BRAINIACS
BRAIN IMAGING AND COMPUTING SCIENCES

2020 Volume 1
Issue 1 Edoc D11DABE6D

Survey, Analysis, and Requirements for Semantic Enhancement to Support Machine Understanding of Scientific Literature*

Adam Craig, Peter Hong, Shreya Choksi, Anousha Athreya, Carl Taswell†

Abstract

Adoption of the proposed recommendations with standards set forth by the World Wide Web Consortium (W3C) regarding the semantic web remains a work in progress, especially with regard to their use in the published research literature. Proponents of the semantic enhancement of scholarly publishing have described it as a visionary breakthrough for the way in which both individuals and machines should be able to obtain meaningful information from data, text, and content management systems. However, the availability and prevalence of useful real-world resources remains limited. In this report, we present a survey of those scholarly research journals that focus on the semantic web and ontology engineering. We highlight noteworthy examples of publishers offering semantic enhancement and markup services in hopes of shedding light on tools that could revolutionize how both academic scholars and the lay public find and understand the published results of scientific research. We then consider the implications of these findings for the growth and development of the semantic web as a whole. We also review proposals for how the semantic web could accelerate the advancement of brain sciences and brain health. Finally, we propose a novel approach to scholarly publishing represented by our planned semantic enhancement workflow process for the Brainiacs Journal. By surveying the current use of the semantic web, we show the need for more motivated and enthusiastic adoption of semantic enhancement in scholarly publishing in order to "stand on the shoulders of giants" and reap the benefits of published research.

Keyphrases

Semantic web, semantic enhancement, semantic markup, ontology engineering, knowledge engineering, semantic publishing.

Contents

*Document created 2020-07-31, updated 2020-12-30.
†Correspondence to ctaswell@bhavi.us.

Introduction

Proponents of the semantic web have long promulgated the vision that internet data and metadata should be machine-readable and that any human-readable text should also have a representation of the content in a format which an automated reasoning agent can use as a basis for inferences. If populated with such representations of the intended meaning rather than just the desired appearance of a corpus of text, the World Wide Web would advance beyond serving content based on simple keyword matches to finding meaningful answers to users' questions. However, this vision has not come to fruition yet. The current version of much of the web remains as it has been as the original lexical web rather than the envisioned semantic web. This lexical web operates in a manner analogous to how a parrot mimics a human's speech and dialogue without understanding the meaning of the spoken language. A partial explanation for this slow transition from lexical web to semantic web remains the fact that anyone wishing to encode knowledge in semantic markup must choose between the labor-intensive process as a human curator of creating it manually, or else using natural language processing algorithms which are advancing but still limited and prone to errors (Rajani and Hanumanthappa 2016).

As of now, researchers in many scientific fields are left with no choice but to accept the heterogeneous conditions of the lexical web for their searching, reading, and understanding needs (Damjanovic et al. 2011). The dilemma here is that most "major scientific findings" are disseminated solely via free-form text formats (Sernadela and Oliveira 2017). Furthermore, the rate at which the scientific community generates new literature makes it impossible for a human researcher to keep track of the latest findings. For example, as of 2020-12-30, PubMed has records of approximately 30,000 journals and over 30 million titles and abstracts. As a result, the need for semantic enhancement of traditional

scholarly research articles has become not only a fundamental necessity for the success of the semantic web's proposed vision, but also for investigators who are trying to further their research agendas by "not re-inventing the wheel", and instead, by building upon and benefiting from the research results already published in the scientific literature.

Methods and Results for Surveys

We surveyed scholarly research journals that utilized semantic enhancements for their published research documents. This survey was based on a list of the top one hundred journals with the highest impact factors listed in the 2020 Journal Citation Reports (Zaman 2020). Devised by Eugene Garfield, the impact factor of a given journal can be calculated by taking the number of citations received in the given year and dividing that number by the number of publications released in the last two years. The formula for the impact factor of a journal is as follows:

$$\mathrm{IF}_y = \frac{\mathrm{Citations}_y}{\mathrm{Publications}_{y-1} + \mathrm{Publications}_{y-2}} \quad (1)$$

The categories for inclusion within the semantic enhancement survey were taken from Shotton (2009). These categories included downloadable XML, downloadable datasets, images, audio, and video, tabbed interfaces, reference management tools, structured digital abstracts, and semantic markup of text . Besides this comprehensive list of categories for inclusion, we also added a category for related third-party linked content. In conducting our survey, we only considered what content was available on the landing page for a sample article, that is, the page to which the DOI of the article resolved. As seen in the subsequent survey of semantic web journals, many publishers provide XML metadata for their articles through a separate REST API. Our results from this survey can be found in Table 1. From the one hundred journals surveyed, ninety nine of them included some form of supplemental semantic enhancement.

Table 1: Enhanced Online Content in Highest Impact Factor Journals

Evaluation Categories	No. of Journals	Percentage
Downloadable XML	0	0%
Downloadable Datasets	0	0%
Audio, Video, Interactive Media	1	1%
Tabbed interfaces	65	65%
Reference Management	37	37%
Structured Digital Abstracts	0	0%
Semantic Text Markup	1	1%
Related Linked Content	34	34%
Journal Sample Size	100	Total

While the types of "semantic enhancement" as described by Shotton *et al.* can serve as a broad description of how effectively publishers are taking advantage of the possibilities of digital publishing, several of Shotton's categories are not relevant to the goal of making the information conveyed in a scholarly research article useful for and usable by automated reasoning agents. To provide a more focused set of categories for knowledge engineering, we propose four levels of *semantic enhancement* denoted by the terms: *None, Metadata, Triples, and Reified* (see Table 2). Category 'None' is the baseline for our interpretation and includes supplementary material that is not machine parsable. Category 'Metadata' includes any kind of machine-parsable metadata format. Category 'Triples' includes machine-parsable metadata formats that support subject-verb-object triple relationships. Finally, category 'Reified' includes machine-parsable subject-verb-object triples that not only represent the metadata but also reify statements and describe the meaningful content, such as the scientific evidence reported and claims discussed in the article. Thus, a Reified semantic representation of an article encodes statements made in the article as entities with properties and relationships to other statements.

