On the Status of Experimental Research on the Semantic Web

Heiner Stuckenschmidt, Michael Schuhmacher,
Johannes Knopp, Christian Meilicke, Ansgar Scherp

Data- and Web Science Research Group,
University of Mannheim, Germany
firstname@informatik.uni-mannheim.de

Abstract. Experimentation is an important way to validate results of Semantic Web and Computer Science research in general. In this paper, we investigate the development and the current status of experimental work on the Semantic Web. Based on a corpus of 500 papers collected from the International Semantic Web Conferences (ISWC) over the past decade, we analyse the importance and the quality of experimental research conducted and compare it to general Computer Science. We observe that the amount and quality of experiments are steadily increasing over time. Unlike hypothesised, we cannot confirm a statistically significant correlation between a paper's citations and the amount of experimental work reported. Our analysis, however, shows that papers comparing themselves to other systems are more often cited than other papers.

1 Introduction

Popper characterizes the nature of science in terms of the falsifiability of claims [1]. Following this statement, careful validation of proposed methods and theories are commonly accepted as the core of reputable research. Over time different scientific disciplines have developed a variety of methodologies for evaluating results ranging from mathematical proofs to use cases and experiments. Semantic Web research and computer science as a whole is a discipline that has a strong formulative research approach [2]: it creates new formalisms, algorithms and systems claimed to be superior to previous proposals. If we follow the idea of reputable science, these claims have to be substantiated by a suitable method of validation, typically formal proofs, controlled experiments or use cases and examples. We claim that Semantic Web research is even more forced to validate scientific claims as it is a rather new area of research that often has to face prejudices of more established disciplines inside computer science and on the other hand faces the dilemma formulated by Wright: *"In [...] dynamic areas, researchers often face the choice: corroborating prior work to strengthen the foundations of the research area or 'pushing the envelope' while relying on prior work that may be less reliable"* [3]. We conclude that experimentation is an important way to validate results of Semantic Web research, especially as it has been argued that it challenges established results in more traditional disciplines [4] and is therefore less accessible to a strictly formal treatment.

Having accepted that experimental research is important on the Semantic Web, we want to investigate the status of experimental research on the Semantic Web with respect to the quantity and the quality of experimental work. In particular, we want to compare the area of Semantic Web with other areas of computer science with respect to the importance given to experimental research. Further, we want to have a closer look at the way experiments are conducted to determine the usefulness of the reported experimental work for validating the claims and a reference for other researchers working on related problems. This question that is linked with the quality of the experimental work is much harder to capture than the pure amount of experimental work. Finally, we are interested in the question, whether doing experiments pays off in terms of research reputation and try to answer this question by analyzing citation statistics for papers with different amounts of experimental work.

Following the design of previous studies on experimental research in computer science (in particular [5] and [6]), we analyzed all papers from the International Semantic Web Conference starting in 2002 with respect to the type of work and amount of experimentation.

Going beyond previous studies, we also took a closer look at experiments with respect to the data used, the parameters measured, and the comparisons conducted.

This paper is structured as follows. In Section 2, we first give an overview of previous studies concerned with experimental work in computer science, summarizing the findings of these studies as a reference we can compare to. Subsequently, we define our research questions and hypotheses concerning the role of experimental work in Semantic Web research, provide more details about the data used, and the steps of the methodology that led us to our results (Section 3). This is followed by a detailed presentation and discussion of the results in Section 4. In Section 5, we conclude with discussing limitations of our study and the reliability of the results.

2 Empirical Studies of Experimental Research in Computer Science

Computer science is mostly regarded as a constructive science concerned with the creation of artefacts that cannot be entirely validated using formal methods [7]. Glass and others compare research approaches in different disciplines related to computer science [2]. Based on a review of major ACM and IEEE journals they conclude that almost 80% of computer science papers propose some new design or method that would actually require evaluation. While the amount of such papers is lower in certain subareas of computer science, like software engineering (55%), still a significant amount of work in computer science is formulative and requires some evaluation.

So far, the most detailed and systematic investigation of experimental research as a means for evaluating formulative research has been carried out by Tichy and others in 1995 [5]. Based on a sample of publications from major computer science journals the authors categorize papers into formal theory, design and modeling, as well as empirical work and others. The papers in the category design and modeling, which correspond to the formulative work in [2] are further analyzed with respect to the importance that is given to experimental work. For this purpose, Tichy and others further classified papers

in this category according to the space devoted to the description of experimental work. It turned out that in computer science literature experimental work is much less prominent than in engineering or natural sciences that were used as a reference. The study was repeated by Wainer and others focussing on a sample of papers published in 2005 by the ACM [6]. The authors used roughly the same setup and compared their results with the findings of Tichy and others, concluding that experimental work had gained importance, but is still behind the level found in other disciplines. We will discuss the results of these studies in more detail and compare them to our findings later.

