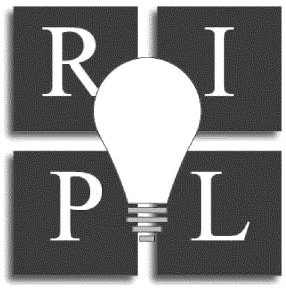
THE JOHN MARSHALL
REVIEW OF INTELLECTUAL PROPERTY LAW



THE NEW ONTOLOGIES: THE EFFECT OF COPYRIGHT PROTECTION ON PUBLIC SCIENTIFIC DATA SHARING USING SEMANTIC WEB ONTOLOGIES

ANDREW CLEARWATER

ABSTRACT

The semantic web is going to become an important tool for scientists who need to accurately share data given context through structured relationships. The structure that defines contextual relationships on the semantic web is known as an ontology; which is a hierarchical organization of a knowledge domain that contains entities and their relations. This paper seeks to answer whether semantic web ontologies are protectable by copyright, and regardless of the outcome, what the best practices are for the scientific community. The best practices for the scientific community should include the adoption of a machine readable ontology license which disclaims copyright protection for publication of public scientific data to assure automation of the integration of ontologies and to maximize easy access to public science materials that can be queried. Sharing and information is essential for scientists and failure to address the possibility of ontologies as a possible constraint to public data access could result in data fragmentation and lost scientific opportunities. The ability of the semantic web to annotate and reuse data relies on the social structure of science supporting data sharing as a norm and as an extension of this norm, open licensing of ontologies should be embraced.

Cite as Andrew Clearwater, *The New Ontologies: The Effect of Copyright Protection on Public Scientific Data Sharing Using Semantic Web Ontologies*, 10 J. MARSHALL REV. INTELL. PROP. L. 182 (2010).

THE NEW ONTOLOGIES: THE EFFECT OF COPYRIGHT PROTECTION ON PUBLIC SCIENTIFIC DATA SHARING USING SEMANTIC WEB ONTOLOGIES

ANDREW CLEARWATER

INTRODUCTION 183
I. DATA SHARING IN SCIENCE, AN UNREALIZED IDEAL 184
A. Social Norms and the Importance of Openness in Science 184
II. THE SEMANTIC WEB 185
A. What is the Semantic Web? 185
1. Resource Description Framework ("RDF") 186
2. Ontologies—Using Ontology Web Language ("OWL2") 187
3. Source Databases (data storage) 188
B. Querying the Semantic Web Using SPARQL 188
III. COPYRIGHT PROTECTION AND THE SEMANTIC WEB 189
A. Copyright Overview 189
B. Copyright of Ontologies 190
C. Copyrightable Subject Matter—Ideas/Expression 191
D. Copyrightable Subject Matter—Titles, Headings, Short Phrases 192
E. Copyrightable Subject Matter—Compilations & Taxonomies 193
F. Copyrightable Subject Matter—Government Works 195
G. Copyrightable Subject Matter—The Merger Doctrine 196
H. Fixation 197
I. Brief Restatement of Copyright Conclusions 199
IV. BEST PRACTICES 200
A. The Panton Principles 200
B. Science Commons: Protocol for Implementing Open Access Data 201
C. Embedded Licensing in Interlinked Data Sets 203
D. Example Semantic Ontology License 204
CONCLUSION 205

THE NEW ONTOLOGIES: THE EFFECT OF COPYRIGHT PROTECTION ON PUBLIC SCIENTIFIC DATA SHARING USING SEMANTIC WEB ONTOLOGIES

ANDREW CLEARWATER*

INTRODUCTION

The current web is a web designed for finding documents.[1] The semantic web is a web designed for finding data.[2] Data is best found through structured relationships when accuracy and context are desired.[3] Data sharing in science is the type of exercise where accuracy and context is required.[4] Scientific patterns of information exchange require standards, and the semantic web can provide useful tools for structuring data according to standard structured relationships.[5] The structure that defines contextual relationships on the semantic web is known as an ontology, which is a hierarchical organization of a knowledge domain that contains entities and their relations.[6]

Is a semantic web ontology protectable by copyright law? Semantic web ontologies may be substantially similar to compilations and taxonomies which are copyrightable subject matter.[7] This paper seeks to answer whether semantic web

---

* Andrew Clearwater is a legal consultant to iCommons, Ltd., a graduate of Suffolk University Law School's Global Law and Technology LLM, and a former Research Assistant at Harvard's Berkman Center for Internet and Society. The author would like to thank Stephen McJohn, his faculty advisor; Stephen Hicks, the Director of the Graduate Law Program in Global Technology; and John Wilbanks, VP of Science at Creative Commons whose outstanding work is the foundation of this article.

[1] *See* Tim Berners-Lee, James Hendler, & Ora Lassila, *The Semantic Web*, SCI. AM., May 2001, at 34, 36–37.

[2] *See id.*; *see also* James Hendler, Tim Berners-Lee, & Eric Miller, *Integrating Applications on the Semantic Web*, 122 J. INST. ELEC. ENG'RS JAPAN 676 (2002) (describing how the semantic web allows information to be more easily linked together).

[3] *See* Berners-Lee et al., *supra* note 1, at 40.

[4] RESEARCH INFO. NETWORK & THE BRITISH LIBRARY, PATTERNS OF INFORMATION USE AND EXCHANGE: CASE STUDIES OF RESEARCHERS IN THE LIFE SCIENCES 7, 37 (2009) [hereinafter PATTERNS OF INFO.], *available at* http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf.

[5] *See id.*

[6] *See* WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY OF THE ENGLISH LANGUAGE 1577 (Philip Babcock Gove, ed., Merriam-Webster, Inc., 2002) (1961); *WordNet Search 3.0*, PRINCETON.EDU., http://wordnetweb.princeton.edu/perl/webwn?s=ontology (last visited Sept. 30, 2010).

[7] *See* 17 U.S.C. § 103(a) (2006) ("[T]he subject matter of copyright as specified by § 102 includes compilations . . . . "); *see also* Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977, 980 (7th Cir. 1997) (holding that the taxonomy is an original work of authorship and qualifies for copyright protection); Oren Perez, *Complexity, Information Overload, and Online Deliberation*, 5 I/S J.L. & POL'Y FOR INFO. SOC'Y 43, 67 (2008–09) (defining ontologies as taxonomies, thus falling within potential copyright protection).

ontologies are protectable by copyright and, regardless of the outcome, determine the best practices for the scientific community.

I. DATA SHARING IN SCIENCE, AN UNREALIZED IDEAL

A. Social Norms and the Importance of Openness in Science

Sharing and information is essential for scientists. A recent study by the Research Information Network and the British Library found that "[m]ost life science researchers spend much of their time searching for and organising information."[8] Given the amount of time spent with data it would seem data access would be a priority among scientists, yet a recent article in *Nature* was subtitled "Most researchers agree that open access to data is the scientific ideal, so what is stopping it happening?"[9] The failure often is the result of data fragmentation, or what James Boyle calls "mutually incomprehensible scientific information."[10] The tools and framework for sharing data have not been made available to scientists and research funding agencies and have not made data preservation and access a priority.[11] The central role of data curation and sharing makes the tools for exchange all the more important.[12] Thus, the ability of the semantic web to annotate and reuse data relies on the social structure of science supporting data sharing as a norm.[13] John Wilbanks, executive director of Science Commons, asserts that:

> We need investment in annotation and curation, in capacity to store and render data, and in shared visualization and analytics. We need open standards for sharing and exposing data. We need the RFCs (Requests for Comments) of the data layer. And, above all, we need to teach scientists and scholars to work in this new layer of data.[14]

Data management has an essential role in science.[15] According to Thomas Kuhn, the most important advances in science come from the "continuing growth of the assembled data that [the scientific community] can treat with precision and detail" and the accumulation of which leads scientists to see the world in new ways

---

[8] PATTERNS OF INFO., *supra* note 4, at 6.

[9] Bryn Nelson, *Empty Archives*, 461 NATURE 160, 160 (2009).

[10] *Id.* at 161.

