In Procs. AAAI 2018 Fall Symposium on A Gathering for AI and Natural Systems,
Washington DC, October 2018.

Extracting Structural Knowledge from Natural Language Documents to Support Biologically Inspired Design

Ashok Goel, Swapnal Acharya, Kimisha Mody, Kaylin Hagopian, Shimin Zhang, and Spencer Rugaber
Design & Intelligence Laboratory, School of Interactive Computing
Georgia Institute of Technology, Atlanta, GA 30332, USA
goel@cc.gatech.edu; sacharya38@gatech.edu; kmody6@gatech.edu; khagopian3@gatech.edu;
shiminzhang@gatech.edu; spencer@gatech.edu,

***Abstract*—**Given a design problem in the context of biologically inspired design, most designers search for natural language documents describing biological cases on the internet, and then construct an understanding of the biological cases for potential transfer to the design problem. IBID is an interactive tool for extracting knowledge of the function, structure, and causal mechanisms of a biological system from its natural language description and organizing this knowledge as a Structure-Behavior-Function model. In this article, we briefly describe how IBID extracts structural knowledge about biological systems from natural language texts and uses the structural knowledge for accessing biology articles relevant to a design problem.

I. INTRODUCTION

Biologically inspired design is a well-known paradigm that uses nature as a source of practical, efficient and sustainable designs to stimulate the design of technological systems (Benyus 1997; Vincent & Mann 2002). The design paradigm consists of two main methods (Helms, Vattam & Goel 2009): problem-driven analogy, in which the designer uses the specification of a design problem to retrieve, adapt and transfer knowledge of a biological system to generate a candidate solution, and solution-based analogy, where the designer uses knowledge of a biological system to find and address new design problems by analogy.

However, most architects, engineers, and designers are not experts at biology (Yen et al. 2014). Thus, most designers neither have knowledge of a large number of biological systems stored in their internal memory, nor a deep understanding of the biological systems in fact available in the memory. Instead, given a design problem, in practice most designers search for relevant natural language documents describing biological cases on the internet, and then construct an understanding of the retrieved biological cases for potential transfer to the design problem (Vattam & Goel 2011, 2013). These observations have led to many efforts to develop computational techniques and tools to support the process of biologically inspired design (Goel, McAdams & Stone 2014).

The first generation of computational tools for supporting biologically inspired design focused on constructing digital libraries of biological cases. This includes the well known and still growing AskNature digital library (https://asknature.org/; Deldin & Schuknecht 2014), an early knowledgebase of biological strategies for design as well as case studies of biologically inspired design. The first generation of computational tools also included our own work on a digital library called DANE (http://dilab.cc.gatech.edu/dane/; Goel et al. 2012) of Structure-Behavior-Function (SBF) models of biological and technological systems, as well as a digital library called DSL (Goel et al. 2015) of case studies of biologically inspired design. While the SBF models (Goel, Rugaber & Vattam 2009) capture deep understanding of how the structure of a system achieved its functions and the SBF language for expressing this understanding is quite general, these libraries were handcrafted and thus limited in size. The question then becomes how may we use AI techniques for automatically acquiring knowledge of biological systems for use in biologically inspired design?

Thus, the current, second generation of computational techniques for supporting biologically inspired design has focused on AI techniques for natural language processing for classifying and accessing natural language documents describing biological systems (Cheong et al. 2011; Glier et al. 2014) Kruiper et al. 2017; Nagel & Stone 2012; Shu 2010; Vandevenne et al. 2016) including our own work on using IBM's Watson tool for accessing biology articles relevant to a design query and answering questions based on the retrieved articles (Goel et al. 2016). It also includes our work on the IBID project described here (IBID for Intelligent Biologically Inspired Design). Previously, in Rugaber et al. (2016) and Spiliopoulou et al. (2015), we described the conceptual architecture of the IBID system for extracting SBF models of biological systems from natural language documents, as well as IBID's techniques for extracting functions of biological systems described in a biology article and locating biology articles in a corpus relevant to a function specified in a controlled vocabulary. In this paper, we provide a brief update on the current status of the IBID project.

