Genuine Semantic Publishing

Tobias KUHN [a,1] and Michel DUMONTIER [b,2]
[a] *Department of Computer Science, Vrije Universiteit Amsterdam, Netherlands*
[b] *Maastricht University, Maastricht, Netherlands*

Abstract. Several approaches and systems have been presented in the context of scholarly communication for what has been called *semantic publishing*. Closer inspection, however, reveals that these approaches are mostly not about publishing semantic representations, as the name seems to suggest. Rather, they take the processes and outcomes of the current narrative-based publishing system for granted and only work with already published papers. This includes approaches involving semantic annotations, semantic interlinking, semantic integration, and semantic discovery, but with the semantics coming into play only after the publication of the original article. While these are interesting and important approaches, they fall short of providing a vision to transcend the current publishing paradigm. We argue here for taking the term *semantic publishing* literally and work towards a vision of *genuine* semantic publishing, where computational tools and algorithms can help us with dealing with the wealth of human knowledge by letting researchers capture their research results with formal semantics from the start. We argue that genuine semantic publications should come with formal semantics as an integral and primary component at the time of publication, that these representations should have essential coverage in the sense that they cover the main claims of the work, that they should be authentic in the sense that they originate from the authors, and that they should be fine-grained and light-weight for optimized re-usability and minimized publication overhead. This paper is in fact not just advocating our concept, but is itself a genuine semantic publication, thereby demonstrating and illustrating our points.

Keywords. semantic publishing, scholarly communication, Linked Data

— Text passages that are new or updated are marked in brown. —

1. Introduction

Many scholars have pointed out that the classical way of publishing scientific articles is ill-suited to deal with the rapid growth of both, volume and complexity, of scientific contributions [1, 2]. To overcome these problems, next generation scientific publishing [3] has to respond to the increasing importance of datasets and

[1] Corresponding Author: Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, Netherlands; E-mail: t.kuhn@vu.nl.
[2] E-mail: michel.dumontier@maastrichtuniversity.nl.

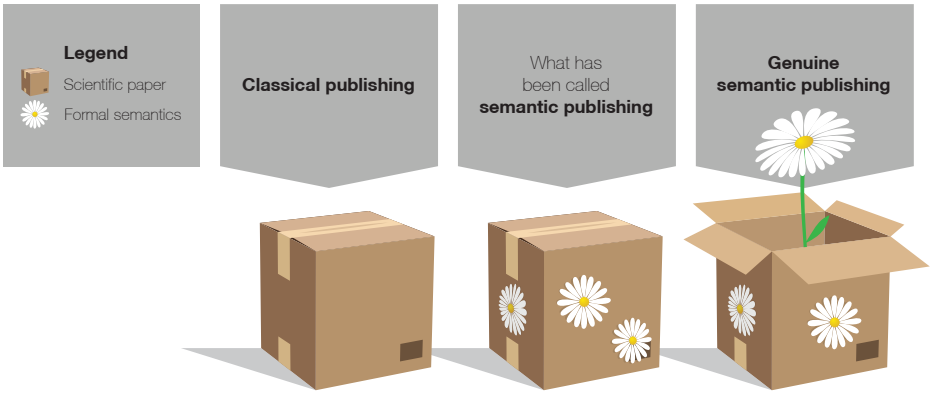

Figure 1. The concept of *genuine semantic publishing* compared to what has been called *semantic publishing*, explained by an analogy where scientific papers are represented by boxes and formal semantics by flowers.

software, and needs to provide methods to automatically organize reported scientific findings. Perhaps the most important shortcoming of the current publication system is that scientific papers do not come with formal semantics that could be processed, aggregated, and interpreted in an automated fashion.

Semantic publishing [4–6] is a general approach to tackle this problem of scholarly communication by using the concepts and tools of the Semantic Web and related fields. This idea was basically born together with the idea of the Semantic Web itself. In 2001, Tim Berners-Lee and James Hendler sketched how they expect researchers in the future to produce machine-readable descriptions of their experiments and findings, in the form of mark-up of their research papers or as independent representations made public on the web [7]. Unfortunately, subsequent work has deviated from this general proposal.

The topic of semantic publishing has received considerable attention during the last few years, most prominently in events that carry the term in their names, specifically the workshop series on Semantic Publishing (SePublica)[3] and the Semantic Publishing Challenges at ESWC conferences [8–10]. However, as we argue below, the presented approaches mostly interpret the term *semantic publishing* in a non-intuitive way. Instead of changing the publishing process, they mostly take existing classical publications as their starting point and simply apply semantic technologies on them, mostly without touching the publishing process or the object that is being published. This leads them to propose solutions that are quite conventional, and fall short of providing a vision for the long-term future. We argue here that we should aim for semantic publishing in the literal sense, which we call *genuine semantic publishing* to distinguish it from the existing term.

2. Semantic Publishing

Semantic publishing has been defined as "anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers" [5], and this definition accurately reflects how the term "semantic publishing" has been used in the recent literature. We argue here, however, that this definition is in one way too restrictive and in another way too inclusive if we want to be faithful to the literal and intuitive meaning of the term and if we aim to follow the spirit of the Semantic Web vision.

In our view, the definition above is too restrictive in the sense that semantic publications according to this definition are required to accompany a "journal article" or a "paper." An entity that contains only a semantic representation of a scientific result, without an accompanying narrative article, could not be considered a semantic publication. On the other hand the definition is too inclusive, in our view, in the sense that it covers very shallow approaches that add little — if anything — to established approaches of publishing. For example, letting authors choose keywords from standardized vocabularies for their paper — as many journals do — in fact "enables its linking to semantically related articles," and therefore by the definition above makes it a semantic publication. As another example, a semantic annotation performed by a third party on an article "enhances the meaning of a published journal article" and therefore would have to be called a semantic publication, even if the semantic annotation is not even made public.