In addition to the survey described above, we highlight here a handful of noteworthy examples of publishers offering semantic enhancement and markup services. To find these publishers, key search terms included, but were not limited to, "semantic enhancement", "semantic markup", and "semantic publishing". Our primary objective in this second survey was to identify and gather a list of journals and publishers offering semantic enhancement services that were not already in the first survey of the top one hundred journals with the highest impact factors. We analyzed journals and publishers using Google Search, Google Scholar, and Semantic Scholar. We were especially cautious when it came to the type of semantic enhancement surveyed. The scope of our search included scientific journals and publishers with semantic enhancement and markup services, but excluded independent software services for semantic enhancement of legal information, business and financial documents, and other fields that are not within the umbrella of scholarly research journals and scientific publishing.

Leading the way, the Pensoft Publishers became the first major scholarly research publishers to implement semantic XML tagging and enrichment of published articles as a routine editorial practice in 2010. This semantic markup service best exemplifies category Metadata of our four semantic enhancement categories. This publishing house has a variety of journals under its name, and one noteworthy achievement is that *ZooKeys*, Pensoft's taxonomic journal, became the first to implement a semantic enhancement workflow practice in a taxonomic journal (Penev et al. 2010). The Royal Society of Chemistry has also taken a major lead in the semantic publishing workflow, as their journals provide enhanced HTML versions of the papers with semantic markup of free-form text by industry experts and provide users with the ability to highlight text terms from ontologies. This type of semantic markup would best fit under category Metadata, because it provides machine-readable metadata but not subject-verb-object triples. The journal, *Acta Crystllagraphica A: Foundations of Crystallography*, has promoted markup of free-form text with links to definitions for many years, and they reference terms from major dictionaries, including the IUCr Online Dictionary of Crystallography (Chapuis n.d.). This journal would fit best under category None because of its inclusion of only non-machine-parsable external, supplemental material. These publishers and journals provide the scholarly research community with a clear demonstration of the benefits and feasibility of semantically enhancing papers, but these practices are rare in the world of scientific publishing. To find out why, we must delve deeper into the previous visions for the semantic web.

In addition to the most influential journals and some notable examples of early adopters of semantic enhancement, we surveyed journals that focus on the semantic web and closely related subjects. To identify these journals, we searched the Clarivate Analytics Web of Science Master Journal List. On 2020-12-24, we searched using the terms "semantic" and "web" without quotes. The search engine reported 1 exact match and 1642 partial matches, but only 6 showed a clear focus on the

Table 2: Categorizing Examples of Semantic Markup

Category	Descriptor	Example
None	no machine-parsable article metadata	The Royal Society of Chemistry publishes multiple formats for exporting citation and supplementary material.
Metadata	article metadata in some machine-parsable format	Pensoft Publishers offers tabbed headings, multiple formats for exporting citations, as well as metadata in XML documents.
Triples	article metadata formatted as machine-parsable subject, verb, object triple relationships	Springer Nature SciGraph publishes RDF subject-verb-object triples representing the bibliographic metadata, but not the content, of the Springer-Nature journals.
Reified	article content reified as identified entities described by machine-parsable subject-verb-object triples	No known journals or publishers provide reified metadata.

semantic web. 10 were more general information science journals, and 5 concerned semantics in human linguistics. It also listed *The Journal of Web Ecology*, which concerns ecology and does not relate to web engineering or semantics. After the first 22 results, it listed seemingly irrelevant journal titles alphabetically. We report results for all the first 22 journals except for *The Journal of Web Ecology*. On the same day, we also searched Oxford Academic Database using "ontology" without quotes. The search engine reported 9 results. Of these, 2 included in their scope the development of formal ontologies for semantic markup, while the remaining 7 were journals of philosophy. We then classified each journal according to the highest of the four levels of metadata it provided (see Table 3). Impact factors reported in this table are for the year 2018 as reported on Clarivate Analytics' profile page for each journal.

Of the journals surveyed, none prominently advertised availability of semantic markup on their home pages, but some publishers maintain separate web services for distributing computer-parsable metadata. IGI Global provides the InfoSci-Databases Platform, which distributes bibliographic data for its articles in XML and in the MARC binary format widely used among libraries (*InfoSci-Databases platform 2020 user's guide* 2020). Elsevier goes a step further by providing a suite of APIs through which a client can search for and download machine-parsable article metadata formatted in JSON or XML using open web standards, such as Atom Syndication Format, Publishing Requirements for Industry Standard Metadata, OpenSearch, and Dublin Core.