II. PROBLEM CONTEXT

Let us consider an expert designer who, like most designers, is a novice in biology. Let us suppose that the designer is interested in designing a system for transporting water to remote regions in her country. Given a corpus of biology articles, how may our designer find biology articles that are relevant to her design problem? How might IBID support her in finding relevant articles?

In previous work (Helms, Vattam & Goel 2010), we found that SBF models help a designer understand a complex biological system so as to better answer questions about its functioning. This led us to develop the Biologue system (Vattam & Goel 2011, 2013) for annotating biology articles based by SBF models, accessing biology articles relevant to a

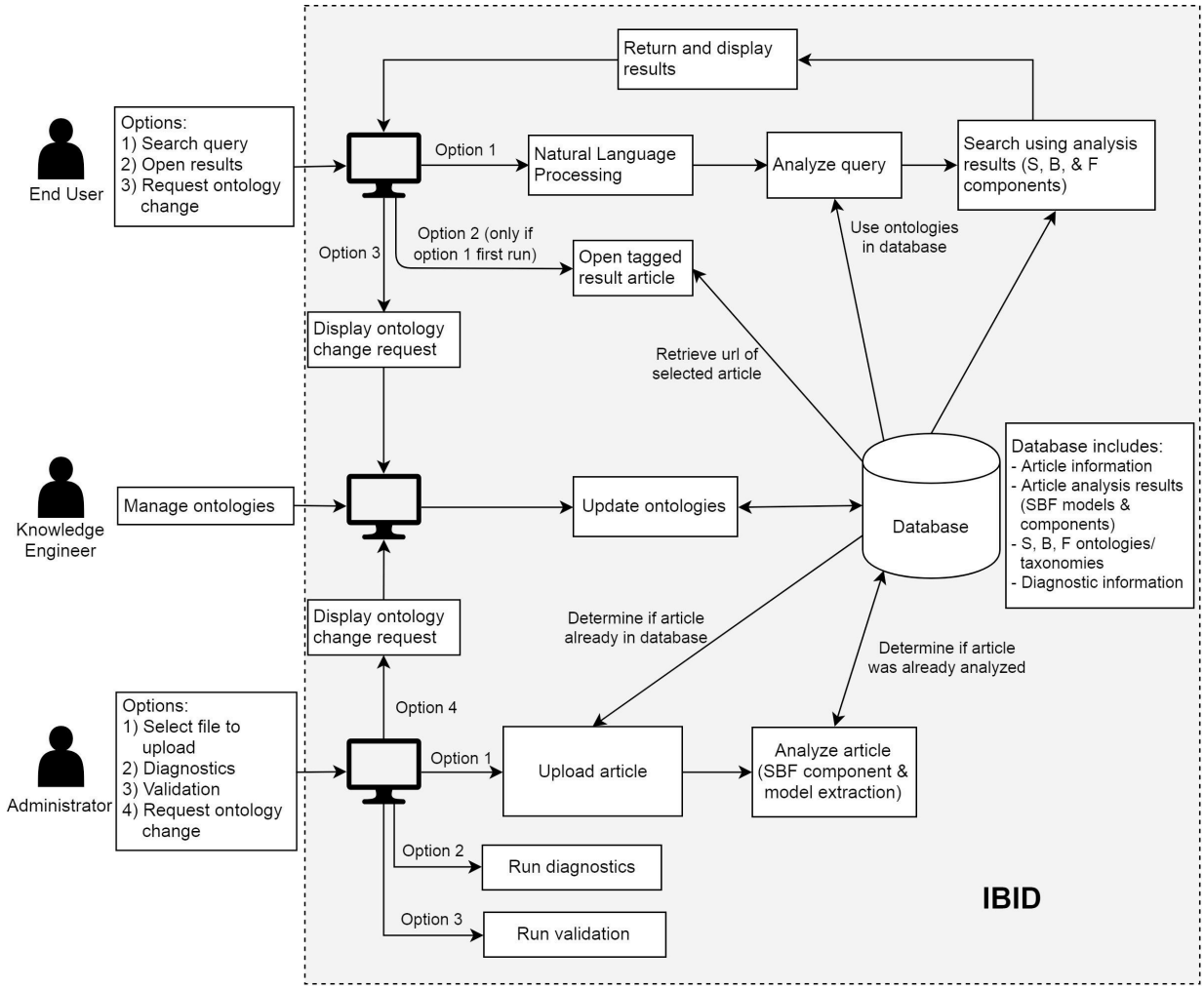

Figure 1: Conceptual architecture for the complete IBID system.

design problem based on the annotations, and using the annotations to help the designer understand the article in terms of SBF models. Biologue showed that if biology articles are annotated with SBF models of the biological systems described in the articles, then many designers are better able to both locate biology articles relevant to their design problem and understand how the biological systems work. The IBID project builds on Biologue. In Biologue, the SBF annotations on the biology articles were handcrafted, while IBID seeks to automate, and thereby generalize and scale, the process.

The IBID system operates in two modes. First, it extracts SBF models of biology articles in a given corpus and annotates the articles with structural, behavioral and functional terms. Given a research article describing a biological system from a journal such as *The Journal of Experimental Biology*, the current version of IBID can extract the function, the structure, and parts of the causal behaviors of the system. Second, given a design query, IBID is intended to locate the biology articles relevant to the query based on the structural, behavioral and functional annotations.

Figure 1 shows the full functionality of IBID as we envision it for its three use cases: engineers and designers (end users) looking for biology articles, knowledge engineers extending IBID's vocabulary, and administrators adding to its repository of analyzed papers. The actions available to each user type are specified and the arrows indicate progression of steps and/or access to/from the database.

III. EXTANT TOOLS