In general, the existing literature seems to interpret the term "semantic publishing" as "adding semantics to something that is published" instead of the more intuitive readings of "publishing something that is semantic" or "publishing in a semantic manner." (We are using the word *semantic* in its narrow technical sense of *carrying a formal logic-based representation of the content's meaning.*)

We argue here for a more intuitive definition of *semantic publishing* that is broader in the sense that no narrative article needs to be present, and that is at the same time narrower in the sense that the semantic representation has to be a first-class object created and published by the authors. We propose the definition that *genuine semantic publishing* occurs when somebody publishes a work that includes authentic and fine-grained representations of its content in a semantic notation, where these semantic representations have essential coverage and are a primary component of the published entity. We explain in more detail below what we mean by *authentic, fine-grained, essential coverage*, and *primary component.*

Figure 1 illustrates our point with a simple analogy. Classical papers are shown on the left hand side as boxes that are closed and hard to access for automated techniques. Existing approaches to what has been called semantic publishing merely adorn this box with formal semantics — represented by flowers in the picture — but leaving it closed. This adornment is very useful, to be sure, but it does not reach to the main content of the box. By only looking at the formal semantics, one can possibly find out the topic of the paper but not what the

[3] `http://sepublica.mywikipaper.org`

paper is actually claiming: the main message is missed. Moreover, the adornment is often attached at a later point, after the box has been shipped so to say, and is therefore not a proper part of it. Speaking in terms of this metaphorical image, we argue that we should open the box and let semantics bloom right from the inside. We should represent the paper's main message with formal semantics. As we see on the right hand side of the figure, this metaphorically turns the box into a flowerpot. Now semantics is the main content, and the scientific paper has become a container for semantics instead of a closed box with a secondary usage as a pin board for semantic annotations.

It seems to be a common unquestioned assumption that the semantic representation of knowledge has to start from a textual representation, and therefore writing a statement down in natural language always needs to be the first step. For example, we can read in a paper on semantic publishing that "learning how the brain creates and decodes meaning from text is essential if we are to provide better tools for scientific inquiry" and that we need to "train computers to help us read scientific text" [6]. While these are certainly interesting and important problems, it is not obvious why they are essential if we take the approach of semantic publishing literally, i.e. if we ensure that the published artifacts come with semantic representations from the start. There is no law of nature that research findings can only be formalized after they have been expressed in a narrative text. It can very well be the other way round, such as a researcher writing a narrative text verbalizing existing formal statements she has come up with. More likely, these two will go hand in hand in an iterative process, much like manuscripts and their content typically being shaped through several rounds of revisions. It has in fact been argued — convincingly in our opinion — that this iterative process of scientific writing contributes in an important way to scientific understanding and discovery [11], and therefore it seems beneficial for the semantic representations to participate in these iterations from the start, and not to come into play only at the point where the text is already finalized. However, many articles in the area of semantic publishing seem to make this implicit text-first assumption, as exemplified by papers presented at semantic publishing workshops claiming that "annotations on all levels pave the way for shared knowledge understanding" [12] and that "semantic publishing [...] can be defined as the activity of enhancing a document" [13], among many others (e.g. [14–16]). The entire approach of *semantic annotation* is based on this text-first paradigm, which is a perfectly valid approach but is surely not the only possibility.

We get a similar picture if we look at the Semantic Publishing Challenge held at the Semantic Web conferences ESWC from 2014 until 2017 [8–10]. There were three tasks defined for each of these three challenges, but none of them actually deals with publishing. Instead they are about automatically extracting and interlinking semantic data from existing publications. Only the "in-use" task of the first challenge was general enough to not exclude publishing ("showcase the potential of Semantic Web technology for enhancing and assessing the quality of scientific production"), but it did not specifically mention the publishing process either. Unsurprisingly then, the approaches presented at these challenges deal with extraction from and annotation of articles that are already written and published

		Research Objects	executable papers	scholarly HTML	Structured Digital Abstracts	Micropublications	Nanopublications
types of formal semantic representations:	– metadata	✓	✓	✓	✓	✓	✓
	– domain (input/output) data	✓	✓				✓
	– program code	✓	✓				
	– arguments / discourse			(✓)		✓	(✓)
	– high-level claims	(✓)		(✓)	✓		✓
publishing unit:	– project level	✓					
	– article level		✓	✓	✓	✓	
	– statement level						✓
necessary components:	– formal semantic data	✓	✓	(✓)	✓	✓	✓
	– narrative text		✓	✓	✓	✓	

Table 1. Properties of existing approaches on the publication of scientific artifacts containing semantic representations

(with the only exception being a paper introducing a publishing platform for Research Objects [17]).

To be clear, we do not mean to deny the value or importance of this body of existing work. To the contrary, these approaches are very important to deal with the wealth of existing publications and those that will become available in the near future. Besides this important work, however, we also need clear and bold visions for the future on how we can improve the form in which such publications are created in the first place.

3. Related Work

Despite the prevalence of approaches that deal exclusively with semantic annotations and semantic interlinking of already published articles, there are a number of existing approaches where the artifacts to be published include from the start semantic representations that originate from the researchers themselves. They cover different aspects that we consider important for genuine semantic publishing. These approaches include Research Objects, executable papers, scholarly HTML, Structured Digital Abstracts, Micropublications, and Nanopublications.

Research Objects[4] [1, 18] are a proposal to extend the notion of a scientific publication beyond a narrative article to include additional resources like data, metadata, code, presentation slides, log files, and workflow definitions. In order to improve reusability and reproducibility of scientific findings, these different

[4] `http://www.researchobject.org`

resources are interlinked and packaged together in zip files called Research Objects. These objects are therefore at a higher level of composition than classical articles in the sense that they can contain articles and other resources, and we can say that their publishing unit is on the level of a *project* rather than the level of an article. The language RDF with its formal semantics is used as the underlying notation. The interpretation of a Research Object's content, however, would typically be performed by a human user, at least according to the main motivating scenario [1]. The authors claim that Research Objects can also include formally represented hypotheses and high-level findings, but no specifics are given about this. Later work by other researchers applied the proposed concepts and used embedded nanopublications (see below) to represent such higher-level claims and hypotheses [19]. A similar approach could probably be followed to express arguments and scientific discourse, but no concrete proposal exists for that.

From a different angle, a number of approaches have been proposed for what has been called *executable papers* (e.g. [20–22]), in particular in response to the Executable Paper Grand Challenge[5] organized by Elsevier in 2011. These approaches follow the same basic idea as the IPython Notebook[6], which has recently become very popular, to interweave narrative text with program code that can be executed in an interactive fashion by the reader to dynamically generate results in the form of tables and diagrams, and thereby to reproduce, verify, and explore the results. Many of the proposed solutions do not explicitly mention semantic interlinking or standardized vocabularies, and they are therefore a borderline case of semantic publishing. However, one can argue that source code is a kind of semantic representation, even though with procedural rather than declarative semantics. Naturally, these approaches focus on algorithms, their source code, and the corresponding input and output data, and do not target the formal representation of high-level claims and hypotheses.