To the best of our knowledge, World Scientific Publishing's *International Journal of Semantic Computing*, the Association for Computing Machinery's *ACM Transactions on the Web*, Emerald Publishing's *International Journal of Web Information Systems*, MDPI's *Future Internet*, Taylor & Francis Books' *Journal of Web Librarianship*, the Linguistic Society of America's *Semantics and Pragmatics*, and Brill's *Syntax and Semantics* do not publish their own repositories of article metadata, but they are members of CrossRef with open references, meaning that automated agents can download and parse bibliographic metadata on their journals and articles through the CrossRef REST API. Similarly, InderScience's *International Journal of Web and Grid Services*, Oxford Academic's *Journal of Semantics*, and de Gruyter's *Journal of Literary Semantics* are indexed in Elsevier's Scopus, so bibliographic metadata records for its articles are available through the Scopus API. Metadata records for *Journal of Literary Semantics* articles are also available through the Semantic Scholar REST API. The Directory of Open-Access Journals (DOAJ) also hosts the metadata for articles from a wide variety of journals through its API, including philosophy journals that the other indices do not cover.

In addition to article metadata services specific to one publisher or federating a broad spectrum of publisher's works, we also found problem-oriented domain-specific services. River Publishers' *Journal of Web Engineering* makes metadata for its articles available not only through both Scopus and CrossRef but also the DBLP Computer Science Bibliography, which has its own REST API. NCBI offers several REST APIs for retrieving article metadata and other information programmatically. The two journals with a biomedical focus, BMC's *Journal of Biomedical Semantics* and Oxford Academic's *Database*, are both part of PubMed Central's Open Access Subset. This means that, not only does the Entrez API provide article metadata and a separate utility to retrieve other articles that each article cites or that cite it, but the OAI-PMH also provides the full text of articles in XML. The client can select text formatted using Dublin Core terms or the Journal Archiving and Interchange tag set (see OAI). Why so many publishers rely on a relatively small number of metadata services remains an open question, but some possible reasons include a lack of interest in making their publications discoverable to automated agents, a lack of technical know-how, or a lack of easily deployed, open-source metadata management software infrastructure that meets their requirements.

Among the publishers of semantic web-focused journals, only Springer-Nature and IOS Press provide semantic descriptions as subject-verb-object triples. Springer-Nature SciGraph provides RDF triple representations of the bibliographic information of all Springer-Nature periodicals and scholarly articles, among other entities (Yaman et al. 2019). However, these descriptions do not include any account of the key claims of the articles. Moreover, they have an inconsistent level of completeness. For example, an inspection using the SN SciGraph Explorer of various articles reveals examples of metadata descriptions without any indication of who authored the articles. However, such provenance information would be valuable for maintaining accountability. IOS Press provides a similar service, LD Connect, which allows the user to navigate from one entity to another in a semantic graph or submit SPARQL queries (Hu et al. 2013). As with SN SciGraph, LD Connect provides only bibliographic information for articles, not descriptions of their contents. Also, although Janowicz and Hitzler report that the IOS Press-published journal *Semantic Web* practices open peer review (Janowicz and Hitzler 2012), only articles since volume 6 issue 4 identify editor names and the percentage of article descriptions that name peer reviewers remains unclear (when last checked 2020-12-28). Furthermore, unlike in SN SciGraph, none of the article descriptions listed the articles' cited references. At this time, *Semantic Web* also hosts a separate platform that allows browsing of linked data records docu-

Table 3: Survey of Journals Related to the Semantic Web

Journal	Publisher	Problem Domain	Impact Factor	Semantic Markup	Service
Journal of Biomedical Semantics	BMC, Springer-Nature	semantic web	1.582	Triples	SciGraph, Entrez, OAI-PMH, DOAJ
Journal of Web Semantics	Elsevier	semantic web	2.429	Metadata	Scopus
International Journal on Semantic Web and Information Systems	IGI Global	semantic web	1.833	Metadata	InfoSci
Semantic Web	IOS Press	semantic web	3.524	Triples	LD Connect
Journal on Data Semantics	Springer-Nature	semantic web	none	Triples	SciGraph
International Journal of Semantic Computing	World Scientific	semantic web	none	Metadata	CrossRef
ACM Transactions on the Web	ACM	web engineering	1.580	Metadata	CrossRef
International Journal of Web Information Systems	Emerald Publishing	web engineering	none	Metadata	CrossRef
International Journal of Information Technology and Web Engineering	IGI Global	web engineering	none	Metadata	InfoSci
International Journal of Web Services Research	IGI Global	web engineering	0.447	Metadata	InfoSci
International Journal of Web and Grid Services	InderScience	web engineering	0.833	Metadata	Scopus
Web Intelligence	IOS Press	web engineering	none	Triples	LD Connect
Future Internet	MDPI	web engineering	none	Metadata	CrossRef, DOAJ
Journal of Web Engineering	River Publishers	web engineering	0.854	Metadata	Crossref, Scopus, DBLP
World Wide Web: Internet and Web Information Systems	Springer-Nature	web engineering	1.770	Triples	SciGraph
Database - The Journal of Biological Databases and Curation	Oxford Academic	knowledge engineering	3.683	Metadata	Scopus, Entrez, OAI-PMH, DOAJ
Journal of Web Librarianship	Taylor & Francis	knowledge engineering	none	Metadata	CrossRef
Applied Ontology	IOS Press	ontology engineering	0.750	Triples	LD Connect
Semantics and Pragmatics	Linguistic Society of America	linguistics	none	Metadata	CrossRef, DOAJ
Journal of Literary Semantics	de Gruyter	linguistics	none	Metadata	Scopus, Semantic Scholar
Journal of Semantics	Oxford Academic	linguistics	1.773	Metadata	Scopus
Natural Language Semantics	Springer-Nature	linguistics	1.381	Triples	SciGraph
Syntax and Semantics	Brill	linguistics	none	Metadata	CrossRef
EPEKEINA: International Journal of Ontology, History, and Critics	The International Center for Philosophical Research	philosophy	none	Metadata	N/A
Metaphysica	de Gruyter	philosophy	none	Metadata	Scopus, Semantic Scholar
Horizon: Studies in Phenomenology	St. Petersburg State University	philosophy	none	Metadata	Scopus
Nuevo Pensamiento: Revista de Filosofía	Universidad del Salvador	philosophy	none	Metadata	DOAJ
Revista de Filosofia Aurora	Pontifical Catholic University of Puerto Rico	philosophy	none	Metadata	Scopus, CrossRef, DOAJ
Revista de Filosofía Madrid	Complutense University of Madrid	philosophy	none	Metadata	DOAJ