Before we describe IBID, we note that it uses several extant tools including the following:

1. Stanford Natural Language Parser[1]. The Stanford Natural Language Parser generates parse trees of input sentences. IBID uses the core Stanford Natural Language Parser tool to find the parse trees for sentences in a biology article (or a natural language design query). IBID uses the parse tree of a sentence to help identify if a part of the sentences refers to the structure, behavior or function of the biological system described in the article.

[1] http://nlp.stanford.edu/software/lex-parser.shtml
[2] https://wordnet.princeton.edu

2. WordNet[2]. WordNet is a large lexical database of English in which different parts of speech are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Given a design query expressed as an English language sentence, IBID uses WordNet to widen the set of search terms.

3. VerbNet[3]. VerbNet is an on-line verb lexicon that includes specific syntactic information and indications of verb class membership. Each verb class in VerbNet is completely described by its frames, thematic roles, and arguments. IBID uses VerbNet to identify and extract function terms from articles and expand its function ontology.

4. Vincent's Ontology of Biological Structure. Julian Vincent has developed a detailed ontology of the structure of biological systems (Vincent 2014). IBID uses a small part of his ontology as a domain-specific controlled vocabulary of biological structures.

5. Domain-Independent Ontologies of Structure, Behavior, and Function. We have developed domain-independent ontologies of the structure, the behaviors, and the function of complex systems that build in part on the extant SBF ontology (Goel, Rugaber & Vattam 2009). Spiliopoulou et al. (2015) and Rugaber et al. (2016) describe IBID's functional ontology. IBID is intended to use these ontologies to capture the structural, behavioral, and functional concepts and relationships in the description of a biological system (or a design query).

IV. Extraction of Structure, Behavior and Function of Biological Systems from Text

As mentioned above, IBID already was capable of extracting functions of biological systems from their English descriptions in research articles (Rugaber et al. 2016; Spiliopoulou et al. 2015). The current version of IBID can also extract the structure and parts of the causal behaviors of a system. Figure 2 illustrates IBID's information pipeline for extracting this information for the biological systems from biology articles.