Yet another angle, with a focus on scientific writing instead of source code and datasets, is taken by approaches on scholarly HTML[7] that propose HTML-based replacements for today's role of PDF as universal file format for scientific articles. Many of these approaches do not explicitly focus on formal semantics, but — in contrast to PDF — semantic markup comes very naturally with HTML via technologies such as RDFa (but with special reader software, similar features can be achieved with PDF [23]). On a higher level, the Linked Research initiative [24] proposes general principles and directions of using HTML for scholarly communication to enable, among other things, the publication of semantic representations. Specific implementations of scholarly HTML include formats and approaches such as RASH [25] and Dokieli [26] provide specific guidelines and tools of how that can be achieved. The semantic markup of such articles is mostly on the meta level, for example about the authors, the structure of the article, and the references to other articles. Furthermore, as the term *markup* indicates, these semantic representations are mostly not independent statements but tightly linked to the narrative text. Some of these scholarly HTML approaches explicitly point to the possibility of using RDFa to formally represent not just metadata

[5] `http://www.executablepapers.com`
[6] `http://ipython.org/notebook.html`
[7] `https://www.w3.org/community/scholarlyhtml/`

but also high-level claims, hypotheses, and arguments, but they do not focus on concrete solutions in this respect. In general, these approaches and technologies have a slightly different focus but — as we will demonstrate below — can play an important role in genuine semantic publishing.

Structured Digital Abstracts [27, 28], having been first proposed ten years ago, are probably the oldest approach of this kind. Their basic idea is to require for articles to come with a machine-interpretable summary of the main claims, besides the classical abstract for human readers. They proposed to let authors themselves capture the claims of their own scientific contributions, such as a newly discovered protein-protein interaction, in a notation with formal semantics. Even though these abstracts are attached to narrative articles, the formally represented findings can be processed and interpreted independently from the narrative text. We proposed a similar approach in previous work with abstracts in controlled natural language [29].

Micropublications [30] are a further approach, which puts the emphasis on the structure and interrelation of scientific arguments and their underlying pieces of evidence. The authors stress that the network of arguments is an essential part of science, of which claims and hypotheses are necessary but not sufficient ingredients. They argue that formal representations of scientific claims are often not practically feasible, whereas the structure among them can be captured more easily and is moreover more important and more valuable to help scientists with computer-aided knowledge management.

Nanopublications are an approach to use the RDF language to represent "the smallest unit of publication" [31] or "core scientific statements with associated context" [32]. This statement-level approach is therefore at a more granular level than most other approaches whose unit of publication is at the article level. Nanopublications consist of three parts, each represented in RDF: an assertion containing the actual content in the form of an atomic small piece of knowledge (e.g. a scientific claim, or a data entry), a provenance part containing metadata about the origin and context of the assertion (e.g. how it was measured), and a publication information part with metadata about the nanopublication as a whole (e.g. when and by whom it was created). Even though the details of their integration into the scientific publishing workflow have remained largely unspecified, nanopublications have received considerable attention during the last few years, with several large dataset having been published in this format [33–35].

Table 1 shows a comparison and summary of the different types of approaches introduced above. We see that they cover different types of semantic representations. We also observe from the table that none of the existing approaches covers all aspects, but a complete coverage could be achieved by a combination of them (and there do not seem to be any fundamental incompatibilities). We will argue below that high-level claims and arguments are the most important parts, because they are — and have always been — the core contributions of a scientific work.

Executable papers, Structured Digital Abstracts, and Micropublications stick to the article as their unit of publication, whereas Research Objects operate at a higher level (at what we call the project level) and nanopublications at a lower one (statement level). We argue here that it is important to cover all these levels,

which means that it has to be possible to publish something as small as an individual statement (i.e. an individual claim, hypothesis, data entry, or link between entities). Once this lowest level is covered, it is then relatively straightforward to compose larger objects out of the low-level ones, like the approach mentioned above of using nanopublications within Research Objects [19].

All these approaches, except some flavors of scholarly HTML, mandate that formal semantic data are part of the published entity, but three out of the five approaches also require or assume that a narrative text accompanies the data. Not coincidentally, these are also the approaches that work on the article level, and therefore stick to the classical unit of publication. We will argue below that narrative text necessarily remains an important part of scientific discourse and communication, but it also has to be possible to publish data that is self-explanatory due to its formal semantics without the need for a narrative. While none of the presented approaches has yet managed to find widespread acceptance, small practical and less intrusive steps have already been successfully implemented, such as the use of unambiguous references to biomedical resources in the form of Research Resource Identifiers (RRIDs) [36].

To summarize, there are a number of existing approaches that cover important aspects of what we think deserves the term *semantic publishing* under its intuitive interpretation, but they are normally overshadowed by the more conventional approaches of semantic annotation and semantic interlinking introduced above. These conventional approaches, in fact, advocate the term *semantic publishing* more visibly, whereas approaches like Research Objects often do not use the term at all. This situation can lead to a general impression that the conventional approaches are the best we can do to apply the ideas of the Semantic Web to scientific publishing. It is our goal to convince the readers of this paper otherwise. While none of the existing approaches covers all aspects that we consider important for genuine semantic publishing, they provide — as we will show below — important and highly valuable building blocks towards that endeavor.

To conclude our discussion of related work, we would like to point out that a large number of general technologies have been developed in the last years that can serve as the basis for approaches on genuine semantic publishing. They come in the form of data formats, ontologies, and software tools. The Semantic Publishing and Referencing Ontologies (SPAR)[8] [37, 38], for example, are a highly valuable suite of ontologies to build semantic models and applications in the domain of scientific publishing. Other examples include the Annotation Ontology [39] and W3C's subsequent Web Annotation recommendations[9], which define how to connect (scientific) text to the respective formal representations, the PROV Ontology [40] to model provenance of digital objects, and the Semanticscience Integrated Ontology (SIO) [41] and the Linked Science Core Vocabulary[10], which cover the different kinds of entities in scientific workflows such as researchers, data, publications, and methods, as well as the relations between them. There is also existing work on modeling arguments and discourse with Semantic Web technology [42], including approaches on formalizing uncertainty in

[8] http://sempublishing.sourceforge.net
[9] https://www.w3.org/annotation/
[10] http://linkedscience.org/lsc/ns/

scientific arguments [43], as well as formal models of scientific experiments [44] and evidence [45].

