Final report for Office of Naval Research Grant N000141010841

David I. Spivak

August 18, 2013

This report will summarize my progress toward the goals of ONR grant N000141010841 (“Categorical Information Theory”), which was in effect from June 2010 to June 2013. Throughout this period I have been employed by the Department of Mathematics at the Massachusetts Institute of Technology (MIT). Until March 2013 I was a Postdoctoral Associate, under the guidance of Professor Haynes Miller, and in March 2013 I was promoted to the position of Research Scientist. The Technical Proposal for this grant can be found online at: http://math.mit.edu/∼dspivak/informatics/technical_proposal2010.pdf.

1 Goals for this grant

My goals for this grant are discussed in Section II.3 and Section III of the Technical Proposal. I will summarize them in a list them below, and in the subsequent sections I will discuss the degree to which I was able to accomplish each of them. My goals for this grant were to:

- Generalize simplicial databases using ”categorical databases”,
- Formulate a connection between databases and ontologies,
- Consider a ”fiber bundle” approach to information architecture and querying,
- Initiate the “mathematical referencing project”,
- Investigate applications outside of mathematics,
- Find a connection between grammars and databases.

I will discuss each of these in turn below, and then I will also discuss some other accomplishments I made during the project period.

1.1 Categorical databases

Before this grant, I had worked on formulating databases using the topological theory of simplicial sets. Shortly before this grant began, I realized that a broad generalization was possible, which would make the connection between databases and categories much more straightforward. In the technical proposal for this grant, I proposed to study this possibility under the name “categorical databases”.

In fact, this idea worked out quite well. I published a paper on the subject, called “Functorial data migration” in *Information and Computation* in 2012. In this paper I showed that database schemas and categories are almost exactly the same thing, and that database instances are functors. I spoke about this subject in the mathematics department and again in the computer science department at MIT. This work also formed the basis for several other papers and collaborations, which I will discuss below.

1.2 Ontology logs

The work on categorical databases fit well with some work I had been doing on understanding the informational structure of natural language. I synthesized these two ideas as something called *ologs*, or ontology logs. These are like concepts maps, involving text-boxes and labeled arrows between them, each of which can be read as an English sentence. The key difference between ologs and concept maps is that ologs must follow certain semantic rules. These rules guarantee that each olog is not only human readable, but also serves as a database schema.

I wrote a paper on this subject and talked about it at seminars in the Math department and the Linguistics department at MIT. This led to a collaboration with Professor Markus Buehler (head of the department of civil and environmental engineering at MIT), who specializes in materials science. We wrote several papers together, using ologs to describe materials at various levels and for various applications. I will say more on this below.

1.3 Lifting problems

I was able to find a simple reworking of my categorical databases idea that fit well with the "fiber bundle" idea, which I had imagined was possible in my technical proposal. Using something called the *Grothendieck construction*, a schema $\mathcal{C}$ and a functor $I\colon \mathcal{C} \to \mathbf{Set}$ could be transformed into a fibration $\int I \to \mathcal{C}$.

What struck me about this, once I investigated it, was its similarity to an existing model of semi-structured data, called the *Resource Descriptive Framework* (or RDF), often used to represent data from the world-wide web. The typical methodology for querying RDF data, namely the SPARQL graph-pattern query, was easily relatable to a common approach to investigating objects in the field of Algebraic Topology, namely the lifting problem approach. After talking with Eric Prud'hommeaux, who works at the W3C at MIT, I wrote a paper about this subject which is set to be published in *Mathematical structures in computer science*. It provides a more structural translation between the RDF and relational data models.

1.4 Mathflow

Although I have not dedicated much time to it, a company called ThoughtCycle has contacted me regarding my proposal to map out the structure of mathematical definitions and theorems. In a project called Math-Flow, which they plan to submit as an SBIR, we propose to put online a more highly-structured version of wikipedia, for math students and researchers.