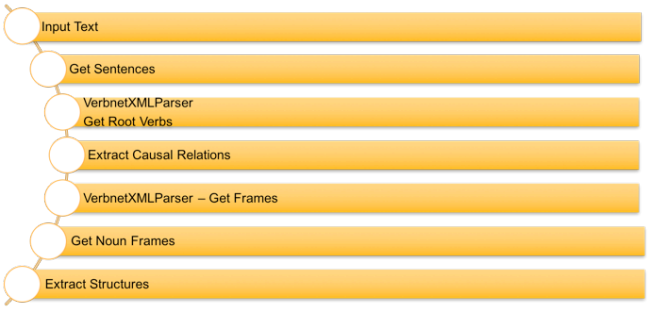

Figure 2: A part of IBID's information pipeline.

For each sentence in a biology article, IBID uses the NLP parser to obtain its phrase structure grammar representation in the form of a tree. Each valid phrase start token in the tree represents the root node of a subtree whose leaf words are combined to create a logical sentence component. For example, one component of "Minute water droplets from fog gather on its wings; there the droplets stick to…" is "Minute water droplets from the fog gather on its wings".

A. Function Extraction

As indicated above, IBID uses a domain-independent controlled vocabulary of functions described in Spiliopoulou et al. (2015) and Rugaber et al. (2016). Each function in this controlled vocabulary is expressed as a frame in VerbNet. The first step of function analysis is to generate a Stanford Dependency (SD) object for a given sentence component, the root of the SD tree is the predicate of the sentence and is then stemmed to produce the root verb. For example, "gather" is the root verb for the component "Minute water droplets from the fog gather on its wings". SD also provides information on whether the root verb is passive by listing any passive nominal subjects. Root verbs for which there are VerbNet records will have their VerbNet syntactic frames matched against the parser's Part-Of-Speech (POS) tags. The best matched predicate, its VerbNet syntactic frame, thematic relations mapping from our sentence component to the frame, and the sentence itself are saved in the database. When retrieved, IBID annotates the article with all this functional information.

B. Behavior Extraction

As with functions, we have developed a domain-independent vocabulary for behavioral concepts and relations. Also as with functions, each behavioral term in this controlled vocabulary is expressed as a frame in VerbNet. As in function extraction, each sentence component is parsed for its predicate and then stemmed. Next, the root verbs are matched against causal verbs in our vocabulary of behavioral terms. If the system finds a causal verb, then IBID replaces it with a verb token and matches it against a list of predefined regular expressions capturing various forms of causal patterns. These causal regular expression patterns will also delineate the sentence component's cause and effect clauses. The causality record, which includes a stemmed predicate, its cause/effect clauses, and the original sentence component are then saved in the database. Finally, IBID annotates the article with this behavioral information.

C. Structure Extraction

IBID searches each sentence in the biology article for terms in Vincent's domain-dependent structure ontology. If it identifies a structural term, it then searches for adjectives that describe the structure/nouns. In addition, IBID uses WordNet is used to find synonyms, hyponyms, and meronyms for each structural term identified. IBID performs this additional search to map the structural terms from the domain-specific ontology into our domain-independent structure ontology. IBID then annotates the article with all this structural information.

V. Faceted Search

IBID uses two kinds of search to locate biology articles in a corpus relevant to a design query: faceted search in which it uses domain-independent controlled vocabularies of structure, behavior and function terms; and search based on design queries stated as English language sentences. In the latter case,

[2] https://wordnet.princeton.edu
[3] https://verbs.colorado.edu/~mpalmer/projects/verbnet.html

the current version of IBID does not use behavioral knowledge for locating biology articles.

Figure 3 shows IBID's interface for faceted search. The three panes on the left show some of high-level terms for search based on function, behavior and structure respectively.