[11] Nature Publishing Group Editorial, *Data's Shameful Neglect: Research Cannot Flourish if Data are not Preserved and Made Accessible*, 461 NATURE 145, 145 (2009) [hereinafter *Data's Shameful Neglect*].

[12] *Id.*

[13] *See* PATTERNS OF INFO., *supra* note 4, at 5–7, 39 (discussing the balance between researchers concerns of unfettered access and data sharing of intellectual capital).

[14] John Wilbanks, *I Have Seen the Paradigm Shift, and It is Us*, *in* THE FOURTH PARADIGM: DATA-INTENSIVE SCIENTIFIC DISCOVERY 213 (Tony Hey et al., eds., 2009), *available at* http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

[15] *Data's Shameful Neglect, supra* note 15, at 145 ("[D]ata management should be woven into every course in science.").

that undermine current theories and cause a paradigm shift.[16] The role of the semantic web, which can provide useful tools for structuring data, is put into a useful context given the needs and values of the scientific community.

II. THE SEMANTIC WEB

A. What is the Semantic Web?

The current web can best be described as a web of documents and not a web of data. When we search the Google search engine, the search results lead us to locations on web sites where terms included or related to our search can be found.[17] For example, typing "color of the sky" into Google leads to a list of web pages that includes "Blue Sky - Why is the Sky Blue?" located at www.sciencemadesimple.com/sky_blue.html.[18]

The semantic web aims to connect data. A search of the web using Wolfram Alpha, a semantic web search engine, gives a very different result.[19] When "color of the sky "is queried by Wolfram Alpha, the result is:

Input interpretation:

What color is the sky?

Result:

sky blue (color)

Color swatch:

. . .

Representations:

fractions | red 0.53 | green 0.81 | blue 0.92

24-bit RGB | red 135 | green 206 | blue 235

HSB | hue 197deg | saturation 43% | brightness 92%

---

[16] T. S. KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS, 169–70 (Univ. Chicago Press, 2d ed. 1970).

[17] *See generally* SERGEY BRIN & LAWRENCE PAGE, STANFORD UNIV., THE ANATOMY OF A LARGE-SCALE HYPERTEXTUAL WEB SEARCH ENGINE (1999), *available at* http://infolab.stanford.edu/pub/papers/google.pdf (presenting Google as a "large-scale search engine" for the very first time).

[18] "Color of the Sky" inquiry, GOOGLE, http://www.google.com.

[19] "Color of the Sky" inquiry, WOLFRAM|Alpha, http://www.wolframalpha.com/input/?i= color+of+the+sky.

hexadecimal | #87CEEB.[20]

The various representations of the blue color, as well as the single page that displays an answer rather than a series of locations, shows some of the important differences introduced by semantic web technologies.

The semantic web is "a general framework wherein syntax is designed to model semantics more closely than conventional online markup languages like HTML currently allow."[21] The semantic web framework has three basic elements: (1) a language that allows descriptions of relationships between concepts; (2) an ontology that describes the relationships and defines interactions within the defined area of knowledge; and (3) a data storage format.[22] Operating together, these basic elements are a powerful tool for data storage and recall.

1. Resource Description Framework ("RDF")

The first element of the semantic web framework is a language that allows statements of relationships between concepts and it is often satisfied by the use of RDF.[23] RDF stands for Resource Description Framework and builds upon the web's use of eXtensible Markup Language ("XML") to tag content.[24] An example of XML in use can be seen by navigating the Creative Commons website.[25] Using the Firefox web browser, right click the mouse to "view page source." The keywords associated with the Creative Commons homepage page are tagged in XML and are now viewable as: "<meta name='keywords' content='creative commons, commons, free culture, open source, attribution, non-commercial, share-alike, no derivatives, lessig' />."[26] These keywords are all terms that describe the Creative Commons but the relationships among the terms are not shown in XML. "The goal of RDF is to enable machines to identify relationships among data at a conceptual level by using XML tags to create 'triples,' much like a subject, verb, object in a normal sentence."[27] As an example, the "attribution" keyword from the Creative Commons page XML described in RDF looks like:

```
<rdf:RDF>
    . . .
    <rdf:Description rdf:about="http://creativecommons.org/ns#Attribution">
    <rdfs:comment xml:lang="en-US">credit be given to copyright holder and/or
```

---

[20] *Id.*

[21] Johnathan Jenkins, *What Can Information Technology Do for Law?*, 21 HARV. J. LAW & TECH. 589, 603 (2008) *citing* Berners-Lee et al., *supra* note 1, at 36–37 (discussing the semantic web and mechanisms for representing knowledge); *see also* Perez, *supra* note 7.

[22] *See* Berners-Lee et al., *supra* note 1, at 36 (discussing the three basic elements in the form of a hypothetical); Perez, *supra* note 7, at 67; Jenkins, *supra* note 21, at 603.

[23] Berners-Lee et al., *supra* note 1, at 38; Michael W. Carroll, *Creative Commons and the New Intermediaries,* 2006 MICH. ST. L. REV. 45, 60 (2006) (explaining the role of RDF to enable machines to identify data relationships).

[24] *See* Berners-Lee et al., *supra* note 1, at 38; Carroll, *supra* note 23, at 59–60.

[25] CREATIVE COMMONS, http://www.creativecommons.org (last visited Sept. 30, 2010).

[26] *Id.*

[27] Carroll, *supra* note 23, at 60.

```
        author</rdfs:comment>
        <rdfs:label xml:lang="en-US">Attribution</rdfs:label>
        <rdf:type rdf:resource="http://creativecommons.org/ns#Requirement"/>
        </rdf:Description>
        . . .
</rdf:RDF>[28]
```

As the example shows, RDF puts information attributes in context. "RDF extends the linking structure of the Web to use [Uniform Resource Identifiers] to name the relationship between things as well as the two ends of the link."[29]

2. Ontologies—Using Ontology Web Language ("OWL2")

The second element of the semantic web framework is an ontology that describes relationships and defines interactions.[30] More formally, an ontology is defined as "a formal, explicit specification of a shared conceptualization."[31] To use a metaphor, keywords in XML are like suggested subjects that publishers print on the inside of book covers.[32] They might give you an idea what you will find within the books contents but you do not know how the book's contents relate to other subjects. When RDF is paired with an ontology it is like finding that same book shelved under the Dewey Decimal Classification. By seeing how the book fits among the surrounding books, you know more about the book as it relates to other subject matter and this allows you to better associate it with other resources available in the library.

Ontologies enable RDF to describe the relationships between data. Tim Berners-Lee describes semantic web ontologies as "[c]ollections of statements written in a language such as RDF that define the relations between concepts and specify logical rules for reasoning about them."[33] Not only can concepts be defined, but equivalence relations can be established to allow more accurate search results.[34] For example, the use of RDF to describe attribution on the Creative Commons web page would allow an ontology to establish that attribution might also refer to credit or acknowledgment but not when used in combination with copyright licensing.

Ontologies are described in a variety of computer languages but a common family of knowledge representation languages endorsed by the World Wide Web Consortium is the Web Ontology Language ("OWL2").[35] OWL2 is compatible with

---

[28] *Document Tree*, CREATIVE COMMONS, http://creativecommons.org/schema.rdf (last visited Sept. 30, 2010).

[29] *Resource Description Framework (RDF)*, WORLD WIDE WEB CONSORTIUM (Feb. 10, 2004), http://www.w3.org/RDF/.

[30] Perez, *supra* note 7, at 67.

[31] Thomas F. McInerney, *Implications of High Performance Production and Work Practices for Theory of the Firm and Corporate Governance*, 2004 COLUM. BUS. L. REV. 135, 176 (2004).

[32] *E.g.*, WAYNE C. BOOTH, GREGORY G. COLOMB, & JOSEPH M. WILLIAMS, THE CRAFT OF RESEARCH (Univ. of Chicago Press, 2d ed. 2003).

[33] Berners-Lee et al., *supra* note 1, at 38.

[34] Carroll, *supra* note 23, at 60.