Table 4: Tracking Measures for Growth and Development of Semantic Web

Key Areas	Type	Definition
Scalability	Indicator	Boolean value for whether semantic databases with inference engines can distribute problems across multiple servers
	Quantitative	Integer count for how many semantic databases with inference engines are distributing problems across multiple servers
Availability of Content	Quantitative	Percentage of existing journals that provide semantic enhancement services
	Quantitative	Integer count of how many network systems are providing semantic metadata
	Quantitative	Integer count of how many organizations are using these network systems
	Indicator	Boolean value for whether the semantic web has a generic data browser
Visualization	Qualitative	Categorical values for which technologies provide capabilities for generating adequate representations for semantic technology
	Quantitative	Integer count for how many current semantic web services are utilizing visualization technologies
Multilingualism	Qualitative	Integer count of how many ontology builders provide multilingual capabilities
	Qualitative	Integer count for how many semantic web annotation services provide other language options

menting its peer review process. But when last checked on 2020-12-28, their web portal did not appear to be operational with the advertised functionality.

Implications for the Semantic Web

Having seen that only a minority of publishers publish semantic descriptions of their articles and that those who do have yet to fully represent the claims made in the reports of research, we ask the following: At what point will the semantic web reach the level of maturity needed for automated reasoning agents to perform meaningful analyses of the content of such articles, including meta-analyses, pre-publication review, and detection of idea plagiarism? To discuss these matters, we must first consider the proposals made in the past involving the future of the semantic web. Seringhaus and Gerstein (2007) proposed their visions for the semantic web indirectly by arguing for the use of an optimal information architecture for biosciences publications that could lead to advancements in the procedures of handling the massive scientific data pool present throughout today's web. They argued for the integration of intelligent markup with free-form text and the production of Structured Digital Abstracts, so that articles could be *computer-readable*. Borgman (2008) discussed the notion that data is exponentially more valuable when it is linked to other relevant publications and resources. Shotton (2009) argued for the use of semantic enhancements in journal articles to increase the intrinsic value within the papers themselves. None of these visions, however, have been implemented at a scale of prevalence. To obtain a better understanding as to why the semantic web has not become more prevalent, we must compare the development of the semantic web to other kinds of networks built in the past.

Although the terms *railroad* and *railway* are oftentimes used interchangeably, the two terms have a slightly different meaning. The railway is defined as the physical infrastructure on which trains travel and can be compared to the lexical web in the sense that both services directly take one person to their desired goal or destination. On the other hand, analogous to the semantic web's goal of serving as a ubiquitous network to make Internet data machine-readable, the railroad is defined as a network of named lines or company routes on which trains could travel. Using this logic, it can be inferred that the development of the railway had to occur before the that of the railroad. Also, the railroads only became prevalent when the railways had a widespread and effective impact on individuals. In a similar manner, the semantic web can only become prevalent if individuals first truly realize the benefits and maximize them to the fullest extent.

Tracking Measures for the Semantic Web

To track the growth and development of the semantic web, we have developed concrete measures as a means to monitor its status in the future (see Table 4). As a whole, we can consider the current evolution of the semantic web with regard to evaluating the number of original W3C standards established for the semantic web that are adopted. In addition to general measures for the entirety of the semantic web, we review a few key areas for which specific measures can be developed relating to its growth. These areas include scalability, availability of content, visualization, and multilingualism (Benjamins et al. 2002).

In information systems, scalability refers to the idea that the resources needed to solve a problem should not grow faster than the size of the problem itself (Hellman 2009). This can be further distinguished into "vertical scaling" where resources needed are linearly related to the size of the problem, or the broader "horizontal scaling" where a problem is distributed across a network of servers, and the number of servers required should scale with the size of the problem (Hellman 2009). With respect to horizontal scaling, measures can be proposed as a boolean indicator of whether there exists a capability for semantic databases with inference engines to distribute problems across multiple servers, and as a qualitative descriptor of how many semantic databases with inference engines are distributing problems across multiple servers.

Availability of content can be measured in terms of how prevalent the use of the semantic web is among both the general public and different industries. We propose quantitative measures such as defining the percentage of existing journals that provide semantic enhancement services. We can also track simple counts of how many network systems are providing semantic metadata and how many organizations are using these network systems. As a method for determining use among the public, we include an indicator of whether the semantic web has a generic data browser similar to lexical web borwsers. Such browsers motivated the growth of the World Wide Web by making it easier for the general public to access web content, greatly broadening the potential readership for website creators. It is important to distinguish here that these generic data browsers must not only be generic in the sense that

they are not tailored to specialized applications but must also provide easily understood views of large amounts of data.