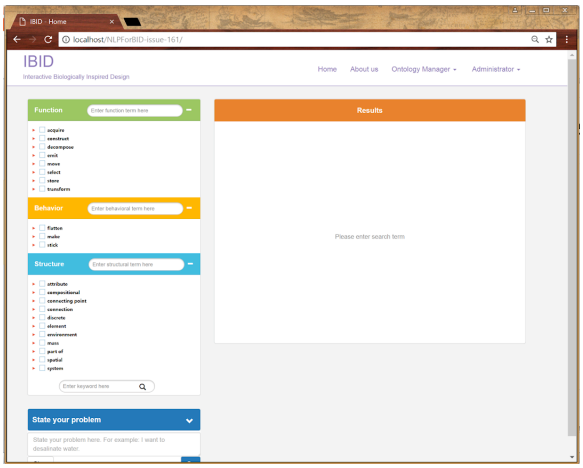

Figure 3: Faceted Search in IBID.

The function facet's controlled vocabulary has eight high-level function terms and multiple sub-level terms (Rugaber et al. 2016; Spiliopoulou et al. 2015). Given a designer's selection of functional terms, IBID searches the functional annotations on articles for the selected verbs.

The behavior facet's controlled vocabulary is made up of a list of causal verbs. When a designer selects causal verbs of interest, IBID searches for articles possessing one or more behavior annotations with the selected verb(s) as their focus.

The structural facet's controlled vocabulary consists of the domain-independent structure ontology we have developed (not shown here). Recall that when IBID extracts structural terms from biology articles, the terms are domain-dependent. In the current version of IBID, we manually map the domain-dependent structural annotations on the biology articles and the domain-independent terms in the faceted search. We intend to automate this process (and IBID already performs automated ontology alignment for natural language search).

VI. Search Based on Natural Language Queries

While faceted search based for functions was already working in IBID (Rugaber et al. 2016; Spiliopoulou et al. 2015), as mentioned earlier, IBID can now also search based on design queries expressed in English sentences. The bottom pane on the left side of Figure 4 indicates the query "I want to create a system for transporting liquid."

IBID first identifies functional and structural terms in the design query (verbs and nouns/noun phrases, respectively) and lemmatizes/stems them. For example, for the input "I want to create a system for transporting liquid", IBID finds *Function: [want, create, transport]* and *Structure: [system, liquid]*.

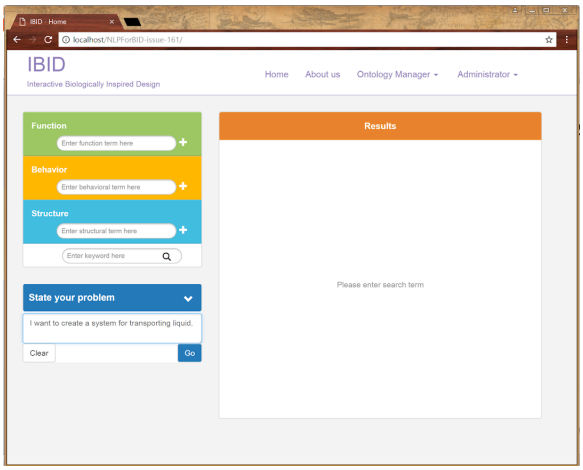

Figure 4: Search based on a design query in IBID.

Next, IBID enlarges this query by adding domain-independent structure terms and high-level function terms. For example, for the input "I want to create a system for transporting liquid", this results in *Function: [acquire, want, construct, create, move, transport]* and *Structure: [system, portion, liquid]*. Finally, IBID uses the same mechanism as in faceted search based on the structural and functional annotations on the biology articles.

VII. Preliminary Testing

Figure 5 illustrates IBID's current state. Greyed-out regions and text indicate steps or functionality yet to be implemented. The unshaded regions indicate steps with functionality currently implemented in IBID. Further refinements to these steps will serve to improve IBID's performance, as will implementing the greyed-out regions.

