Information is Free as in Beer, and We're All Drunk

Benjamin Pollack
bmp5@duke.edu

Andrew Todd
amt13@duke.edu

May 2, 2005

1 Introduction

The Web has grown tremendously over the past decade. Since 1993, when the only website in existence was CERN's phonebook directory [Conc], the Web has become a symbol of the information age. Ten years after its inception, the word "Web" has a new definition in all major dictionaries, all major operating systems come with web browsers, and economic commerce has grown from nothing in 1993 to $12.1 billion at a growth rate of over 40% per year [Kes03]. From banking to email to home photos to weather to news to driving directions, computer users in hundreds of nations can do it on the Web. By nearly any measure, the World Wide Web has proven to be an incredible success.

Yet despite its incredible growth, the Web still leaves much to be desired. Computers have little concept of what the data on the Internet *means*. Although search engines have gone a long way towards finding the information that web users want, they have done little or nothing to present that information in a way meaningful to the viewer. A user looking for a group of statistics on economic growth still must manually sort out that information from the results provided by a search engine such as Google. A person looking for information on weather patterns might get many links to individual cities' weather, but would still have to extract and compile that information himself in order to meaningfully interpret it.

The Semantic Web promises to change all of that, turning the Web into a repository of data to be parsed by aggregators and compilers to produce customized content for each Web surfer. Unfortunately, such a world raises tremendous questions of copyright and fair use. Until now, the Internet in general, and the World Wide Web in particular, have managed to work under the existing copyright framework with no real problems. The Web in its current form resembles existing media closely enough that copyright holders have had to make no real paradigm shift. Although the threshold for copying is very low, the copying that does take place has until now generally involved reuse of large unaltered portions of the original, and thus has clearly been either legal or illegal under existing IP law. In the future, however, the divide between legal and illegal copying will not be so clear. The rapidly encroaching Semantic Web maintains the near-zero cost replication of any publicly posted data on the Internet and adds to it the ability to automatically and transparently create meaningfully unique derivative works. The Web today therefore finds itself at a crossroads.

It can either become a Xanadu-like system with built-in copyright tracking provisions for all data, or it can recognize that widespread content reuse is common and make acceptable changes in copyright law to recognize and accommodate this new reality.

2 Multiple Roads to the Semantic Web

2.1 Where We Are Today

When Tim Berners-Lee invented the World Wide Web in the early 1990s, he intended it to be a "universal space of information." In order to transform the Web from a "glorified television channel" to this vision, he emphasized the intrinsic necessity early on of pervasive metadata. Metadata would provide a means for machines to understand the vast amounts of information being distributed; with understanding would come the ability to collate and reprocess information [BLa]. Since that time, the World Wide Web Consortium (W3C) has worked through their Semantic Web Working Group to make this metadata scheme a reality by developing standards to bring metadata to the Web.

Technically speaking, their Semantic Web Working Group relies on a plethora of different specifications, starting with XML as a foundation and then extending it into less familiar acronyms such as RDF and OWL [W3Cb]. Their theory has been that a universal language for classifying data can be built and put into place much the same way that HTML was, gradually supplanting existing content by running parallel to existing solutions—but creating the Semantic Web specification has taken years more than the first version of HTML.

Nevertheless, portions of the Semantic Web are already coming online. Rich Site Summary (RSS) feeds, for example—a subset of the Resource Description Framework, or RDF [Conb]—contain news stories' headlines and partial content and have gradually been gaining presence as a way for blogs and news agencies to disseminate their stories. Most major web publishing platforms, including Blogger[1], WordPress[2] and Movable Type[3] automatically generate RSS feeds, and many major news sites, including CNN, the New York Times, the BBC, and CBS also provide RSS feeds

[1] `http://www.blogger.com/`
[2] `http://wordpress.org/`
[3] `http://www.movabletype.org/`

of their stories. Entire portals, such as My Yahoo[4], My Netscape[5] and Bloglines[6] have been built around around the concept of "aggregating" these various RSS feeds into a single page custom-tailored to the needs of each individual web user. Users can also use programs, such as Safari[7] or NetNewsWire,[8] to provide a live news-feed on their desktop, outside their web browser.

