University of Zurich[UZH]

Abraham Bernstein
Natasha Noy

Is This Really Science? The Semantic Webber's Guide to Evaluating Research Contributions

TECHNICAL REPORT – No. IFI-2014.02

2014

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

Abraham Bernstein, Natasha Noy
Is This Really Science? The Semantic Webber's Guide to Evaluating Research Contributions
Technical Report No. IFI-2014.02
Dynamic and Distributed Information Systems
Department of Informatics (IFI)
University of Zurich
Binzmuehlestrasse 14, CH-8050 Zurich, Switzerland
http://www.ifi.uzh.ch/ddis/

Is This Really Science? The Semantic Webber's Guide to Evaluating Research Contributions

Abraham Bernstein*and Natasha Noy†

April 7, 2014

Abstract

The Semantic Web is an extremely diverse research area. Unlike scientists in other research fields, we investigate a diverse set questions using a plethora of methods. The goal of this primer is to provide context for scientists in the Semantic Web and Linked Data domain about the purpose of research questions and their associated hypotheses, the tension between rigor and relevance thereof, possible evaluation approaches typically used, and pitfalls in terms of reliability and validity.

For example, where is the scientific problem in developing a system or a tool? How do we frame the discussion of generating linked data from a given corpus such that others will actually care about our work? When is it a good idea to use our (Semantic Web) technology when the problem has already been successfully attacked by other means?

We strive to make this primer as practical as possible. Hence, after a short more theoretical introduction we will pick up a series of examples from our research domain and use them to exemplify the implications of our introductory theoretical treatment. We hope that this text will help the reader to explore the scientific basis of their research more systematically.

*bernstein@ifi.uzh.ch, University of Zurich, Zurich, Switzerland
†natashafn@acm.org, Stanford University, Palo Alto, CA, USA

1 So Why Should I Read This?

> It was Genevieve's first day in graduate school and she had a great idea of how to improve the Semantic Web. Indeed, she had been attracted to the idea of sharing data on the web ever since she first put together a spreadsheet of her favorite songs and had wanted to share it with a friend. Having had lots of science courses such as Biology, Physics, and Chemistry during her undergraduate studies, she was sure that making a scientific contribution would be easy: wasn't a scientific contribution 'just' adding understanding of how something worked in the real world?
>
> But this is where the problems started. What was the real world in a domain such as the Semantic Web (or Computer Science)? She read through the proceedings of the International Semantic Web Conference and plowed through a number of journals. It seemed to her that many of these papers did not actually describe how to understand the world better—whether real or abstract—but rather described solutions or approaches on how to attack a certain problem. The papers did not describe natural phenomena, but rather analyzed properties of the systems that authors or their colleagues have created. This approach was distinctly different from the approach in the traditional sciences. Furthermore, a solution in one set of papers seemed sometimes completely different from a solution in another set of papers. Everything started to look like a huge mess...

Genevieve's is not alone with her problem. Indeed, the Semantic Web is a confusing domain to do research in. The field, if it can be called a field, has gone through a number of iterations of what is the 'right' way to do things and what questions should be asked. Even for senior members of the field these questions crop up repeatedly. At program committee meetings and in reviews, we discuss those issues at length and often disagree. In science, we are supposed to "stand on the shoulders of giants"[1] and to allow others to build on our work. For this continuity to be possible *we need to develop a notion of what a valid contribution is*, such that others can build on stable foundations. Hence,

> *the goal of this primer is to provide context for scientists in the Semantic Web and Linked Data domain about the purpose of research questions and their associated hypotheses, the tension between rigor and relevance, possible evaluation approaches, and pitfalls in terms of reliability and validity.*

Note that we are not the only field that is at odds with understanding what constitutes a scientific contribution. In 1989, Banville and Landry (1989) wrote a very interesting paper about the field of Management Information Systems, or MIS, called "Can the Field of MIS be Disciplined." They were asking themselves whether MIS, which was what business-school research about computers was called at the time, was actually a discipline. They provided three dimensions for defining the cohesion of a research field. The first dimension was:

[1] attributed to John of Salisbury, see MacGarry (1955) and `http://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants`

Functional dependence "... the extent to which researchers have to use the specific results, ideas, and procedures of fellow specialists in order to construct knowledge claims which are regarded as competent and useful contribution."

We focus on this dimension in this text. The reason for writing this primer is that both when writing our own papers and when reviewing papers written by others we sometimes feel that this question of reliability and validity of research results seems undervalued in our discipline.

We wanted to be as practical as possible in this text. We think of it as a primer with lots of examples for people like Genevieve, who are starting their research career and would like some practical help in developing a sense of what is a valid approach to convince others of one's results and where they should be weary to follow in others' footsteps.

