# Report for ONR Grant N000141310260

David I. Spivak

October 19, 2013

## Contents

## 1 Introduction

This report, submitted to Predrag Neskovic at the Office of Naval Research, summarizes my past-year's progress toward the goals of ONR grant N000141310260. Broadly speaking, my objective is to understand the fundamental nature of information and communication, and to express it mathematically. Claud Shannon's *Information theory* deliberately avoids the question of meaning; in my research, on the other hand, meaning is of primary importance.

In order for data to be meaningful, it must carry some structure. The relatively young mathematical subject of *Category theory* is the study of structure, and I have identified it as a particularly relevant field for the describing the nature of information. The idea is that there exist many different notions of "informativity", i.e. many types of information-bearing structure, including databases, ontologies, computer programs, neural networks, etc. Each such notion includes an infinite array of conforming structures (e.g. an infinite variety of different database schemas are possible), and each such structure is instantiated by an infinite array of conforming data (e.g. an infinite variety of instances for a given database). This layering of abstraction is a major strength of category theory. Another (related) strength is the ability to translate between different paradigms (e.g. from one database to another, or from ontologies to databases).

Below I will summarize three advances I have made in the past year of research. These include

**Section 2:** a new database query language (called FQL) and its implementation;

**Section 3:** a new approach to wiring diagrams, which captures their self-similar nature; and

**Section 4:** a book I wrote to disseminate my research.

I will then include a section (Section 5) that lists my papers, presentations, and transitions.

## 2 FQL

Functorial query language (FQL) was developed with Ryan Wisnesky, a Computer Science graduate student at Harvard, who will work with me as a postdoc next year at MIT. It is a new database query language based on a joint paper (submitted to the ICDT 2014 database conference), called On The Relational Foundations Of Functorial Data Migration.

### 2.1 Main contribution

This work advances my 2010 paper Functorial data migration in three ways. First, it makes that work accessible to a broader audience of computer scientists, by describing it logically. Second, it adds a rigorous description of “attributes”, which can be seen as the translation of a database’s machine-readable linking structure into the “arbitrary” language of human users. Third, it implements all this in a user-friendly program, which is freely available and open-source.

### 2.2 Importance

The relational model of databases, invented by Codd, is based on the somewhat outdated mathematics of set theory. Category theory and its close cousin Homotopy Type Theory, greatly extend set theory. Based entirely in the category-theoretic framework, FQL provides a new approach that dovetails with programming languages and modern mathematical theory. Thus it opens up a broad spectrum of research possibilities. In June 2013 Val Tannen, a well-known database theorist, wrote to Ryan and myself about our paper, saying, “I find the paper *very* interesting!”

### 2.3 Limitations of existing approaches

Current database *practice* is quite far from current database *theory.*This is unfortunate, and in my opinion it is caused by the inherent brittleness of the relational model, which has become entrenched.

### 2.4 Plan

Ryan and I plan to pursue this project, in particular benchmarking it and comparing its strengths and weaknesses to existing approaches. Ryan also hopes to start a company in a few years with FQL as its core IP.

I have also been included, as an “unfunded collaborator”, on an AFOSR FY 2014 MURI white paper entitled “Homotopy type theory: unified foundations of mathematics and computation”. Team members include Fields medalist Vladimir Voevodsky, and Steve Awodey, both of whom are major contributors to this new area of research. I am excited to work with this team because its application to my informatics works looks quite promising.

## 3 Wiring diagrams

Wiring diagrams are used to describe circuits and serve as a standard for specifying manufacturing processes. It is often noted that wiring diagrams have a self-similar nature: one can be “plugged into” another, thus adding a layer of detail that was previously hidden from view. Thus wiring diagrams have a hierarchical structure, which has never been well-described mathematically. I have written two papers on this topic.

### 3.1 Main contribution

This work presents an *operadic* framework for wiring diagrams, which has two advantages. First it provides a concrete specification language for wiring diagrams that elegantly expresses their self-similar nature. Second, it makes an abstraction barrier between the diagram itself and its possible functional meanings. This barrier is formally understood as the distinction between an operad and its algebras.

### 3.2 Importance

By using category theory to formalize wiring diagrams, their structure can be understood by a wider audience, lowering the barrier to high-quality open source implementations. One contribution of our approach (joint work with Dylan Rupel) is the temporality, or causal nature, of directed wiring diagrams. That is, we solved a long-standing problem in the field: we showed that the data emitted from a wiring diagram (even one containing feedback loops) is based only on past inputs, never on inputs which come in the future. As obvious as this should be, it has been difficult to prove using past approaches.