Different additional studies have been performed in subdisciplines of computer science. Most notable in Software Engineering [8, 9] and Computer-supported cooperative work [10, 11]. Zelkowitz [8] identifies different forms of validation that can be found in the area of Software engineering and investigates the use of these different forms of validation in the Software Engineering literature in a quantitative study. In 2009 Zelkowitz repeated the study and reports the development over time [9]. He concludes that the amount of papers with a real evaluation has risen from only about 30% in 2000 to over 60% in 2009 moving towards the level that Tichy and others have presented for Computer Science as a whole. Pinelle and others [10] look at evaluation in papers on computer-supported collaborative work. The findings are in line with the above mentioned studies with a fraction of about 70% of the papers containing some kind of evaluation. On the other hand, only 30% of the papers used controlled experiments in a laboratory setting. Wainer and Barsottini performed a follow-up study on papers submitted to the ACM CSCW conference over a period of six years [11]. They found out that while overall the amount of experimental work has not increased, there was a significant increase in papers that performed an evaluation in terms of field experiments. Some smaller studies have been carried out in narrower fields. Prechelt performed a quantitative study of experimental approaches in the field of neural networks [12]. Like Machine Learning as a whole this area heavily depends on experimentation as a form of evaluation. Therefore the study is less concerned with the amount of experimentation, but with the specific setting of the experiments. As a central point of study, Prechelt looks at the nature and the number of datasets used in the experiments, discovering that most papers only use one single dataset as a basis for controlled experiments.

In summary, previous studies identified design as the dominant research methodology in Computer Science while empirical work is less important. Further, the studies showed that the importance of systematic experiments as a means of validating design research has gained importance over the last decades.

3 Research Questions and Method

The goal of this paper is to investigate the status of experimental research in the area of Semantic Web. In particular, we aim at investigating whether the importance of experimental work is comparable to the one in computer science in general as it has been identified in the previous studies discussed above. This question has two aspects: we need to identify work that can be characterized as Design and Modeling and therefore asks for an experimental evaluation. Having identified this work, we want to know whether experimental work has the same importance as in computer science in general.

As Semantic Web research is a rather young discipline, we are specifically interested in the development of the role of experimental research over time. Beyond these purely descriptive aspects, we also want to analyze the factors influencing the importance of experimental work. With respect to this, we look at the relation between amount and quality of experimental work and impact of a paper in terms of citations.

H1. Like in computer science in general, Design and Modeling work is the dominant form of research on the Semantic Web.
H2. The importance of experimental work on the Semantic Web is comparable with computer science in general.
H3. The importance of experimental work on the Semantic Web is increasing over time.
H4. The quality of experimental work on the Semantic Web is increasing over time.
H5. Strong experimental work increases the impact of a paper.

We conducted an empirical study for testing these hypotheses. For this purpose, we took the papers published at ISWC since 2002 and manually classified them according to the scheme proposed by [5]. In addition, we had a closer look at papers containing descriptions of experimental work with respect to the data used and the claims made. In the following, we describe the study design and the data used in detail and discuss the results of the study as well as the implications for the hypotheses stated above.

3.1 Data

As the goal of the study is to make valid assertions about the area of Semantic Web as a whole, the dataset used in the study has to be representative for the work conducted in the area. Making a good selection is complicated by the fact that the area of Semantic Web is not as well defined as more established research areas. Today, many conferences and journals contain work relevant for the Semantic Web. On the other hand, many researchers active in Semantic Web research also publish in other scientific disciplines such as artificial intelligence or database systems. Instead of trying to identify relevant work in different scientific outlets, we decided to focus on the International Semantic Web Conference as the major community event assuming that the work published there is representative for the whole area. Therefore, we included all full papers from the main research track of the ISWC conferences since 2002 instead of taking samples from different outlets. There are other potential sources of publications in particular, the ESWC and ASWC conference series as well as the Journal of Web Semantics and the Semantic Web Journal. Concerning ESWC and ASWC, we can safely assume that the ISWC conference series is the leading outlet and thus a more representative source of data. We explicitly decided against including journals, because conferences better reflect developments in young and dynamic fields such as the Semantic Web. The Journal of Web Semantics, however, might be included in future studies to compare the different kinds of publication outlets.

The dataset used in this study thus includes 500 papers from the following conference editions:

– 11. ISWC 2012: Boston, MA, USA (41 papers[1])
– 10. ISWC 2011: Bonn, Germany (50 papers[2])
– 9. ISWC 2010: Shanghai, China (51 papers[3])
– 8. ISWC 2009: Chantilly, VA, USA (43 papers, Research Track[4])
– 7. ISWC 2008: Karlsruhe, Germany (43 papers, Research Track[5])
– 6. ISWC / 2. ASWC 2007: Busan, Korea (50 papers, Research Track[6])
– 5. ISWC 2006: Athens, GA, USA (52 papers, Research Track[7])
– 4. ISWC 2005: Galway, Ireland (53 papers, Research Track[8])
– 3. ISWC 2004: Hiroshima, Japan (48 papers, Research Track[9])
– 2. ISWC 2003: Sanibel Island, Florida, USA (42 papers, Research Track[10])
– 1. ISWC 2002: Chia, Sardinia, Italy (27 papers, Research Track[11])

In order to measure the impact of papers in the dataset, we use citation statistics from Google Scholar (http://scholar.google.de/) and Microsoft Academic Search (http://academic.research.microsoft.com/). We use two different sources of citation statistics because it is well known that citation counts can differ significantly between different sources depending on the coverage of sources and the counting policy. Google Scholar has a very liberal counting policy that typically leads to a very high number of citations. In particular, as pointed out in [13], Google Scholar also covers grey literature citing a publication. Microsoft Academic Search is more conservative and counts fewer citations on average.