Here is the piece of text taken from a biology research article (Chrispeels & Maurel 1994) that we have used for evaluating parts of IBID:

```
Bulk flow of water across a membrane occurs in
response to an osmotic or hydrostatic gradient.
Osmotic water permeability is readily measured in
small vesicles or cells by the stopped-flow light-
scattering technique, a method that relies on the
dependence of light scattering on vesicle or cell
volume, and is used to quantitate the time course of
net water flow that occurs in response to
transmembrane osmotic gradients. The osmotic
gradients are established by adding an impermeant
solute to the external solution. With the help of
other chemical and physical methods to measure
diffusional and osmotic water transport across
biological membranes, biophysicists and cell
physiologists have obtained evidence for the
existence of facilitated or channel-mediated water
transport in several membranes (see Macey, 1984;
Finkelstein, 1987). Membranes with facilitated water
transport share a number of properties (Macey, 1984;
Verkman, 1992) that generally support but do not
prove the presence of water channels. For example,
water transport across these membranes is inhibited
by mercurial sulfhydryl reagents, demonstrating the
existence of proteinaceous components in water
channels, and the functional unit of the water
channel in kidney tubules and red blood cells is 30
kD, as determined by radiation inactivation (van
Hoek et al., 1991).
```

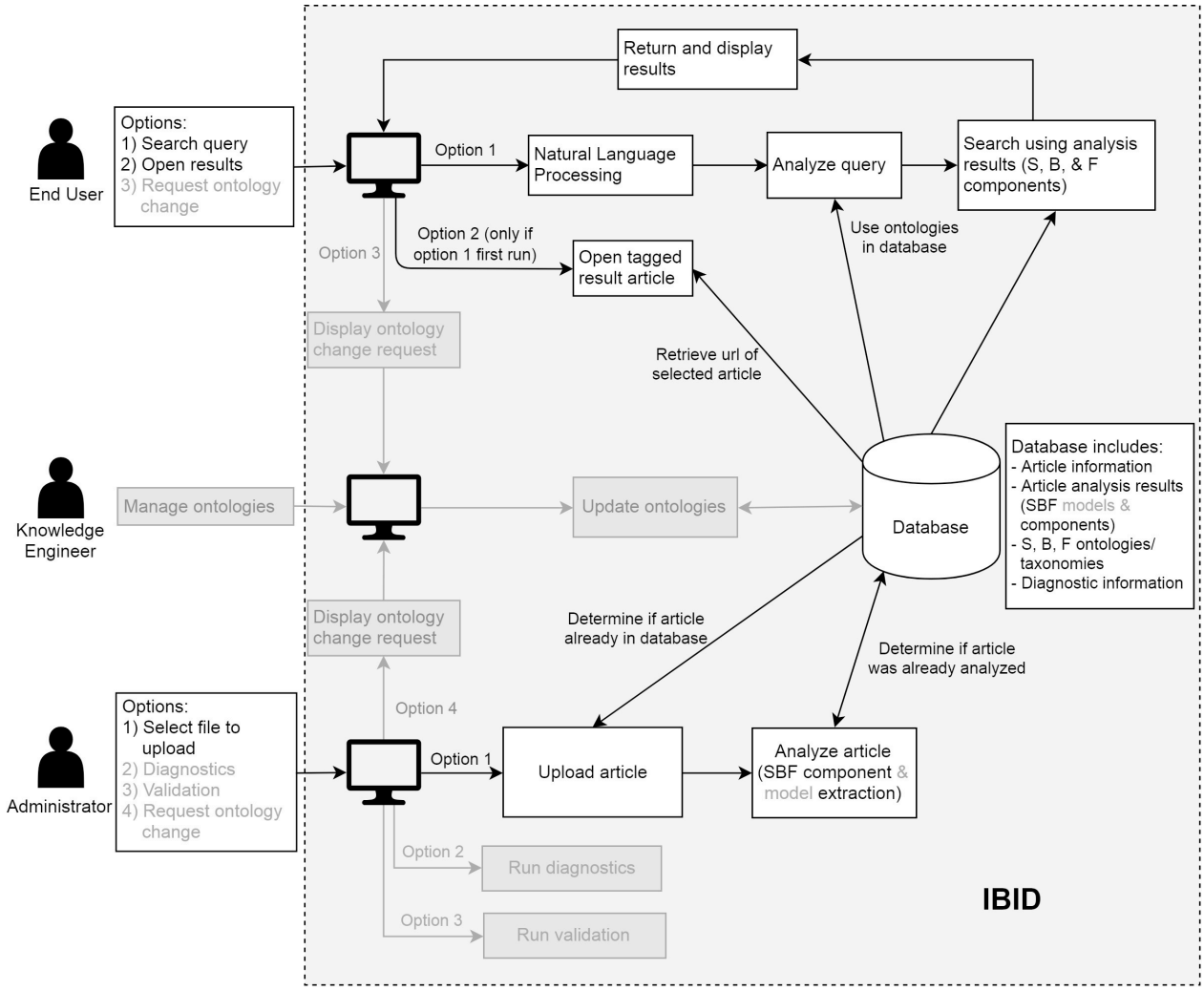

Figure 5: The current state of the IBID system.

We selected seven participants for our study, where the participants were not experts in biology (as with most biologically inspired designers). We asked the participants to

> List the structure terms in the above paragraph. Structure terms refer to the components, substances and connections of a system. For instance, in the following sentence: Trees can transport water from the ground by their vascular system. "water", "ground" and "vascular system" are the structure terms of the system "tree".

We gave exactly the same text and problem to IBID and compared the results with the human participants. We used the commonly used F_1 metric for the comparison, as it captures both the fraction of relevant terms that were retrieved as well as the fraction of the retrieved terms that were relevant. We found F_1 for identifying structural terms in the above experiment to be 79%. An interesting observation is the recall was higher than the precision, meaning that there were a larger number of false positives as compared to false negatives. While promising, these results are very preliminary; we still need to test IBID on a large corpus of full biology articles.

