

EKOSS – An Ontology-based Semantic Web System for Knowledge Sharing

Steven Kraines

Science Integration Programme – Human, the University of Tokyo

sk@scint.dpc.u-tokyo.ac.jp

The EKOSS (Expert Knowledge Ontology-based Semantic Search) system has been developed for supporting the sharing, discovery, and integration of expert scientific knowledge using semantic web and AI technologies. The current EKOSS system supports two ontologies that are founded on description logics and implemented in OWL-DL: the SCINTENG ontology extending the Epistle Core Model and the SCINTHUMAN ontology extending the GALEN ontology. The EKOSS concept of providing a system for empowering domain experts to author their own A-Boxes describing their knowledge resources is described and details regarding the development of the SCINTHUMAN ontology for biosciences are discussed.

1. Introduction: the Expert Knowledge Ontology-based Semantic Search System

EKOSS (Expert Knowledge Ontology-based Semantic Search) has been developed for sharing expert scientific knowledge using semantic web technologies (Kraines *et al.* 2006a, Kraines *et al.* 2006b). The EKOSS system web site provides a collaborative knowledge sharing environment where knowledge experts can submit computer-readable descriptions of the knowledge resources that they have developed, such as research papers, databases, computer simulation models, and even *curricula vitae*. It then allows other users to make requests for knowledge resources related to some knowledge resource requirement, such as a particular condition or system problem that they are studying. The semantic search system matches user requests to computer-readable knowledge resource descriptions using inference based on description logics and rules that are provided for each of the domain ontologies that are registered to the system.

2. Leveraging semantic web technologies and DL ontologies for EKOSS

The primary development goal of the EKOSS system is to provide a deployable web portal software package for empowering scientific researchers to publish their own computer-readable semantic descriptions, for example as a part of the scientific paper publication process (Kraines *et al.* 2006a, Kraines *et al.* 2005). However, as a part of this development, we have been using the EKOSS system prototype as a test bed to explore the potential for adding value to published knowledge resources based on the logical formalism underlying the semantic descriptions created for each of the knowledge resources on the system. In particular, development of ontology T-Boxes with sufficient semantic richness and logical structure has been an important topic in our research.

Early in the project, we elected to use description logics (DL), specifically the *SHOIN(d)* description logic supported by OWL-DL, as the logical formalism for the EKOSS system, mainly because of the relatively extensive support for DL in the form of reasoning software such as KAON, pellet and RacerPro, and ontologies, such as the Epistle Core Model and GALEN ontologies. We envision three levels of applications that could be realized through the accumulation of computer-readable semantic knowledge descriptions authored by the creators of the knowledge resources themselves:

1. *Individual resource level*: Creators of knowledge resources can benefit from the computer-interpretable semantic descriptions of those resources because computers can use the descriptions to provide high level semantic services such as specialized views, automatic language/concept translation, and inference of embedded “facts”.
2. *Semantic search level*: A repository of computer-interpretable semantic descriptions of knowledge resources can be established that can be searched against using a DL reasoner. This kind of semantic search could increase search precision by letting searchers specify relationships between terms, and increase search recall because the search engine can draw on embedded relationships in both the knowledge description and the search conditions.
3. *Knowledge structuring/mining level*: Overall relationships and semantic “motifs” can be extracted from large repositories of semantic knowledge descriptions through application of graph algorithms and semantic technologies, etc.

In the past three years, we have gone through several cycles of an iterative process involving searches for usable DL technologies, establishment of competency questions and system specification goals, implementation of software elements of the EKOSS framework, and case studies and beta-testing by research assistants (mainly undergraduate and graduate students doing studies in one of the targeted domains). From the beginning of the EKOSS system development, we have chosen the upper level ontologies of the Epistle Core Model (European Process Industries STEP Technical Liaison Executive. 2005) and the GALEN ontology (Rector *et al.* 1997) to provide the basis for knowledge models that act as formalized languages for creating semantic descriptions in the domains of engineering and life sciences, respectively. We then developed the domain ontologies SCINTHUMAN and SCINTENG by extending these upper level ontologies with concepts and properties for the targeted domains.