These aspects of the Semantic Web, while mildly interesting, are hardly revolutionary. Yet as users and developers gradually gain an increasingly better idea of how to leverage technologies such as RSS, the applications become increasingly unique. A large number of services now focus on providing categorization of data in ways that make finding the information and finding similar content significantly easier. del.icio.us,[9] for example, on its surface is nothing more than a bookmark repository, but in actuality is a reasonably advanced Semantic Web application. For each URL that a user bookmarks, they associate an arbitrary number of *tags*—one-word identifiers that explain what topics a particular URL discusses. These tags are dynamically created and shared between all bookmark repositories, enabling users to get a higher-level view of what content may actually be contained within a URL. For example, a URL with the tags "blog," "software," and "publishing" may prove far more informative than simply seeing a link on Google to a product called WordPress. Because the tags are associated with URLs from many users, they are constantly updated re-categorized. Users can even subscribe with an RSS client to URLs with specific groups of tags—for example, one could ask to get an update whenever a user posts a new URL that has the tags "political" and "speech," and view such updates with their existing RSS aggregation client. Such a system goes a long way towards cataloging the web, and allows web browsers to present customized views of related web pages in a semi-intelligent manner right now.

Flickr[10], recently acquired by Yahoo, implements comparable functionality for personal photographs. A user uploads photographs, attaches relevant tags, and can then view groups of photos, both site-wide and personal, based on those tags. Then, as in del.icio.us, users can subscribe with an RSS aggregator to a feed of new images in user-specified categories. Software is already emerg-

[4] `http://my.yahoo.com`
[5] `http://my.netscape.com/index2.psp`
[6] `http://www.bloglines.com`
[7] `http://www.apple.com/macosx/features/safari/`
[8] `http://www.ranchero.com/`
[9] `http://del.icio.us`
[10] `http://www.flickr.com`

ing that leverages both del.icio.us and Flickr to provide fascinating new functionality that is much closer to the Semantic Web in vision than the basic functionality of either site by itself. Jonathan Deutsch, for example, has modified WordPress to automatically link all words in his blog to appropriate tags at del.icio.us[11] so that users can instantly learn more about a topic, while extisp.icio.us provides viewers a symbolic representation of someone's interests based on their del.icio.us tags. Flickr and del.icio.us have even been combined into a format that basically results in a user-specified computer-generated weblog based on recent user tagging actions.[12]

Technorati[13] has leveraged these concepts and combined them with blog aggregation to go one step further. Technorati has specified a method for tagging individual URLs that a weblog links, and allows users to view those links by tag. Because Technorati is itself a blog aggregator, end-users are presented with a comparatively fine-grained view of what types of information different websites have, and, like del.icio.us or Flickr, can subscribe to certain tags or combination of tags so as to be presented with a live feed of Internet updates of material that they may find interesting [Tec].

2.2 The W3C's Way

Yet these sites and programs only fulfill a small portion of the vision for the Semantic Web as laid out by Tim Berners-Lee himself. Even fully combined, they provide little more than basic aggregation of monolithic data; RSS feeds provide the equivalent of individual newspaper columns instead of entire pages, but still cannot break up those columns into the individual news tokens they contain. In the case of the global tagging services provided by del.icio.us and Flickr, users can have customized views of the web, but no processing is performed on the content. Users essentially get a disorganized collection of links to complete pages, whose relevance may not be immediately obvious. In other words, we have much of the same problem as a typical search engine, with only very minor improvements.