What this text is not about: This text is not a complete primer to research in our field. Indeed, we focus only on the purpose of research questions and their associated evaluation. We do not focus on approaches for gathering such research questions such as grounded theory[2] or Case Study research (Eisenhardt, 1989). We also completely disregard the other dimensions that Banville and Landry (1989) focused on: mainly, what and how stable the right questions to be investigated are.[3]

The next steps: We organized this text as follows: first, we dive into a bit of theory and ask ourselves what research really is. But don't worry, we are not going to enter meta-theoretic discussions about ontology, epistemology, and the human nature (we refer the so inclined to Burell and Morgan (1979)). We will just provide sufficient context to dive into our main question. Next, we explore the question of validity by example. We explore a number of typical research questions of our field and discuss their different incarnations in terms of usefulness and validity. We close with a list of a few things that we believe you should absolutely remember at the end. So if you are in hurry, then go on and jump directly to Section 4—but you will miss all the fun stuff in the middle.

2 Laying Some Foundation (What Others Call Definitions and Theory...)

Before you continue reading, we need to tell you something important. This text is not intended to replace any course in research methods. Indeed, it is a quick and dirty introduction to some of these concepts. So don't expect any miracles

[2] See Glaser and Strauss (1967) or `http://en.wikipedia.org/wiki/Grounded_theory`

[3] For those curious about the exact formulation of the other dimensions, here they are:

Strategic dependence " ... the extent to which researchers have to persuade colleagues of the significance and importance of their problem and approach to obtain a high reputation from them."

Strategic task uncertainty " ... the stability of problem formulations, and of hierarchies of problems according to their importance and significance, varies across fields.. "

It would be a fascinating exercise to ask oneself, if the field of the Semantic Web could be disciplined. But, alas, we can focus only on one thing in this text.

here! What you will get is a quick introduction that will help you understand the points that we are trying to make later on. If you want to know more, take a class in research methods or immerse yourself in a book such as Judd et al. (1991).

2.1 Of Physics and Stamps

Ernest Rutherford,[4] a physicist and Nobel laureate in chemistry, exclaimed at some point that "all science is either physics or stamp collecting." What he meant by this statement is that there are two kinds of science: the one that studies a phenomenon and creates hypotheses about it and the other that catalogues and categorizes phenomena. It turns out that his distinction is quite useful.

Physics in this framework represents an approach of developing a theory that explains our observations. Consider the sketch in Figure 2.1. On the lower left we depicted the "real world"—in our case, a set of people interacting. We, as scientists, observe this world via observation depicted in the sketch as a crystal ball. Note that the picture in the crystal ball is somewhat hazy. Observation only allows us to see a limited part of "The World" (we will talk more about that soon).

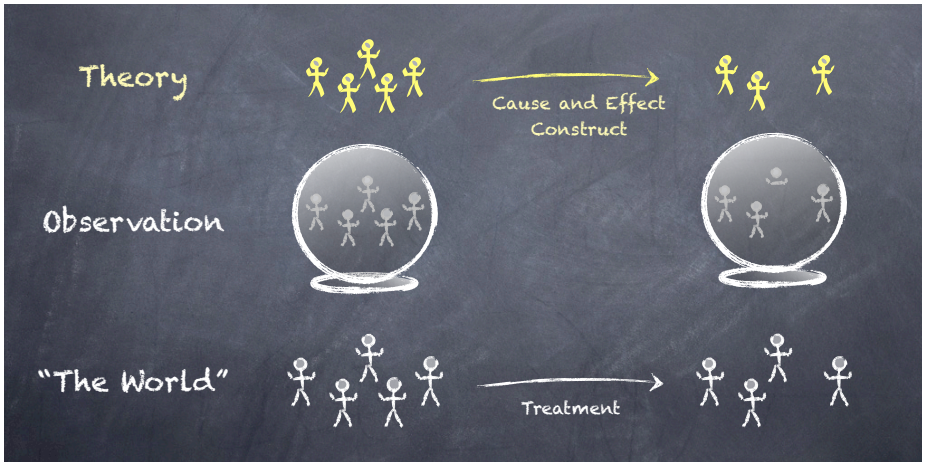

Figure 1: Theory, Hypotheses, and the "The World"

When we study the world, we often develop ideas about the phenomenon that we are looking at. In our sketch, we may think that when we have five funny figures and something specific occurs then the result should be that only three of those figures are left. Essentially, we develop a system of ideas—in scientist lingo *a theory*—that is intended to help us understand how the "The World" works, explain interactions between elements in "The World," and make predictions about what might happen when certain events occur in "The World." We often express these ideas as cause and effect constructs: If we have five figures

[4] See `http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1908/rutherford-bio.html` or `http://en.wikipedia.org/wiki/Ernest_Rutherford`

standing on a platform and the train stops, then we will be left with three figures on the platform. Note that we have now essentially moved in the top (yellow) part of the sketch from left to right. So far, however, the theory is pure mental construct and we have no idea whether it is correct. In the rest of this section, we will discuss several ways of determining whether our theory is correct and, more important, of convincing ourselves and others that our study really demonstrates what we think it does.

