

Christopher Thomas

PhD Candidate in Computer Science

Email: topher@knoesis.org

Phone: +1-937-687-8674

<http://knoesis.org/researchers/topher>



377, Joshi Research Center.
3640 Colonel Glenn Highway,
Dayton OH 45435-0001.
Phone: (937) 775-5203.
Fax (937) 775-5133 .
<http://knoesis.org>

Research Interests and Applications (short)

Broad research areas I am working in:

- Knowledge representation.
 - Ontology design paradigms.
 - Analyzing and agreements between human and formal conceptualizations.
- Information extraction
 - NLP methods, based on analysis of parse trees and dependency graphs.
 - * Efficiently find multi-word expressions.
 - * Disambiguate concept mentions.
 - Based on statistical, non-NLP methods.
 - Based on a combination of both.
- Machine Learning
 - Statistical techniques applicable for information extraction.
 - Combining informed and uninformed methods.
 - Tackling the problem of data sparsity

- Computer Vision

Specific application areas include:

- Computational Semantics.
 - Semantic Web.
 - Analogical Reasoning.
- Knowledge extraction from community-generated content
 - Automatic domain-model generation.
 - Information extraction from large corpora and the web in general.
 - * Extracting semantic relations.
 - * Finding syntactic information.
- BioInformatics for Glycan Expressions
 - Formal ontology creation.
 - Constrained population of the ontology with formal representations of molecules mined from textual representations.
- Image Processing
 - Image Segmentation.

- Image(segment) annotation.

Research Interests and Projects (extended)

In my PhD research I have focused on approaching formal semantics from different angles. For a machine to easily draw deductive conclusions, it needs to have access to a formal representation of the information it is fed. I first explored what the limitations of a computational understanding of semantics are, even in the face of a perfect formal representation. For example the so-called symbol-grounding problem is a fundamental limitation of computing machines. Subsequently I researched ways to increase the expressiveness of formalisms to account for uncertainty in the represented domains. With this background, I focused my research on the question of how to make a computer understand (written) language in particular. The understanding paradigm is here a Wittgensteins *meaning-as-use*. We cannot teach a machine a comprehensive understanding of language and human communication, but we can a)algorithmically define domain-specific actions on relations and b)give a limited vocabulary to users for them to show us how the concepts and relations are used in a specific domain. The specific issues I am concerned with are:

- How to automatically extract more meaningful information from unstructured data
 - HPCO: Automatically create an ontology for the domain of human cognitive performance using Doozer domain model creation, pattern-based information extraction and NLP-based information extraction.
 - * Create a domain taxonomy.
 - * Connect domain concepts with semantic relations using macro-reading/pattern-based techniques.
 - * Find mentions of concepts and relations in research literature (MedLine) by parsing, disambiguating and mapping.
 - Develop a distributed parsing and analysis framework to efficiently process text.
 - Recognize syntactic and semantic types, complex (multi-word) expressions and map them to known concepts.
 - Use syntactic structures to infer semantic relatedness.
 - * Use the created knowledge base of mentions, relations and the taxonomy to allow for a focused search of the research literature that also allows browsing the extracted relations as links that were not previously present in the literature.
 - Doozer: automatically creating domain models/thesauri based solely on a user query.
 - Pattern-based Information Extraction: use statistical techniques on surface patterns to find instances of relationship types in free text.
 - TaxaMiner: semiautomatic taxonomy extraction from biomedical abstracts using clustering techniques and NLP.
- How to model formal ontologies and how to meaningfully represent information
 - e.g. Modeling of GlycO, an Ontology for the Biochemistry domain. Representations of molecular structure and metabolic pathways.
 - How to (semi-)automatically populate ontologies from structured and semistructured sources, such as databases and database driven websites.
 - * e.g. Extraction of complex molecule representations from multiple sources
 - How to transform and disambiguate the extracted data to gain more meaningful representations