In Berner-Lee's vision, the Semantic Web is nothing less than the development of search engines which treat the Web "as though it were one giant database, rather than one giant book" [BLb]. Web pages would completely vanish, and even existing Semantic-web–like structures, including

[11] `http://www.tumultco.com/blog/index.php?p=21`
[12] `http://oddiophile.com/taggregator/index.php?tag=`
[13] `http://www.technorati.com/`

RSS, would disappear. In their place, the web would become nothing but a collection of facts and ways to view these facts. According to the current roadmap, the Semantic Web would have six major layers: an assertion layer; a schema layer; a conversion layer; a logical layer; and a proof layer. The assertion layer would be a collection of facts that, according to a given resource, were veritable and accurate. The schema layer would describe what aspects of facts ("properties") exist for a particular kind of datum. The conversion layer would be charged with combining and reformatting facts based on the rules specified in the schema layer. The logical layer would reason from these layers to attempt to deduce truth from a collection of facts. Finally, the proof validator would ensure that the logical layer's deductions are correct. Users would leverage these layers by making a query—for example, "How can a ship hang in a sky in exactly the way that a brick doesn't?"—and then these layers would generate a coherent answer from all of the data available on the World Wide Web. To generate machine-usable queries and manipulate the results, users would rely on third-party applications that could present coherent and human-readable result summaries [BLb].

Although this design sounds wonderful on paper, there exist two major issues: finding someone willing to build up the data-store, and figuring out how to build the system in the first place. Existing Semantic-Web–like systems, such as RSS, generate their content automatically from more traditional forms, such as news articles, diary entries, or even bookmark collections,[14] yet the Semantic Web by definition would require humans to enter nothing but strings of facts and descriptions of interpretations [BLb]. Who would be willing to do so? In effect, the W3C's vision of the Semantic Web puts the cart before the horse, attempting to rigorously define how an information database could be built without concerning itself about who will build such a database or who would be interested in using it. The W3C's vision for the Semantic Web, by their own admission, ends up far closer to natural language query systems such as OpenCyc than to the existing Web [BLc]. Also problematic, the creative content that has driven the Web thus far cannot belong in this space: novels, diaries, blogs, artwork, and audio recordings that are not collections of facts and postulates do not have a place in this system. The W3C's semantic web cannot do anything for *art*.

The Semantic Web seems to be trying to solve problems far removed from the comparatively simple issue of presenting users with customized versions of the data they want. Instead, it tries

[14] `http://del.icio.us`

to be a utopian dream, a repository for a kind of global knowledge network. Yet despite all of its flaws, the Semantic Web defined by the W3C unquestionably does solve the problem of how users could be presented with just the information the want. Some aspects of the W3C's vision likely will eventually see widespread adoption, but fortunately, their vision is not the only answer. Computers may instead help us deconstruct the data on the Web themselves.

2.3 Google's Way

Google was originally founded in 1996 by two graduate students at Stanford who saw it as a way to test their theory that a search engine that ranked results' relevance using an analysis of page interrelationships would work much better than the simple keyword searches then in use [Wik]. Although the founders still lead the company, Google's stated objective is now much wider. Their goal is no less than the ability "to organize the world's information and make it universally accessible and useful" [Goob]. Google is moving towards this objective by employing a number of methods that work *with* existing data, not against it.

The first and perhaps most innocuous method Google employs is simply to buy content and display it on the Web for free in a human-readable form, as they have done for Google Maps/Satellite.[15] In this way, they are participating in the first revolution of Web design that Tim Berners-Lee described several years ago: the idea that existing databases, with proprietary data, could be adapted to have dynamically generated Web interfaces [BLb]. In these cases, though the data internally is richly described, Google does not in any real way help to make this proprietary data more easily and universally machine-readable by third-party applications.[16]

Google itself has attempted to provide an official channel for its basic text-search functions through the Google API, which provides a standard method, through the SOAP protocol, to query and access search results through an XML format. It has been in beta testing for several years now, with developers able to perform 1,000 free queries a day on their license key. No subscription service to obtain a higher query volume has been announced, and the API service has not been extended to other places where it could make data "accessible and useful," like Google Maps [Good].

[15] `http://maps.google.com`

[16] This has, however, certainly happened on an ad-hoc basis as programs such Gmail Drive (`http://www.viksoe.dk/code/gmail.htm`) have simply pretended to be web surfers, relying on Google's consistent presentation, extracted the information manually.