4. Genuine Semantic Publishing

As we have shown above, most approaches that go under the label *semantic publishing* are not actually about publishing, and the approaches that *do* target the publication of semantic representations cover different aspects thereof that only partly overlap. We therefore think that there is a need for clear criteria of *genuine* semantic publishing. As we started to argue above, important aspects of semantic publishing concern the authoritativeness and essential coverage of semantic representations, as well as their status in relation to narrative articles and their granularity level.

The first aspect we would like to discuss here is what we call *essential coverage* of semantic representations with respect to the entity to be published. A representation has essential coverage with respect to a work if it covers (at least) the essence of the work. The essence of a work is its main message, which for scientific articles normally consists of the main claims, findings, and arguments. A semantic representation may not cover all aspects discussed and described throughout a scientific work, but for it to have essential coverage it has to cover the main points: If you had to summarize a paper in one sentence, the content of this sentence has to be present in the semantic representation too. In other words, it is not sufficient to focus on what is *easy* to represent; we have to focus on what is *important*. One can also see it as a kind of democratization process of making automated agents first-class citizens: English-speaking agents (e.g. human researchers) get the main content of the work in their English-based representation (i.e. the narrative text); so RDF-speaking agents (e.g. Linked Data aware software) should also get the main content of the work in their RDF-based representation. As we will see below, this perspective aligns very well with the well-established Web technique of content negotiation.

Another important aspect is the authoritativeness of the source of the semantic representations, which determines the authenticity of these representations. Semantic representations can only be considered authentic if they originate from an agent that is authoritative in the given situation. In the case of the publication of a scientific result, the only authoritative source are the researchers (who are called *authors* in this context). In other words, semantic representations of scientific results are only authentic if they are provided by the researchers themselves, and this relation can be made explicit with a precise provenance representation. It has long been known in the area of knowledge engineering that the process of formalizing expert knowledge is not merely a process of "transferring" or "converting" knowledge from existing representations inside the heads of experts to formal representations of a form that can be stored in a knowledge base. Rather it has to be seen as a creative *modeling process* [46, 47] where formal structures are generated that existed only in an incomplete, implicit, and unconscious form in the experts' heads. Explaining a result in a narrative is simpler than formally modeling it, in the sense that natural language allows the writer to remain vague

and even ambiguous. Accurately modeling knowledge only from such a narrative text with its inherent vagueness and ambiguity is therefore in general not just difficult but strictly impossible without a further connection to the authoritative source. Genuine semantic publishing requires the authors of scientific results to perform the modeling task themselves, because they are — by definition — the only authoritative source. We claim that — contrary to many existing approaches — we should *not* try to relieve the authors of this burden (though we should of course try to help them). Otherwise, the semantic representations cannot be considered authoritative and should therefore not be considered part of the publication's content (unless the person who produces the formal representations becomes a co-author), and we would end up with a situation where semantic representations are disconnected from what is published, which is against the essence of the semantic publishing idea. As Tim Berners-Lee and James Hendler made clear when the Semantic Web was just about to come into existence as a research field, it "involves asking people to make some extra effort" [7]. In the case of scientific publishing, it involves asking authors to make the extra effort of providing formal semantic representations of their findings.

To make the semantic representations first-class citizens, they furthermore need to have an existence in their own right. We cannot call something a genuine semantic publication if the semantic representations are attached to an already published article at a later point, or if they can only be interpreted in the context of the narrative article. Neither should these semantic representations be considered just another type of supplementary material, listed somewhere at the very end of the article as an noncommittal extra file. In fact, one of the defining properties and one of the big advantages of declarative and monotonic semantic notations like RDF is that statements are in an important sense self-explanatory and independent. Such a formal statement can be taken out of its context and stripped from natural language explanations attached to it, and it still means exactly the same thing, as far as the formal semantics are concerned.

In turn, this self-explanatory and independent nature allows for publications of semantic representation to be very light-weight and fine-grained. More so than narrative texts, formal representations with declarative and monotonic semantics can be easily broken down into independent pieces, and it therefore seems foolish to propose semantic publishing solutions that would not allow people to exploit this nice property. Such light-weight semantic publications might consists of just a single statement (like "X is related to Y"), and for larger chunks of semantic representations we should make it possible to refer to such individual statements in a fine-grained way (e.g. refer explicitly to the statement "A causes B" within a larger set of statements).

Based on these arguments, we define that *genuine semantic publishing* needs to comply with the following criteria:

1. A scientific work needs to come with formal representations that are semantic, in the sense that they are not just machine processable but *machine interpretable*, and that are linked so they add to the existing formal body of knowledge.

2. These semantic representations might be underspecified but need to have *essential coverage* in the sense that they cover (at least) the core of the main claims of the given work.
3. They need to be *authentic* in the sense that the respective authoritative persons create or approve the semantic representations. Domain data can only come from the researchers, and metadata has to come from the people responsible for the form of the published work, i.e. the researchers and/or the editors.
4. The semantic representations need to be a *primary* component of the published work, made available together with everything else at the time of publication. They must furthermore have an independent existence in their own right and not merely be appended or attached to the main entity as noncommittal extra data.
5. The semantic representations and their containers need to be *fine-grained and light-weight*. Even though such semantic representations might often be published in larger collections, the publication of minimal additions and corrections needs to be possible without a large overhead.

Most, maybe all, existing approaches on what has been called semantic publishing comply with the first criterion, but only a few of them propose or support representations that comply with the others. We illustrate below that these criteria are in fact not difficult to achieve with existing technologies.

Here, we should briefly discuss an aspect that we deliberately left out of our criteria. Several of the related approaches introduced above (in particular executable papers and scholarly HTML) have a specific focus on how semantic representations can enhance the user experience in the form of interactivity. While we think such interactivity can be highly valuable, we argue for a clear distinction between publication and use, where interactivity belongs to the latter. It is precisely the benefit of formal semantic representations that they facilitate all kinds of subsequent (interactive) use but are agnostic about the precise circumstances and technology. Genuine semantic publications may therefore come with specific interactive features, but it is not appropriate to make that a strict requirement.

