# BDEI: Biodiversity Information Organization using Taxonomy (BIOT)

Dr. Susan Gauch
sgauch@ku.edu
Electrical Engineering and Computer Science and
Information Technology and Telecommunications Center
University of Kansas

**Abstract**
Biological resources are the basis of much of this Nation's prosperity. The wise stewardship of these riches affects current and future generations and, therefore, must be rooted in decisions based on the best information available. Because of the rapid growth of all content on the Internet, there is a corresponding increase in online biological information of all types. However, the sheer volume of data, and its widely varying quality and physical distribution make it impossible for decision makers to make critical decisions without the benefit of the existing body of information. This project aims to investigate and develop intelligent knowledge management tools and deploy them in this important domain. In particular, we will investigate the use of a taxonomy as a reference ontology to provide a conceptual framework to allow intelligent agents to browse biodiversity information in a variety of formats widely distributed on the World Wide Web.

**1. Introduction**The World Wide Web (WWW) offers the promise of unlimited access to electronic information. Unfortunately, the actuality is electronic access to unlimited anarchy. A tremendous amount of effort has gone in to creating valuable online collections of biodiversity information. The Partnerships in Enhancing Expertise in Taxonomy (PEET) project [1] is collecting information on organisms for which information is disappearing and is required to put it online. The National Biological Information Infrastructure (NBII) [2] is maintaining a clearinghouse of meta-information on biological information. The Tree of Life project [3] is providing an infrastructure for collaborative production of information across the biological taxonomy. This information is available in a wide variety of formats from a wide variety of locations. However, using search engines to locate biodiversity information in the midst of the billions of collected Web pages on every conceivable topic makes looking for a needle in a haystack seem trivial in comparison. The search engines fail for three primary reasons: 1) Biological information is a very small subset of a very large collection; 2) Search engines only index HTML pages linked from other HTML pages; and 3) Biological nomenclature is enormous and unwieldy for all but the specialist.

This project is developing intelligent access to online biodiversity information. Because of the difficulties posed by searching enumerated above, we will concentrate initially on providing intelligent browsing support. Navigating online resources via the taxonomy solves the problems listed in the previous section. First, because only biodiversity information will be added to the taxonomy, the information presented will be high quality and on topic. Second, because the user will be directed to the original sources, they can access information in any format. Finally, because the user will be navigating an online taxonomy, they need only recognize the nomenclature (presented in multiple forms).

## 2. Related Work

The Semantic Web represents a growing desire to make information access on the Web more knowledge-based so that humans and intelligent software can work together [8]. One approach to this is to annotate online resources with their topics or categories in an ontology, or graphical arrangement of categories. An early of subject hierarchies for query routing was developed in HyPursuit [9]. More recently, *OntoSeek* uses the Sensus ontology [10], a simple taxonomy structure of approximately 50,000 nodes [11] to provide designed for content-based information retrieval from online yellow pages and product

catalogs. The system developed by Labrou and Finin [12] uses *Yahoo!* [7] as an ontology to semantically annotate Web using *Telltale* [13][14][15] an n-gram based classifier. A richer and more powerful representation is provided by *SHOE* [17][18] (Simple HTML Ontology Extensions). Ontology-based systems require some method to classify online resources with the appropriate categories, generally based on probabilistic formulae [18][19], vector space methods [16][21] and structural information [20].

Increasingly, research is attempting to provide users with individually tailored information. *WebWatcher* [22][23] and. *Syskill & Webert* [24] use explicit feedback to recommend links to follow to users as they browse the web. *Personal WebWatcher* [25] (an extension of *WebWatcher*), *WebMate* [26], *Amalthaea* [27], and *Letizia* [28][29] also provide browsing assistance, but they learn about the user by watching their actions rather than asking them for information. The system relies on implicit feedback including links followed by the user or pages and/or bookmarked pages. *Alipes* [30], on the other hand, attempts to gather and filter news articles and on behalf of the user. All of the above systems represent the user profiles as vectors of weighted keywords. In contrast, OBIWAN [31][32] uses a more semantically oriented profile based on weighted categories.

## 3 Approach

We will employ intelligent agents to organize and present biodiversity information on the Web. The system will rely on the provision of a *reference ontology*, or canonical organization scheme. For many applications, no such reference ontology exists, however the DARPA Agent Markup Language (DAML) initiative [4] collected 156 ontologies for different domains. We are fortunate in this project that the reference ontology, the biological taxonomy, already exists, is well understood and widely adopted.