VIII. Discussion

The current, second generation of computational tools for supporting biologically inspired design focuses on AI techniques for automatically locating biological knowledge relevant to design problems. IBID is an interactive system presently under development for helping biologically inspired designers locate biology articles that describe biological systems relevant to a design query as well as to produce SBF models of the systems.

IBID first extracts structural, behavioral and functional terms in biology articles and annotates the articles with the terms. Then, given a natural language design query, IBID locates the biology articles relevant to the query based on the articles' annotations. IBID uses two kinds of search to locate biology articles: faceted search based on domain-independent controlled vocabularies of structures, behaviors and functions; and natural language query search currently for function and structure. Preliminary results based on faceted search using structural terms appear promising.

ACKNOWLEDGMENTS

We are grateful to Julian Vincent for sharing his ontology for describing biological systems (Vincent 2014); IBID uses Vincent's domain-specific ontology of biological components and connections.

REFERENCES

Benyus, J. (1997) *Biomimicry: Innovation Inspired by Nature.* William Morrow.

Cheong H, Chiu I, Shu L. Stone R., McAdams, D. (2011) Biologically Meaningful Keywords for Functional Terms of the Functional Basis. *Journal of Mechanical Design* 133:21007.

Chrispeels, M., & Maurel, C. (1994). Aquaporins: The Molecular Basis of Facilitated Water Movement Through Living Plan Cells? *Plant Physiology*, 105(1), 9-13.

Deldin, J-M., Schuknecht, M. (2014) The AskNature database: enabling solutions in biomimetic design. In A. Goel, D. McAdams & R. Stone (Eds.), *Biologically inspired design:* , Springer-Verlag, London (2014), pp. 17-27.

Glier, M., McAdams, D., & Linsey, J. (2014) Exploring Automated Text Classification to Improve Keyword Corpus Search Results for Bioinspired Design. *Journal of Mechanical Design*, 136(11).

Goel, A., Awasthy, P., Creeden, B,, Kumble, M., Salunke, S., Sarathy, S., Shetty, A., Vijayaraghavan, D., & Wiltgen, B. (2016). Using Watson for Supporting Design Creativity. In *Procs. Fourth International Conference on Design Creativity,* Atlanta, Georgia, October 2016.