Like wikipedia it will have a page for each article, but unlike wikipedia, the links will carry context. The idea is that when one is looking at a proof which references another theorem, it is often unclear how that theorem applies in the current context. How do the variables in the proof fit with the variables in the other theorem? The idea of mathflow is that in order to construct a link between, say, a proof to another theorem, one must include information describing precisely how the theorem is being applied in the current context, i.e. how the variables from one are being attached to the variables of the other.

If successful, this may do more than offer a more straightforward user experience in understanding subjects within mathematics. It may also make the mathematical canon more digestible by computers, as well as give insight into the very fabric of mathematics. Like the "friendship graph" in facebook is one of that company's most valuable assets, the network of interaction between various theorems, definitions, and proofs within mathematics could become quite valuable.

1.5 Applications

My efforts to apply my research have led to two successful directions, namely computer science and materials science. The latter is a bit surprising – why materials science? My guess is that it's not so much because of the immediate applicability of my research to materials science. Instead, I believe the work is more broadly applicable but the contingency of history led to a meeting between myself and Markus Buehler, each of whom was interested in the collaboration.

In computer science I have worked with a Harvard graduate student named Ryan Wisnesky to reformulate my paper on Functorial Data Migration into the language and style of relational database theorists (rather than that of category theorists, for whom it was written). During that project I made substantial additions, greatly clarifying the notion of *attributes*. I now understand these not as structurally information-bearing, but instead as an interface with the human being. In other words, suppose a single database was tasked with handling the an entire mail-delivery operation. In this case, it can change the names and IDs of addresses, senders, receivers, routes, etc. at will, as long as the connection pattern between these things is unchanged.

But as soon as the addresses or routes or senders refer to something not within the database but within the *outside world*, the database is given an additional task. While it can continue to shuffle around and change values in its internal naming system for various addresses, etc., it now must maintain a method for making connections between those names and those used in the outside world. These are the attributes.

In materials science, I have worked with Professor Markus Buehler (department head of Civil and Environmental Engineering at MIT), and closely with his graduate student Tristan Giesa. We've published four papers, using ologs to formalize analogies between hierarchical protein materials and social networks and between spider silk and music. We also coined the term *building block replacement problem*, in which ones goal is to find ways to replace an expensive or rare material building block with one of less environmental impact or cost, while still retaining the same functionality. We showed how ologs may be useful in that endeavor.

1.6 Kleisli database instances

As discussed above, one of my hopes was to understand how grammars and natural languages could be captured mathematically. While I did not end up finding anything new to say about naturally integrating context-free grammars into the field of information management, I did work on relaxing some of the constraints of relational (or more precisely categorical) databases. While not directly related to grammars, one can see the various relaxations as changes of the pragmatic linguistic context, like saying "for the remainder of this conversation, whenever I say X is a Y, I always will mean that some human H observes X to be a Y".

Using monads, I showed how one can allow the data in a given column to be non-atomic, while still having many good properties. For example one could allow bags, lists, sets, and other data structures in each column, rather than simple atomic elements. The categorical model of databases already fits nicely with both the functional programming language paradigm and the the relational model of databases, which is in computer applications, though often felt to be a bit rigid. By using monads to relax some of this rigidity, the approach continues to fit nicely with the functional paradigm which has been highly invested in monads for the last 20 years. The relationship between this work and linguistic context was spelled out in my book, *Category theory for scientists*, Chapter 5.3.

1.7 Category theory for scientists

During the winter of 2012 / 13, I wrote a book called *Category theory for scientists* which I used to teach a class in the spring 2013 semester. The purpose of this class was to disseminate my research, to hear student ideas on how well it fit with their own subject (i.e. to continue my research through student challenges to the applications I was proposing), and to create the potential for more collaborations in the future. During the course, the book grew from 214 pages to 261 pages, because the students would ask for more explanation on various topics and I would use that to improve the book.

I believe this book will be useful to many people in non-mathematical fields, who are beginning to hear tell of the power of category theory. Since putting the book online, I have received a large increase in the amount of unsolicited email from academics and practitioners in various fields.

The book has been accepted for publication by MIT press.