On the other hand, Google itself has also begun to harvest data from others' websites using what are effectively advanced screen-scrapers. Google News is perhaps the most prominent example of this, and also the one that has caused the most controversy. Unlike a typical news page like `http://www.cnn.com` or `http://news.bbc.co.uk`, Google News uses automated artificial intelligence algorithms to determine what the most important, current, and relevant news stories are and displays them in a customized order to the end-user [Gooc]. None of the content displayed on Google News' homepage was originally written by Google; instead, Google's servers harvest news stories from other websites automatically. The headlines, lead paragraphs, and thumbnails of accompanying images are displayed, along with a link to the original article. More often than not, Google's servers do this story extraction without the express permission of the original content producer. This lack of permission became apparent when Agence France Presse, one of the largest news story and photo producers in the world, sued Google over the use of their stories and images in Google News.

According to AFP's complaint, filed on March 17, 2005, Google was reposting headlines, story leads, and thumbnails of images—"the 'heart' of AFP's stories"—on their website [AFP]. Fundamentally, the problem arises from AFP's business model. AFP licenses the content they produce to a number of outlets such as newspapers and news websites, which then provide the content through their websites to what they assume are human readers.[17] AFP charged that Google's aggregator effectively published stories in the same manner as their normal customers, but without paying royalties. Although Google has since taken steps to ensure no AFP stories appear on Google News, none dispute that AFP stories used to appear [Per]. No further action has yet been taken, but should the case go to court, it will provide an important precedent for copyright policy on the Semantic Web.

Interestingly, no other major news organization, such as the Associated Press, has taken similar action, despite an equal or even larger number of stories syndicated through Google News. It may be impossible to tell whether this is because they do not feel they have a legal basis, or simply because they feel that Google's services provides a net economic benefit to them regardless of possible copyright infringement. Nearly one hundred years ago, however, the AP was involved in

[17] Yahoo's AFP top stories page (`http://news.yahoo.com/news?tmpl=index&cid=1504`) provides an example of how their content is usually licensed.

a court case, *International News Service v. Associated Press*, which brought up some of the same issues presented by the AFP case today. The AP brought suit against the International News Service, alleging that the INS had been copying and rewriting news stories written by the AP and passing them off as their own, thereby bypassing the tremendous network of reporters that the AP had to maintain to continue to report on international news. The INS was able to receive these news stories in a timely matter both by bribing news workers at AP-affiliated newspapers in order to gain access directly to the AP wire and also simply by copying the stories from early-edition newspapers and public news-post bulletin boards.

International News Service argued that the Associated Press had given up the rights to the news they had gathered as soon as they "post[ed] it upon bulletin boards so that all may read... that when it thus reaches the light of day it becomes the common possession of all to whom it is accessible; and that any purchaser of a newspaper has the right to communicate the intelligence which it contains to anybody and for any purpose, even for the purpose of selling it for profit to newspapers published for profit in competition with complainant's members" [INS]. The court rejected this argument, saying that *although news did not itself enjoy copyright protection*, it did certainly represent the end product of a large amount of effort on the part of editors, reporters, and transmitters, and therefore was a commercial product that could be sold and owned like any other.

The argument presented by International News Service very clearly echoes the argument that would probably be presented by Google should the Agence France Presse case go to court, yet there are some important differences as well. First of all, all of Google's news aggregation is currently done by computer with no direct human input to rewrite or recombine the stories. Second, Google does not in any way claim to have written the stories and headlines they create themselves; Google clearly labels and links to the stories' original sources. Finally, although this has no bearing on a copyright-infringement case, Google does not currently make money on their news service; unlike most other areas of their website, there are no advertisements, and Google does not offer premium or syndication services.

Unfortunately, other current legal precedents also seem to come strongly against Google. In CCC Information Services v. Macclean Hunter Market Reports, CCC attempted to publish en-

hanced and extended version of information originally obtained from APSA systems. The court held that such an action was not fair use and sided with Macclean in finding CCC's actions illegal [CCC]. The Supreme Court denied petition for certiorari.[18] Similarly, the Supreme Court ruled in Harper and Row vs. Nation Enterprises that publishing content that would normally qualify as fair use may not qualify if such an act deliberately detracts from another publication, causing direct financial damage [Har]. Such a precedent seems to imply that, were a case similar to AFP to occur in the United States and the courts were to agree that Google excerpts the "heart" of news stories, their aggregation service would be ruled illegal.