[35] *OWL 2 Web Ontology Language Document Overview*, WORLD WIDE WEB CONSORTIUM, 1 (Oct. 27, 2009), http://www.w3.org/2009/pdf/REC-owl2-overview-20091027.pdf.

RDF and acts as a vocabulary extension of RDF.[36] "Ontologies are critical for applications that want to search across or merge information from diverse communities" and OWL2 provides the ability to describe and represent an area of knowledge with context.[37]

3. Source Databases (data storage)

The third element of the semantic web framework is data storage.[38] The RDF information needs to be stored and available for quick analysis via the OWL2.[39] There are many storage infrastructure systems in use including "Jena2, Sesame, rdfDB, Redland, Kowari, and FORTH RDF Suite."[40] These systems are often referred to as RDF stores, and they provide both data storage and access.[41] It is not critical to understand data stores in order to understand the copyrightability of semantic ontologies so only this brief introduction is needed.

B. Querying the Semantic Web Using SPARQL

One of the great benefits of Semantic Web technologies is the ability to query data at "different levels of granularity and specificity" made possible by the hierarchical representation of data.[42] For instance, Alan Ruttenberg suggests there are three levels of representing scientific knowledge: record level (database records), statement level (what researchers say), and domain level (the best understanding of consensus).[43] This is possible if data relationships in the ontology are described in sufficient complexity to represent the range of relationships that represent scientific knowledge.

The SPARQL Protocol and RDF Query Language ("SPARQL") enables RDF searches of the collection of asserted statements and is supported by the World Wide Web Consortium.[44] Tim Berners-Lee, the World Wide Web Consortium Director, probably states the importance of SPARQL best when he said, "Trying to use the Semantic Web without SPARQL is like trying to use a relational database without

---

[36] *Id.* at 8.

[37] *OWL Web Ontology Language Use Cases and Requirements,* WORLD WIDE WEB CONSORTIUM, § 1.1 (Feb. 10. 2004), http://www.w3.org/TR/2004/REC-webont-req-20040210/.

[38] *See* Berners-Lee et al., *supra* note 1, at 36.

[39] Andrea Harth & Stefan Decker, Conference Paper at Proceedings of the Third Latin American Web Congress: Optimized Index Structures for Querying RDF from the Web (2005), *available at* http://sw.deri.org/2005/02/dexa/yars.pdf.

[40] *Id.*

[41] *See id.* at 7, 9.

[42] *See* Berners-Lee et al., *supra* note 1, at 36–37 (illustrating examples of the different levels of "granularity and specificity"); Alan Ruttenberg, Introduction to Science Commons and the Neurocommons: Building an Information Framework for Neuroscience 16 (Oct. 19, 2007) (on file with The John Marshall Review of Intellectual Property Law).

[43] Ruttenberg, *supra* note 42, at 17.

[44] *See SPARQL Protocol for RDF,* WORLD WIDE WEB CONSORTIUM, § 1 (Jan. 15, 2008), http://www.w3.org/TR/rdf-sparql-protocol/; *SPARQL Query Language for RDF*, WORLD WIDE WEB CONSORTIUM, § 1 (Jan. 15, 2008), http://www.w3.org/TR/rdf-sparql-query/.

SQL."[45] While SPARQL is recommended, it is a formal language which is important for accuracy but can be detrimental to ease of use.[46] A spectrum of available query languages ranges from easy natural language tools, to mid-complexity semi-structured languages, to formal languages such as SPARQL.[47]

Perhaps the best way to understand the formality of SPARQL is through an example. The Science Commons Neurocommons Text Mining Pilot describes how to use SPARQL to query neuroscience-related PubMed abstracts.[48] The Science Commons example asks: "What are all the CNS-related PubMed abstracts that mention that Entrez Gene 5999?"[49] The example query is shown below:

```
prefix nc: <http://sw.neurocommons.org/2007/annotations#>
    SELECT distinct ?pmid  WHERE
    {       ?pubmed nc:has-id ?pmid.
            ?pubmed nc:has-abstract ?abstract.
            ?span nc:has-context ?abstract.
            ?phrase nc:has-context ?span.
            ?phrase nc:has-nc0.0-interpretation ?ggp.
            ?ggp nc:if-gene-described-by <http://sw.neurocommons.org/2007/entrez-gene/5999>.
    }[50]
```

The trade off, when choosing query languages, is often between ease of use and accuracy.[51] SPARQL is well suited to scientific inquiries because accuracy of the results is critical for scientific data use.

III. COPYRIGHT PROTECTION AND THE SEMANTIC WEB

A. Copyright Overview

Copyright is a statutory system of property protection the authority of which originates in the United States Constitution under Article I, Section 8, Clause 8.[52]

---

[45] *SPARQL is a Recommendation*, WORLD WIDE WEB CONSORTIUM (Jan. 15, 2008), http://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation; *see also* Phil Spector, Introduction to SQL (Mar. 19, 1999), *available at* http://www.stat.berkeley.edu/~spector/sql.pdf (explaining that "SQL" is the abbreviation for Structured Query Language which is a standard language for accessing or manipulating data in a relational database).

[46] Valentin Tablan et al., Conference Paper at Proceedings of the 5th European Semantic Web Conference: A Natural Language Query Interface to Structured Info, at 2–3 (Mar. 12, 2008), *available at* http://gate.ac.uk/sale/eswc08/clone-ql/clone-ql.pdf.

[47] *Id.* at 1–2; Abraham Berstein, *Making the Semantic Web Accessible to the Casual User*, YOUTUBE (June 26, 2008), http://www.youtube.com/watch?v=ayym9jJFIgQ.

[48] *See Neurocommons Text Mining Pilot*, SCIENCECOMMONS, http://sciencecommons.org/projects/data/nc_technical_overview/textmining/ (last visited Sept. 30, 2010).

[49] *Id.*

[50] *Id.* (showing an example query using a GET of the SPARQL endpoint to return all the CNS-related PubMed abstracts that mention that Entrez Gene 5999).

[51] *See* Tablan, *supra* note 46, at 3.

Copyright law grants exclusive rights for authors of original works fixed in a tangible medium of expression for limited times to promote the progress of science.[53] Essentially, copyright builds a property boundary where there is no natural barrier because it applies to a non-rivalrous good that otherwise contains no power to exclude.[54] The rationale for such a boundary is to act as an incentive to create original works, to promote mass distribution of original works, and to allow authors recuperation of costs.[55] Additional rationales include personality or natural rights.[56] These additional rationales are primarily used outside the United States,[57] yet they are part of the philosophical underpinnings of copyright.[58] It is important to note that under copyright law expression is protected, while ideas are not.[59] This important dichotomy helps to lessen the negative impacts copyright may have such as second generation underproduction or the permission culture that develops to accommodate the required exchange of rights.[60]

B. Copyright of Ontologies

An ontology, in a philosophical context is "the metaphysical study of the nature of being and existence" but in computer science it is "a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations."[61] Ontologies are critical to the technology of the semantic web and they allow for more accurate and flexible data retrieval even in queries that request unexpected combinations of data.[62] Given the significance of ontologies to the semantic web, it is important to know whether an ontology that describes the relationships and defines data interactions on the semantic web is protectable by copyright law.

---

[52] *See* U.S. CONST. art. I, § 8, cl. 8 (giving Congress the right to grant exclusive rights to the author of a writing).

[53] *Id.*; 17 U.S.C. § 102 (2006).

[54] Derek E. Bambauer, *Faulty Math: The Economics of Legalizing* The Grey Album, 59 ALA. L. REV. 345, 357 (2008).

[55] Stanley M. Besen & Leo J. Raskind, *An Introduction to the Law and Economics of Intellectual Property*, 5 J. ECON. PERSP. 3, 4 (1991).

[56] *E.g.*, Berne Convention for the Protection of Literary and Artistic Works, art. 6 bis, Sept. 9, 1886, amended on Sept. 28, 1979, 828 U.N.T.S. 221.

[57] *See* Matt Jackson, *Harmony or Discord? The Pressure Toward Conformity in International Copyright*, 43 IDEA 607, 608 (2003).