Furthermore, it is probably helpful to discuss here the different types of claims that a scientific work can make. A large part of the body of scientific work deals with what has been called "normal" or "puzzle-solving" science [48]. In this type of science, known kinds of relations and properties are discovered for objects of known kinds, such as a statement that a given mutation of a given gene can be the cause for a given disease. Such types of statements are relatively straightforward to formalize, for example by connecting a concept identifier for the given gene mutation with the concept identifier for the given disease by the use of a relation denoting the causal relationship, possibly augmented with the needed qualifications and contexts (such as the species to which it applies). In a next step, such a statement as a whole can be formally linked to its authors and to the study from which they derived it (such as a clinical trial and its properties). If the authors represent these formula in a specific language like RDF (assuming existing established vocabularies cover all needed terms), save them in a file, and share and archive them on the Web, then we have perfect case of a genuine semantic publication. The authors may want to add a narrative to it, but they do not need

to, as the semantic representation speaks for itself. More disruptive and more abstract kinds of scientific contributions involve the criticism of existing concepts or arguments, and the advocation of new ones. In the most extreme case, this can consist of proposing a paradigm shift that can lead to a scientific revolution [48]. By their nature, these types of contributions are harder to formalize, but it is always possible to at least make the action of criticizing or advocating explicit and to position the objects in the space of related concepts, arguments, or paradigms.

Finally, before we move on to demonstrate in detail how advocating a new concept can be achieved with a genuine semantic publication, let us reflect for a moment on the impact of this proposal. The machine-interpretability of publications' main claims entails that software could automatically connect, aggregate, and reason about the body of published scientific work. For example, we could automatically answer complex questions or produce interactive science maps, not only at the meta-level of papers, authors and their relations, but also on the domain level of tangible and abstract concepts and objects of study. This will allow scientists (and others) to acquire a more accurate and more complete picture of the current state of science with much less effort, which in turn can accelerate scientific work and improve its quality. The support for small fine-grained publications can further speed up scientific discovery, as researchers no longer need to wait for a larger body of work to assemble, but can publish smaller findings as they come in. Results from such software solutions will never be error-free, but due to our authenticity requirement we can find out which authors are to blame for mistakes we find in the semantic representations, instead of some anonymous software component or human annotator. This in turn can put strong incentives on authors to provide good formal representations for their works. It is hard to foresee how all the involved technical — let alone social and institutional — aspects would unfold, but it is not hard to imagine that such technology could have a profound positive impact on the communication of science.

5. Genuine Semantic Publishing in Action

It turns out that all the technologies needed for applying genuine semantic publishing are already available and most of them are very mature and reliable. There are no technical obstacles preventing us from releasing our results from today on as genuine semantic publications (though more work is needed on ontologies that cover all relevant aspects and areas, and on nice and intuitive end-user interfaces to make this process as easy as possible).

The paper that you are reading is in fact a genuine semantic publication. It has different representations for different types of usage. You might be reading these lines while sitting on a beach and reading from a sheet of paper printed from the article's PDF version, or you might be reading it in your office from a web page in HTML format within your browser window. In either case, these representations contain the narrative text, which we carefully wrote to explain and motivate our ideas to human readers. But we also make our work available to automated agents (i.e. any kind of software programs), for which we have different representations that consist of formal RDF statements instead of narrative text. Importantly,

these RDF statements convey the same main message as the narrative text: They are different representations of the same work!

To formally represent the main content of the paper, we can make use of existing ontologies and vocabularies, such as CiTO [37] and SKOS [49]. Specifically, our paper's main message is the advocacy of the new concept of genuine semantic publishing, which can be expressed as follows in the Turtle RDF notation [50]:

```
p:paper cito:describes p:GenuineSemanticPublishing ;
        cito:supports p:GenuineSemanticPublishing .
```

There is to our knowledge no existing ontology that would exactly capture the relation of a publication *advocating* a given concept, but the combination of the two relations `describes` and `supports` from the CiTO ontology comes close. We as authors should of course say a bit more about this new concept, most importantly that it is related to the existing concept of semantic publishing:

```
p:GenuineSemanticPublishing skos:related dbpedia:Semantic_publishing .
```

And we can express our critical position on that concept:

```
p:paper cito:critiques dbpedia:Semantic_publishing .
```

Next we can formally represent the five criteria based on which we define our new concept:

```
p:GenuineSemanticPublishing skos:definition
  p:GenuineSemanticPublishingCriteria .

p:GenuineSemanticPublishingCriteria dct:hasPart
  p:GenuineSemanticPublishingCriterion1 .

p:GenuineSemanticPublishingCriterion1
  dct:title "First criterion of genuine semantic publishing: machine
    interpretability" ,
  dct:description "A scientific work needs to come with formal representations
    that are semantic, in the sense that ..." .

p:GenuineSemanticPublishingCriteria dct:hasPart
  p:GenuineSemanticPublishingCriterion2 .

...
```

We can try to capture part of the content of these criteria in RDF as well, but at some point we have to stop and be content with an informal description in natural language (at the latest when we hit the symbol grounding problem). However, we believe that it is always possible to build a formal representation of the main content at the highest level, such as introducing and advocating a new concept, even though we will mostly not be able to provide a complete formal definition. In this sense, such a representation is underspecified but has essential coverage.

We would like to note here that — while we are confident in declaring that our own representation complies with our criteria — we do not intend to claim that it achieves them to the highest degree possible. It is, to the contrary, still a quite crude representation that leaves many details and aspects of our main

claims and arguments untouched. For example, we state that our paper critiques the concept of semantic publishing, but we do not say why and in what way, namely that we claim its interpretation to be not intuitive and not visionary. We are not aware of any ontology that would allow us to express this, and we restricted ourselves for this demonstration to existing resources. More work will be needed on establishing such ontologies and best practices to facilitate more precise and more inclusive formal models of scientific findings and arguments, but the currently existing vocabularies already allow — at least in our case — to achieve a basic level of genuine semantic publishing.