- * e.g. Transforming textual representations of complex molecules into a canonical formal representation used in GlycO; after the transformation, molecules from different sources could be compared and incorrect molecules could be discarded (many databases, even though manually curated, carry legacy errors from older databases. Small percentages of incorrect data can be detrimental for automatic reasoning over the database contents)
- How to build formal representations that still match human conceptualizations of the domain
- How to find more powerful knowledge representations formalisms that allow for uncertainty and contradictions or are aware of the degree of truth/certainty/subjectivity of the knowledge
 - see e.g papers on Semantics for the Semantic Web: the Implicit, the Formal and the Powerful or On the Expressiveness of the Languages for the Semantic Web - Making a Case for “A Little More”
- How to harvest individual and symbiotic human intelligence as computing resources, especially how to gain formal knowledge representations.
 - In the process of recent web developments, e.g. what is known as Web 2.0, individuals and communities have started to accumulate quasi-formal knowledge in the form of tags and annotations.
 - This information can be analyzed, processed and formalized
 - * e.g. augmenting a folksonomy to a taxonomy
 - The contributors can be coaxed into more carefully reviewing the information provided
 - * e.g. community review in Wikipedia
 - My focus here is on how to provide a framework that does not only allow users to provide more formal representations, but that a) makes the formalism transparent, so the user does not have to deal with it and b) make it desirable to participate and to provide highest quality. See e.g. Luis von Ahn’s work on Human Computation and my Essay on Web Wisdom.

In my Masters work , the focus of my research was Semantics in general and the question of how to make a computer understand (written) language in particular. I concentrated on automatic ontology learning using hierarchical clustering, natural language processing and analogical reasoning.

- Other topics I have been working on during my AI and CS studies were:
 - Document clustering and Taxonomy learning using Hierarchical k-means or Naïve Bayes Classifier and NLP techniques
 - Automatic texture recognition and ontology-based image annotation. This tool trained on several textures and was given a spatial ontology of how objects consist of different textures
 - Genetic algorithm to build Hidden Markov Models for predicting protein secondary structures
 - Genetic algorithm that compose music.
 - Neural Networks that can track a ball for the RoboCup robot soccer.
 - Automatic maze generation for the game Mummy Maze, involving algorithms to find solutions for the mazes which used heuristics and forward/backward chaining procedures.
 - Artificial Agents whose behavior was dependent on
 - * Rule Systems
 - * Decision Trees
 - * Neural Networks
- to demonstrate the strengths and shortcomings of different forms of representations in AI.

- Other interests
 - Logic:
 - * First Order Logic and its subsets, such as Description Logics, Frame logic.
 - * Nonmonotonic logics, such as defeasible logics.
 - * Uncertain reasoning using probabilistic or fuzzy logics.
 - * Analogical Reasoning
 - Language: Processing of language in the human brain, especially Metaphors. This was the first step towards my interest and active research in analogical reasoning. There is much evidence that humans see their worlds metaphorically, i.e. we try to see new discoveries in the light of known ones, we tend to describe concepts in technical domains with words that are very close to our everyday experience (e.g. parent-child relations in tree data structures, etc.). Based on this I believe that the ability to do analogical reasoning will assist a computer in learning concepts.

Education

Wright State University (Relocated in Jan 2007) Dayton, OH
Ph.D. Computer Science Jan 2007- June 2010

- **Ph.D Dissertation:** Manual and Automated Ways for Ontology Creation (*Working title*).

University of Georgia Athens, GA
Ph.D. Computer Science Aug 2003- Dec 2006

- Relevant Courses: Semantic Web, Statistical Language Models (BioInformatics), Scientific Computing (distributed computing), Advanced Information Systems, Compilers, Automata and Formal Languages, Advanced Algorithms

University of Georgia Athens, GA
M.S. Program Artificial Intelligence Aug 2001- Aug 2003

- Relevant Courses: NLP, Artificial Intelligence, Logic Programming, Computational Intelligence, Machine Learning, Genetic Algorithms, Philosophy of Language, Epistemology, Cognitive Psychology

Universität Koblenz Germany
BS in Computervisualistik Oct 1998- May 2001

- Relevant Courses: Computational Linguistics and Semantics, Software Engineering, Computer Vision, HCI, Cognitive Psychology, Aesthetics

Universität zu Köln Germany
Studies in Mathematics, German Linguistics/Literature and Philosophy Feb 1995 - June 1998

- Relevant Courses: Calculus, Linear Algebra, Computational Linguistics, German Linguistics

Research Experience

Research Assistant and Student Coordinator Semantic Web Research Lab
kno.e.sis center Jan 2007-Present.