3. Development of the SCINTHUMAN ontology for life sciences

The following are some of the competency questions that we have come up in the development of the SCINTHUMAN ontology based on GALEN:

- “Is there a paper about a protein located in the plasma membrane of a unicellular organism that activates the phosphorylation of a protein whose name contains the text string ‘ypd’?”
- “Is there a paper about a protein that has a basic non-polar amino acid following a lipid attachment site and that is the substrate of a molecular modification process regulated by some enzyme?”
- “Is there a paper about a biomolecule that is transported to a cell nucleus and that contains a semi-metallic chemical element from the third row of the periodic table?”
- “Find all papers that describe regulation mechanisms for post-translational molecular modifications of mitochondrion proteins in the brain cells of primates”
- “Find the commonly occurring collocation and regulation patterns involving retroviruses”
- “Find all evidence that CK20 is involved in a disease process in epithelial cells”
- “Find and integrate all protein regulation processes occurring in green algae”

In order to meet these competency goals, the SCINTHUMAN ontology provides elements for formalizing knowledge from the life sciences domain in the following areas:

- 1) biomolecular processes regulated by specific biological entities or other processes
- 2) structural composition of physical entities in biology from biomolecules to organisms
- 3) sequences of structure components in biomolecules, such as nucleic acids in genes or domains in proteins
- 4) taxonomical classification of organisms (from the NCBI taxonomy at <http://www.ncbi.nlm.nih.gov/sites/entrez?db=Taxonomy>)
- 5) functional characteristics of biochemical building blocks, including chemical elements, nucleic acids, and amino acids
- 6) secondary structures of proteins (from the Tambis ontology)
- 7) qualitative and quantitative properties using the feature/state mechanism from GALEN as explained in Rector *et al.* (1997)

Through the iterative process of the EKOSS system development outlined above, we identified several issues concerning the use of the original GALEN ontology for the envisioned applications and competency requirements of our system:

1. a mechanism to limit the degrees of freedom in property choice for describing particular relationships between instances is needed in order to make the interface for creating semantic descriptions more usable
2. the DL axioms and property attributes are not semantically rich enough to support the kind of inferences that we hope to achieve with the system
3. the conceptual knowledge model for process regulation and sequences of processes was inadequate for supporting semantic inferences in the targeted domain

In order to address these issues, the SCINTHUMAN ontology modifies and extends on the GALEN ontology in the following ways.

- Added constraints to **Process – Regulation** model
 - Reification of the concept of process regulation and other modifications to align ontology with conceptualizations of GO, BioPax, and SBML
 - **Regulations** and **SpecificProcesses** are disjoint
 - Process **Actors** must operate through a **Regulation**
 - **Regulations** cannot be domains of **hasProduct** or **hasSubstrate**

- Added model for describing **SequenceStructures**
 - For describing protein and nucleic acid sequences
 - **occursBefore/After, occursDirectlyBefore/After...**
- Reconsideration of **Classifications** and **Roles**
 - **OrganismClassification** holds species information
 - **StructureClassification** to designate groups of structures, e.g. proteins
 - **Roles** generally only applied to structures (role of a process defined by the role of its structures)
- “Disabled” high-level properties in semantic knowledge description authoring tools
 - **hasRole, hasFeature, hasActor, hasActee, ...**

In addition, we have made several major expansions to the ontology, including the following:

- Added classification trees for organisms (from NCBI taxonomy)
- Added complex class definition axioms for elements, nucleic acids, and amino acids

Using the property type identifiers that are provided by the OWL-DL specification, we define properties as being functional, inverse functional, transitive and/or symmetric where appropriate, and we define inverse property relationships for all of the major properties given in the SCINTHUMAN ontology. A schematic diagram showing the allowed property connections between the main classes in the SCINTHUMAN ontology is given in figure 1.