Now what about stamps? Stamp collecting has a great tradition in science. Even though some interpret Rutherford's statement as somewhat derogatory toward stamp collecting, the latter has its purpose. In "stamp collecting," we gather observations about "The World" and we usually organize and structure them. These structures can be extremely helpful. They can, for example, help people communicate about things. Just consider the International Classification of Diseases (ICD),[5] which is being used by many medical suppliers to classify symptoms and diseases or the Standard Industrial Classification (SIC) Codes[6] used to categorize companies and products in trade. So stamp collecting can be useful, but how do we know that we ordered our collection "correctly"? How do we demonstrate that our organization is not only "correct," but also helps solve some problem in the real world? Can we tease out research contributions in projects that, at least on the surface, appear to be stamp collecting? Many examples in this text address precisely this problem. If we are collecting stamps not only for the goal of collecting them (i.e., we are observing the world not out of mere curiosity), then there are many potential research contributions along the way.

When doing either "physics" or "stamp collecting," our overarching goal is always to enable other people to build on our work. With physics-style activities, we want others to be able to use our theories and maybe to build upon them. The idea of the Turing machine, for example, was also a theory for conceptualizing and making statements computation. Many have relied on this theory since. Similarly, the ICD codes have been useful in many healthcare applications.

In both approaches, there is a need to establish a notion of correctness of our results so that others will believe that our results are useful. To establish correctness, we typically gather evidence supporting our cause and effect constructs or the usefulness of our stamp collection in some use case.

So how can we convince others that our results are trustworthy? First, we need to demonstrate that we followed the right steps to get to our results. In other words, we didn't just dream them up in an afternoon but we followed a well established procedure. Second, we need to demonstrate that the results that we achieved following this procedure were correct. In other words, the results did no just come out of a random number generator or were produced by pure co-incidence. These are questions of reliability and validity. Let's delve into these points for just a little bit.

[5] `http://www.who.int/classifications/icd/revision/en/index.html`

[6] See `https://www.osha.gov/pls/imis/sic_manual.html` or `http://en.wikipedia.org/wiki/Standard_Industrial_Classification`

2.2 The Procedure: Anatomy of a Research Project

While this guide is not intended to be a paper about scientific method—there are many a monograph on that subject—we provide some context for the discussion that follows by briefly describing the key steps in a research project. Any research study typically proceeds along the following set of steps:

Define research questions: Research questions will determine where and what kind of research we will be doing and will define specific objectives of our study. A research question helps to frame what the study actually answers.

Formulate a hypothesis (or several): Remember that a hypothesis must be falsifiable. The goal of our research will be to test whether our hypothesis is true or false. The hypothesis essentially refines the research question into something we can prove or provide ample evidence for.

Design evaluation to test the hypothesis: Here is where we devise the test for the hypothesis. We will talk about many threats to validity in such evaluation. The key is to make sure that our evaluation indeed tests the hypothesis that we set out to validate.

Run the study: Here we finally do it! This part is often "easy," even though it will likely take the bulk of our time and that is where the majority of the effort will go.

Analyze the results: This part is where we sit back and think. We go back to the hypothesis and frame the results around that. Negative results are also useful. We think carefully about what the results actually mean both for our study and in the grand scheme of things.

Report and publish: Remember to highlight what specific contributions the reported work brings.

Now comes the dirty little truth of science: just like the waterfall model[7] is an idealized model of how software engineering projects work very few research projects o follow this neat model. We stumble, go back, iterate. So this plan is not a definitive plan to be followed blindly step-by-step. Yet, these steps are useful signposts in the journey and will help us to orient ourselves (see Suchman (1987) for a fuller treatment of the differences between plans as instructions and resources for action).

2.3 'See, I am right!' or How to Convince Others of Our Results

To make others believe that they can trust our contribution, we need to provide some sort of evidence. Evidence is mostly nothing else but an additional set of observations geared towards establishing the correctness of our system of ideas or the usefulness of our categorization. Hence, these additional observations are there to establish that our findings are *reliable* and *valid*.

[7] See Royce (1970) or `http://en.wikipedia.org/wiki/Waterfall_model`

Reliability[8] is a measurement of the consistency of the results. Judd et al. (1991, p. 51) describe the reliability of a measure as "the extent to which it is free from random error components."

In other words, are the results that we get consistent with our observations? Even in the best of all worlds the results may vary due to some element beyond our control. For example, if we have two observers looking at the same situation, they may describe it differently. The degree of difference is a measurement of inter-rater reliability.