1.8 Functorial query language

As mentioned above, I have been working with a Harvard graduate student named Ryan Wisnesky on bridging the divide between my work on categorical databases and the language and tools most prevalent in computer science research. In the process, he developed a language called *Functorial Query Language* (FQL), which is open-source and freely available online.

FQL offers a simple syntax by which one can: input a finitely presented category $\mathcal{C}$ (i.e. database schema), input a set-valued functor $I\colon \mathcal{C} \to \mathbf{Set}$ (i.e. a database instance), input a functor $F\colon \mathcal{C} \to \mathcal{D}$ between schemas (i.e. a schema mapping). All of this can be displayed in a variety of ways, including tabular, graphical, and JSON. One can also pullback an instance along any functor ($\Delta_F I$), as well as finding the left or right Kan extensions ($\Sigma_F I, \Pi_F I$).

These form the basis for a new conception of databases that has not been seen before. Based on feedback by experts, this new formulation seems to be worthwhile (e.g. Val Tannen wrote to us in an email, “I find this *very* interesting.”).

2 Papers, presentations, collaborations, and outreach

Below I will list some publications, presentations, collaborations, and outreach I have been involved with in connection with the ONR grant.

2.1 Book

Spivak, D.I. (2013) *Category theory for scientists. Accepted for publication by MIT Press.* 261 pages. Available online: http://arxiv.org/abs/1302.6946

2.2 Papers

- Spivak, D.I.. (2013) “Database queries and constraints via lifting problems.” To appear in *Mathematical structures in computer science*. ePrint available: http://arxiv.org/abs/1202.2591
- Spivak, D.I. (2012) “Functorial Data Migration”. *Information and Communication*. Vol 217, pp. 31 – 51. ePrint available: http://arxiv.org/abs/1009.1166
- Giesa, T.; Spivak, D.I.; Buehler, M.J. (2012) “Category theory based solution for the building block replacement problem in materials design”. *Advanced Engineering Materials.* DOI: 10.1002/adem.201200109
- Spivak, D.I.; Kent, R.E. (2012) “Ologs: a categorical framework for knowledge representation”. *PLoS ONE* 7(1): e24274. doi:10.1371/journal.pone.0024274.
- Wong, J.Y.; McDonald, J.; Taylor-Pinney, M.; Spivak, D.I.; Kaplan, D.L.; Buehler, M.J. (2012) “Materials by design: Merging proteins and music”. *Nano Today* 7, issue 6, pp. 488 – 495.
- Giesa, T.; Spivak, D.I.; Buehler M.J. (2011) “Reoccurring patterns in hierarchical protein materials and music: The power of analogies”. *BioNanoScience* 1 no. 4, pp. 153-161.
- Spivak, D.I.; Giesa, T.; Wood, E.; Buehler,M.J. (2011) “Category Theoretic Analysis of Hierarchical Protein Materials and Social Networks.” *PLoS ONE* 6(9): e23911. doi:10.1371/journal.pone.0023911

Preprints

- Spivak, D.I. (2013) “The operad of wiring diagrams: Formalizing a graphical language for databases, recursion, and plug-and-play circuits.” Submitted to . *Applied categorical structures.* Available online: http://arxiv.org/abs/1305.0297
- Spivak, D.I.; Wisnesky, R. (2012) “On the relational foundations of functorial data migration.” Submitted to *ICDT*. ePrint available: http://arxiv.org/abs/1212.5303.
- Spivak, D.I. (2012) “Kleisli database instances”. ePrint available: http://arxiv.org/abs/1209.1011.