4. An example of application of the SCINTHUMAN ontology

Figure 2 shows the semantic description that was created from the following abstract text (Essers *et al.* 2006):

“Beta-catenin is a multifunctional protein that mediates Wnt signaling by binding to members of the T cell factor (TCF) family of transcription factors. Here, we report an evolutionarily conserved interaction of beta-catenin with FOXO transcription factors, which are regulated by insulin and oxidative stress signaling. Beta-catenin binds directly to FOXO and enhances FOXO transcriptional activity in mammalian cells.... Association of beta-catenin with FOXO was enhanced in cells exposed to oxidative stress.... These results demonstrate a role for beta-catenin in regulating FOXO function that is particularly important under conditions of oxidative stress.”

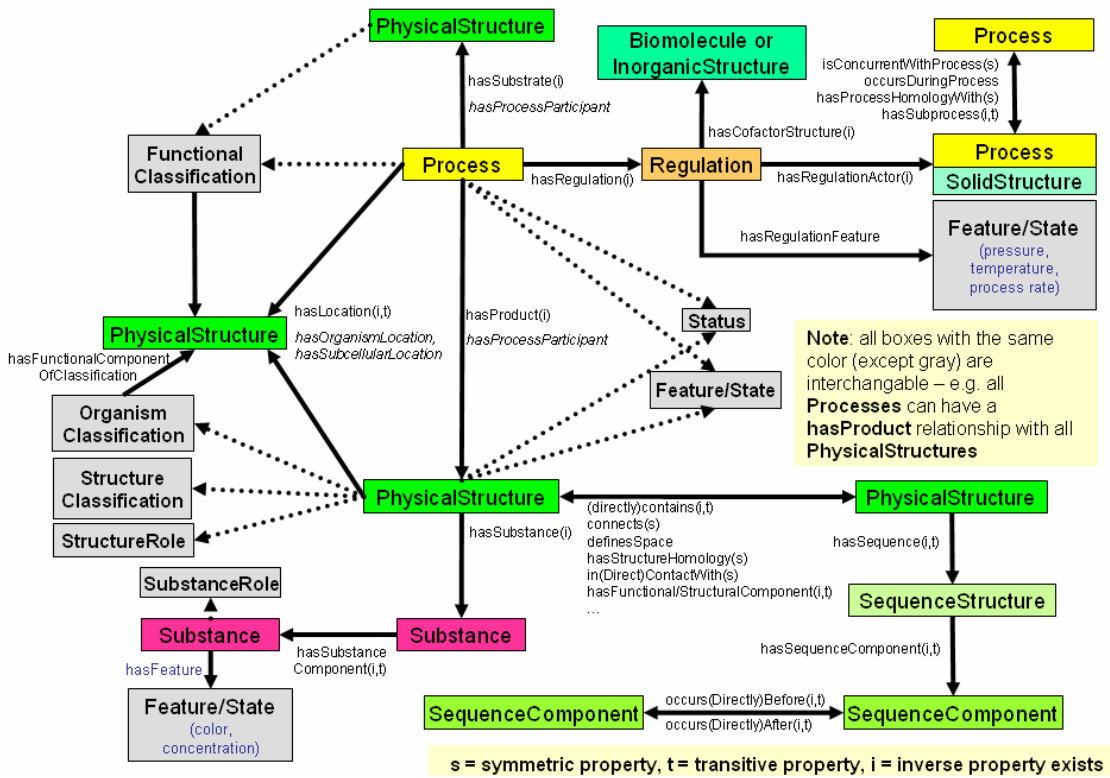


Figure 1: the SCINTHUMAN knowledge model diagram

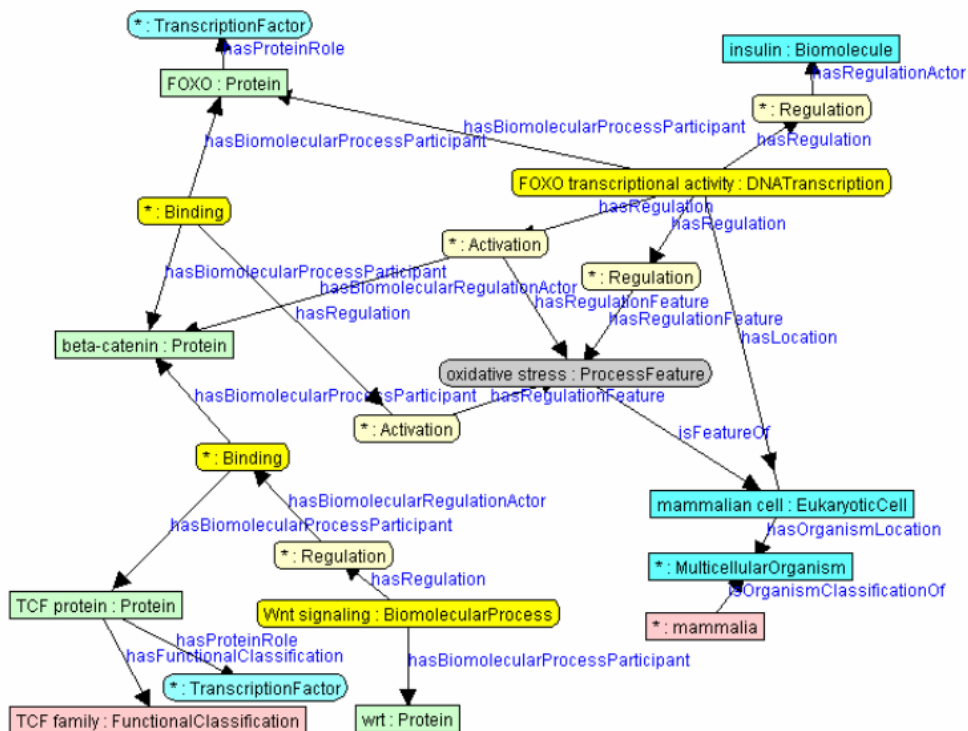


Figure 2: example of A-Box constructed according to the SCINTHUMAN ontology

5. Summary

We are developing the EKOSS system with the goal of providing an effective means for computer-supported sharing, discovery and integration of expert scientific knowledge. Several technologies from the domains of artificial intelligence and the semantic web are being leveraged towards the development of the EKOSS system. In particular, we have been developing two ontologies, the SCINTENG and SCINTHUMAN ontologies, for domains of engineering and of life sciences, respectively. We will continue the iterative process described above through at