As another example consider a method that does some probabilistic reasoning to infer the `sameAs`-relationship in the linked open data cloud (LODC). If the method does not constantly provide results of some quality but sometimes provides good results and in other runs generates awful results—even on the same data—then the method is not very reliable, as you don't know when to trust it.

Validity[9] measures the degree to which a conclusion is well founded. Validity measures to what degree the evidence that we provide can be used to confirm (or reject) our hypotheses. There are numerous types of validity defined in the literature. Here, we will briefly touch upon a small set of validity types that we will use in the rest of this primer:

Construct validity refers to the degree at which the things that we measure and relate to one another (i.e., the variables that we measure in our observations) actually represent the constructs of our theory. Does measuring the number of times our subject laughs out loud during the use of our tool correlate with her satisfaction in using it?

Internal validity refers to the degree to which we can draw conclusions from our observations. If our theory assumes that A influences B, can we really infer this causal relationship from our observation or is there some other variable C that is confounding our findings? For example, is our subject happy because she is really satisfied with our tool or because she happened to hear a funny story just before entering the room?

External validity refers to the degree to which findings generalize beyond the setting in which we tested our theory. If we tested our tool with one dataset, can we truly generalize our findings towards a second? For example, given that our ontologist friend finds our ontology-development tool easy to use, can we generalize that a novice will also find the tool quite usable?

Each of these validities has various aspects that can threaten them. It is good research practice to discuss these threats—often called *limitations*—when reporting your results.

[8] See Judd et al. (1991, p. 51) or `http://en.wikipedia.org/wiki/Reliability_(psychometrics)`

[9] See Judd et al. (1991, pp. 27-36) or `http://en.wikipedia.org/wiki/Validity_(statistics)`

2.4 Science is a risky business

There is on oft-repeated joke that any field that has "science" in its name is not really science. Indeed, computer scientists need to question perhaps more often than others whether what we are doing is science or engineering. If we know from the start that we will succeed—build a system, develop a formalism, describe a certain phenomenon—is it really science? Or is it a "mere matter of programming"?

In most cases, PhD theses in computer science will have a healthy dose of both, science and engineering. Yet, too often, we tend to go light on the science part. However, when we approach computer science (and semantic web) investigations the way a scientist would—formulating research questions and hypotheses and then testing the hypotheses—we gain critical advantages. Most important, as we will show through our section on specific problems (Section 3), once we think carefully about our hypothesis, evaluation plan comes out directly out of that. It becomes extremely clear what we need to evaluate and what the metrics should be. And that by itself helps "cure" considerable number of research projects in our field. Furthermore, we learn to think critically as scientists and to understand what it is exactly that our contribution is? What specific piece of the puzzle have we put in? What problem have we addressed?

2.5 So what? Who cares?

In addition to the above, any study needs to pass the test of relevance. Throughout our examples, we will keep coming back to these two questions: so what and who cares? Why is it important to translate a particular corpus into RDF? Why should we publish a survey of features that are used in ontologies on the Web? There are often good answers to these questions, but just as often we forget to ask them. Understanding *why* we are doing what we are doing will help us motivate our research and formulate the hypotheses in a way that will address the core problem that we are trying to solve. In more practical terms, it will also help us convince those who should fund our research on why our research is important. Many of us have chosen careers as scientist because we are simply curious to understand how the world works and because we find it thrilling when we figure out a solution to an intellectually complex problem. Yet, wouldn't it be nice to know that someone else will benefit from what we discover and that someone else cares whether we succeed or fail?

3 Practice: Building houses (hopefully on a stable foundation)

In this section, we go through a number of examples and try to analyze them with the expressions discussed above. Note that if you don't find your own study here then it is not because you shouldn't be thinking about the issues raised here but because we ran out of paper (or is it electrons?). So this is not an exhaustive but an illustrative collection of examples.

Throughout the section we will try to categorize research questions or hypotheses into one of three categories: *the good*, *the bad*, or *the ugly*, and some-

times *excellent*. These three categories are, obviously, oversimplifying the spectrum of possible issues, but they cad serve as an initial, simple compass.

3.1 "My Semantic Web System Is Better Than Your Semantic Web System" or what does "better" mean?

We start with an example set of problems that initially do not seem to cause significant issues with designing evaluations. This class of problems comprises development of methods or algorithms that improve the performance of existing systems. For instance, we might have an idea how to perform reasoning faster, or have a method for extracting structured data from text with higher accuracy than the current methods.