Classifying Online Resources The system will spider and classify the information identified by the Tree of Life [3] to form the training data for the classification algorithms. The National Biological Information Infrastructure, and the PEET project sites will be used as the initial information to be classified. Most classification techniques work by classifying individual documents in a vacuum. We will work on developing a contextual classification technique that considers not just the individual document, but also the context of the document as represented by the other documents collected from its original site. The use of link structure in searching has lead to great improvements in search quality [5][6]. We will investigate whether link structure can also be exploited to improve classification. We will extend this by developing a domain-specific component that looks for specific taxonomic names within the documents.

Browsing Online Resources The Tree of Life project [3] provides a way to browse informational pages arrange via the taxonomy. This system provides an example of the power of the collaborative nature of the Web. Each category in the taxonomy is individually maintained by a volunteer expert. Due to its distributed nature, there is no global overview and no sophisticated way to navigate among the sites. Generally, each category in the taxonomy contains only locally developed content, with few links to the many other resources available on the topic. We will provide a global overview of the resources, a framework, within which the content from many sites can be viewed. This browsing hierarchy will always be displayed, placing the information being viewed in context and also providing quick navigation from topic to topic.

Adapting Information Presentation The system described so far provides a useful way to browse biodiversity information. However, we wish to investigate how we can we adapt the information presentation for a multiplicity of users and tasks. The automatic categorization of the information must identify more than just the location in the taxonomy for a particular resource, it must also analyze and store (using XML tags) a variety of features for the resource as well: the level of expertise, the format, the level of detail, the quality of the information, the timeliness of the information, etc. Based on the user's level of expertise the taxonomy viewed and the specific resources displayed will be adapted. The

category labels should vary - for a sixth grader the common names for the families and species will be used but for a domain expert, the scientific names will be the default. The number of taxonomic categories displayed will also be varied. For the sixth grader, many of the resources associated with particular species will be combined and promoted up the taxonomy to be presented as information about the entire family. The particular resources shown for a category, and the order of presentation, will also vary based on a user-supplied profile indicating their age, level of expertise, and the type of information they are seeking.

**4. Conclusions and Future Work**When completed, we will have an effective tool for use by others to effectively heterogeneous, distributed biodiversity information on the Web. Equally importantly, we will have learned: 1) how to classify biodiversity information automatically; 2) how to navigate a huge, distributed, online taxonomy; 3) how to present summaries of subtrees of a taxonomy; and 4) how to select information for each category appropriate to individual user's needs and expertise. Future work will concentrate on providing access to distributed biodiversity information via search as well as browsing.

**Bibliography**


[1] NSF PEET Project Homepage, http://www.nhm.ukans.edu/peet/, July 4, 2001.

[2] NBII - Metadata: Clearinghouse, http://www.nbii.gov/datainfo/metadata/clearinghouse/, July 4, 2001.

[3] The Tree of Life Root Page, http://phylogeny.arizona.edu/tree/life.html, July 4, 2001.

[4] DAML Ontology Library, http://www.daml.org/ontologies/, July 4, 2001.

[5] Brin, S., Page, L. "The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks* 30(1-7), 1998, 107-117.

[6] Kleinberg, J. M. "Authoritative Sources in Hyperlinked Environment", *Journal of the ACM* 46, 1999.

[7] Yahoo!, http://www.yahoo.com/, July 4, 2001.

[8] Berners-Lee, T., Hendler, J., and Lassila, O., "The Semantic Web," *Scientific American*, May 2001.

[9] Weiss, R., Velez, B., Sheldon, M. A., Nemprempre, C., Szilagyi, P., Duda, A., and Gifford, D. K., *Proc 7th ACM Conf. on Hypertext*, Washington, DC, March 1996.

[10] Knight, K. and Luk, S. Building a Large Knowledge Base for Machine Translation. *Proc. Amer. Assoc. Artificial Intelligenc Conf. (AAAI),* pp. 773-778, 1999.

[11] Guarino N., Masolo C., and Vetere G., OntoSeek: Content-Based Access to the Web, *IEEE Intelligent Systems* 14(3), May/June 1999, 70-80.