3.2 Annotation Scheme

In order to be able to compare our findings to previous studies on the role of experimental work in computer science as a whole, we used the classification scheme proposed in [5] with the modifications described in [11], i.e. the merge of the two categories 'Empirical Work' and 'Hypothesis Testing'. This allows us to relate our results to the finding reported in both papers. In particular, we classified papers according to the following four major categories.[12]

1) Formal Theory Papers whose main contributions are formal propositions, e.g. lemmata and theorems and their proofs.
2) Design and Modeling Papers whose main contributions are systems, techniques (e.g. algorithms) or models whose claimed properties cannot formally be proven.

[1] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2012-1.html
[2] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2011-1.html
[3] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2010-1.html
[4] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2009.html
[5] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2008.html
[6] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2007.html
[7] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2006.html
[8] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2005.html
[9] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2004.html
[10] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2003.html
[11] http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2002.html
[12] Descriptions of categories are taken from [5].

3) Empirical Work / Hypothesis Testing Papers that collect, analyze and interpret observations about known designs, systems, models or hypotheses.
4) Other Papers that do not fit the other categories (e.g. surveys).

Further, we annotated all papers in Category 2 with additional information about the experiments conducted. Following Tichy et al., we use the number of pages devoted to the description of the experiment and its outcome as an indicator for importance of the experimental work and therefore annotate every paper with the number of pages describing experiments and the fraction of the overall paper they constitute.

Further, we annotate all papers of Category 2 with the following information about the nature of the experiments.

Standard used for Comparison Does the paper report about *different settings* or the system or method? Are results compared against *existing baselines*? Are results compared against the results of *other systems*? The latter includes both indirect comparisons against results reported in other papers and direct comparisons obtained by executing the other system as part of the experiments.
Datasets used Has *one dataset* been used or have *several datasets* been used within the experiments? Has the dataset been *self-created* by the authors for the purpose of conducting the experiments or is it *externally provided*?

We use this information as an indication of the quality of the experimental design, assuming that an ideal experimental design will compare a proposed system against other leading systems or at least sensible baselines using several datasets with different characteristics. One can argue about whether externally provided datasets should be preferred over self-created ones, in many cases externally provided datasets are publicly accessible benchmarks that support the comparison with other systems, which we consider desirable.

3.3 Study Design

Annotation Process The classification of papers into the four categories was performed manually by a group of five annotators, three of which were senior and two junior level researchers. One of the senior researchers acted as a judge, while the other four were annotators. We started with the 2012 papers which were annotated by all four annotators to get a feeling for the level of agreement and discuss difficult cases to reach a common understanding of the category definitions and typical problems. In a second round, the remaining papers were annotated by two groups consisting of one senior and one junior annotator. One group performed the annotation of papers from even years, the other of papers from odd years. All papers where the annotators disagreed on the correct category were forwarded to the judge who made a final decision on the category.

In the same way, the number of pages devoted to experimental work was annotated at a granularity of half pages. First all annotators determined the number of pages for the 2012 conference. Subsequently, all remaining papers were annotated in two groups. Papers with a disagreement were forwarded to the judge. In case of consensus on the

category and a disagreement of just half a page, the judgment of the senior annotator was used. The detailed analysis of the experimental setting was carried out by two senior researchers where one annotated the papers from odd and another annotated the papers from even years.

In order to check how hard it is to decide on the classification of each paper, the inter-annotator agreement for both annotator pairs in the second round was computed using Cohen's Kappa [14]. The result for each pair is a number between zero and one, where zero means that the agreement between both annotators cannot be distinguished from chance, while one means perfect agreement. The annotators of the odd years reached a kappa of $\kappa = 0.63$ while the group annotating the even years scored $\kappa = 0.47$. There is no universally accepted value range defined for Cohen's Kappa, but there are interpretations of Cohen's Kappa in the literature that say these results can be considered to be *moderate* (even years) and *substantial* (odd years) agreement [15]. It is safe to say that the kappa values easily exceed an agreement by chance which means that the classification task was well defined. Most disagreement result from confusions between Category 1 and 2, i.e. 32 out of 77 disagreements. That means it is often unclear whether a paper should be considered as a theoretical paper or a modeling paper without experiments. All disagreements were finally resolved by the decision of the judge.