To start, it might be very tempting to form the following hypothesis:

The Bad: *"My reasoner is very fast and efficient"*

Yet, there is nothing we can measure here. This hypothesis is not falsifiable: without having a way to define what "fast" and "efficient" means, we cannot prove or disprove this hypothesis. We can try a falsifiable hypothesis:

The Ugly: *"My reasoner can classify SNOMED CT in 3 minutes"*

or a similar one:

The Ugly: *"I can achieve a precision of 80% on entity extraction from a corpus X"*

While these hypotheses sound reasonable, they do not actually provide any useful information. Is 3 minutes good or bad? Is 80% good or bad? These numbers are useful only in comparison to a similar method or algorithm. Just reporting the precision and recall or the time numbers by themselves is not particularly useful: it lacks the context that the reader needs in order to assess the advance that the reported work is providing. Thus, we must reformulate the hypothesis in a way that will give us a way to falsify it. For example:

The Good: *"My reasoner is faster than another reasoner X on a class of ontologies Y"*

It is rare that we will be able to make a universal statement saying that our reasoner is *always* faster than another reasoner (or that our method is *always* better than another method, whatever our metric for "better" is). However, we can usually make such a statement for a particular class of problems. Note that here comes a potential threat to external validity: How important is this class of problems? For instance, perhaps, our reasoner works very well for one particular ontology. Is this improvement sufficient? Most of the time, the answer will be no. However, if this ontology is, say, SNOMED CT—an extremely important and

large ontology in the field of medical informatics—then maybe just improving reasoning on that one ontology constitutes a significant contribution.

An even better hypothesis would be the one that not only identifies a class of problems where our reasoner works best, but also highlights what it is about our reasoner than makes the key contribution:

> **Excellent**: *"Using X will improve the efficiency of reasoning on the class of languages Y, compared to the state of the art."*

First, such a hypothesis provides a clear evaluation path: try to reason without our "secret sauce" X, and then compare the results to what we get when we mix X in.

3.1.1 Evaluation methods

Once we formulated our hypothesis in a way that makes the scope of the task clear and identifies what we are comparing to, evaluation becomes reasonably straightforward: we run our method on a representative set of examples defined by our scope and we compare the performance to that of other methods. If the other methods report results on the same set of inputs, we don't need to rerun them ourselves.

Note the word *representative* in the paragraph above: because we obviously cannot test our reasoner on every ontology that has ever been published, how do we test that it actually is indeed faster for the class of ontologies that we gave identified? Selecting the test set randomly, or running our method over all ontologies in an independent repository would usually be sufficient to prove external validity.

3.1.2 What to watch out for?

We have seen a number of problems with these types of evaluations, mostly threading internal and construct validity.

First, we need to establish that the effects that we measure are actually a result of our "secret sauce" and not the result of some other artifact of our test (internal validity). If, for example, our approach interacts especially well with some idiosyncrasies of the disk-cache or the internal JVM optimizer then the conclusion that our approach was better is not valid. It is really the disk cache or the JVM that makes the difference.

Second, it might be that our way of measuring could be misleading (construct validity). Suppose we want to measure the quality of an ontology and we report the ratio of axioms to the number of classes in the ontology. There is likely a problem with construct validity here as this ratio might indicate the complexity of an ontology but is unlikely to be a proxy for its quality.

Third, our hypothesis should almost always define the *scope* for the class of problems that we are addressing, as almost no technique is universal: the class of ontologies where our reasoning method is particularly good or our ontology-mapping produces the results that are better than the state of the art, or the domain where our knowledge extraction works well and where we tested it. Note

that the scope may limit external validity. Indeed, whenever we limit the scope of an investigation the following two questions arise:

1. We need to ensure that this class of problems is indeed important (a question of relevance). Does anybody care about these ontologies? this domain?

2. We need to establish that conclusions drawn from reasoning on the chosen representative sample generalize to other settings. Essentially, we need to establish external validity.

Fourth, we need to make sure that we really are comparing to the state of the art—and to the tools that were designed for a similar class of problems. For example, if we compare our ontology-mapping algorithm that we designed for ontologies with rich lexical information to another algorithm that was designed to rely primarily on structure, but we use an ontology with very little structure and lots of lexical information, our approach will definitely win. But the comparison will not be a valid one. When comparing to the "state of the art," we must justify to ourselves that we are indeed comparing to the methods that were designed to operate on a similar class of problems.

Fifth, it is easy to fall into the trap of over-generalizing our conclusions: if we have tested our reasoner only on a particular class of ontologies, we might hypothesize that it will be faster for other ontologies, but we cannot conclude that.

Finally, as the last of our hypotheses indicates, the most useful results are the ones that not only show that our approach is better, but also give an indication of *why* it is better.

3.2 "Look, Ma, no hands!" or where is the a scientific problem in building an application?

Much of research in the Semantic Web is applied research, and in many cases our main contribution is building a computer program that others can use. However, the following is not really a research hypothesis:

The Ugly: *"I have developed a system. It works!"*

There is no research question here. More important, there is nothing to measure and no way to fail. So, where is the science? What can we evaluate and measure?