[58] *See generally* Jackson, *supra* note 57, at 613–16 (explaining the origins and development of copyright internationally).

[59] *E.g.*, Harper & Row, Publishers, Inc. v. Nat'l Enters., 471 U.S. 539, 556–57 (1985) ("No author may copyright his ideas or the facts he narrates. . . . [But] they may at least enjoy the right to market the original expression contained therein as just compensation for their investment."); *see* Baker v. Selden, 101 U.S. 99 (1879).

[60] Thomas F. Cotter, *Fair Use and Copyright Overenforcement*, 93 IOWA L. REV. 1271, 1288, 1317 (2008).

[61] *WordNet Search 3.0*, *supra* note 6; *accord.* WEBSTER'S, *supra* note 6, at 1577.

[62] *See* 2 WILLIAM F. PATRY, PATRY ON COPYRIGHT § 4:51 (2010); Berners-Lee et al., *supra* note 1, at 43.

C. Copyrightable Subject Matter—Ideas/Expression

A threshold issue that must be determined is whether ontologies are copyrightable subject matter. This argument can be broken down into five sub-arguments, the first of which addresses the idea/expression dichotomy recognized by copyright law. The copyright statute states that "[i]n no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work."[63]

The definitive case defining the idea/expression dichotomy is *Baker v. Selden*.[64] This case involved a book authored by Selden titled, *Selden's Condensed Ledger, or Book-keeping Simplified*, which described and taught a method of bookkeeping.[65] The court held that "a claim to the exclusive property in a peculiar system of book-keeping cannot, under the law of copyright, be maintained by the author of a treatise in which that system is exhibited and explained."[66] The system of book-keeping fell outside of copyright protection because the system was an idea or process that is not protected.[67] The public policy that supports this conclusion makes sense. If the process is absolutely necessary, then providing copyright protection prohibits public use, which is in the domain of patent law.[68] One significant review of the principles espoused in *Baker v. Selden* was by *Mazer v. Stein* which essentially reaffirmed *Baker*'s holding.[69]

A leading computer software case which may illuminate the applicability of the idea/expression dichotomy to ontologies concerns a computer program that performed accounting functions.[70] In *Lotus Development v. Borland International*,[71] the menu command hierarchy used by the program was found not to be protectable under copyright law.[72] The arrangement of the commands was a method of operation, and therefore, was considered an unprotected idea rather than a protectable expression.[73]

A semantic web ontology exhibits some of the qualities of a system or method of operation. In a sense, the ontology is a system for finding data.[74] It is a structure that serves the function of leading a user to other content.[75] Like an accounting ledger and its associated system, the ontology is a functional writing which limits the

---

[63] 17 U.S.C. § 102(b) (2006).

[64] 101 U.S. 99 (1880).

[65] *Id.* at 99–100.

[66] *Id.* at 99.

[67] *Id.* at 102.

[68] *Id.* ("To give to the author of the book an exclusive property in the art described therein, when no examination of its novelty has ever been officially made, would be a surprise and a fraud upon the public. That is the province of letters-patent, not of copyright.").

[69] Mazer v. Stein, 347 U.S. 201 (1954) (finding statuettes copyrightable "in so far as their form but not their mechanical or utilitarian aspects are concerned"). *But see* Pamela Samuelson, *Why Copyright Law Excludes Systems and Processes from the Scope of Its Protection*, 85 TEX. L. REV. 1921, 1923, 1956–58 (2007) (questioning the use of *Baker* and *Mazer* as the root of the idea-expression dichotomy).

[70] Lotus Dev. Corp. v. Borland Int'l, Inc., 49 F.3d 807, 809–10 (1st Cir. 1995).

[71] *Id.*

[72] *Id.* at 815.

[73] *Id.*

[74] *See* Hendler et al., *supra* note 2, at 676.

[75] *See* Berners-Lee et al., *supra* note 1, at 36–38.

scope of possible copyright protection. On the other hand, a semantic web ontology is unlike an accounting ledger and its associated system because it is both more specific and more broad. The ontology is much more specific in that it covers in much greater detail the many relationships between concepts. These concepts, unlike the accounting system, may be described differently by different people while still remaining useful.[76] The ontology is much broader in that it is not restrained to a single topical area, such as accounting, but may comprise any number of knowledge domains.[77]

The best modern analogy to semantic web ontologies may be found in XML schemes through which the function is to "express shared vocabularies and allow machines to carry out rules made by people."[78] They are both creative and systematic, but they are also untested by the courts. Given the current discussions in law review articles, it appears that thin copyright protection may be possible.[79] The best support for this conclusion is that an XML schema is used within a system but is not itself systematic.[80] "XML is a syntax that allows computer users to create their own sets of markup tags, also known as 'schemas.'"[81] These sets of tags, like semantic ontologies which are, in a sense, contextual sets of markup tags, are a step in the process of identifying data but they are a separate instrument with separate creative expression.

The key confusion as to whether copyright applies to a semantic web ontology lies in that the ontologies have mechanical or utilitarian aspects which allow searches to be carried out but they also contain expressive aspects in that it takes creativity to define a knowledge domain and its structure. It is possible that very thin protection may be provided but current case law is insufficiently developed to be conclusive on the applicability of the idea/expression dichotomy to a semantic ontology.

D. Copyrightable Subject Matter—Titles, Headings, Short Phrases

Copyright protection requires only a minimal level of creativity.[82] This requirement of originality is often a very low bar, but "[t]he Register of Copyright

---

[76] *See* Hendler et al., *supra* note 2, at 678.

[77] *Id.*

[78] *XML Schema,* WORLD WIDE WEB CONSORTIUM, http://www.w3.org/XML/Schema (last visited Sept. 30, 2010).

[79] *Compare*, Douglas E. Phillips, *XML Schemas and Computer Language Copyright: Filling in the Blanks in Blank Esperanto*, 9 J. INTELL. PROP. L. ASS'N 63, 76–78, 83, 105–07 (2001) (arguing that XML Schemas are systematic and therefore not copyrightable), *with* Trotter Hardy, *The Copyrightability of New Works of Authorship: "XML Schemas" as an Example*, 38 HOUS. L. REV. 855, 911–18 (2001) (arguing that *Baker* does not apply and copyright protection for XML schemas is possible).

[80] Hardy, *supra* note 79, at 911–13.

[81] Phillips, *supra* note 79, at 67 (citing Brian E. Travis, *XML and SOAP Programming for BizTalk Servers xiii* at 1, 40–42 (2000) (stating that SOAP is the Simple Object Access Protocol that uses a standard set of XML tags)).

[82] Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 348 (1991) (stating that even factual compilations can meet the requisite level of originality if they possess a "minimal degree of creativity").

refuses to register mere forms, titles, column headings, simple check lists, common information (standard calendars, height and weight charts, event schedules) and the like."[83] Additionally, the Code of Federal Regulations in Title 37, § 202.1(a) defines material not subject to copyright and states that copyright does not protect names, titles, slogans, or some types of short phrases.[84] Copyright is not a reward for labor, originality and independent creation are required.[85]

Case law has developed the minimal level of creativity requirement to allow protection to include business telephone listings which, in one case, were arranged in categories most useful to Chinese Americans.[86] Only the most mundane and ordinary phone book categories have been denied protection because alphabetical listing of information without creative section headings fell below the minimal level of creativity.[87] "[A] work may be original even though it closely resembles other works so long as the similarity is fortuitous, not the result of copying."[88] Semantic ontologies require a description of a domain of knowledge which may presumably contain several creative choices.[89] Ontology protection will likely be based on a fact specific inquiry that looks at creative choices made in defining the scope of the domain of knowledge. If the choice of scope is similar to choosing to create a list of businesses that interest Chinese Americans, meaning that it does not simply rely on political or geographic boundaries but instead serves a novel set of data, the ontology is more likely to receive protection.[90] In other words, the focus of copyright protection as it relates to originality will look to whether creativity was used in selecting the scope of the domain of knowledge.