- Partially funded from Semdis (NSF) and HP Incubation/Innovation grants
- Investigated the quality of community generated Web content.
- Used the identified high quality content as text corpus for domain thesaurus/ontology extraction.
- Currently leading a project on Information Extraction and Use in the domain of human cognitive performance.
- Published (and under review) 6 peer-reviewed conference and journal papers, including one paper in AAAI and two papers in ACM/IEEE Web Intelligence, a competitive conference with an acceptance rate of 16% (2007) and 17% (2008).

Research Assistant

LSDIS Lab, University of Georgia

GlycO
Aug 2003-Dec 2006.

- Funded by Bioinformatics for Glycan Definitions Grant
- Investigated the role of Ontologies in Bioinformatics.
- Published 7 peer-reviewed conference, workshop and journal papers and 2 invited papers, including one paper in AAAI and one paper in WWW (top conferences in the area of Semantic Web)

Research Assistant

Universität Koblenz

Image Recognition Lab
1999-2000.

- Developed camera calibration software
- Image Segmentation algorithms and filters.

Professional Experience

Visiting Researcher

Max Planck Institute Informatics, Saarbrücken, Germany

Database group
Jun 2009 - Oct 2009

- Developed statistical methods to deal with incomplete data.

Summer Research Intern

Hewlett Packard Labs, Palo Alto, CA

Storage Group
May 2007 - Sep 2007

- Developed a system to automatically create domain thesauri based on keyword specifications.
- The algorithm was patented and served as the prototype for the Taxonom.com service.

Summer Research Intern

Amazon.com Seattle, WA

Amapedia Group
May 2005 - Sep 2005

- Developed algorithms to automatically determine missing product specifications.
- Exploratory research into creating the Amapedia service.

Grant Proposals

I have contributed extensively to the grant proposals listed below during my Ph.D. In addition to actively participating in the writing process, the results from my research have been used in funded proposals.

- [1] My work on pattern-based fact extraction was initially funded by an HP Incubation grant (2008).
- [2] The work got further funding from an HP Innovation grant (2009).
- [3] A spin-off of the work in Domain-Model creation and pattern-based fact extraction, extended with NLP-based entity and relationship recognition/extraction was funded by the US Air Force Research Labs, Wright Patterson Air Force Base, Dayton, OH. I am currently leading the research efforts in this project
- [4] Significant contribution to NSF proposal on Information Extraction using large text corpora and fact corpora in the form of Linked Open Data. The idea is to have a hermeneutic circle of knowledge. Extracted and verified facts are automatically put back into the collection of Linked Open Data and can serve as future training data. Submitted December 2009, under review

Conference Publications

- [1] C. Thomas, P. Mehra, W. Wang, A. Sheth, G. Weikum, and V. Chan. A Pertinence Measure for the Extraction of Named Relations. *Under preparation for ICDM*, 2010.
- [2] C. Thomas, R. Kavuluru, W. Wang, C. Ramakrishnan, D. Cameron, P. Mendes, A. Sheth, P. Parikh, V. Chan, P. Fultz, and A. Jadhav. Model-Based Focused Literature Search. *Under preparation for ISWC*, 2010.
- [3] C. Thomas, P. Mehra, W. Wang, A. Sheth, G. Weikum, and V. Chan. Automatic domain model creation using pattern-based fact extraction. *Submitted to CIKM Conference*, 2010.
- [4] C. Thomas, W. Wang, P. Mehra and A. Sheth. What Goes Around Comes Around Improving Linked Open Data through On-Demand Model Creation. *Web Science Conference*, 2010.
- [5] C. Thomas, P. Mehra, R. Brooks, and A. Sheth. Growing Fields of Interest - Using an Expand and Reduce Strategy for Domain Model Extraction. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:496–502, 2008.
- [6] C. Thomas and A. Sheth. Semantic Convergence of Wikipedia Articles. In *Proceedings of the 2007 IEEE/WIC International Conference on Web Intelligence*, pages 600–606, Washington, DC, USA, November 2007. IEEE Computer Society.
- [7] S. S. Sahoo, C. Thomas, A. Sheth, W. S. York, and S. Tartir. Knowledge Modeling and its Application in Life Sciences: A Tale of two Ontologies. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 317–326, New York, NY, USA, 2006. ACM Press. (Equal contribution from first 2 authors, the paper was in equal parts about their contributions to the Glycomics project with UGA-CS and CCRC)
- [8] C. Thomas, A. Sheth, and W. York. Modular Ontology Design Using Canonical Building Blocks in the Biochemistry Domain. In *Proceeding of the 2006 conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, pages 115–127, Amsterdam (NL), 2006. IOS Press.
- [9] P. Doshi and C. Thomas. Inexact matching of ontology graphs using expectation-maximization. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pages 1277–1282. AAAI Press, 2006.