2.4 Economic Implications of a Semantic Web

Assuming that all of the legal and technological hurdles of a Semantic Web could be surmounted, a fundamental problem remains: who would be willing to pay for the new system? Both Google's vision and the W3C's vision deny content producers a realistic way of recuperating their expenditures. Currently, a large portion of content is funded by advertising. Advertisements on the web generated $9.6 billion in revenue in 2004 and are expected to hit $12.7 billion in 2005 [Ger05]. In the first quarter of 2005 alone, Yahoo made $1.02 billion of $1.14 billion in from their marketing services division [Yah]. In a world powered by the Semantic Web, this industry, which provides the backbone for companies such as Google [Gooa], Yahoo, and others, would vanish.

The incentive to publish information is further reduced by the fact that the W3C's Semantic Web in particular is nothing but a compendium of facts. As recently as 1991, the Supreme Court ruled clearly that a compendium of facts cannot be copyrighted [Fei]. This means that even subscription-based models such as those used by Slate Magazine or the Wall Street Journal would have no hope of surviving. To some extent, those who make the information aggregators might be willing to defer part of the cost. A number of existing news organizations have begun assembling their own RSS browsers, for example, in order to raise money and encourage users to use their stories over the competition [Gar05]. One might hypothesize that those who produced the aggregation clients would begin funding content generation. Yet these products, though able to collate from a variety of sources, have as one of their primary features favoring the content of their creators over other

[18] 516 U.S. 817

content [Gar05]—a feature that would not make sense in a purely factual Semantic Web. Such a system seems far more likely in the long run to result in separate databases with proprietary content than to work as a sustainable model for economic activity.

The solution to these problems requires a radical rethinking of copyright law, its implementation, or both.

3 Xanadu: Where Intellectual Property Acts Like Real Property

Over twenty years ago, some already recognized that a global information network had much to offer to everyday people if they were allowed to reuse and modify content, yet those same visionaries also recognized that such a system would not be possible within our existing copyright model. Ted Nelson's Xanadu Project stands out as one of the most complete proposals for how to properly deal with this problem. Ted Nelson's life vision was for Xanadu to implement a system that strongly supported reuse, yet would not do so at the expense of content authors. His book *Literary Machines* describes the major components of his vision. In Nelson's own words, Xanadu was designed "to re-kindle the freedoms of yesterday and extend them into the electronic future of tomorrow, a computer program intended to tie everything together and make it all available to everyone" [Nel92].

Abstractly, Xanadu at first appears to be nothing more than a document database. Users pay to use the Xanadu system and pay to view each document within the system. All content published within a Xanadu system would be published permanently; once a user had published a document, there would be no way to remove it. Moreover, documents would be versioned. Changes to a document could be tracked across time, and even when a document was modified, old versions could still be recalled [Nel92].

The full description of Xanadu, however, goes far beyond any previous or currently existing document publishing and archival system. Because so many documents were available, Nelson devised various systems of cross-linking content to ease navigation and searching. Users would then be able to jump quickly between related documents. Nelson described documents in this system as "hypermedia," and described the links between documents as "hyperlinks"—creating both terms and introducing them into popular usage [Wol95]. The World Wide Web has partially realized this vision; its subset of Xanadu's hyperlinks have revolutionized online media. Yet Xanadu's true

impact on copyright stems both from its permanently stored documents and a feature it has that the World Wide Web lacks: *transclusion*.

Transclusion describes a mechanism by which content may be excerpted, repackaged, and displayed, automatically resolving royalties on-the-fly and compensating the authors of the original content appropriately. Transclusion, described in chapter 2.6, works by allowing authors of new content to directly quote a specific part of a specific version of existing content in their new document. The Xanadu system tracks exactly what content an author has transcluded in this manner, storing what fraction of the work is original and what fraction belongs to other authors. When a user calls up a document with transcluded content, the price for viewing that document is split proportionately between any author whose content appears in any way in the document [Nel92]. For example, a news story that was 1000 words long and include a 50 word quote from MSNBC, a 50 word quote from the BBC, a photo from CNN, and 200 word paraphrase of an idea taken from the New York Times might receive 600 "credits" with the remaining 400 credits split between the other sources. In this way, the problem of reusing and redistributing someone else's work would be resolved automatically: redistribution could be actively encouraged because the computer network would ensure that no one's content would be misappropriated, stolen, or plagiarized.