Test of Hypotheses Based on the classification of papers according to the four main categories, we compare the distribution of papers from ISWC to the distributions reported in previous studies for general computer science and other disciplines (H1). Further, we look at papers from Category 2 (Design and Modeling) in more detail. In particular, we analyze how the papers distribute across the subcategories defined by the fraction of the pages devoted to the description of experimental work (0%, (0% - 10%], (20% - 50%], $>$ 50%) and compare the distribution with previous studies (H2). We then look at the development of experimental work over time by plotting the distribution of papers across all categories over the past eleven years. We also look at the average number of pages devoted to experimental work in the different years and compute the correlation between year of publication and number of pages (H3). In a similar way we look at the experimental setting in more detail. For papers from Category 2 we analyze the standards used for comparisons and the datasets used as input to the experiments. We interpret these features and their characteristics as indicators for experimental quality in terms of significance and validity and analyze whether the experimental quality has increased over the past eleven years (H4). Finally, we used statistical models to test for correlation between the pages devoted to experimental work and the features that are indicators for experimental quality on the one hand and the impact of the paper on the other hand. We control the influence of other variables, such as the year of publication, to avoid spurious correlations, that do not appropriately reflect the dependencies between experimental work and its influence on a papers impact (H5).

4 Results

In the following, we discuss our findings regarding the different hypotheses in more detail. In particular, we present descriptive statistics of the ISWC paper collection and results of investigating possible correlations with research impact.

4.1 H1. Like in computer science in general Design and Modeling work is the dominant form of research on the Semantic Web

We investigated the first hypothesis by comparing the distribution of papers across the four main categories 'Formal Theory', 'Design and Modeling', 'Empirical Work' and 'Other', with the results of the previous studies conducted by Tichy et al. and Wainer et al. respectively. The results are shown in Table 1.

	ISWC 2002-2012	[6]	[5]
1) Formal Theory	11.2% (56)	4.1% (6)	18.7% (48)
2) Design / Modeling	**80.8% (404)**	**70.1% (103)**	**64.1% (164)**
3) Empirical Work	5.4% (27)	22.4% (33)	10.2% (26)
4) Other	2.3% (13)	3.4% (5)	7.0% (18)
	100% (500)	100% (147)	100% (256)

Table 1. Comparison of the relative share of papers in each of the four research method categories. While the figures from [5] refer to papers published in 1995, and [6] used papers from 2005, our study covers 11 consecutive years from 2002 – 2012. This also explains the comparably high number of papers (500) included in our study. Noteworthy is that all three studies found a similar pattern, revealing Category 2) Design and Modeling as being the domination method of research.

Looking at the results, we see that like in previous studies, most of the work, namely 80.8% falls into the category 'Design and Modeling' while 11.2% of the work is of theoretical nature and only 5.4% is empirical work in the sense of our classification, leaving 2.3% other papers. This confirms our hypothesis that Design and Modeling is the dominant form of research on the Semantic Web. Comparing this to the results of the previous study, we can see that the dominance of design and modeling work is even more visible than in the previous studies, where 64.1% and 70.1% of the work was classified as Design and Modeling. Partially this difference can be explained by the general trend to more practical work in computer science and the different periods the studies were carried out: While Tichy and others only considered papers published in 1993 and Wainer and others analyzed papers published in 2005, our study includes papers published between 2002 and 2012. This means that our results should at least be comparable with the results from Wainer and others that fall into the period covered by our study.

Another noticeable observation is the lack of a significant amount of empirical work on the Semantic Web. With only 5.4% of all papers, the fraction of empirical work is only half as large as in the 1995 study and only a quarter of the amount found by the 2005 study. In fact, besides some papers that investigated the amount and nature of linked data and ontologies found on the web, there is no empirical work concerned with Semantic Web technologies. This could be explained by the fact that the Semantic Web

is still a very young area of research where the focus is still on creating new technologies rather than on analyzing the impact of the new technologies on the Web.

4.2 H2. The importance of experimental work on the Semantic Web is comparable with computer science in general

We investigate the claim that experimental work has the same importance in the Semantic Web area as in Computer Science in General based on the criterion of importance proposed by Tichy and others in their original study. In particular, Tichy and others propose to use the fraction of the paper devoted to the description of experimental work. We follow this suggestion and compare the distribution of papers in the relevant category of Design and Modeling Papers across the different subcategories proposed by Tichy and others.

Fig. 1. Comparison between three studies reporting on the relative share of pages of Category 2 papers dedicated to experiments. While the figures from [5] refer to papers published in 1995, and [6] used papers from 2005, our study covers 11 consecutive years from 2002 – 2012.