Visualization refers to the organization of semantic content in ways that are intuitive and easily recognizable to users. This will involve categorical descriptors of which technologies provide capabilities for generating adequate representations for semantic technology as well as quantitative descriptors for how many current semantic web services are utilizing visualization technologies (Benjamins et al. 2002). Multilingualism standards are also important in creating a truly global semantic web used by users all over the world, and so capabilities must be put in place for semantic web technologies and resources to be written in a variety of different languages. This will include measurements of how many ontology builders provide multilingual capabilities, and how many semantic web annotation services provide other language options for annotation of content (Benjamins et al. 2002).

Applications to Brain Sciences and Brain Health

While evaluating the growth of the semantic web, it is important to consider its potential for applications in particular problem-oriented domain-specific fields. Assuming that the semantic web should be built with data integration, interoperability, and connecting diverse information systems using intelligent machine computing, semantic web technologies should facilitate the management and sharing of knowledge among a diversity of multi-disciplinary and trans-disciplinary fields (C. Taswell 2008; Karami and Rahimi 2019). The semantic web should provide a means to overcome the lack of interoperability among the numerous databases that contain the data and metadata which researchers need in order to answer complex questions (Lam et al. 2006). As an example, we survey some applications of the semantic web to brain research and neuroscience.

One example of how the semantic web can aid integration of heterogeneous data is the Semantic SenseLab (Samwald, Chen, et al. 2010). This platform combines anatomical and neurophysiological information from NeuronDB, descriptions of the actions of pathological and pharmacological agents on the brain from BrainPharm, and computational models of neurons and brain regions from ModelDB to facilitate simulation of brain states under a wide variety of healthy and disease conditions. One of the major benefits Samwald, Chen, et al. (2010) observed was the ability to use semantic reasoning agents to identify both errors in data entry and contradictions between claims in the scientific literature, but they also found that OWL ontologies they used did not always provide adequate tools for representing uncertainty, evidence, and data provenance. These efforts built on the earlier Semantic Synapse Project, an attempt to develop a suite of ontologies representing knowledge about the biology of neuronal synapses (Samwald and Adlassnig 2005).

Sahoo et al. (2008) similarly demonstrated the value of using semantic markup and ontologies to integrate databases by combining gene information from the Entrez Gene and HomoloGene databases with pathway information from the KEGG, Reactome, and BioCyc databases. To do this, they built a new information model in OWL, called Entrez Knowledge Model (EKoM), designing it specifically to provide semantic relationships among fields in the Entrez Gene database and to interoperate with the existing BioPAX pathway ontology. This allowed them to translate three complex questions about the role of genetics in nicotine dependence into SPARQL queries and derive answers from information in the integrated databases: "Which genes participate in a large number of pathways?" "Which genes (or gene products) interact with each other?" and "Which genes are expressed in the brain?". One major hurdle Samwald, Chen, et al. (2010). encountered was identifying when the same entity was represented in multiple databases, such as when IDs 00620 in KEGG and 71406 in Reactome both referred to the pyruvate metabolism pathway. They worked around this by adding an 'owl:sameAs' assertion whenever wo pathways had the same 'SHORT-NAME' property (Sahoo et al. 2008). That they resorted to this *ad hoc* approach suggests the need for more methodical approaches to identifying equal or equivalent entities.

Iyappan et al. (2016) took this approach a step further by developing the NeuroRDF framework, integrating curated data from protein-protein interaction databases such as Bind and IntAct, gene and protein associations mined from article on PubMed using their own custom named entity recognition software, and gene expression resources, including GEO and ArrayExpress. Their main use of this system was to identify genes that play a causal role in Alzheimer's Disease by identifying candidates with a causal influence on expression of numerous other genes that are disregulated in Alzheimer's patients. Mechouche et al. (2009) demonstrated a very different application of semantic web technology: a hybrid system that used numerical image segmentation to form initial guesses about the locations of specific gyri and sulci in an MRI, then decided on the final annotations of those regions by using a semantic reasoning engine and anatomical ontology to determine which guesses were consistent with existing knowledge about their positions in a typical human brain. This innovative approach shows the power of semantic web technologies to bridge the gap between textual knowledge and numerical sensor data in order to draw conclusions about the physical world. Beyond these individual, application-specific uses of semantic web technology, Ruttenberg et al. (2007) envision a more general process of using the semantic web to review different hypotheses, identify evidence for or against them in the literature, and propose new experiments that could fill gaps in understanding of a problem domain.

The PORTAL-DOORS Project (PDP) began with the development of the ManRay ontology for radiopharmaceuticals and nuclear medicine, which C. Taswell et al. (2006) intended to support the use case of automated meta-analyses of clinical trials involving PET brain scans. The need for interoperable metadata repositories that identify, describe, and locate resources on the internet, web, and grid lead to the creation of the PORTAL-DOORS System and, later expanded to the Nexus-PORTAL-DOORS System, a messaging protocol and REST API that organize diverse lexical metadata properties and semantic descriptions and associate them with a resolvable URI entity label (C. Taswell 2008; Craig, Bae, et al. 2016). To support the progress of research in the domain of brain health, we developed the BrainWatch Nexus combined directory and registry (diristry) as a working example of a repository for semantic descriptions of a wide variety of brain sciences related resources, including articles, journals, data sets, people, and organizations. By populating this diristry with high-quality records, we hope to create a valuable information repository for resource discovery and analysis. As described in Choksi and C. Taswell (2020), the principle of Garbage In, Garbage Out asserts that, in order to provide quality output, an algorithm must receive quality input. With this in mind, Choksi and C. Taswell (2020) describe metrics of metadata record quality that Brain Health Alliance will use to evaluate and monitor all of the PDP web services and data repositories that it maintains.