least one more cycle, and we welcome comments and suggestions for making the SCINTHUMAN ontology a more effective formalized language for creating computer-readable semantic descriptions of knowledge resources from the domain of life sciences, and for making EKOSS system a more effective platform for supporting the sharing, discovery, and integration of expert knowledge.

References

- Essers, MAG, de Vries-Smits, LMM, Barker, N, Polderman, PE, Burgering BMT, Korswagen, HC. 2005. Functional interaction between β -Catenin and FOXO in oxidative stress signaling. *Science* 308: 1181-1184
- European Process Industries STEP Technical Liaison Executive. 2005. EPISTLE Core Model (www.btinternet.com/~Chris.Angus/epistle/specifications/ecm.html)
- Kraines S.B., R. Batres, M. Koyama, D. R. Wallace, H. Komiyama. 2005. Internet-based integrated environmental assessment: using ontologies to share computational models. *Journal of Industrial Ecology* 9(3):31-50
- Kraines, S. B., W. Guo, B. E. Kemper, and Y. Nakamura. 2006a. EKOSS: A knowledge-user centered approach to knowledge sharing, discovery, and integration on the Semantic Web. *Proceedings of the 2006 International Semantic Web Conference*
- Kraines, S. B., W. Guo, B. E. Kemper, and Y. Nakamura. 2006b. A semantic web application for expert knowledge sharing, discovery, and integration. *Proceedings of the 2006 International Semantic Web Conference*
- Rector, A., Bechhofer, S., Goble, C., Horrocks, I., Nowlan, W., Solomon, W. 1997. The GRAIL concept modeling language for medical terminology. *Artificial Intelligence in Medicine* 9: 139-171