Nelson repeatedly failed to create a working version of Xanadu, and the spectacular nature of the failures he has left in his wake have caused some to call a Xanadu-like system *de facto* impossible [Wol95]. Yet despite Nelson's failure to realize his vision, others have implemented the heart of Xanadu. Jason Rohrer, a faculty member of the computer science department at the University of California, implemented a functional tranclusion and micropayment system called token_word in ten days, building it on top of existing web technologies such as PayPal, Perl, and HTML [Roh02]. In token_word, users begin with 50,000 "tokens," and are able to add new tokens through PayPal. When users go to view a document, they are charged one token for each word appearing in the document. The tokens are split proportionally based upon all transcluded material. If a user wishes to transclude a portion of an existing document in a new document that he himself is publishing, he need merely go to that document, select the relevant region, put it on a clipboard to get a transclusion ID, and then paste that ID into a new document that he creates. The system automatically takes care of the rest [Roh02].

token_word proves that the Xanadu vision can be implemented quickly with current technology, and the fact that it relies on existing web technologies (HTML, Perl, and HTTP) to deliver its content might make it seem like the perfect solution. Unfortunately, although token_word serves as an excellent technology demonstration, such a system cannot be implemented within the existing Internet infrastructure.

4 Why Xanadu Won't Work on the Internet

4.1 Inability to Secure Data

token_word is not identical to Xanadu for a number of reasons. token_word does not and cannot prevent users from simply copying and pasting text, thereby avoiding the transclusion system. When users can avoid splitting royalties for their work, they likely would, collapsing the system. token_word is not *universal*, nor can it be within the existing Internet framework, meaning that there would be multiple token_word websites rather than one single repository of documents and that transclusion of existing Internet material would be impossible. To enforce a universal deployment would by definition require a total redesign of the World Wide Web and its underlying protocol to support not only the security necessary to enable transclusion, but to enforce it—essentially a digital rights management (DRM) system for the Web. Such an action would break all existing applications and likely meet stiff consumer and industrial resistance. Further, even this would not truly be sufficient to realize Xanadu. To guarantee the exact vision specific by Xanadu, including guaranteed availability of all versions of a document, would require a tightly managed centralized data-store so large that many experts believe running it efficiently would be impossible [Wol95]. Yet the persistence that such a store would provide, transclusion falls apart.

Even if such a system were devised, users *still* would not necessarily be prevented from copying and pasting text to circumvent the transclusion system, or even from pasting entire reams of text extracted from the Xanadu system. The incredibly low bandwidth required to host copied text would likely make such piracy far more common than existing music and movie piracy. DRM solutions likely would prove ineffective, as most existing major DRM systems have been broken.

JHymn[19], for example, breaks iTunes' encryption scheme in an easy-to-use Java Swing application; DeCSS-inspired media players, including VLC[20] and MPlayer[21] can play encrypted DVDs; and although it is no longer available, Elcomsoft briefly had a product available that circumvented Adobe's PDF duplication security systems. We may therefore hypothesize that a DRM system that attempted to secure the Internet at large implemented in the same general manner as any of these existing systems would likely be broken, returning us basically to where we are now. Any Xanadu implemented this way would not solve the problem; it would simply add more components that could go wrong. In order for a properly functioning Xanadu system to exist, it would have to employ a full control structure to ensure not only that data within the system only supported duplication through the official transcluding mechanism, but also to prevent any data from leaving that system. The only solution to this problem is secure hardware.