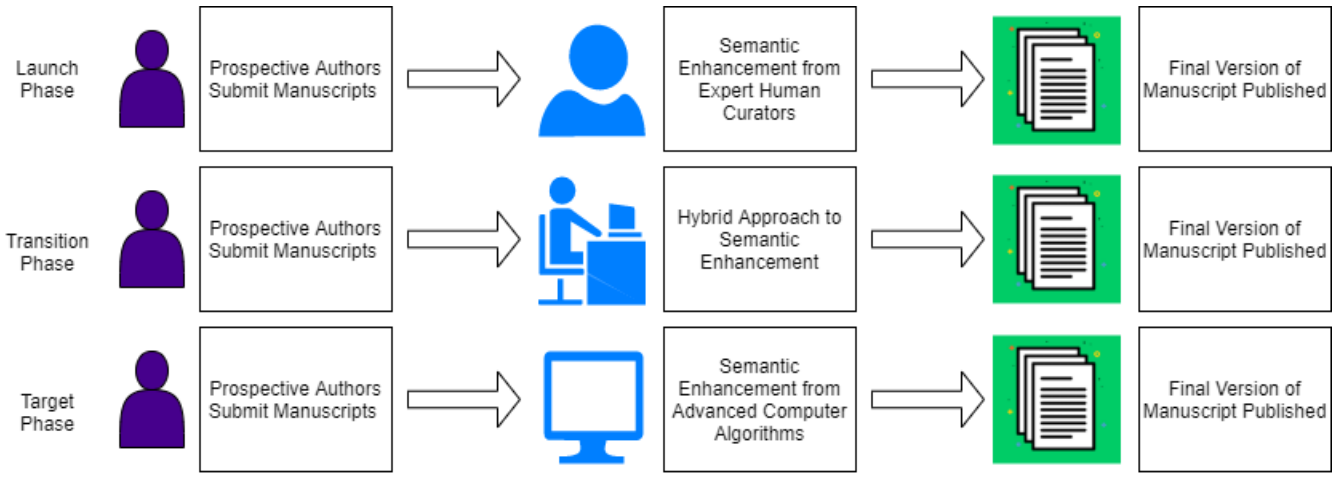

Figure 1: Semantic Enhancement Workflow Outline for Brainiacs Journal

Semantic Enhancement for the Brainiacs Journal

To support the potential of the semantic web as a platform for making better use of the information in scholarly research articles, Brain Health Alliance is launching the Brainiacs Journal of Brain Imaging And Computing Sciences. All documents submitted for publication in the Brainiacs Journal must adhere to specific standards and requirements set forth by the peer review workflow process found at www.brainiacsjournal.org. Prospective authors must follow a list of formatting standards for their manuscript and provide a non-technical summary that a wide range of readers can understand. Furthermore, every article submitted to the Brainiacs Journal must have a corresponding entry in the Brainiacs Nexus diristry, which must include an RDF representation of the key claims of the article. In the launch phase of the journal, expert human curators will create these semantic descriptions during this peer review workflow process. In the subsequent transition phase, as we learn from this experience which tasks are rote and repetitive and which require human judgement, we will continue the development of our software tools and algorithms to expedite those steps of the process that we can most reliably automate. In the final target phase, we plan to have developed the software algorithms sufficiently to the point where they can generate the semantic descriptions of the natural language text in a fully automated manner (see Figure 1 on the workflow process).

While we call this the target phase, it does not represent a point of stagnation. Brain Health Alliance (BHA) will continue to iterate through improvements to make the semantic representations more thorough and accurate. For example, even after natural language processing algorithms are able to fully convert the text of the article to a semantic description, we will most likely need the authors to include written descriptions of the figures. However, as image processing improves, we plan to incorporate a module that automatically provides semantic descriptions of figure contents. Generating semantic descriptions of tables may also require a separate algorithm, but we expect to receive tables in a machine-readable format that can be easily converted to RDF and only requires additional triples for context.

However, the Brainiacs Journal will not merely generate semantic descriptions of its articles but will also use them to make inferences about how well a paper adheres to standards for citing prior work. To do this, human reviewers and, in later issues, automated algorithms will compute the FAIR Attribution to Identified Reports (FAIR) Metrics defined in Craig, Ambati, et al. (2019). This approach classifies claims according to whether the authors present them as pre-existing or novel to the work. Pre-existing claims are then classified as valid if the authors cite a source that contains an equivalent claim, invalid otherwise. Novel claims are classified as valid if no known source contains an equivalent claim, invalid otherwise. We then count how many of the claims fall into each of the four categories and compute a family of ratios from them that reflect different aspects of good citation practices (Craig, Ambati, et al. 2019): Is the amount of properly cited background proportionate to the number of novel claims? How prevalent is misattribution? What fraction of claims do the authors seem to have plagiarized from another work? This assessment process will facilitate verification that each article upholds the scholarly integrity requirements proposed and described by S. K. Taswell et al. (2020).

