ADAPTATION OF AN OPEN SOURCE SEMANTIC AND CONCEPTUAL RETRIEVAL FRAMEWORK TO THE ASTROBIOLOGICAL DOMAIN. Lisa Miller[1] and Rich Gazan[1], [1]University of Hawaii Department of Information and Computer Sciences, 1680 East-West Road, Honolulu, HI 96822, ljmiller@hawaii.edu, gazan@hawaii.edu

Introduction: Astrobiology is by nature a system-level science, meaning that it is concerned with complex, multidisciplinary, multi-phenomena behaviors of large physical and biological systems. Due to the breadth of the undertakings in astrobiological inquiry, researchers in the field must rely heavily on information technology to consolidate and represent knowledge and data from across many disciplines [1]. AIRFrame, the Astrobiology Integrative Research Framework, is an integrative knowledge framework under active development by the University of Hawaii NASA Astrobiology Institute. By leveraging the power of new standards and technologies developed for the Semantic Web, AIRFrame allows diverse research concepts to be related and discovered as a high-level activity. Using a combination of state of the art semantic retrieval methods and pattern classification algorithms, AIRFrame permits users be they researchers, students, or the general public to search for a *meaning* rather than attempting to guess the correct combination of keywords to get search results on a desired topic. By eliminating the need to search for data using various combinations of specific keywords that have no set standard, differ across disciplines, and change over time and by easing the burden of sifting through vast amounts of rapidly accumulating data, AIRFrame can reduce the cognitive load on researchers and other users and allow discovery of otherwise unfindable resources, embodying many of the recommendations on computer-supported collaborative work coming from research in the social informatics field [2].

Discussion: AIRFrame is intended to be an integrated discovery framework which is able to show users both conceptual and functional relationships between diverse research documents. By employing an ontology and semantic markup of documents, AIRFrame allows discovery of related concepts which might be invisible to users of standard search engines unfamiliar with key-terms used by diverse disciplines. We have based the underlying functionality of AIRFrame on Textpresso [3], a freely available, open-source information retrieval and extraction system which has been successfully implemented on 17 different sets of literatures within the biology community, from *C. elegans* genome research to pharmagenomics.

The AIRFrame/Textpresso system consists of two major components, a database of full-text scholarly documents and an ontology. The system's ontology is a catalog of types of objects, abstract concepts, and their relationships within the domain of interest. For example, in the "regulation" category some words and phrases included are: "elevate", "life extending", and "truncate", while the "comparison" category includes terms such as "dissimilar", "equal sized", and "related". The system database is built by converting all documents in a corpus of text into plain-text and separating out each sentence. Then every word or phrase in each document which occurs in the ontology is marked-up with eXtensible Markup Language (XML) tags. The XML tags enable the system to handle semantic queries such as a search for the keyword "amino" along with the category "regulation". A query of the system searches the entire database looking for individual sentences containing the specified keyword(s) paired with terms tagged with the desired category(s). The search results are returned as ranked references to documents that contain one or more matching sentences, the ranking based on the number of matches.

One result of a search of the AIRFrame database using the keyword "amino" and category "regulation" will be the journal article *Amino-acid-dependent signal transduction* [4] because it contains sentences such as: "This *antagonism* between *amino acids* and cAMP with regard to mTOR activation would nicely explain the opposing effects of *amino acids* and glucagon [...]." The word *antagonism* is listed in the ontology under the regulation category and thus matches the search criteria. As can be seen from this example, a user can perform a semantic search for relevant literatures based on a desired meaning without domain specific knowledge of terms with this system.

For AIRFrame and the astrobiological domain, we have been developing the text database and most of the ontology independently from previous Textpresso implementations. Our preliminary corpus of documents consists of both full-text and abstract-only documents retrieved from keyword searches of ISI's Web of Science, works written by our fellow UHNAI teammembers, and documents culled from the bibliographies of these works. We hope to create a more exhaustive database of documents as work on the system progresses.