4.2 Xanadu's Reliance on "Secure" Hardware

The basic feature enabling these DRM solutions to be broken is that their all-software nature makes them easy to observe and modify. One of the solutions that has been proposed to prevent this form occurring is to rely on trusted computing. Trusted computing is a complex term that has a variety of meanings. For the purposes of this discussion, we will rely on Microsoft's definition, which is also the definition used by the Electronic Frontiers Foundation, a group dedicated to protecting computing rights. By that definition, trusted computing refers to computers that support "memory curtaining," secured input and output, sealed storage and "remote attestation" [Fou]. Of these four technologies, three are of interest to our discussion of Xanadu: memory curtaining, which refers to preventing programs from reading other programs' memory at the hardware level if they lack "proper" authorization from a designated authority; secured input and output, which refers to ensuring that only "authorized" devices are allowed to send and receive data from the CPU; and remote attestation, which refers to the ability to detect changes to software and to react appropriately [Fou].

These last three technologies offer the potential to make computing more secure in general, but

[19] `http://jhymn.sourceforge.net`
[20] `http://www.videolan.org/vlc/`
[21] `http://www.mplayerhq.hu/homepage/design7/news.html`

they also have the potential to severely restrict legitimate computer usage. The EFF notes that such a system has the ability to restrict your ability to copy your own data around even to the limited degree that would normally qualify as fair use. If a content producer wishes to ensure that their you only be allowed to play a given song on a specific computer cannot be broken, they can use remote attestation in particular to ensure that you have not tampered with the software or replaced their software with one that will decrypt the music [Fou]. Admittedly, these problems have little to do with Xanadu *per se*. Yet because Xanadu could only ensure that everything worked within its transclusion framework if it had such an ironclad DRM system, all of the security issues and personal rights issues of trusted computing foist themselves onto a Xanadu.

4.3 Economic Problems

The economic hurdles presented by Xanadu also make adoption of such a system highly unlikely. For example, both Xanadu as described by [Nel92] and token_word rely heavily upon micropayments to make themselves work. Micropayments, unfortunately, have repeatedly failed to succeed. Although subscription services have been tremendously successful for organizations such as the Wall Street Journal [WSJ05] and Salon [Car05], micropayments have not met with their predicted success. The W3C even decided to abandon their micropayment activity and their micropayment working group, despite having issued a public working draft [W3Ca], due to the profound lack of interest and hostility of consumers [Cona]. Xanadu's transclusion compensation model unfortunately ties it to such a micropayment system; a monthly subscription is impossible if the number of articles that a user will view is unknown and everyone must be compensated. The chance of Xanadu magically succeeding with micropayments where others have failed seems dim.

Not all data belongs in such a system, either. The Web's e-commerce sites, for example, which are growing on the web at 40-50% per year [Kes03], have no place in such a system, since Xanadu has no mechanism to provide for interactivity. (And, indeed, if it did, since all documents in Xanadu must be held indefinitely for transclusion to function, vendors would have to determine how to handle out-of-date catalogs and order-forms that would nevertheless remain available.) A movement away from the current Web to a more Xanadu-like system has the potential to hurt the e-commerce industry severely. Likewise, private corporate applications on intranets and the Internet

would have to remain on the World Wide Web, even though many features, such as event calendars, meeting agendas, and more would benefit from the technologies promised by the Semantic Web. Given that these applications include such nontrivial functionality as banking and email, their exclusion cannot be ignored. Even if Xanadu could become widespread, then, a parallel system would have to exist anyway that accomplished exactly the same thing, but without micropayments and transclusion—in other words, the Web in its present form.

5 Creative Commons: A Copyright Revolution

As has been demonstrated, Xanadu, the ultimate system of DRM, cannot work in the real world and is incompatible with the current end-to-end structure of the Internet. However, as the continuing success of media sales—whether through iTunes, physical CDs, books, software, or DVDs—have demonstrated, most people are willing to respect some aspects of intellectual property. Lawrence Lessig of the Stanford Law School in fact went a step farther, complaining during a lecture at Duke Law School that "we live in an era when the idea of property is... such a non-thought; when the importance and value of property is taken for granted... when to question the universality and inevitability of complete propertization is to mark yourself as an outsider" [Les02].