E. Copyrightable Subject Matter—Compilations & Taxonomies

Compilations are copyrightable subject matter under 17 U.S.C. § 103.[91] This protection "extends only to the material contributed by the author of such work, as distinguished from the preexisting material employed in the work."[92] Compilations are defined by the copyright statute as "a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original

---

[83] Elizabeth Carpentier & Henry Parr, *Fixation and Subject Matter*, *in* 20 S.C. JUR. INTELL. PROPERTY § 14 (2010) (quoting a letter from the Copyright Office); *accord. Feist*, 499 U.S. at 348.

[84] 37 C.F.R. § 202.1(a) (2010).

[85] *See* Int'l News Serv. v. Associated Press, 248 U.S. 215, 250 (1918).

[86] *See* Key Publ'ns, Inc. v. Chinatown Today Publ'g Enters., Inc., 945 F.2d 509, 514 (2d Cir. 1991).

[87] *E.g.*, *Feist*, 499 U.S. at 348–49, 362.

[88] *Feist*, 499 U.S. at 345.

[89] *See* Evan D. Brown, *Copyright on the Semantic Web: Divergence of Author and Work*, 19 WIDENER L.J. 829, 830–32, 840 (2010).

[90] *See Key Publ'ns*, 945 F.2d at 513–14.

[91] 17 U.S.C. § 103 (2006).

[92] *Id.*

work of authorship."[93] It is the arrangement or grouping of what might otherwise be uncopyrightable subject matter that becomes copyrightable as a compilation.[94]

Within the broader theory of compilations, there is the more specific category of taxonomies. "A taxonomy is a collection of Controlled Vocabulary terms organized into a hierarchical structure."[95] Several cases have developed the copyright protectability of taxonomies. In a Seventh Circuit case in 1997, a dental association's taxonomy was found to be an original work of authorship entitled to copyright protection, and its taxonomy was not an uncopyrightable "system."[96] The dental association's taxonomy is protected as a relationship between the numbers and the descriptions which means that copying the code is considered infringement, but taking and using parts of the taxonomy is allowed. On the other hand, a Third Circuit case in 2001 held that part numbers did not constitute a copyright protectable taxonomy.[97] Unlike the dental association's taxonomy where "the number 04267 is assigned to the short description 'guided tissue regeneration-nonresorbable barrier, per site,' per tooth (includes membrane removal),"[98] the part number system was highly mechanical and there wasn't any choice about how a fastener should be classified.[99] The degree of creativity and choice seems to be the controlling test for taxonomy copyright protection. In fact, a 2009 case solidifies this understanding of the test. In *Want Ad Digest v. Display Advertising*,[100] a New York District Court held that a classified advertisement publisher's arrangement of subheadings was protectable by copyright because the arrangement of subheadings demonstrated a minimal degree of creativity.[101]

William F. Patry, in his treatise *Patry on Copyright*, mentioned the issue of ontologies used by the semantic web in the context of taxonomies.[102] Interestingly, Patry believes that though taxonomies, compilations, and ontologies appear to be used as interchangeable terms, this is not truly the case.[103] Judge Easterbrook, in *American Dental Ass'n v. Delta Dental Plans Ass'n*,[104] stated that the American Dental Association ("ADA") code "could be a compilation only if its elements existed independently and the ADA merely put them in order."[105] Describing and defining the relationships in a body knowledge is referred to by Judge Easterbrook as a "taxonomy,"[106] but it is more accurately deemed an ontology as the term is defined in computer science.[107] This ambiguity in meaning leaves copyrightable subject matter

---

[93] *Id.* § 101.

[94] Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 499 (1991); David E. Rigney, *What Constitutes a "Compilation" Subject to Copyright Protection*, 88 AM. L. REP. FED. 151, § [2a] (2010).

[95] *Ethnographic Thesaurus*, AM. FOLKLORE SOC'Y, http://et.afsnet.org/glossary.html (last visited Sept. 30, 2010).

[96] Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977, 980–81 (7th Cir. 1997).

[97] Southco, Inc. v. Kanebridge Corp., 258 F.3d 148, 156 (3d Cir. 2001).

[98] *Id.* at 155 n.10 (quoting *Am. Dental*, 126 F.3d at 977).

[99] *Southco*, 258 F.3d at 156.

[100] 653 F. Supp. 2d 171 (N.D.N.Y. 2009).

[101] *Id.* at 179.

[102] 2 PATRY, *supra* note 62, § 4:51.

[103] *See id.*

[104] Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977 (7th Cir. 1997).

[105] *Id.* at 980.

[106] *Id.*

[107] 2 PATRY, *supra* note 62, § 4:51.

for ontologies a confusing matter. Possibly the strongest argument for semantic web ontology copyright protection can be found in *American Dental* when the court states: "Facts do not supply their own principles of organization. Classification is a creative endeavor."[108] Where an ontology is created in a way that is not highly mechanical, and allows for any choice of categories and relationships, it is likely that the ontology will be protectable by copyright.

F. Copyrightable Subject Matter—Government Works

United States Government works, which are works "prepared by an officer or employee of the United States Government as part of that person's official duties," are excluded from copyright protection under 17 U.S.C. § 105.[109] The public policy behind this is that as a country we do not want to restrict use or access to public records.[110] Additionally, the incentive that copyright provides to produce the work is not needed when a government employee is already being paid and directed to produce the work.[111] Much of the science funded by the United States government is carried out by non-governmental entities,[112] but there are several government labs.[113] There is a potential for the government to participate in or separately create semantic ontologies to facilitate data sharing between the many projects that it funds. For these relatively rare situations, it is worth briefly considering the protection of government works.

The case law illustrates situations in which government works will not fall under copyright. *Banks v. Manchester*[114] involved a court reporter's claims to copyright and the court excluded copyright protection for the syllabi, statements of the cases, and opinions because these were created by the government and therefore could not be works of authorship by the reporter.[115] While adopting the government's work did not lead to copyright protection, government adoption of a private expressive work, such as a standard in agency claim filling, will not lead to loss of copyright protection.[116] In *Practice Management Information Corp. v. AMA*,[117] adoption of a private work as a government standard did not render the privately held copyright invalid and the work, as a result, did not enter the public domain.[118] A final nuance in the relationship between government works and private works is

---

[108] *American Dental*, 126 F.3d at 979.

[109] 17 U.S.C. §§ 101, 105 (2006).

[110] *See* H.R. REP. NO. 94-1477, at 56 (1976); *see also* 2 PATRY, *supra* note 62, § 4:58 (discussing origins of demands for public access to printed laws).

[111] *See* H.R. REP. NO. 94-1477, at 56.

[112] *See, e.g.*, *About Funding*, NAT'L SCI. FOUND., http://www.nsf.gov/funding/aboutfunding.jsp (last visited Sept. 30, 2010) (listing types of organizations that compose the approximately 11,000 projects funded by the NSF).

[113] *E.g.*, *About Argonne*, ARGONNE NAT'L LAB., http://www.anl.gov/Administration/index.html (last visited Sept. 30, 2010) (stating that Argonne is "one of the . . . oldest and largest national laboratories").

[114] 128 U.S. 244 (1888).

[115] *Id.* at 247.

[116] H.R. REP. NO. 94-1477, at 60.

[117] 121 F.3d 516 (9th Cir. 1997).

[118] *Id.* at 521.

established in *Veeck v. Southern Building Code Congress International.*[119] In *Veeck*, the court held that a private code enacted as law results in a loss of copyright protection that the code previously enjoyed.[120] The court found that this loss of copyright protection was true primarily under the merger doctrine,[121] but this is also true under the government works exclusion.[122]

A semantic web ontology may be created by an officer or employee of the United States government as part of that person's official duties, in which case the resulting ontology would not be protected by copyright law.[123] The more interesting question is whether a semantic ontology that was privately developed then adopted by the government would be treated more like an adopted standard as in *Practice Management Information Corp.* or more like a private code as in *Veeck*. The standard analogy seems to work best because the ontology is not a set of laws to be followed but rather a way to make all data fit the same form as with the standardization of claims in *Practice Management Information Corp.*, therefore adoption by the government would not result in a loss of copyright protection.[124]

G. Copyrightable Subject Matter—The Merger Doctrine

The merger doctrine is a judicial creation that limits copyright protection of creative expression when there is a limited number of ways in which an idea can be expressed.[125] This can be a powerful doctrine because it removes all copyright protection as it relates to the expression of the idea which, unlike fair use, allows copying in any context as long as the use remains within the scope of the doctrine.[126] In the context of compilations and organizational systems the merger doctrine seems to only apply when ideas and expression are not creative enough to receive protection absent the application of the doctrine.[127] In other words, the merger doctrine is unlikely to be a critical part of the semantic web ontology protection analysis but it is important to provide some background information because the doctrine is often raised in this context.