In any case, the benefits of such a representation of the main message of a paper might not seem obvious at this point. One of the main benefits comes when *subsequent* papers or nanopublications start referring to these formal representations. As a fictitious example, a subsequent paper might propose the concept of "advanced semantic publishing" that includes our criteria 1 to 4, but criticizes number 5 and suggests to replace it with a different one:

```
p2:anotherPaper cito:describes p2:AdvancedSemanticPublishing ;
                cito:supports p2:AdvancedSemanticPublishing .
p2:AdvancedSemanticPublishing skos:related p:GenuineSemanticPublishing .
p2:anotherPaper cito:critiques p:GenuineSemanticPublishingCriterion5 .
p2:AdvancedSemanticPublishing skos:definition
  p:AdvancedSemanticPublishingCriteria .
p2:AdvancedSemanticPublishingCriteria dct:hasPart
  p:GenuineSemanticPublishingCriterion1 ,
  p:GenuineSemanticPublishingCriterion2 ,
  p:GenuineSemanticPublishingCriterion3 ,
  p:GenuineSemanticPublishingCriterion4 ,
  p2:AdvancedSemanticPublishingCriterion .
p2:AdvancedSemanticPublishingCriterion
  dct:title "Criterion for advanced semantic publishing" ,
  dct:description "..." .
```

This example shows how we can formally capture the high-level relation of papers' content, and thereby place them in the wider context of the literature on the respective topic.

The above RDF representations are interpretable by machines, and thereby automated software agents of all sorts can read and process them. Human readers, of course, normally prefer a natural text representation of a paper's content. To account for such different demands, resources on the web can in general have different equivalent representations for different types of agents. *Content negotiation* can then be used in the background to find a suitable representation based on the agent's request (mediated by the browser) and the available representation formats on the server side. Alternatively, we can use special kinds of hyperlinks on a landing page to achieve the same effect within HTML. We will use this landing page approach here for demonstration purposes because it makes the different representations more explicit, but the presence of a landing page is not required.

We can create a simple landing page with links to the different (classical and semantic) representations of the work. With just a few lines of HTML code, we can define a canonical URL and some minimal metadata (such as title and authors of the work; more metadata is available in the actual representations):

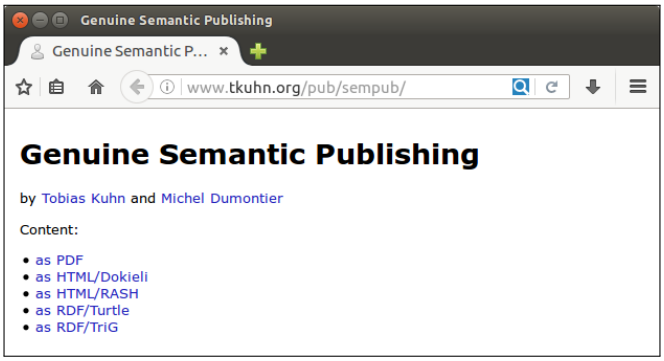

Figure 2. The landing page pointing to different versions of the work.

```
<!DOCTYPE html>
<html>
<head>
<link rel="canonical" href="http://www.tkuhn.org/pub/sempub/">
<title>Genuine Semantic Publishing</title>
</head>
<body>
<h1>Genuine Semantic Publishing</h1>
<p>by
<a href="http://orcid.org/0000-0002-1267-0234" rel="author">Tobias Kuhn</a> and
<a href="http://orcid.org/0000-0003-4727-9435" rel="author">Michel Dumontier</a>
</p>
```

And then we can link to different representations of the content of the given work:

```
<p>Content:</p>
<ul>
<li><a rel="item" href="sempub.pdf"
  type="application/pdf">as PDF</a></li>
<li><a rel="item" href="sempub.dokieli.html"
  type="text/html">as HTML/Dokieli</a></li>
<li><a rel="item" href="sempub.rash.html"
  type="text/html">as HTML/RASH</a></li>
<li><a rel="item" href="sempub.ttl"
  type="text/turtle">as RDF/Turtle</a></li>
<li><a rel="item" href="sempub.trig"
  type="application/trig">as RDF/TriG</a></li>
</ul>
</body>
</html>
```

Specifically, we link to the PDF version of this work, two flavors of HTML (Dokieli and RASH), and RDF representations in Turtle (without provenance information and metadata) and TriG (with provenance information and metadata in the form of nanopublications). We hereby also demonstrate how existing technologies from related work, such as Dokieli, RASH, and nanopublications, can help us to achieve genuine semantic publishing.

Figure 2 shows what such a minimal landing page looks like in a browser, and the respective data can be found online[11] and in the supplemental material. Importantly, these list items point to different representations of the *same work*, each covering the work's main points and thereby satisfying the second requirement of genuine semantic publishing with respect to essential coverage. The RDF representations are machine interpretable, which satisfies our first criterion, and the fact that they appear on the same level as the narrative papers shows that they are a primary component of the published work, satisfying the fourth criterion. The fact that we as authors created and approved all these representations moreover satisfies the third criterion of authenticity.

To illustrate the last criterion of being fine-grained and light-weight, let us assume that somebody wanted to add at a later point just a single triple to assert the connection between our first criterion and the concept of Linked Data:

```
p:GenuineSemanticPublishingCriterion1 skos:related dbpedia:Linked_data .
```

We can save this triple in a file and create a bare minimum landing page that could look as follows:

```
<!DOCTYPE html>
<html>
<head><title>Genuine Semantic Publishing and Linked Data</title></head>
<body>
<p>
by <a rel="author" href="http://orcid.org/0000-0002-1267-0234">Tobias Kuhn</a>
</p>
<p>Content:</p>
<ul>
<li><a rel="item" href="sempubld.ttl" type="text/turtle">as RDF</a></li>
</ul>
</body>
</html>
```

Together, these two files, containing fewer than 500 bytes, form a complete publication according to our criteria. This demonstrates that fine-grained contributions down to single triples can be published in a very light-weight manner with an overhead of just a few hundred bytes.

6. Conclusions

The downsides and limitations of the current scientific publishing paradigm have become apparent in many ways, from the researchers unable to deal with the avalanche of new papers published in their fields to the struggles of elevating scientific datasets to the level of appreciation they deserve. We argue that we need both, grand visions and small practical steps, to move forward and advance science communication, to make sure that the benefits of future breakthroughs are not offset by our inefficiency in communicating them.

We have to make sure, however, that we do not confuse our grand vision with the small practical steps towards it. *Semantic publishing* was once a grand vision

[11] http://www.tkuhn.org/pub/sempub/

but the term was then hijacked by approaches implementing small practical steps. These small steps are certainly important, but they also made us lose sight of the longer-term vision.