The way we usually approach these types of problems is by stepping back and analyzing what was the type of problem that our system was supposed to solve? Were we trying to design an interface that makes it easier for domain experts to edit ontologies? Were we designing a search interface that helps finding the information more efficiently? Understanding answers to these questions, usually leads to formulating a hypothesis that we can test—and evaluate our system in the process. For instance:

Better: *"Domain experts can use our system effectively to accomplish a task X (e.g., map between large ontologies)"*

The question here is what does "effectively" mean? Before embarking on an experiment (and in this case, the experiment will likely be a user study), we must define what it is that we will measure. It might be usability, and thus standard usability metrics might be important. It might be task completion: have the experts completed all the tasks successfully? Note, however, that without anything to compare our system to, the evaluation seems rather weak. For example, is 70% completion rate for the tasks a good or a bad thing? What about 80% error rate? Without meaningful comparisons, these numbers do not provide any useful information at all.

Thus, a much better hypothesis would be

Excellent: *"Domain experts can use our system more effectively than another system Z to accomplish a task X (e.g., map between large ontologies)"*

Of course, in order to have such a hypothesis, there must be another system Z that was designed for a similar task. Indeed, if the authors of the system Z have stated that it was designed to assist in mapping ontologies with up to 1,000 classes and we compare it to our system on ontologies with 10,000 classes, the comparison won't be any better than not comparing to Z at all.

We recognize that it might often be difficult, if not impossible, to find another tool that was designed to work under similar conditions. Thus, the hypotheses would often be about comparing our system to another version of our own system. For example, we may want to evaluate whether adding a graph-based navigation to our ontology-browsing tool improves users' comprehension of ontologies (e.g., measured by the time it takes users to complete a set of tasks, and the error rate on task completion). The key idea to remember in this case is to draw the conclusions appropriately: it is tempting to claim that our system is fantastic at enabling users to complete certain types of tasks. However, what we have demonstrated is only that adding a graph-based navigation improves the users' completion of tasks. It is a relative measure, not an absolute one. We cannot claim that our system is "good" or "bad"; we can only claim that a particular feature of our system (which may be generalizable to other systems) leads to certain improvements. It is a more modest claim, but a valid—and often a very useful—one.

3.2.1 Evaluation methods

Once we have identified a hypothesis that we are testing, there is a variety of evaluation mechanisms that we can use to evaluate our tool. It can be a controlled experiment, with a usability study. We can identify a set of tasks and measure how fast and how accurately users complete the tasks. When reporting on these studies, it is important to discuss who the subject in the user study were, how we selected them, how many subjects we had, what the

protocol for the study was. Without such information, it might be impossible for readers to assess the validity of the study. For instance, a study that asked two of the authors' colleagues to evaluate a tool is different from a study that invited 10 students "off the street." Both are valid studies, but the external validity of conclusions drawn from the two studies is different.

3.2.2 What to watch out for?

Building a system, a particular a system that addresses an important problem, and a system that many users will rely upon, is in itself a technology contribution. Indeed, our field could not succeed if we did not have the key infrastructure tools that allow us to build ontologies, store our data, use ontologies in our applications, and so on. These contributions are extremely important. Yet, by itself, these are engineering contributions. For a PhD thesis, there often needs to be a research question that we are answering and that is what we were trying to identify in this section.

As a side note, if we indeed build a tool that becomes widely used, analysis of this usage and lessons learned could themselves become important research contributions as they would help us test hypotheses on how and why users deploy such tools. But we must to remember to formulate those hypotheses and to analyze the usage of the tool as a way of testing these hypotheses.

As for the other warnings for these types of systems, they are exactly the same as for the system that provide comparisons to other semantic web systems (Section 3.1): we need to make sure that we are evaluating the tool using the tasks that are in scope for the tool, using the tasks that matter, and checking the threats to internal and construct validity of our evaluation.

3.3 "We will put everything in RDF and the world will be a better place" or Who should care if we succeed?

When the Semantic Web research was in its infancy, just getting *any* data to work with was a luxury. Today of course there are billions of facts available in the linked open data cloud; reams of information can be accessed in RDF. Yet, lots of information is not yet available in structured form. And thus, when confronted with a new corpus that we would like to analyze, we may start with the following hypothesis for our information-extraction approach:

The Bad: *"I can convert this corpus of abstracts into RDF."*

By now, we already know what is wrong with this hypothesis: we cannot falsify it, as we cannot really fail. A better version might be:

The Ugly: *"My conversion process produces better linked data than conversion process X."*

While this hypothesis sounds better than some of the "bad" ones that we have identified above, it is still problematic. And the problem here is the external validity of the argument—the infamous "so what?" Why should anybody

care if we succeed in creating this corpus of RDF data (or similar)? What would be better then and what makes one conversion better than another? Hence, to answer these questions, it is useful to think in terms of *who* we think should care and what task they will be able to perform that they couldn't perform before. For instance, if we are extracting structured data from a corpus in order to improve the performance of search on that corpus, we might formulate the following hypothesis:

The Good: *"Using extracted linked data will improve search performance on the corpus compared to existing methods."*

Note that such a hypothesis also gives us a very specific success metric that we can use: does our extraction improve the search performance for users or not? Or perhaps, we have nothing to compare it to and may need to evaluate a broader claim:

The Good: *"Using extracted linked data will enable advanced querying that was not possible before"*

Proving such a hypothesis is harder, but not impossible: (1) we will need to show convincingly that we really are providing a capability that was not available before; and (2) we will need to convince ourselves, and our readers, that there are people who actually care about having such a capability. In other words, we will still need to remember the "who cares?" question. Proving (1) can indeed be tricky, and will require us to convince the readers that we really have looked at all the published and available methods and indeed none of them can do what we are proposing. Use-case scenarios can answer the second question: who and when would want to ask these types of queries.

Note that while it is important to have a core hypothesis that focuses on the external validity of the argument, it might also be useful to have an auxiliary hypothesis that highlights the technical contribution of our method:

The Good: *"My method for extracting structured data has better accuracy/ coverage/precision/recall/etc. than the state of the art."*

But such a hypothesis should not be the only one that we are investigating. Actually, in more empirically inclined (social) sciences you will often find a large list of hypothesis of increasing specificity. In our discipline this is less common but could increasingly occur as our discipline is becoming increasingly empirical.

3.3.1 Evaluation methods

Because the types of problems that we describe in this section would often have at least two hypotheses—one aiming at establishing the external validity and

another one testing the technical merits—evaluation methods might differ as well. The external-validity hypothesis will likely require evaluation of the performance of the specific task that we are trying to improve (e.g., the search performance). The technical-merits hypothesis is similar to the types of evaluation that we discussed in Section 3.1.

3.3.2 What to watch out for

Note that even though we focused on external validity when discussing this class of problems, we did so mainly because these are the problems that tend to suffer the most from the "so what?" question. However, we should be thinking about any of the other types of problems that we discuss throughout this paper in a similar vein: why are we trying to improve the technical characteristics of a particular process? Who will benefit?

A last word of caution: external validity of almost all empirical evaluations will almost always have limits. Almost no finding is universally true. Hence, it is important to think how general our claim should be and maybe even to state clearly—if possible—where our claim fails. Newtonian physics, for example, has a limited external validity. But in most life situation we find that it is still extremely useful.

3.4 "Using our hammer for every nail" or does the use of semantic web technology actually improve anything?

As Semantic Web researchers, we are excited about the toolkit that our technology provides: ontologies, queries, linked data, and typed graphs. Naturally, we want to apply this technology to solve problems that seemed hard or intractable before. Sometimes, we try to apply this technology to solve problems that have been solved just fine by other means. We must carefully approach such tasks to make sure that we are not just trying to use semantic web technology for the sake of using semantic web technology. Consider, for instance, a project that uses linked data to improve movie recommendations.

The Ugly: *"We can use linked data to make movie recommendations."*

While it sounds reasonable on the surface, formulated as such, this problem is a perfect example of "using our hammer for every nail": why should using semantic web technologies to solve a problem where machine learning seems to work just fine be an advancement by itself? However, if—and this is a critical *if*—we can show that we can do movie recommendations *better* than the current state-of-the-art, and it happens that linked data is our "secret sauce", then we have a winner. In other words, people outside the semantic web community do not care which technology we use to solve the problem, they just want the problem solved. Thus, using a particular technology in the solution is not itself a contribution; rather, providing a solution—whatever the technology—is the solution.

Excellent: *"We will significantly improve the quality of recommendations by using linked data technology."*

3.4.1 Evaluation method

The key in such evaluation is to benchmark the performance of the system that uses our new technology with a system that does not. In the movie-recommendation example, compare the quality of recommendations using traditional techniques and using linked data. If the latter provides better recommendations, then indeed we have a significant contribution. If the results are the same in terms of time, cost, precision, then we really have not achieved that much—even if we have demonstrated that our technology is as good as some other technology.

3.4.2 What to look out for

The real test in such work is convincing our peers from other communities (not semwebbers) that we have indeed made a contributions. Researchers outside of the semantic web community will not really care whether or not we used a particular technology. But can we publish our paper in a conference outside of our field that deals with the problem area. For instance, can we publish our improvements to the movie recommendations in the ACM RecSys conference? Will researchers in that field care? If we improved the performance, they sure will!

3.5 Stamp Collection: "Looking where the light is" or what is the value in surveying what is out there?

As mentioned above, Lord Rutherford once said: "All science is either physics or stamp collecting." With lots of linked data now available on the Web, it is extremely tempting to start doing "stamp collecting." For instance, we can survey what types of ontologies exist? What types of constructs are prevalent in linked data? How many links across different datasets can we find? The danger with such surveys—in particular as all of our constructs are man-made—is that we will be "looking where the light is," as in the joke about a person searching for his lost car keys not where he remembers dropping them, but under a lightpost, because "that's where the light is."