Figure 1 compares the distribution of papers across classes between our study and the two previous studies looking at Computer Science in general. The first observation is that in the study of Tichy conducted in 1995 the fraction of Design and Modeling papers that contained no description of experimental work at all is significantly larger (45% vs. 31% and 33%) while the fraction of papers with more than 20% of the pages devoted to the description of experiments is significantly smaller (approx. 26% vs. 34% and 36%) than in the other two studies. This visible difference, again can be explained by the general increase of importance of experimental work since the early Nineties. Comparing our results to the Study of Wainer et al., we can see that the difference between the distribution is very small. Except for the category of papers with more than 50% of the pages devoted to experimental work, the differences between the classes are always within two percentage points. This seems to suggest that the importance of experimental work on the Semantic Web is comparable with General Computer Science literature published by the ACM.

Being aware of the general tendency that experiments become more important over time, we take another look at the papers from the study of Wainer and others and the papers from ISWC 2005 to be able to directly compare papers published in the same year. The results are summarized in Table 2.

	ISWC 2005	ACM Sample 2005 [6]
0%	40%	33%
(0% - 10%]	11.1%	9.7%
(10% - 20%]	15.6%	21.4%
(20% - 50%]	28.9%	30.1%
$>$ 50%	4.4%	5.8%

Table 2. Comparison of the relative share of pages of Category 2 papers dedicated to experiments. For both studies, ours and [6], we report only papers from 2005 here. We observe a generally lower amount of pages for experiments when comparing ISWC to general Computer Science.

Here, we observe a slightly different picture. When only looking at papers from 2005, we see that the fraction of papers without any experimentation is higher (40%) than the figure reported by Wainer and others (33%) and also higher than the average fraction across all ISWC conferences. On the other hand, the fraction of papers with more than 10% of the pages describing experiments is lower (49%) compared to the study by Wainer (57%) and also much lower than the average across all ISWC conference (also 57%). We conclude that at least in 2005, experimental work did not yet have the same level of importance in Semantic Web research than in general Computer Science, while averaged across all ISWC conferences, the importance is comparable to general Computer Science in 2005.

4.3 H3. The importance of experimental work on the Semantic Web is increasing over time

The inconclusive result of comparing the number of pages as an indicator for the importance of experimental work across the different studies asks for a deeper analysis of the development of the indicator over time. We explain the observation that, while in 2005 experimentation was not as prominent in ISWC papers than in general computer science, the results measured across all ISWC conferences was comparable with the results of the 2005 study by Wainer et al. by hypothesizing that the importance of experimental work was rather low in the early years of the ISWC conference. This is not uncommon for new fields of research, as first, the principled ideas have to be laid out and basic ideas have to be tested in prototypical form. Only later, when the field is more established and the problems are better understood, systematic experiments become the standard way of validation. As the first ISWC conference took place in 2002, the field was still in a rather early stage in 2005. According to our hypotheses H3, we expect the importance to have significantly increased since then, which would explain the result over all conferences.

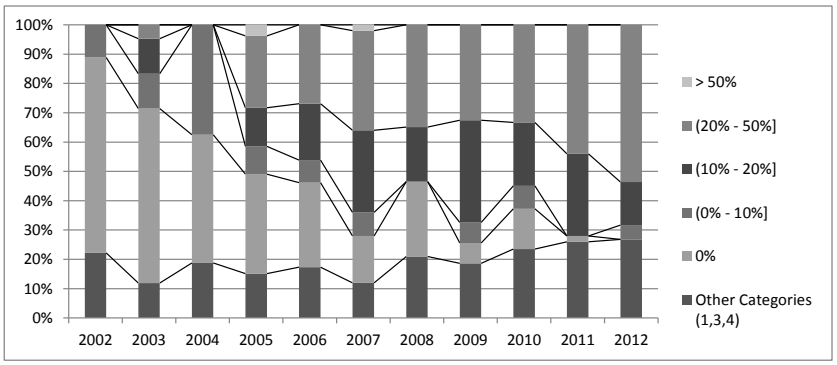

Fig. 2. Barchart showing the relative share of papers of Category 2 (Design and Modeling), grouped by the relative number of pages dedicated to experiments per year: 0%, $(0\% - 10\%]$, $(10\% - 20\%]$, $(20\% - 50\%]$, $> 50\%$. All other categories $(1, 3, 4)$ are summarized in one class. Most noteworthy is the decrease over time for papers without any experiments (0% pages), while the group of $(20\% - 50\%]$ is growing.

We test this hypotheses by looking at the development of the different categories over the years 2002 to 2012. In particular, we look at the development of the different subcategories under Design and Modeling to get an impression, whether the importance of experiments is increasing in this category. The results are summarized in Figure 2. The first observation to be made is, that the overall amount of papers in Design and Modeling stays roughly the same - around 80% - with a slight decrease to about 75% in the last two years. Inside this category, however, we can see a radical shift in the classification from 2002 to 2012. The shift can best be observed when looking at the subcategory of papers with 0% of pages describing experimentation and the subcategory of papers with 20% to 50% of the pages devoted to experimentation. While the former category contained about 70% of the papers in 2002 it completely disappeared by 2012, showing that today Design and Modeling papers without experimentations are not any more considered to be adequate. On the other hand, the amount of papers with 20-50% experimentation show a constant increase and represents more than 50% of the papers in 2012. In 2005 there were still more papers without experiments (about 35%) than papers with 20-50% (about 25%), which explains the results reported above.