Goel, A., McAdams, D., & Stone, R. (Editors, 2014) *Biologically inspired design: Computational methods and tools*. Springer-Verlag, London.

Goel, A., Rugaber, S., & Vattam, S. (2009) Structure, Behavior and Function Models of Complex Systems: The Structure-Behavior-Function Modeling Language. *AIEDAM* 23: 23-35, April 2009..

Goel, A., Vattam, S., Wiltgen, B., & Helms, M. (2012) Cognitive, Collaborative, Conceptual and Creative - Four Characteristics of the Next Generation of Knowledge-Based CAD Systems: A Study in Biologically Inspired Design. *Computer-Aided Design* 44(10): 879-900, May 2012.

Goel, A., Zhang, G., Wiltgen, B., Zhang, Y., Vattam, S., & Yen, J. (2015). On the Benefits of Digital Libraries of Analogical Design: Documentation, Access, Analysis and Learning. *AIEDAM* 29(2), May 2015.

Helms, M., Vattam, S., & Goel, A. (2009) Biologically Inspired Design: Products and Processes. *Design Studies* 30(5):606-622, September 2009.

Helms, M., Vattam S., & Goel, A. (2010) The Effects of Functional Modeling on Understanding Complex Biological Systems. In *Proc. 2010 ASME Conference on Design Theory and Methods,* Montreal, Canada, August 2010.

Kruiper, R., Vincent, J., Chen-Burger, J. & Desmulliez, M. (2017). Towards identifying biological research articles in computer-aided biomimetics. In *Procs. Conference on Biomimetic and Biohybrid Systems*, 242–254. Springer.

Nagel J., & Stone R. (2012). A computational approach to biologically inspired design. AIEDAM 26 (2): 161 -176..

Rugaber, S., Bhati, S., Goswami, V., Spiliopoulou, E., Azad, S., Koushik, S., Kulkarni, R., Kumble, M., Sarathy, S., Goel. A. (2016) Knowledge Extraction and Annotation for Cross-Domain Textual Case-Based Reasoning in Biologically Inspired Design. In *Procs. 24th International Conference in Case-Based Reasoning (ICCBR 2016).* 342-355, Atlanta, USA, October 2016; pp. 342-355.

Shu L. (2010) A Natural-language approach to biomimetic design. *AIEDAM* 24:507–519.

Spiliopoulou, E., Rugaber, S., Goel, A., Chen, L., Wiltgen, B., & Jagannathan, A. (2015) Intelligent Search for Biologically Inspired Design. In *Proc. 20th ACM Conference on Intelligent User Interfaces,* Atlanta, Georgia, March 2015; pp. 77-80.

Vandevenne, D., Verhaegen, P., Dewulf, S., & Duflou, J. (2016). SEABIRD: Scalable search for systematic biologically inspired design. *AIEDAM* 30(01), 78-95.

Vattam, S., & Goel, A. (2011) Foraging for inspiration: Understanding and supporting the information seeking practices of biologically inspired designers. In *Proc. 2011 ASME DETC Conference on Design Theory and Methods*, Washington DC, August 2011.

Vattam, S., & Goel, A. (2013) Seeking Bioinspiration Outline: A Descriptive Account. In *Proc. 19th International Conference on Engineering Design (ICED13*), Seoul, Korea, August 2013, pp. 517-526.

Vincent, J. (2014). An ontology of biomimetics. In *Biologically Inspired Design: Computational Methods and Tools,* A. Goel, D. McAdams & R. Stone (eds.), pp. 269-285, London: Springer.

Vincent, J., & Mann, D. (2002). Systematic technology transfer from biology to engineering. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,* 360(1791), 159-173.

Yen, J., Helms, M., Goel, A., Tovey, C., Weissburg, M. (2014) Adaptive Evolution of Teaching Practices in Biologically Inspired Design. In *Biologically Inspired Design: Computational Methods and Tools*, A. Goel, D. McAdams & R. Stone (editors), Chapter 7, pp. 153-200, London: Springer-Verlag.