Journal Publications

- [10] C. Thomas and A. Sheth. Web Wisdom - An Essay on How Web 2.0 and Semantic Web can foster a Global Knowledge Society. *Submitted to Computers in Human Behavior*, Elsevier.
- [11] P. Doshi, R. Kolli, and C. Thomas. Inexact matching of ontology graphs using expectation-maximization. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(2):90–106, 2009.
- [12] V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth. Taxaminer: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266, 2005.

- [13] A. P. Sheth, C. Ramakrishnan, and C. Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *Int. J. Semantic Web Inf. Syst.*, 1(1):1–18, 2005.
- [14] S. Sahoo, C. Thomas, A. Sheth, C. Henson, and W. York. GLYDEan expressive XML standard for the representation of glycan structure. *Carbohydrate research*, 340(18):2802–2807, 2005.

Workshop Publications

- [15] A. Sheth, W. York, C. Thomas, M. Nagarajan, J. Miller, K. Kochut, S. Sahoo, and X. Yi. Semantic Web technology in support of Bioinformatics for Glycan Expression. In *W3C Workshop on Semantic Web for Life Sciences*, pages 27–28, 2004.
- [16] N. Oldham, C. Thomas, A. Sheth, and K. Verma. *METEOR-S Web Service Annotation Framework with Machine Learning Classification. Semantic Web Services and Web Process Composition*, pages 137–146, 2005, Springer.

Book Chapters

- [17] C. Thomas and A. Sheth. On the expressiveness of the languages for the semantic web - making a case for a little more. *Fuzzy Logic and the Semantic Web*, pages 3–20, 2006.

Patent

- [18] P. Mehra, R. Brooks and C. Thomas. ONTOLOGY CREATION BY REFERENCE TO A KNOWLEDGE CORPUS.

Professional Activities

• Project Leader

- GlycO - An ontology for Complex Carbohydrate expression - led the ontology creation, 2004-2006
- HPCO - An ontology for human cognitive performance - currently leading the research efforts. I am guiding 2 junior PhD students and one programmer. 2008 and ongoing

• Program Committee memberships

- Extended Semantic Web Conference (ESWC) 2010, Social Web Track
- EKAW 2010
- KREAM'2010 Knowledge Representation and Applied Decision Making
- Social Semantic Web 2009
- CORE 2008
- COMBEK 2008
- WISM 2008
- External Reviewer: WWW 2009, ISWC 2008, CIKM 2007

Professional Skills

Languages: Java, C, C++, Prolog, PHP, XHTML, CSS, Javascript, SQL

Operating Systems: Linux; Windows 9x, XP, Vista; Mac OS X

Knowledge Representation: RDF/RDFS, OWL, XML, Different Logics Formalism

Other Skills and Extracurricular Activities

Spoken Languages: English, German, French

Music: I play several musical instruments, e.g. Piano, Guitar, Clarinette

Awards and Accomplishments

- Awarded a rare undergraduate research assistantship at the Image Recognition Lab at the University of Koblenz, Germany
- Stipend from the DAAD (German Academic Exchange Service) to study Artificial Intelligence at the University of Georgia, Athens, GA, USA. (2001)
- My Research on domain model creation and fact extraction was awarded an HP Incubation grant in 2008

References

Amit P. Sheth
Director, kno.e.sis center
Wright State University

Dayton, OH
+1-937-239-0625
<http://knoesis.org/amit>
amit@knoesis.org

Pankaj Mehra
Distinguished Technologist
HP Labs Palo Alto

Palo Alto, CA
+1-408-858-0916

dr.pankaj.mehra@gmail.com

Other references available on request.