In this position paper, we aimed to focus again on the grand vision, which we propose to call *genuine semantic publishing* to distinguish it from the existing approaches. We argued that genuine semantic publications should not only come with representations that are machine interpretable, but that these representations also need to have essential coverage of the work's main claims, that they need to be authentic in the sense that they are approved by the authors, that they should form a primary component of the work, and that they should allow for fine-grained and light-weight contributions.

By explaining how this very paper was written as a genuine semantic publication, we demonstrated that — as far as technology is concerned — the vision is not that grand after all. Technically, genuine semantic publications are at a basic level already feasible nowadays with established and mature technologies. Of course, many grand challenges remain, including many details of the overarching formal models, reliable domain models in many fields, intuitive user interfaces, data publishing infrastructures, attribution and recognition of scientific efforts, and effective incentive structures. All these challenges can only be addressed, however, with a clear vision of how scientific publishing should develop in the future.

Acknowledgment

We would like to thank Silvio Peroni and Tim Clark for discussions on the topic, and the reviewers and Herbert van de Sompel for their very valuable suggestions to improve the article. Figure 1 was designed by Germán Barboza, from Cordero Producciones.[12]

— NEW REFERENCES: [11, 23, 26, 36, 41, 44, 45, 48] —
— UPDATED REFERENCES: [9, 12, 16, 28, 33, 35, 38] —
— REFERENCE REMOVED: T. Kuhn, P. E. Barbano, M. L. Nagy, and M. Krauthammer, "Broadening the scope of nanopublications," in *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*. Springer, 2013. —

References

[1] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop *et al.*, "Why linked data is not enough for scientists," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, 2013.

[2] J. Priem, "Scholarship: Beyond the paper," *Nature*, vol. 495, no. 7442, pp. 437–440, 2013.

[3] T. Clark, "Next generation scientific publishing and the web of data," *Semantic Web*, vol. 5, no. 4, pp. 257–259, 2014.

[12] `http://designers.designcrowd.com/designer/32548/cordero-producciones`

[4] D. Shotton, K. Portwin, G. Klyne, and A. Miles, "Adventures in semantic publishing: exemplar semantic enhancements of a research article," *PLoS computational biology*, vol. 5, no. 4, p. e1000361, 2009.

[5] D. Shotton, "Semantic publishing: the coming revolution in scientific journal publishing," *Learned Publishing*, vol. 22, no. 2, pp. 85–94, 2009.

[6] A. de Waard, "From proteins to fairytales: directions in semantic publishing," *Intelligent Systems, IEEE*, vol. 25, no. 2, pp. 83–88, 2010.

[7] T. Berners-Lee and J. Hendler, "Publishing on the semantic web," *Nature*, vol. 410, no. 6832, pp. 1023–1024, 2001.

[8] C. Lange and A. Di Iorio, "Semantic publishing challenge — assessing the quality of scientific output," in *Semantic Web Evaluation Challenge*. Springer, 2014, pp. 61–76.

[9] A. Di Iorio, C. Lange, A. Dimou, and S. Vahdati, "Semantic publishing challenge — assessing the quality of scientific output by information extraction and interlinking," in *Semantic Web Evaluation Challenge*. Springer, 2015, pp. 65–80.

[10] S. Vahdati, A. Dimou, C. Lange, and A. Di Ioro, "Semantic publishing challenge: bootstrapping a value chain for scientific data," 2016. [Online]. Available: http://cs.unibo.it/save-sd/2016/papers/html/vahdati-savesd2016.html

[11] F. L. Holmes, "Scientific writing and scientific discovery," *Isis*, vol. 78, no. 2, pp. 220–235, 1987.

[12] P. Smrz and J. Dytrych, "Towards new scholarly communication: A case study of the 4A framework," in *First Workshop on Semantic Publication (SePublica 2011)*. CEUR-WS, 2011. [Online]. Available: http://ceur-ws.org/Vol-721/paper-07.pdf

[13] A. Ruiz-Iniesta and O. Corcho, "A review of ontologies for describing scholarly and scientific documents," in *4th Workshop on Semantic Publishing (SePublica 2014)*. CEUR-WS, 2014. [Online]. Available: http://ceur-ws.org/Vol-1155/paper-07.pdf

[14] A. C. S. Croset, S. Kafkas, M. Liakata, and A. Oellrich, "Exploring the generation and integration of publishable scientific facts using the concept of nano-publications," in *First Workshop on Semantic Publication (SePublica 2011)*. CEUR-WS, 2011. [Online]. Available: http://ceur-ws.org/Vol-721/paper-02.pdf

[15] C. H. Marcondes, "A semantic model for scholarly electronic publishing," in *First Workshop on Semantic Publication (SePublica 2011)*. CEUR-WS, 2011. [Online]. Available: http://ceur-ws.org/Vol-721/paper-06.pdf

[16] A. Di Iorio, S. Peroni, F. Vitali, and J. Zingoni, "Semantic lenses to bring digital and semantic publishing together," in *Proceedings of the 4th International Workshop on Linked Science (LISC 2014)*. CEUR-WS, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2878586

[17] R. Palma, P. Hołubowicz, O. Corcho, J. M. Gómez-Pérez, and C. Mazurek, "ROHub — a digital library of research objects supporting scientists towards reproducible science," in *Semantic Web Evaluation Challenge*. Springer, 2014, pp. 77–82. [Online]. Available: 10.1007/978-3-319-12024-9_9

[18] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan, "Research objects: Towards exchange and reuse of digital knowledge," *The Future of the Web for Collaborative Science*, 2010. [Online]. Available: http://precedings.nature.com/documents/4626/version/1/files/npre20104626-1.pdf

[19] A. González-Beltrán, P. Li, J. Zhao, M. S. Avila-Garcia, M. Roos, M. Thompson, E. van der Horst, R. Kaliyaperumal, R. Luo, T.-L. Lee *et al.*, "From peer-reviewed to peer-reproduced in scholarly publishing: the complementary roles of data models and workflows in bioinformatics," *PLOS one*, vol. 10, no. 7, p. e0127612, 2015.

[20] P. Nowakowski, E. Ciepiela, D. Harężlak, J. Kocot, M. Kasztelnik, T. Bartyński, J. Meizner, G. Dyk, and M. Malawski, "The collage authoring environment," *Procedia Computer Science*, vol. 4, pp. 608–617, 2011.