The increase in importance can also be observed well when directly looking at the number of pages instead of the categories. Figure 3 shows a standard box-plot for the relative number of experiment pages for Category 2 (Design and Modeling) papers. We identify a trend of growing importance of experiments over time. With the exception of 2010, the median is constantly rising up to 25% in 2012. Measuring this trend in figures, the Spearman Correlation Coefficient is statistically significant ($r_S(402) = .49, p < .000$).

4.4 H4. The quality of experimental work on the Semantic Web is increasing over time

With respect to H4, we decided to focus on four binary variables as indicators for experimental quality. Our choice is based on the following assumptions.

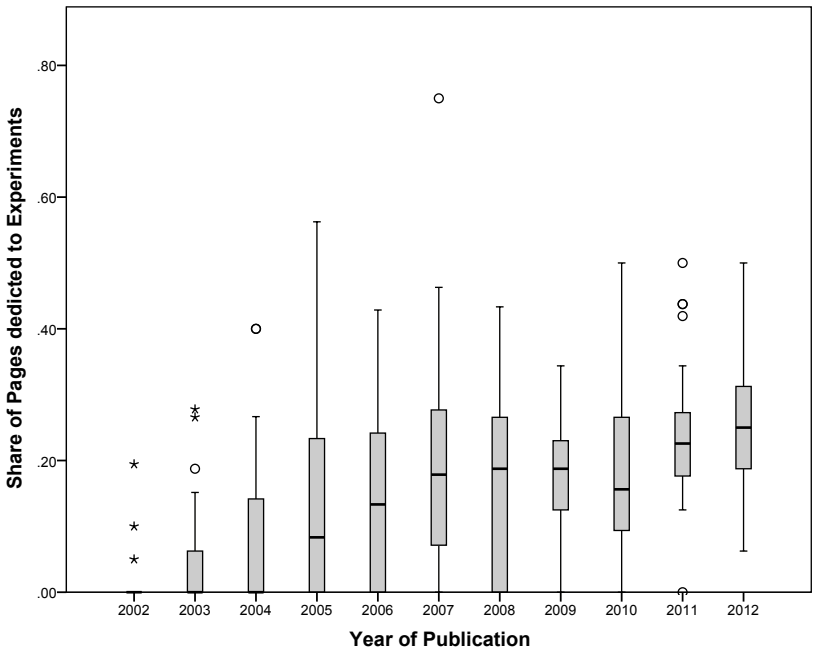

Fig. 3. Box-plot showing the relative number of pages of Category 2 (Design and Modeling) papers by year of publication. The median starting at 0% in 2002 increases constantly (with the exception of 2010) over time, reaching its top of 24% in 2012. The second/third quartile, denoted by the box, varies, but is since 2009 clearly above zero. Outliers are displayed as circles/stars.

- Using several datasets is better than using only one dataset (SEVERAL).
- Using an already existing dataset is better than using a dataset that has been created for the purpose of conducting the experiments (OTHER).[13]
- Comparing the proposed approach against a baseline or comparing different settings against each other is better than no such comparison (BASEDIFF).
- Comparing the proposed approach against other algorithms/systems is better than no comparison (SYS).

The variables SEVERAL and OTHER can be interpreted as indicators for the universal validity of the reported results. The variables BASEDIFF and SYS indicate whether the authors informed the reader on the performance (e.g. runtimes), quality (e.g. precision), or usability compared to alternative approaches. Without such a comparison, it is hardly possible to draw any conclusions related to the improvements made.

The results of our analysis are shown in Figure 4, where we depicted the countings for all four variables with respect to Category 2 papers. Figure 4 reveals a clear trend.

[13] During our annotation phase, we observed problems in deciding whether a simple setting should or should not be treated as a baseline. For that reason we did not distinguish between comparisons against a baseline and comparisons of different settings and counted each such comparison in the same variable.

The quality of experimental work is increasing over time with respect to each variable. In 2003 only a minor share of all papers had a positive characteristic in one of the four variables, while in 2012 more than 50% of all papers had a positive characteristic in three of four variables. However, only 33% of all papers in 2012 compared their results against other systems (SYS). While this is an improvement compared to the previous years, there are still many papers that do not compare their results against other systems. We computed also the correlation between the year of publication and the four quality measures using Spearman's rank correlation coefficient. We find that all variables show positive and statistically significant correlations with the year of publication (r_S between .36 and .46, $p \leq .000$).