The merger doctrine often fails to limit copyright. In a 1995 case in Texas, the court rejected the application of the merger doctrine as a reason to deny copyright protection for a compilation of threshold values used to determine when a computer's hard drive was about to fail.[128] The values were discretionary choices, not facts, which lead the court to find copyright protection and sufficient freedom to express the idea such that the merger doctrine need not apply.[129] A Second Circuit opinion in

---

[119] 293 F.3d 791 (5th Cir. 2002).

[120] *Id.* at 800 ("'[T]he law,' whether it has its source in judicial opinions or statutes, ordinances, or regulations, is not subject to federal copyright law.").

[121] *See* discussion *infra* Part III.G.

[122] *See Veeck*, 293 F.3d at 800–01.

[123] *See* 17 U.S.C. § 101 (2006).

[124] *See* Practice Mgmt. Info. Corp. v. AMA, 121 F.3d 516, 521 (9th Cir. 1997).

[125] Mason v. Montgomery Data, Inc., 967 F.2d 135, 138 (5th Cir. 1992).

[126] *Id.*

[127] *See id.* at 140 n.7.

[128] Compaq Computer Corp. v. Procom Tech., Inc., 908 F. Supp. 1409, 1419 (S.D. Tex. 1995).

[129] *Id. But see id.* at 1419 n.13 ("Merger would be implicated if Compaq's threshold values were solely predictions of when the hard drives would fail.").

1994 resulted in a similar conclusion when it held that copyright protection for a compilation of valuation information for used vehicles stored in a computer database was not precluded by the merger doctrine because the expression of the idea was separable from the idea.[130] The court rejected the concept that each entry in the database was the idea of the value of the particular vehicle and that the idea could only be communicated as the related dollar figure associated with it in the database.[131]

The merger doctrine has been recently applied to limit copyright protection but the effect of the ruling is of limited consequence. In a 2005 case, the Sixth Circuit found a lack of copyright protection under the merger doctrine for an automobile transmission parts seller's classification scheme because the structure of the numbering system was the only form of expression possible.[132] It is important to note, however, that the numbering scheme was also not protectable by copyright because the choice of the numbers used did not show any evidence of creativity so it would not have been proper subject matter for copyright under 17 USC § 102(a).[133] This lack of originality which excludes the numbering scheme from proper subject matter is clearly stated by the court when it writes: "The mere fact that numbers are attached to, or are a by-product of categories and descriptions that are copyrightable does not render the numbers themselves copyrightable."[134] In many ways, the merger doctrine is an extraordinary remedy and courts are much more comfortable finding lack of protection on other grounds, such as lack of originality.[135]

In determining the applicability of the merger doctrine to semantic web ontologies it seems unlikely that the doctrine will have any limiting effect. The creation of ontologies often involves discretionary choices and descriptions and both of these attributes help to prevent it from being classified as merely a rigid classification scheme with insufficient alternatives to express the idea.[136] Ontologies are a reflection of the values of the organizer, some things must be included, some things must be left out, and this discretion distances semantic ontologies from the more rigid numbering schemes that fell to the merger doctrine.

H. Fixation

Fixation is the act of putting a work into a tangible form.[137] "'Copies' are material objects, other than phonorecords, in which a work is fixed by any method

---

[130] CCC Info. Servs., Inc. v. Maclean Hunter Mkt. Reports, Inc., 44 F.3d 61, 72–73 (2d Cir. 1994).

[131] *Id.* at 67, 72.

[132] ATC Distrib. Grp., Inc. v. Whatever It Takes Transmissions & Parts, Inc., 402 F.3d 700, 707 (6th Cir. 2005).

[133] *Id.* at 709; *see also* 17 U.S.C. § 102(a) (2006).

[134] *ATC Distrib.*, 402 F.3d at 708–09; *see also* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 345 (1991) ("To qualify for copyright protection, a work must be original to the author. Original, as the term is used in copyright, means only that the work was independently created by the author . . . and that it possesses at least some minimal degree of creativity.") (citation omitted).

[135] *See, e.g.*, *ATC Distrib.*, 402 F.3d at 707 (addressing the question of originality and disposing of portions of the work before moving to merger).

[136] *Id.*

[137] 17 U.S.C. § 101 (2006).

now known or later developed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[138] For example, if a guest lecturer arrived at a law school class without notes and gave a brilliant lecture that contained many new and interesting ideas, the lecture would not be protected by copyright if there was no sound recording, video recording, PowerPoint slides, or notes fixing the lecture in a tangible medium.[139] The policy behind the fixation requirement is that fixation serves as both evidence and notice of the scope of the work and what is being claimed by the author.[140]

With digital copies, there are many complexities introduced that require a somewhat sophisticated understanding of how computers or other digital devices function. In *MAI Systems Corp. v. Peak Computer, Inc.*,[141] an infringing copy was made when software was transferred to a computer's Random Access Memory ("RAM") for use by Peak in diagnosing computer problems.[142] The copy in RAM was considered sufficiently permanent or fixed.[143] Congress responded by exempting the computer services industry,[144] but the technology specific approach still stands and sufficient permanence is still the fixation test.[145]

The application of fixation to a semantic web ontology should be investigated under two scenarios: (1) when the system is at rest and (2) when the system responds to a query. First, when the system is at rest the semantic web ontology is stored in a standard format within a database which is stored in a computer's permanent memory, most likely on a hard drive.[146] Unless the ontology has recently been used by the system or the system is designed to fetch information when it is started, the ontology or parts of the ontology are unlikely to have been loaded into RAM by the computer. This situation is very similar to *MAI Systems'* protectable fixation which protected both RAM and hard drive copies as fixed in a tangible medium of expression.[147] Second, when the system responds to a query there are new relationships created or defined by the query.[148] The question then becomes: are these new relationships protectable and if so, by whom?

There are many different ways to build semantic systems and variations in system design will have important implications as to how an ontology responds to queries including whether those queries bring about new fixed copies. It is possible to design a system where inferred results are cached for complex answers so that intermediate answers are set in temporary tables which are active in RAM and the final answer retrieval is then created by the union of the remaining query results and

---

[138] *Id.*

[139] *See id.*

[140] Douglas Lichtman, *Copyright as a Rule of Evidence*, 52 DUKE L.J. 683, 730 (2003).

[141] 991 F.2d 511 (9th Cir. 1993).

[142] *Id.* at 518–19.

[143] *Id.* at 519.

[144] Digital Millenium Copyright Act, Pub. L. No. 105-304, § 202, 112 Stat. 2860, 2878 (1998) (codified as amended at 17 U.S.C. § 512 (2006)).

[145] 17 U.S.C. § 101 ("A work is 'fixed' in a tangible medium of expression when it[] . . . is sufficiently permanent . . . .").

[146] *See* discussion *supra* Part II.A.3.

[147] *See MAI Systems*, 991 F.2d at 518–19.

[148] *See* discussion *supra* Part II.B.

information pulled from temporary tables.[149] Are there any copyright implications for this intermediate table? Probably not. When a system deduces a specific answer, it does so according to the relationships and hierarchy of the stored ontology.[150] These temporary tables are partial copies, and if *MAI Systems* can be used as an indicator, even these temporary tables are sufficiently fixed under § 102.[151] Temporary intermediate tables created in response to queries in some ways represent the structure of the ontology they are drawn from, but they are limited in duration and only subsets of a much larger contextual set of relationships.