[21] P. Van Gorp and S. Mazanek, "Share: a web portal for creating and sharing executable research papers," *Procedia Computer Science*, vol. 4, pp. 589–597, 2011.

[22] M. Kohlhase, J. Corneli, C. David, D. Ginev, C. Jucovschi, A. Kohlhase, C. Lange, B. Matican, S. Mirea, and V. Zholudev, "The planetary system: Web 3.0 & active documents for STEM," *Procedia Computer Science*, vol. 4, pp. 598–607, 2011.

[23] T. K. Attwood, D. B. Kell, P. McDermott, J. Marsh, S. Pettifer, and D. Thorne, "Utopia documents: linking scholarly literature with research data," *Bioinformatics*, vol. 26, no. 18, pp. i568–i574, 2010.

[24] S. Capadisli, R. Riedl, and S. Auer, "Enabling accessible knowledge," *CeDEM15: Conference for E-Democracy and Open Government*, p. 257, 2015. [Online]. Available: http://csarven.ca/enabling-accessible-knowledge

[25] A. Di Iorio, A. G. Nuzzolese, F. Osborne, S. Peroni, F. Poggi, M. Smith, F. Vitali, and J. Zhao, "The RASH framework: enabling HTML + RDF submissions in scholarly venues," in *In Proceedings of the ISWC 2015 Posters & Demonstrations Track*, 2015. [Online]. Available: http://ceur-ws.org/Vol-1486/paper_72.pdf

[26] S. Capadisli, A. Guy, R. Verborgh, C. Lange, S. Auer, and T. Berners-Lee, "Decentralised authoring, annotations and notifications for a read-write web with dokieli," in *International Conference on Web Engineering*. Springer, 2017, pp. 469–481.

[27] M. R. Seringhaus and M. B. Gerstein, "Publishing perishing? towards tomorrow's information architecture," *BMC bioinformatics*, vol. 8, no. 1, p. 17, 2007.

[28] A. Ceol, A. Chatr-Aryamontri, L. Licata, and G. Cesareni, "Linking entries in protein interaction database to structured text: the FEBS letters experiment," *FEBS letters*, vol. 582, no. 8, pp. 1171–1177, 2008.

[29] T. Kuhn, L. Royer, N. E. Fuchs, and M. Schroeder, "Improving text mining with controlled natural language: A case study for protein interations," in *3rd International Workshop on Data Integration in the Life Sciences 2006 (DILS'06)*. Springer, 2006.

[30] T. Clark, P. Ciccarese, and C. Goble, "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications," *Journal of Biomedical Semantics*, vol. 5, no. 1, p. 28, 2014.

[31] B. Mons, H. van Haagen, C. Chichester, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond *et al.*, "The value of data," *Nature genetics*, vol. 43, no. 4, pp. 281–283, 2011.

[32] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nano-publication," *Information Services and Use*, vol. 30, no. 1, pp. 51–56, 2010.

[33] C. Chichester, P. Gaudet, O. Karch, P. Groth, L. Lane, A. Bairoch, B. Mons, and A. Loizou, "Querying neXtProt nanopublications and their value for insights on sequence variants and tissue expression," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014.

[34] J. M. Banda, T. Kuhn, N. H. Shah, and M. Dumontier, "Provenance-centered dataset of drug-drug interactions," in *Proceedings of the 14th International Semantic Web Conference (ISWC 2015)*. Springer, 2015, pp. 293–300.

[35] N. Queralt-Rosinach, T. Kuhn, C. Chichester, M. Dumontier, F. Sanz, L. I. Furlong *et al.*, "Publishing DisGeNET as nanopublications," *Semantic Web–Interoperability, Usability, Applicability*, 2016.

[36] A. E. Bandrowski and M. E. Martone, "RRIDs: A simple step toward improving reproducibility through rigor and transparency of experimental methods," *Neuron*, vol. 90, no. 3, pp. 434–436, 2016.

[37] S. Peroni and D. Shotton, "FaBiO and CiTO: ontologies for describing bibliographic resources and citations," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 17, pp. 33–43, 2012.

[38] S. Peroni, D. Shotton, and F. Vitali, "Scholarly publishing and Linked Data: describing roles, statuses, temporal and contextual extents," in *Proceedings of the 8th International Conference on Semantic Systems*. ACM, 2012, pp. 9–16.

[39] P. Ciccarese, M. Ocana, L. J. G. Castro, S. Das, and T. Clark, "An open annotation ontology for science on web 3.0," *Journal of biomedical semantics*, vol. 2, no. 2, p. 1, 2011.

[40] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, "PROV-O: The PROV ontology," *W3C Recommendation*, 2013. [Online]. Available: https://www.w3.org/TR/prov-o/

[41] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath *et al.*, "The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery," *Journal of biomedical semantics*,

vol. 5, no. 1, p. 14, 2014.

[42] J. Schneider, T. Groza, and A. Passant, "A review of argumentation for the social semantic web," *Semantic Web*, vol. 4, no. 2, pp. 159–218, 2013.

[43] A. De Waard and J. Schneider, "Formalising uncertainty: An ontology of reasoning, certainty and attribution (ORCA)," *Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine*, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2887634

[44] L. N. Soldatova and R. D. King, "An ontology of scientific experiments," *Journal of the Royal Society Interface*, vol. 3, no. 11, pp. 795–803, 2006.

[45] M. H. Brush, K. Shefchek, and M. Haendel, "SEPIO: A semantic model for the integration and analysis of scientific evidence," in *ICBO/BioCreative*. CEUR-WS, 2016. [Online]. Available: http://ceur-ws.org/Vol-1747/IT605_ICBO2016.pdf

[46] W. J. Clancey, "The knowledge level reinterpreted: Modeling how systems interact," *Machine Learning*, vol. 4, no. 3-4, pp. 285–291, 1989.

[47] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data and Knowledge Engineering*, vol. 25, no. 1–2, pp. 161–197, March 1998.

[48] T. S. Kuhn, *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[49] A. Miles and S. Bechhofer, "SKOS simple knowledge organization system reference," *W3C recommendation*, p. W3C, 2009. [Online]. Available: https://www.w3.org/TR/skos-reference/

[50] D. Beckett and T. Berners-Lee, "Turtle — terse RDF triple language," *W3C Recommendation*, 2011. [Online]. Available: https://www.w3.org/TeamSubmission/turtle/