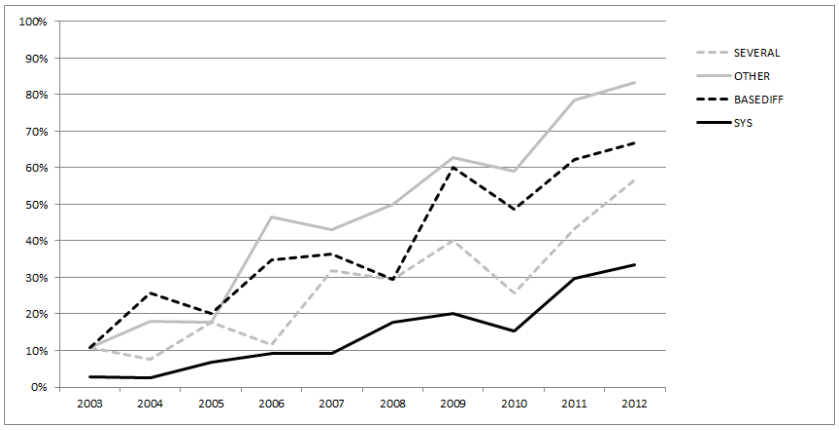

Fig. 4. Development of relative share of Category 2 (Design and Modeling) papers complying to different evaluation quality indicators over time. While all indicators start at a low level of $\leq 11\%$ in 2003 and rise with the years, we found the usage of externally provided datasets (OTHER) to increase the most. Nevertheless, even in 2012 only about on third of all papers compare themselves to other existing systems (SYS).

Our observations can be explained by two factors. One factor might be an increasing awareness of the importance attributed to experimental work. Another factor might be the general development of the community. What has been a novel area of research 10 years ago, might have become an established research area associated with well-defined problems, commonly accepted formats, well-known datasets and accepted benchmarks. Obviously, both factors go hand in hand, resulting in the positive trend that we reported in our evaluation.

4.5 H5. Strong experimental work increases the impact of a paper

For analyzing the potential relation between the amount of experimental work of a paper and its impact, we employ a generalized linear model (GLM). As described above, we take the relative number of pages describing experiments (RELPAGES) as a general proxy for the importance of the experimental part within a paper. Following [5], we

thus make the assumption that the better the experiments in a paper are, the more the paper reports about the experiments.

For measuring the impact of a paper, we divide the citation count by the age of a paper in years (RELCITATIONPA). We use only Google Scholar data, as the Microsoft Academics citation figures are strongly correlated with the Google Scholar data, the statistically significant Pearson's correlation is 0.969, and would thus result in similar findings.

	Correlation	Sig. (2-tailed)	N
Scholar Citation Count per Year (RELCITATIONPA)	1.000	.000	500
Experimental Pages Count (RELPAGES)	-0.175	.000	404
Year of publication (YEAR)	-0.458	.000	500

Table 3. Spearman's Correlation Coefficient for the citation count per year, the relative number of pages for experiments, and the year of publication. All correlations are statistically significant and negative with respect to the year of publication. The experimental pages count was surveyed only for the 404 Category 2 papers.

Our first analysis reviews the simple pairwise variable correlation. For all correlations here $df = 402$. We find that a statistically significant Spearman's correlation ($r_S = -0.175$) between RELCITATIONPA and RELPAGES exists. This would indicate that a decrease in the number of citations goes hand in hand with an increase of the amount of pages spend on experiments. But as shown in Table 3, we also observe a strong correlation of -0.458 between RELCITATIONPA and the year of publication (YEAR). This is consistent with the temporal development reported above in Section 4.3. Thus, from this data it cannot be concluded if RELCITATIONPA effects the impact of a paper, or if both developments just coincided with the general development over time. For a more fine-grained analysis of the effects, we thus use a GLM for regression analysis.

The GLM takes a log-link function to explain the outcome of our dependent variable RELCITATIONPA as the result of a $Tweedie(p = 1.5)$-Distribution[14] taking into account RELPAGES and AGE, as well as the binary quality measures BASEDIFF, SYS, OTHER, and SEVERAL. We choose a Tweedie distribution function as it suits citation count data, where several papers have zero citations, well. In addition, we find that the model shows better goodness of fit values, measured by the Akaike Information Criterion (AIC), for our data, compared to a linear regression as well as to a loglinear GLM with a Poisson distribution (Poisson-Regression). The Omnibus likelihood-ratio Chi-Square test of our model ($df = 18$, cf. Table 4) versus the intercept-only model confirms a significant difference ($p \leq .000$). We use the 404 Category 2 (Design and Modeling) papers and group the values for RELPAGES into the five classes described above, see e.g. Figure 2, and denoted them in the following by the variable RELPAGESCLASS.