As opposed to the traditional paid subscription model, the open access (OA) publishing model allows individuals to access articles openly and freely in the field of scholarly academic research given that they have the ability to go online and retrieve the articles (Laakso et al. 2011). This publishing model offers widespread visibility of manuscripts at a global scale. This, however, raises the question of who is bearing the cost of maintaining the publishing service. Currently, OA can be accomplished by following either the gold or green route. The gold route allows for immediate open access to the final published paper immediately after publication (made possible because the publisher charges the author a publication fee), and copyright for the article is retained by the authors. Gold OA articles can be published in one of two types of journals: fully OA journals (all papers are published OA) or hybrid journals (authors are given an option to chose either OA publishing or traditional subscription-based publishing). On the other hand, green OA is made possible because of the payments made from subscribers of the journal who pay for early access to each issue. At a later date, the articles become available for free to the general public. At the publisher's discretion, the public archive may include the final version, a preprint, or both, and the copyright for the manuscript is retained by the publisher as well. Most fully OA and hybrid journals offer green OA.

The Brainiacs Journal follows neither the green nor the gold OA model. As a 501-c-3 nonprofit organization, Brain Health Alliance will support its publication service for the Brainiacs Journal with the pool of funds donated to BHA, and will not charge any fees to either authors or readers. However, similar to the gold OA model, there will be never be a

time delay imposed for access to the manuscripts after publication. Furthermore, peer review will never be blinded (neither single-blinded nor double-blinded), nor will peer review be restricted in time. Peer reviewers may contribute their remarks and reviews at any time, but must provide proof of identity for authorship of their reviews. A record of all peer reviews, including semantic descriptions of reviewers' comments, will be published in association with and linked to the primary published document that has been reviewed by the peer. Authors of both primary documents and secondary reviews will always be identified and known to readers in a manner consistent with the principles for promoting scientific truth and research integrity in our hitchhiker's guide to scholarly research integrity (S. K. Taswell et al. 2020).

Conclusion

The semantic web has not yet matured and remains in the early stages of its development. Even with admirable proposals and endeavors underway to enhance the web semantically, there remains much more work to be done in the semantic publishing realm to build a foundational cyberinfrastructure. The semantic web should not be deemed mature until and unless the presence of semantic enhancement and markup becomes mainstream with a much greater prevalence. In order for this proposed vision to be realized and become successful, individual scholarly research journals must first play their part in the widespread adoption of semantic enhancement services analogous to how dominos play their part in the 'domino effect'. As the scientific community achieves more advances in basic and applied research, we must build an equally advanced and sophisticated semantic web with information systems capable of comprehending the results from this scientific progress with knowledge engineering tools and technologies.

Citation

Brainiacs 2020 Volume 1 Issue 1 Edoc D11DABE6D
Title: "Survey, Analysis, and Requirements for Semantic Enhancement to Support Machine Understanding of Scientific Literature"
Authors: Adam Craig, Peter Hong, Shreya Choksi, Anousha Athreya, Carl Taswell
Dates: created 2020-07-31, updated 2020-12-30, published 2020-12-31, reprinted 2026-03-07

Contact: ctaswell@bhavi.us
NPDS: LINKS/Brainiacs/Craig2020SARSE
DOI: 10.48085/D11DABE6D

Affiliations

Brain Health Alliance Virtual Institute, www.BHAVI.us, Ladera Ranch, California, USA.

References

[1] V. R. Benjamins, J. Contreras, O. Corcho, and A. Gomez-Perez. "The six challenges of the semantic web" (2002) (cited pp. 5, 6).

[2] C. Borgman. "Data, disciplines, and scholarly publishing." *Learned publishing* 21.1 (2008), pp. 29–38 (cited p. 5).

[3] G. Chapuis. *IUCr Online Dictionary of Crystallography*. URL: https://dictionary.iucr.org/Main_Page (cited p. 2).

[4] S. Choksi and C. Taswell. "The Nexus-PORTAL-DOORS-Scribe (NPDS) Learning Intelligence aNd Knowledge System (LINKS)." *Brainiacs Journal of Brain Imaging And Computing Sciences* 1.1, B61CA3D89 (1 Dec. 30, 2020), pp. 1–9. DOI: 10.48085/B61CA3D89 (cited p. 6).

[5] A. Craig, A. Ambati, S. Dutta, P. Kowshik, S. Nori, S. K. Taswell, Q. Wu, and C. Taswell. "DREAM Principles and FAIR Metrics from the PORTAL-DOORS Project for the Semantic Web." In: *2019 IEEE 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (June 28, 2019). Pitesti, Romania: IEEE, June 2019, pp. 1–8. DOI: 10.1109/ECAI46879.2019.9042003. URL: https://portaldoors.org/pub/docs/ECAI2019DREAMFAIR0612.pdf (cited p. 7).

[6] A. Craig, S. H. Bae, T. Veeramacheneni, S. K. Taswell, and C. Taswell. "Web Service APIs for Scribe Registrars, Nexus Diristries, PORTAL Registries and DOORS Directories in the NPD System." In: *Proceedings 9th International SWAT4LS Conference*. Amsterdam, Netherlands, 2016. URL: https://ceur-ws.org/Vol-1795/paper4.pdf (cited p. 6).

[7] V. Damjanovic, T. Kurz, R. Westenthaler, W. Behrendt, A. Gruber, and S. Schaffert. "Semantic enhancement: the key to massive and heterogeneous data pools." In: *Proceeding of the 20th international IEEE ERK (electrotechnical and computer science) conference*. 2011, pp. 413–6 (cited p. 1).

[8] E. Hellman. "Is semantic web technology scalable?" (June 2009). URL: https://go-to-hellman.blogspot.com/2009/06/is-semantic-web-technology-scalable.html (cited p. 5).

