisted by the semantic markup
- Create clusterings unsupervised and then use a Support Vector Machine to classify a document input by the user
- Display connections with topic maps by making Textpresso output in ISO Topic Map form
Personas

- We envision AIRFrame being able to present information in dynamic ways such as this:


- Would an interface like this be valuable?
LDA

- Used by Personas
- Returns document as a distribution across topics
  - Topics are clusters of single word terms
  - Manually labeled for Personas
- Personas uses an SVM trained with the LDA groupings for on-the-fly classification
Snappy Words

- WordNet visualizer

http://www.snappywords.com/?lookup=nasa
Open Questions

- How best to use our semantic markup to present our data?
- How to visualize our information to allow discovery and understanding of connections?
- How best to search for and represent
  - Category links between terms?
  - Authorship linkages between projects?
  - NASA/NAI goal links to research?
- Ontology building
  - Can it be at least partially automated?
Issues

• Speed
  – Users need to have feedback if they need to wait.
  – Slow unsupervised methods can be run offline during database building

• Users need to be able to do novel searches, not just a few predefined ones.

• Complexity and size of ontology
Thank you!

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

To test out our textpresso version:
www.ifa.hawaii.edu/airframe/textpresso
References and Resources


“Omnigator Vizigator.”
http://www.ontopia.net/omnigator/plugins/viz/viz.jsp?
tm=opera.ltm&id=puccini&redirect=%2Fomnigator%2Fmodels%
2Ftopic_complete.jsp%3Ftm%3Dopera.ltm%26id%3Dpuccini

“Ontologies Come of Age.”
http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-
mit-press-%28with-citation%29.htm

M. Uschold and M. Gruninger, “Ontologies: Principles, methods and
applications,” To appear in Knowledge Engineering Review, vol. 11,
1996.

N.F. Noy and D.L. McGuinness, Ontology development 101: A guide to
creating your first ontology, Citeseer, 2001.

PANDORA: Programs for AstroNomical Data Organization Reduction
and Analysis. http://cosmos.iasf-milano.inaf.it/pandora/

Textpresso: An Ontology-Based Information Retrieval and Extrac
tion System for Biological Literature.
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0020309
References and Resources


References and Resources


- “What is an ontology and why we need it”

References and Resources


- Han, Jiawei, and Kamber, Micheline: Data mining: concepts and techniques, Morgan Kaufmann, 772, 2006