[12] Labrou, Y. and Finin T. Yahoo! As an Ontology - Using Yahoo! Categories to Describe Documents. *Proc. 8th Intl. Conf. on Information and Knowledge Management*, 180-187, 1999.

[13] Chowder, G. and Nicholas, C. Resource Selection in Café: an Architecture for Networked Information retrieval. *Proc. SIGIR'96 Workshop on Networked Information Retrieval*, Zurich, 1996.

[14] Chower, G. and Nicholas, C. Meta-Data for Distributed Text Retrieval. *Proc. 1st IEEE Metadata Conference*, 1996.

[15] Pearce, C. and Miller, E. The Telltale Dynamic Hypertext Environment: Approaches to Scalability. *Advances in Intelligent Hypertext,* Springer-Verlag, 1997.

[16] Zhu, X., Gauch, S., Gerhard, L., Kral, N. and Pretschner, A. Ontology-Based Web Site Mapping for Information Exploration, *Proc. 8th Intl. Conf. on Information and Knowledge management (CIKM '99)*, Kansas City, MO, Nov. 1999, 188-194.

[17] Heflin, J., Hendler, J. and Luke, S. SHOE: A Knowledge Representation Language for internet Applications. *Technical Paper, Institute for Advanced Computer Studies*, University of Maryland, College Park.

[18] Luke, S., Spector, L., Rager, D. and Hendler, J. Ontology-Based Web Agents. *Proc. First International Conference on Autonomous Agents (AA'97),* 1997.

[19] Gover, N., Lalmas, M., and Fuhr, N. A Probabilistic Description-Oriented Approach for Categorising Web Documents. *Proc. 8th Intl. Conf. on Information and Knowledge management (CIKM '99)*, Kansas City, MO, Nov. 1999, 475-482.

[20] Larkey, L. Automatic Essay Grading Using Text Categorization Techniques. *Proc. 21st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[21] Hsu, Wen-Lin and Lang, Sheau-Dong. Classification Algorithms for NETNEWS Articles. *Proc. 8th Intl. Conf. on Information and Knowledge management (CIKM '99)*, Kansas City, MO, Nov. 1999, 114-121.

[22] Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T. Web Watcher: A Learning Apprentice for the World Wide Web. *Proc. AAAI Spring Symp. on Information Gathering from Heterogeneous, Distributed Environments*, March 1995.

[23] Joachims, T., Freitag, D., Mitchell, T. Web Watcher: A Tour Guide for the World Wide Web. *Proc. IJCAI'97,* August 1997.

[24] Pazzani, M. Muramatsu, J. Billsus, D. Syskill & Webert: Identifying Interesting Web Sites. *Proc. 19th Natl. Conf. on Artificial Intelligence*, 1996.

[25] Mladenić, D. Personal Web Watcher: Design and Implementation. *Technical Report IJS-DP-7472, J. Stefan Institute*, Department for Intelligent Systems, Ljubljana, Slovenia, 1998.

[26] Chen, L. and Sycara, K. A Personal Agent for Browsing and Searching. *Proc. 2nd Intl. Conf. on Autonomous Agents*, New York, USA, 1998, 132-139.

[27] Moukas, A. Amalthaea: Information Discovery and Filtering Using a Multiagent Evolving ecosystem. *Proc. 1st Intl. Conf. on the Practical Application of Intelligent Agents and Multi Agent Technology*, London, 1996.

[28] Liebermann, H. Letizia: An Agent that Assists Web Browsing. *Proc. Intl. Conf. on AI*, Aug. 1995.

[29] Liebermann, H. Autonomous Interface Agents. *Proc. ACM Conf. on Computers and Human Interaction (CHI'97)*, Atlanta, USA, May 1997.

[30] Widyantoro, D., Ioerger, T. Yen, J. An Adaptive Algorithm for Learning Changes in User Interests. *Proc. 8th Intl. Conf. on Information and Knowledge Management*, 1999, 405-412.

[31] Pretschner, A. and Gauch, S. "Ontology-Based Personalized Search," *Proc. 11th IEEE International Conf. on Tools with Artificial Intelligence (ICTAI '99)*, Chicago IL, November 1999, 391-398.

[32] Chaffee, J. and Gauch, S. "Personal Ontologies for Web Navigation," *Proc. 9th Intl. Conf. on Information and Knowledge Management, (CIKM '00)*, McLean VA, Nov. 2000, 227-234.