[9] Y. Hu, K. Janowicz, G. McKenzie, K. Sengupta, and P. Hitzler. "A linked-data-driven and semantically-enabled journal portal for scientometrics." In: *International Semantic Web Conference*. Springer. 2013, pp. 114–129 (cited p. 3).

[10] *InfoSci-Databases platform 2020 user's guide*. 2020th ed. IGI Global. 701 E. Chocolate Avenue, Hershey, PA 17033, USA, 2020 (cited p. 3).

[11] A. Iyappan, S. B. Kawalia, T. Raschka, M. Hofmann-Apitius, and P. Senger. "NeuroRDF: semantic integration of highly curated data to prioritize biomarker candidates in Alzheimer's disease." *Journal of biomedical semantics* 7.1 (2016), pp. 1–15 (cited p. 6).

[12] K. Janowicz and P. Hitzler. "Open and transparent: the review process of the Semantic Web journal." *Learned publishing* 25.1 (2012), pp. 48–55 (cited p. 3).

[13] M. Karami and A. Rahimi. "Semantic web technologies for sharing clinical information in health care systems." *Acta Informatica Medica* 27.1 (2019), p. 4 (cited p. 6).

[14] M. Laakso, P. Welling, H. Bukvova, L. Nyman, B.-C. Björk, and T. Hedlund. "The development of open access journal publishing from 1993 to 2009." *PloS one* 6.6 (2011), e20961 (cited p. 7).

[15] H. Lam, L. Marenco, T. Clark, Y. Gao, et al. "Semantic web meets e-neuroscience: an RDF use case." In: *Proceedings of International Workshop on Semantic e-Science, ASWC*. 2006, pp. 158–70 (cited p. 6).

[16] A. Mechouche, X. Morandi, C. Golbreich, and B. Gibaud. "A hybrid system using symbolic and numeric knowledge for the semantic annotation of sulco-gyral anatomy in brain MRI images." *IEEE Transactions on Medical Imaging* 28.8 (2009), pp. 1165–1178 (cited p. 6).

[17] L. Penev, D. Agosti, T. Georgiev, T. Catapano, et al. "Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples." *ZooKeys* 50 (2010), p. 1 (cited p. 2).

[18] S. Rajani and M. Hanumanthappa. "Techniques of semantic analysis for natural language processing – a detailed survey." *International Journal of Advanced Research in Computer and Communication Engineering* 5 (Oct. 2016), p. 4. ISSN: 2278-1021. DOI: 10.17148/IJARCCE (cited p. 1).

[19] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, et al. "Advancing translational research with the Semantic Web." *BMC bioinformatics* 8.3 (2007), pp. 1–16 (cited p. 6).

[20] S. Sahoo, O. Bodenreider, J. Rutter, K. Skinner, and A. Sheth. "An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence." *Journal of biomedical informatics* 41.5 (2008), pp. 752–765 (cited p. 6).

[21] M. Samwald and K.-P. Adlassnig. "Bringing neuroscience to the Semantic Web: the Semantic Synapse Project." In: *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*. Vol. 1. IEEE. 2005, pp. 145–150 (cited p. 6).

[22] M. Samwald, H. Chen, A. Ruttenberg, E. Lim, L. Marenco, P. Miller, G. Shepherd, and K.-H. Cheung. "Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience." *Artificial intelligence in medicine* 48.1 (2010), pp. 21–28 (cited p. 6).

[23] M. Seringhaus and M. Gerstein. "Publishing perishing? Towards tomorrow's information architecture." *BMC bioinformatics* 8.1 (2007), p. 17 (cited p. 5).

[24] P. Sernadela and J. L. Oliveira. "A semantic-based workflow for biomedical literature annotation." *Database* 2017 (2017) (cited p. 1).

[25] D. Shotton. "Semantic publishing: the coming revolution in scientific journal publishing." *Learned Publishing* 22.2 (2009), pp. 85–94 (cited pp. 2, 5).

[26] C. Taswell. "DOORS to the Semantic Web and Grid with a PORTAL for Biomedical Computing." *IEEE Transactions on Information Technology in Biomedicine* 12.2 (2 Mar. 2008). In the Special Section on Bio-Grid published online 3 Aug. 2007, pp. 191–204. ISSN: 1089-7771. DOI: 10.1109/TITB.2007.905861 (cited p. 6).

[27] C. Taswell, B. Franc, and R. Hawkins. "The ManRay Project: Initial Development of a Web-Enabled Ontology for Nuclear Medicine." In: *Proceedings of the 53rd Annual Meeting of the Society of Nuclear Medicine, San Diego, CA*. Vol. 47. suppl 1. Soc Nuclear Med, June 2006, p. 1431 (cited p. 6).

[28] S. K. Taswell, C. Triggle, J. Vayo, S. Dutta, and C. Taswell. "The Hitchhiker's Guide to Scholarly Research Integrity." In: *2020 ASIS&T 83rd Annual Meeting* (Oct. 22, 2020). Vol. 57. Wiley, 2020, e223. DOI: 10.1002/pra2.223. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.223 (cited pp. 7, 8).

[29] B. Yaman, M. Pasin, and M. Freudenberg. "Interlinking SciGraph and DBpedia datasets using link discovery and named entity recognition techniques." In: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2019 (cited p. 3).

[30] W. Zaman. *JCR, SCI complete list of journal reports 2020*. June 2020 (cited p. 2).