I. Brief Restatement of Copyright Conclusions

It is important to quickly recap what may be protectable before best practices are discussed. Whether or not a semantic web ontology will qualify as copyrightable cannot be predicted with a great degree of certainty given the lack of case law applying to this technology. Despite this lack of certainty, it is important to know, if copyright protection of an ontology is possible, what aspects of the ontology are most likely to be protected. Overall, it appears that semantic ontologies will receive limited (or thin) copyright protection. If copyright is denied, it will likely be under the theories of idea/expression[152] or insufficient creativity similar to a title, heading, or short phrase.[153] Semantic web ontologies otherwise appear to meet the qualifications of compilations and taxonomies,[154] they will not fall to the merger doctrine,[155] and they are sufficiently fixed.[156]

1. **Idea/Expression**: Copyright protection for an original work of authorship does not extend to any idea or system,[157] but the best argument in favor of copyright protection for a semantic web ontology is that the ontology is used within a system but is not itself systematic.[158]

2. **Titles, Headings, Short Phrases**: Copyright protection is contingent upon a showing of at least a minimal level of creativity and ontology protection that will likely be based on a fact specific inquiry that looks at the creativity involved in choosing the scope of the domain of knowledge.[159]

3. **Compilations & Taxonomies**: Where an ontology is created in a way that is not highly mechanical and allows for any choice of categories and

---

[149] *See, e.g.*, Jing Mei et al., *Ontology Query Answering on Databases*, 4273 LECTURE NOTES COMPUTER SCI. 445 (2006) (explaining ontology query answering on databases by means of Datalog programs such as a SPARQL query using an OWL ontology).

[150] *Id.*

[151] *See* 17 U.S.C. § 102 (2006); *MAI Systems*, 991 F.2d at 518–19.

[152] *See* discussion *supra* Part III.C.

[153] *See* discussion *supra* Part III.D.

[154] *See* discussion *supra* Part III.E.

[155] *See* discussion *supra* Part III.G.

[156] *See* discussion *supra* Part III.H.

[157] 17 U.S.C. § 102(b) (2006).

[158] *See* discussion *supra* Part III.C.

[159] *See* discussion *supra* Part III.D.

relationships it is likely that the ontology will be protectable by copyright.[160]

4. Government Works: A semantic ontology privately developed then adopted by the government will likely remain protectable by copyright to the same extent it previously was as is evidence by other standards adoption case law.[161]

5. The Merger Doctrine: It seems unlikely that the doctrine will have any limiting effect on semantic web ontologies because they are a reflection of the values of the organizer and sufficient discretion is exercised in their creation.[162]

6. Fixation: Semantic ontologies will likely be considered sufficiently fixed to qualify as protectable under § 102.[163]

IV. BEST PRACTICES

Scientists need to be aware of the copyrightable aspects of ontologies. If data sharing is more accurately and easily accomplished on semantic web, then copyright protected ontologies, which would require payment for queries, will be a possible barrier to data access. Rights to these ontologies will give the rights holder the ability to control copies.[164] Even if semantic web ontologies are not copyrightable, it may be advisable, as a best practice, to affirmatively disclaim ownership when publishing an ontology that is to be used in conjunction with the publication of public scientific data.

A. The Panton Principles

When public money is used for science there is, at the very least, an obligation of information access that should come with this enablement by the public.[165] This almost self-evident statement has been much more elegantly and specifically stated in the form of the "Panton Principles."[166] These principles state that public science should come with no licenses, no innovation controls, and have the ability to be globally reused.[167] The short form of these principles is as follows:

---

[160] *See* discussion *supra* Part III.E.

[161] *See* discussion *supra* Part III.F.

[162] *See* discussion *supra* Part III.G.

[163] *See* discussion *supra* Part III.H.

[164] *See* 17 U.S.C. § 106 (2006) (granting exclusive rights in a copyrighted work to the holder of its copyright). *But see* 17 U.S.C. §§ 107–22 (providing limitations on the exclusive rights of § 106).

[165] *See* discussion *supra* Part III.F and accompanying notes.

[166] Peter Murray-Rust et al., *Principles for Open Data in Science*, PANTON PRINCIPLES, http://pantonprinciples.org/ (last visited Sept. 30, 2010).

[167] *See id.*

1. "When publishing data make an explicit and robust statement of your wishes."[168]

2. "Use a recognized waiver or license that is appropriate for data."[169]

3. "If you want your data to be effectively used and added to by others it should be open as defined by the Open Knowledge/Data Definition."[170]

4. "Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended and ensures compliance with both the Science Commons Protocol for Implementing Open Access Data and the Open Knowledge/Data Definition."[171]

Of course, these are data specific principles but the concept could easily be opened up to include the ontology that describes the data.

The method of ontology creation is critical for defining how an ontology should best be licensed. An ontology can be created in a distributed or centralized effort. Is one method preferable? Should the ontology be created in a distributed fashion by each public science data contributor and then licensed to the public under a recognized waiver or license that disclaims copyright ownership? Should there be a centralized effort, with the government as the author, which would cause all of the work to enter the public domain and avoid the copyright problem altogether? The Internet has shown us the power of distributed innovation and it would be foolish to turn our backs on this lesson and centralize the development of a web technology.[172] The lack of formality of the Panton Principles has appeal, in that people may more broadly accept the principles, but another more normative version of these ideas can be seen in the Science Commons protocol for open access to data.[173]

B. Science Commons: Protocol for Implementing Open Access Data

The Science Commons' *Protocol for Implementing Open Access Data* directly addresses the need for an "open access" structure for distributing data or databases

---

[168] *Id.*

[169] *Id.*

[170] *Id.*

[171] *Id.* "PDDL" is short for "Public Domain Dedication and License." *Id.* "CCZero" is short for "Creative Commons Zero Waiver," which is Creative Commons' method of releasing material to the public domain. *Id.*

[172] *See generally, e.g.*, *About the Apache HTTP Server Project*, APACHE SOFTWARE FOUND., http://httpd.apache.org/ABOUT_APACHE.html (last visited Sept. 30, 2010) ("The Apache HTTP Server Project is a collaborative software development effort aimed at creating a robust, commercial-grade, featureful, and freely-available source code implementation of an HTTP (Web) server.") Netcraft's September 2010 Web Server Survey lists Apache as the number one server with over fifty-seven percent of the market. *September 2010 Web Server Survey*, NETCRAFT (Sept. 17, 2010), http://news.netcraft.com/archives/2010/09/17/september-2010-web-server-survey.html.

[173] John Wilbanks, *Reaching Agreement on the Public Domain for Science*, SCIENCEBLOGS (Feb. 19, 2010 12:24 PM), http://scienceblogs.com/commonknowledge/2010/02/reaching_agreement_on_the_publ.php.

and it will be submitted to the World Wide Web Consortium for consideration as an Internet standard.[174] With a focus on interoperability of scientific data, this more formal protocol as compared to the Panton Principles is a good fit for adding ontology-specific licensing terms for the sharing of public science.[175] The risk of not taking such an approach is to leave things as they are where "[t]here are too many databases under too many terms already, and it is unlikely that any one license or suite of licenses will have the correct mix of terms to gain critical mass and allow massive-scale machine integration of data."[176]

Keeping things simple but exact is important. The protocol uses a data mark and metadata for use with databases and data.[177] Ideally, the licensing information would be machine-readable to assure automation of the integration of ontologies and to maximize easy access to public scientific data that can be queried. The database protocol should be expanded to apply to both the data and the ontologies. This way, ontologies cannot act as barriers to data sharing in a system that is designed to share data.[178] Arguably, Part 4.1 of the protocol already waives any ontology copyright when the protocol requests that the licensor "waive all rights necessary for data extraction and re-use;"[179] but in the interests of clarity and simplicity, copyright protection in the ontology should be specifically waived.