### 3.3 Limitations of existing approaches

Current approaches do not formally explain how substitution of wiring diagrams into other wiring diagrams should work, nor do they prove the temporality theorem described above.

### 3.4 Plan

I plan to continue this work with collaborators Dylan Rupel and Nat Stapleton. We expect to find applications to spreadsheets (which are currently considered morasses of complication). We also expect to find applications in cognitive neuroscience or artificial intelligence.

## 4 Category theory for the sciences

I wrote a book, called *Category Theory for the sciences*, which has been accepted for publication by the MIT Press. I also taught a class at MIT on this subject in the Spring 2013 semester. It had a diverse enrollment of 18 students = 7 undergraduates + 11 graduate = 5 math + 4 EECS + 3 physics + 3 engineering + 3 other.

### 4.1 Main contribution and importance

The main contribution of this book is to disseminate to a broader scientific audience the possible advantages of a category-theoretic approach to information modeling. This book has led to a few new collaborations, including with industry (Honeywell) and a neuroscience graduate student at MIT.

### 4.2 Limitations of existing approaches

Category theory as a language of science is a new possibility; in some sense, that is, there are no existing approaches. Zooming out a bit, we could say that the existing approach to communicating science and scientific research is via English prose. The limitation of prose is that it is difficult to be precise, especially if one also wishes to also be concise. This difficulty is unfortunately exploited to hide details and caveats in research. It also prevents one lab from easily reproducing the methods of another. By mathematically formalizing scientific exposition, one can hope to lower the barrier to and thus greatly increase the dissemination of knowledge.

### 4.3 Plan

While this work is in some sense complete, I do plan to continue to disseminate my research as broadly as possible. I may write another book, e.g. on wiring diagrams, at some point. I also plan to teach the "Category theory for scientists class" again in Fall 2014 or Spring 2015.

## 5 Papers, presentations, and transitions

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

### 5.1 Book

Spivak, D.I. (2013) *Category theory for the sciences.* Accepted for publication by MIT Press. 261 pages. Old version available online: `http://arxiv.org/abs/1302.6946`

## 5.2 Papers

- Spivak, D.I.. (2013) "Database queries and constraints via lifting problems." *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. Available online: http://math.mit.edu/~dspivak/informatics/BuildingBlockReplacement.pdf

### Preprints

- Rupel, D.; Spivak, D.I. (2013) "The operad of temporal wiring diagrams: formalizing a graphical language for discrete-time processes". ePrint available: http://arxiv.org/abs/1307.6894
- 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.* ePrint available: 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.

## 5.3 Invited Presentations

NIST 2013/06/12;
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;

## 5.4 Transitions

I have been told of three transitions of my research.

The first was a transition of ologs by a collaboration between Amgen and Microsoft. This transition occurred more than a year ago, but I'm including it here anyway. Dave Balaban, a former VP at Amgen said this to me (in an email from September 2013):

> As for your influence, it has been substantial. Our gold standard for specifying anything now is that it must be categorical. We are beginning to trust nothing else. Our current simulation specifications owe much to you. Ologs have influenced us significantly as well. I now understand that knowledge representation can be rigorous and extendable. You send us down this path and have helped us along the way. I am very grateful that we have been able to work together. Hopefully more collaboration will come.

The second was by an emerging company called MJBA-Tech. In September 2013, a representative wrote me saying:

> We spoke briefly this afternoon about a possible working or advisory relationship with our company, MJBA-Tech. The company is based in the Washington, DC area. We have developed a working operational prototype of a categorical database. Now we are seeking external funding from investors to go into full development.

The third was by a developer named Tom Ellis in the UK, who implemented my operadic approach to databases. He wrote (coincidentally also in September 2013):

> I am writing to let you know that I have implemented a relational query language (as an embedded domain specific language (EDSL) in Haskell) which is based on your operad formulation of queries in ”The operad of wiring diagrams”.
>
> For several years I have been trying to understand the essence of relational queries but it wasn't until I read that paper that everything clicked into place.
>
> [SNIP]
>
> The EDSL compiles to SQL and has been used in production for the last three months to run queries against my clients existing PostgreSQL database.
>
> I would summarize this language as an unqualified success. It is easier to write relational queries in it than it is to use SQL, partly because true nature of the relational queries is laid bare, and partly because it is embedded in Haskell, a rich supporting language. Personally I think this approach is revolutionary, although it may only appear revolutionary to those who already consider Haskell revolutionary, a small but accomplished group of programmers.
>
> Since it has been developed as part of a work contract the code is not available for now. I hope in time it will be open sourced.