Our primary focus has been the creation of a thesaurus and ontology for the astrobiology field and the adaptation of the Textpresso searching mechanism to a much broader linguistic domain. The original devel-

opers of Textpresso had the advantage of building their ontology directly from existing biological resources such as the Gene Ontology and all previous implementations of Textpresso have kept a narrow focus within specialized subsets of biology. In our case, we are developing a much broader-based astrobiology ontology and must modify the searching system to accommodate it. Due to the breadth astrobiology research and our desire to keep our work standards-based for future integration, we have determined that we need to give greater structure and formality to our ontology in order to maintain coherence and usability.

To this end, we are in the process of integrating work done on the Vocabulary Explorer to retrieve semantic concepts from astronomy vocabularies [5] with parts of the existing Textpresso biological ontology. The Vocabulary Explorer along with the International Astronomical Union Thesaurus and the International Virtual Observatory Alliance are all utilizing the Simple Knowledge Organization System (SKOS) a new standard data model recommended by W3C in August, 2009 for sharing and linking thesauri, taxonomies, classification schemes and subject heading systems on the Semantic Web [6]. We are working to adapt the Textpresso system so that as AIRFrame it will access and use SKOS formatted ontologies directly in order to make it more standards compatible, able to handle a deeper hierarchical classification, and easily updatable in the future.

Future Work: One area under active but experimental development for AIRFrame is the automatic classification of documents into phrase-based clusters based on the work done by [7] on an existing Textpresso implementation. Using Support Vector Machines and a subsumptive phrase-based clustering method assisted by the Textpresso XML markup to classify documents, this method has been shown to facilitate quick navigation through a hierarchy of subjects to find documents belonging to a specific concept without ever entering a keyword or search term. This addition also leads directly to one of the primary goals of AIRFrame which is to provide novel data visualizations of the contents of our corpus. We intend the final implementation of AIRFrame to display search results not only in text-based format but also as dynamic concept maps showing relationships that are difficult to grasp with text-only results. Category linkages between terms and across disciplines, authorship linkages between documents or research projects, and connections to NASA/NAI goals are some of the concepts that will be greatly enhanced by visualizations provided by AIRFrame. To that end, we are also exploring the standardization of the output from the search engine into ISO topic maps [8] so that visualization software such as the open source Vizigator [9] might be directly compatible.

Conclusions: We are actively developing AIRFrame, the Astrobiology Integrative Research Framework, to support astrobiological research and discovery through standards based leveraging of Semantic Web and pattern classification technologies. AIRFrame will encompass an ontology and thesaurus for astrobiological terminology, a semantic search engine, and concept visualization when completed. Our current focus is on creating the ontology and thesaurus, representing it in the new W3C standard, SKOS, for open and easy access, and adapting the existing Textpresso system to our ontology requirements. We are also working on gathering astrobiology related scholarly documents for incorporation into our searchable database and on developing clustering techniques for improved navigation and connection discovery in this vast interdisciplinary corpus. Looking towards the future we are focused on developing this system in the format which will enable incorporating our results into novel data visualizations to assist researchers, data curators and the general public in accessing astrobiological data.

Acknowledgements: This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science.

References: [1] Cummings, J., *et al.* (Eds.) (2008) *Beyond being there: A blueprint for advancing the design, development, and evaluation of virtual organizations*. Building Effective Virtual Organizations, An NSF Workshop. [2] Kaptelinin, V., *et al.* (2006) *Acting with Technology: Activity Theory and Interaction Design*. MIT Press. [3] Muller, H.M., *et al.* (2004) *PLoS Biol*, *2*(11), e309. [4] van Sluijters, D.A., *et al.* (2000) *Biochemical Journal*, *351*(Pt 3), 545. [5] Gray, A., *et al.* (2009) *Information Processing and Management*. [6] Miles, A., *et al.* (2009) *SKOS Simple Knowledge Organization System Reference*. W3C. http://www.w3.org/TR/skos-reference/. [7] Chen, D., *et al.* (2006) *BMC Bioinformatics*, *7*(370). [8] Biezunski, M., *et al.* (2002) *ISO/IEC 13250, Topic Maps*. http://www.y12.doe.gov/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf. [9] Ontopia *Omnigator: The Topic Map Browser*. http://www.ontopia.net/omnigator.