Consider the following project: "We will create a set of features that ontologies have and will describe ontologies from our catalog according to these features." Of course this statement is not a hypothesis, but rather a research plan. Even so, before we embark on such a plan, we should ask ourselves: "Why would this information be useful? What will drive our selection of features?" While it may be counter-intuitive at first to start with such questions, in reality, having the answers to these questions will make such a survey not only much more focused, but also will enable us to scope it properly. We no longer will need to look at *every* possible feature, for example. Thus, the following hypothesis might drive our survey of ontologies that exist on the web:

The Good: *"Only a small number of OWL constructs are used in the publicly available ontologies."*

Why would anybody care? For instance, we might be designing an ontology editor and want to make sure that we make it extremely easy to edit the most common constructs and don't crowd out the user interface with the ones that are never used. With such a setting, we know exactly what questions to ask. Our hypothesis may or may not turn out to be true; indeed, our findings may be used in other settings (e.g., to drive new standards or to focus research on optimizing reasoning), but having the scope helps us in guiding our exploration.

3.5.1 Evaluation methods

When performing a survey to test a hypothesis about some features of a corpus, we need to make sure that our corpus is representative. For instance, when we try to determine which OWL constructs ontologies use more frequently, taking a corpus of ten ontologies developed by a group that focuses on high-performance reasoning might not give us a representative sample. Studying ontologies in a particular domain (e.g. biomedical ontologies in BioPortal) will enable us to draw conclusions of what constructs scientists in that domain use. However, we need to be careful in over-generalizing the conclusions (e.g., do earth scientists develop ontologies in a way that is similar to the way scientists in biomedical informatics do it?). Yet, trying to analyze all ontologies that are available on the semantic web might not be practical. Thus, convincing ourselves that a sample that we study is representative of the types of ontologies that are critical to our hypothesis is the most tricky question here.

3.5.2 Stamp Collection for Theory Development

Note that as we mentioned before, stamp collection can also be used for theory development using methodologies such as case studies and/or grounded theory. In field such approaches are, sadly, rather seldom used for theory development. The discussion of these approaches is, therefore, way beyond the scope of this primer.[10]

4 Closing

The above examples provide a collection of possible issues that arise when searching for an appropriate research questions and synthesising appropriate hypotheses. They are intended as exemplars and we simplified many of the practicalities in order to highlight the issues that we wanted to convey. What all the examples show is that it is imperative for any scientific study to *present a clear question* (or set of questions), to *argue clearly the question's relevance (remembering to address the "so what?" and "who cares?" questions)*, and *to provide evidence to support its conclusion.*

[10] As mentioned you might want to look at Glaser and Strauss (1967) or `http://en.wikipedia.org/wiki/Grounded_theory` for grounded theory and Eisenhardt (1989) for Case Study research

Given the nature of the scientific enterprise, some of the above arguments are going to have limitations—either inherently connected to the methods or due to the impossibility to investigate complete data. Consequently, every study should *clearly state any threats to validity* to highlight what kind of problems curb the generality of the conclusions.

At the beginning, we stated that

> *the goal of this primer is to provide context for scientists in the Semantic Web and Linked Data domain about the purpose of research questions and their associated hypotheses, the tension between rigor and relevance, possible evaluation approaches, and pitfalls in terms of reliability and validity.*

We hope that we were able to provide this context and, hence, contribute to improve the clarity and methodological quality of Semantic Web and Linked Data research studies—a goal we think is worthy of pursuing.

References

Banville, C. and Landry, M. (1989). Can the field of mis be disciplined? *Communications of the ACM*, 32(1):48–60.

Burell, G. and Morgan, G. (1979). *Sociological Paradigms and Organizational Analysis - Elements of the Sociology of the Corporate Life*. Heinemann.

Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review*, 14(4):pp. 532–550.

Glaser, B. G. and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter, New York, NY.

Judd, C., Smith, E., and Kidder, L. (1991). *Research Methods in Social Relations*. Harcourt Brace Jovanovich College Publ.

MacGarry, D. (1955). *The Metalogicon of John of Salisbury: A Twelfth-century Defense of the Verbal and Logical Arts of the Trivium*. University of California Press.

Royce, W. W. (1970). Managing the development of large software systems: concepts and techniques. *Proc. IEEE WESTCON, Los Angeles*, pages 1–9. Reprinted in *Proceedings* of the Ninth International Conference on Software Engineering, March 1987, pp. 328–338.

Suchman, L. A. (1987). *Plans and situated actions - the problem of human-machine communication.* Learning in doing: social,cognitive,and computational perspectives. Cambridge University Press.