Open ontology efforts already exist, and the best example may be the Open Biological and Biomedical Ontology ("OBO") Foundry site hosted at the Berkeley Bioinformatics Open Source Project.[180] This open community of ontology authors shares the task of ontology development in its field and adheres to the OBO Foundry Principles as defined on April 24, 2006.[181] Interestingly, despite their sophisticated effort to create controlled vocabularies for shared use and their clearly stated principles,[182] copyright protection or licensing is not explicitly addressed by the project.[183] This is true despite contributions to the National Center for Biomedical Ontology which is one of the National Centers for Biomedical Computing supported by the National Institutes of Health ("NIH"), a government agency.[184] Use of ontologies on the BioPortal site is governed by The National Center for Biomedical Ontology terms of use which states that the ontologies will be "freely available for

---

[174] *Protocol for Implementing Open Access Data*, SCI. COMMONS, http://sciencecommons.org/projects/publishing/open-access-data-protocol/ (last visited Sept. 30, 2010).

[175] *See id.*

[176] *Id.*

[177] *Id.*

[178] *See generally* Thomas R. Gruber, *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, 43 INT'L J. HUM.-COMPUTER STUD. 907 (1993), *available at* http://tomgruber.org/writing/onto-design.pdf (describing the role of ontologies in supporting knowledge sharing activities).

[179] *Protocol for Implementing Open Access Data*, *supra* note 174, § 4.1.

[180] OPEN BIOLOGICAL & BIOMEDICAL ONTOLOGIES, http://www.obofoundry.org/ (last visited Sept. 30, 2010); *see also* Barry Smith et al., *The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration*, 25 NATURE BIOTECH. 1251 (2007).

[181] *OBO Foundry Principles*, OPEN BIOLOGICAL & BIOMEDICAL ONTOLOGIES, http://www.obofoundry.org/crit.shtml (last visited Sept. 30, 2010).

[182] Smith et al., *supra* note 180, at 1251–52.

[183] *See OBO Foundry Principles*, *supra* note 181.

[184] *NCBO BioPortal: Help and About*, NAT'L CENTER FOR BIOMEDICAL ONTOLOGY, http://bioportal.bioontology.org/home/release (last visited Sept. 30, 2010).

public use."[185] It is encouraging to see an open community of ontology creators but it would be even better to see them using a standard legal license as a tool to make clear their implicit intentions about their possible copyright in the ontology. Indeed, given the complexity of copyright protection of ontologies, explicit and clear licensing terms are needed.

C. Embedded Licensing in Interlinked Data Sets

Best practices for publishing and connecting structured data on the Web recommend machine readable licenses.[186] Machine readable licenses are important when dealing with semantic web ontologies and open data because a primary function of structuring data for use with the semantic web is to connect external data sets.[187] These re-combinations of data need to be done in a way that honors the license for both the data and the ontology.[188] Machine readable licenses allow interoperability while respecting the author's intentions.[189] To fully enable public access to public scientific data it is best to automate a licensing preference for data and ontologies where copyright has been waived. This helps to prevent restrictively licensed ontologies from being used to limit access to scientific data that has been released to the public.

A vocabulary and a set of instructions already exist for enabling discovery and usage of linked datasets. One popular option is Vocabulary of Interlinked Datasets ("voiD") which is an RDF based schema to describe linked datasets.[190] This can be used to indicate the OWL ontologies used by a dataset.[191] For example to express the Science Commons ontology and public domain data license as a statement in voiD, the void:vocabulary and dcterms:license properties can be used.[192] For example:

```
:ScienceCommons a void:Dataset;
    void:vocabulary <http://sw.neurocommons.org/2007/kbsources/sciencecommons.owl>
    dcterms:license <http://creativecommons.org/licenses/publicdomain/>
```

The vocabulary statement defines the Science Commons ontology and its location and the dcterms statement defines the license for the data.[193] An important limitation of this method of license declaration in its current form is that

---

[185] *Terms of Use*, NAT'L CENTER FOR BIOMEDICAL ONTOLOGY, http://www.bioontology.org/terms (last visited Sept. 30, 2010).

[186] *See* Glenn Otis Brown, *Creative Commons Unveils Machine-Readable Copyright Licenses* (Dec. 16, 2002), http://creativecommons.org/press-releases/entry/3476; *see also* Christian Bizer et al., *Linked Data—The Story So Far*, 5 INT'L J. ON SEMANTIC WEB & INFO. SYS. 1, pt. 1–2. (2009).

[187] *Bizer*, *supra* note 186, pt. 1–2.

[188] *See id.* pt. 7.

[189] *See id.*

[190] Richard Cyganiak et al., *voiD Guide—Using the Vocabulary of Interlinked Datasets* 2 (Jan. 29, 2009), http://void-impl.googlecode.com/svn/trunk/guide/voiD-guide_v63.pdf ("[V]oiD aims to provide a vocabulary to bridge data publishers and data users, so that users can find the right data for their tasks more easily . . . .").

[191] *Id.* at 9.

[192] *Id.* at 7, 9.

[193] *See id.*

"dcterms:license" refers to "the object container and serializers" which means that the license refers only to the stored data and not to the ontology that describes the data.[194] I propose that the best practice should be to create a new instruction that allows a statement to be made in voiD which defines the ontology license. This will make the licensing terms of the ontology clear and machine readable.

D. Example Semantic Ontology License

RadLex® is a reference ontology for the domain of radiology which is licensed under the RadLex® Ontology License.[195] BioPortal describes the ontology as "a controlled terminology for radiology—a single unified source of radiology terms for radiology practice, education, and research."[196] The RadLex® Ontology License permits public access and: "clinical, research, educational and commercial activities without charge."[197] Interestingly, RadLex® claims copyright ownership in the ontology and grants a copyright license:

> Subject to the terms and conditions of this License, [Radiological Society of North America] and each Contributor hereby grants Licensee a perpetual, worldwide, non-exclusive, no-charge, royalty-free, copyright license to reproduce, publicly display, publicly perform, prepare Modifications, and distribute the Work with or without Modifications, subject to the Limited Use of Identifiers in Section Four (4) of this License.[198]

This permissive license allows many freedoms to both end users or possible system builders and it is a good example of a collaboratively written ontology that has been built by scientist to be used by scientists for the publication of public scientific data.[199] If this license is to be used as a guide for best practices the only consideration for improvement might be a requirement to make sure the license is machine readable. It might also be worthwhile to try to reduce the length and formality of the license to make the license as accessible as possible to users of varied legal sophistication.

---

[194] *Class DCTerms*, ADORE FEDERATION, http://african.lanl.gov/aDORe/projects/DIDLTools/docs/modules/did-adore/javadoc/org/adore/didl/content/DCTerms.html (last visited Sept. 30, 2010).

[195] *RadLex® Ontology License*, RADIOLOGICAL SOC'Y N. AM., 1, http://www.rsna.org/RadLex/upload/radlex_public_license_version_1-0-1.pdf (last visited Sept. 30, 2010).

[196] *RadLex Metadata,* NCBA BIOPORTAL, http://bioportal.bioontology.org/ontologies/21275 (last visited Sept. 30, 2010).

[197] *RadLex® Ontology License*, *supra* note 195.

[198] *Id.*

[199] *Id.* (stating that the ontology has been developed by the Radiological Society of North America, the National Institute of Biomedical Imaging and Bioengineering, and the National Institute of Health's cancer Biomedical Informatics Grid project).

CONCLUSION

The semantic web is going to become an important tool for scientists who need to accurately share data given context through structured relationships. Information use and exchange requires standards and the semantic web is beginning to solidify around technologies which will help to standardize these structured relationships. Ontologies define contextual relationships on the semantic web and it is likely that a semantic web ontology may only be thinly protected by copyright law. Given the likelihood of copyright protection of semantic web ontologies, the best practices for the scientific community should include adopting a machine readable license which disclaims copyright protection for publication of public scientific data and assures automation of the integration of ontologies which will maximize easy access to public science materials that can be queried. Sharing information is essential for the progress of science and failure to address the possibility that ontologies might pose a constraint to public data access could result in data fragmentation and lost scientific opportunities. The ability of the semantic web to annotate and reuse data relies on the social structure of science supporting data sharing as a norm, and as an extension of this norm, open licensing of ontologies should be widely embraced.