2.3 Invited Presentations

Courant Institute 2012/12/03;
U. Oregon 2012/11/12;
Brown U. 2012/09/19;
Mathfest (Madison, WI) 2012/08/04;
Stanford Center for Biomedical Informatics Research (Colloquium) 2012/07/27;
Office of Naval Research 2012/06/13;
U. Texas (Special geometry seminar) 2012/01/31;

Amgen Inc. 2012/01/24–25;
Carnegie Mellon U. (POP seminar) 2012/01/18;
UIUC (MSS Colloquium) 2011/11/29;
Agent-based complex systems conference (IPAM) 2009/10/13. Johns Hopkins (Topology seminar) 2011/11/21
Amgen Inc. 2011/02/17–18;
New England Database Summit (poster, joint with Carlo Curino) 2011/01/28;
Boston Haskell 2011/01/20;
Harvard U. 2010/11/03 (EECS seminar);
Galois Inc. 2010/10/22 (Tech talk);
MIT 2010/09/20 (Topology seminar);
MIT 2010/09/16 (CSAIL seminar);
MIT 2010/09/15 (Linguistics – semantics reading group);
Foundational methods in computer science conference (U. Calgary) 2010/06/12;
Galois Inc. 2010/06/03 (Tech talk).

2.4 Collaborations

Below is a list of collaborations that have successfully led to papers.

I collaborated with:	**resulting in:**
Robert Kent	A published paper in which we announce ologs. A first version of this paper was written by myself alone—here I discussed rules for putting English phrases on objects and arrows in a category to yield structure that both has clear semantic content (i.e. is easily understood by users) and also serves as a database schema. Based on a request by the referee, Kent wrote a section describing a tight connection between this work and that of Sowa on semantic nets and the information flow framework, as well as a connection to the *institutions* of Goguen and Burstall.
Markus Buehler, Tristan Giesa, Elizabeth Wood, Joyce Wong, John McDonald, Micki Taylor-Pinney, David Kaplan	Four published papers, with various subsets of the seven people to the left as coauthors. In two of these papers we used category theory, and ologs in particular, to make formal analogies between materials science and other domains: western music and social networks. In another we wrote about using category theory as a formal design language. In another we coined the term "building block replacement problem" to describe the holy grail of materials science: to replace expensive or rare material building blocks with cheap and available ones, while retaining the functionality. To do that, one has to understand how fine points in form and structure lead to the incredible functionality we see in biological materials such as spider silk. Enunciating these complex structures and their functionality can be done using ologs.
Ryan Wisnesky	A paper submitted to a database conference (ICDT). In it we clarify the relationship between my work on categorical databases ("Functorial Data Migration") and classical relational database theory. We also added significantly to my previous work by introducing attributes, which do not hold information internal to the database but are significant to users. We also worked together to produce a working application called FQL which implements almost all of my theory of functorial data migration.

2.5 Outreach

I am interested in disseminating my research more widely to a variety of audiences. To that end, I have participated in the following outreach activities.

I was involved with:	**in which I:**
Harvard dance program	Was a performer in a 4-person dance at the Harvard Dance Center on November 29, 30, and December 1, 2012. While the other 3 were dancing, I did a "TED-talk" style lecture (complete with introductory "TED" music) involving my theory of ologs. That is, I used ologs to spell out, with mathematical precision, a critique of the dance currently being performed by the other three cast-members.
Boston math circle	Gave two lectures to a group of about 15 gifted math students, whose ages ranged from about 9 to about 16. Both were about category theory. One of them focused on higher categories using tiling diagrams, the other was about operads and wiring diagrams.
MIT UROP program	I led nine (9) UROPs (Undergraduate Research Opportunity Program) with six different undergraduate students at MIT. All of these were about categorical information theory in a various forms, ranging from probabilistic ologs, to neuroscience, to programming with circuits, to Petri nets and reaction networks.
Johnson and Johnson and Amgen	Gave several talks about my research on categorical databases to each company's informatics team. In each case we worked long past the end of the talk, first to make clear how what I had done could be useful in their work, and second to explain to me the outstanding problems they were most concerned with.
Science News	Was interviewed for a Science News (web edition) piece called One of the most abstract fields in math finds application in the 'real' world, published on May 20, 2013. Here I talk about the potential societal significance of using category theory to model real-world phenomena.
Teaching at MIT	I taught a graduate-level course in the mathematics department at MIT, called "Category Theory for Scientists". I had 18 students enrolled, nine of which were from the math department, and the other nine were from various science and engineering disciplines. Each student did a final project relating category theory to their field of expertise.