Under this model, RELPAGESCLASS has a negative parameter and thus indicates a negative effect of the number of pages on the likelihood to be cited. But as the model

[14] For $1 < p < 2$, this is a compound Poisson-Gamma distribution with a point mass at zero.

effects test given in Table 4 reports a non-significance ($p = .239$), we cannot conclude that RELPAGESCLASS has an effect on RELCITATIONPA. On the other hand, SYS and SEVERAL both have a statistical significance ($p = .050$ and $p = .053$). SEVERAL has a positive parameter (0.250), thus indicating a positive effect of reporting experimental results for different datasets and not only for one dataset. We make a similar observation for papers comparing their own system to other systems, as SYS has also a positive parameter (0.301). Because both variables, SYS and SEVERAL, are strong indicators for

	Wald Chi-Square	Deg. of Freedom	Significance
(Intercept)	281.760	1	.000
RELPAGESCLASS	5.510	4	.239
AGE	164.634	10	.000
BASEDIFF	3.139	1	.076
SYS	3.835	1	.050
OTHER	0.205	1	.651
SEVERAL	3.750	1	.053

Table 4. Test of Model Effects with Wald-Chi-Square for the GLM with RELCITATIONPA as depended variable. Variable RELPAGESCLASS is not statistically significant for the model, but the two quality indicators SYS and SEVERAL have significance test values of $p \leq .05$ and $.053$

profound experimental work, we may take this finding as a support for our hypothesis and conclude, that comparing oneself to others increases the likelihood to get cited. Nevertheless, if this correlations reveals a causality remains somehow doubtful, as the positive correlation may also originate from the fact that papers with an active and large research community have more opportunities to cite other systems, while papers on isolated topics simply do not have peer papers to related to. Regarding all other variables, our model can currently not give a statistically reliable explanation whether they have any effect or not.

5 Conclusion

After more than 10 years of Semantic Web conferences, we believe it has been time to conduct a study like this. It serves a basis for a backward analysis of what has happened so far alike as actuates fruitful future discussions that will help steering the kind of research conducted in our community. Our main aim was to learn how the field of Semantic Web research is doing compared to general computer science and to show that the field is on its way to become an established scientific discipline with high standards concerning experimental evaluation of work.

Our results confirm that Semantic Web, as other emerging fields, has undergone a significant change with respect to the importance and quality of experimental work. We found that the amount of experimental work done is comparable to Computer Science in general and that the quality of experiments in terms of the use of publicly available datasets and comparison to other systems and benchmarks has continuously increased

over the last ten years. In particular, we see that today it is virtually impossible to get design and modeling work accepted in the main track of ISWC without having experimental results. Further, our results show that papers that relate their contribution to existing datasets or other systems are more often cited than others.

As next steps, we will add more Semantic Web conferences like ESWC and journals such as Journal of Web Semantics and Semantic Web Journal. In addition, we will conduct more detailed analyses such as investigating the influence of (co-)authorship with respect to citations. We also plan to conduct analyses of the citations based on the single ISWC conferences. To this end, we are looking at the papers published in a specific year of a conference only and investigate whether the papers of a specific category are statistically more cited than papers in other categories. Finally, we would like to automatically obtain the topic of the articles by extracting topic models from the abstracts and analysing the citation distribution over these topics.

References

1. Popper, K.: The logic of scientific discovery. Routledge (2002)
2. Glass, R., Ramesh, V., Vessey, I.: An analysis of research in computing disciplines. Communications of the ACM **47** (2004) 89–94
3. Wright, D.: Motivation, design, and ubiquity: A discussion of research ethics and computer science. arXiv preprint arXiv:0706.0484 (2007)
4. van Harmelen, F.: Where Does It Break? or: Why the Semantic Web Is Not Just "Research as Usual". The Semantic Web: Research and Applications (2006) 1–1
5. Tichy, W., Lukowicz, P., Prechelt, L., Heinz, E.: Experimental evaluation in computer science: A quantitative study. Journal of Systems and Software **28** (1995) 9–18
6. Wainer, J., Novoa Barsottini, C., Lacerda, D., Magalhães de Marco, L.: Empirical evaluation in computer science research published by acm. Information and Software Technology **51** (2009) 1081–1085
7. Ramesh, V., Glass, R., Vessey, I.: Research in computer science: an empirical study. Journal of systems and software **70** (2004) 165–176
8. Zelkowitz, M., Wallace, D.: Experimental models for validating technology. Computer **31** (1998) 23–31
9. Zelkowitz, M.: An update to experimental models for validating computer technology. Journal of Systems and Software **82** (2009) 373–376
10. Pinelle, D., Gutwin, C.: A review of groupware evaluations. In: Enabling Technologies: Infrastructure for Collaborative Enterprises, 2000.(WET ICE 2000). Proeedings. IEEE 9th International Workshops on, IEEE (2000) 86–91
11. Wainer, J., Barsottini, C.: Empirical research in CSCW a review of the ACM/CSCW conferences from 1998 to 2004. Journal of the Brazilian Computer Society **13** (2007) 27–36
12. Prechelt, L.: A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. Neural Networks **9** (1996) 457–462
13. Bauer, K., Bakkalbasi, N.: An examination of citation counts in a new scholarly communication environment. D-Lib Magazine (2005)
14. Cohen, J., et al.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20** (1960) 37–46
15. Viera, A.J., Garrett, J.M.: Understanding interobserver agreement: The kappa statistic. Fam Med **37** (2005) 360–363