The question, then, and the subject of Lessig's talk, is how to integrate the two systems that currently coexist on the Internet, namely, the intellectual-property concept of complete information control and the Internet's design philosophy as a place for the neutral exchange of data. As has been described by various commentators, including [Tod] and [Les02], those who control the access hardware for the Internet can control the way that content moves across it. This makes it quite possible that, even if Xanadu cannot work, Internet service providers may attempt to limit access patterns to try to protect their property rights. Lessig gives the example of early residential broadband providers, who were also cable companies, limiting their users to ten minutes of streaming video at a time because they believed it was a threat to their business model. That business model had three levels of control. The first, physical, deals with control over the hardware of content delivery. In the second and third, logical and content, the cable company could control both what is shown and who owns what is shown. This contrasts greatly to the Internet, where such companies only have direct control over the physical layer [Les02].

The ramification here is that the concept of universal property holds dangers for the future of the Web and of the Internet itself. If everything written and published on the Web or Semantic Web by default continues to be automatically protected by the full rights and privileges of copyright, content producers will continue to attempt to protect their property using whatever level of means are available to them, whether physical or logical. Cable companies may be the only group able to restrict their users' view of the Internet, but others have tried to protect through various means of obfuscation. Examples include simply displaying text as an image, making it much harder to copy in a useful manner, abusing the Flash plugin as a design element in webpages, and forcing users to download proprietary browser plugins that take advantage of system features in the operating system to prevent the downloading or screen-capturing of images. None of these are seen very often on the open Web, mainly because most of them cannot work very well unless they can collaborate with the operating system kernel or specific Web browsers.[22]

These problems stem from the conception of intellectual property as exactly that — property. Attempts to control data in these ways, without a complete restructuring of computing as we know it, will always fail in some way. The solution lies in Creative Commons, which seeks to replace much of the control associated with property in this system with collaboration.

5.1 How It Works

Creative Commons was founded by many of the players in the failed challenge to copyright extension laws *Eldred v. Ashcroft*, including Lawrence Lessig and Eric Eldred himself [CCF]. It seeks to lay out clear, definitive licenses which compromise between the public domain and complete copyright control and which are suitable for use with creative works, including text, audio, video, and images. Unlike most open-source or "free" licenses to date, these licenses not intended for use with software.

The requirements placed on producers and consumers of Creative Commons-licensed content are slightly different than with most software licenses. There are four identifying aspects which constitute the whole range of these "some rights reserved" agreements, creating, in total, six different licenses which content producers can apply. The first aspect is attribution, which has become

[22] `http://www.artistscope.com/secure_image3/`
`http://www.invisitec.com/`
`http://www.antssoft.com/htmlprotector/`

mandatory for all Creative Commons licenses as of version 2.0. Anyone who uses an author's work, in whole or in part, derivative or verbatim, must give proper credit in their own final product [CCT]. There is also a version being developed for wikis, in which there may be too many authors to list. In the wiki version, all contributing authors to the collaborative group project give up their individual rights to credit, knowing instead that anyone who uses part or all of the wiki they contributed to will have to attribute the work to the group [CCW].

An attribution-only license is only one step above the public domain; it offers authors and content producers no control over how their work is used. The next step away from public domain, if desired, can be in one of two directions: ShareAlike or No Derivative Works. ShareAlike acts much like the more "viral" open source software licenses, like the GPL. Effectively, it requires that all derivatives that might be created using the originally licensed work also be released under the same conditions. The No Derivative Works proviso, on the other hand, requires that work be only printed verbatim, and, of course, under the same license. Since ShareAlike only applies to derivatives, the two are incompatible [CCT].

Finally, and perhaps most controversially, it is possible to add a "non-commercial" provision to CC-licensed work. In other words, people can use, redistribute, and, unless No Derivative Works has been selected, make new things using an author's work, but they cannot make money off of any of these activities. In order to sell any of these works, they must contact the original author and get permission; the original author presumably wants some form of royalties. Since Creative Commons has said that they will not help track intellectual property and collect royalties [CCF], this part is left up to the involved parties.

The non-commercial provision has not been the only part of Creative Commons to draw criticism. There have been many objections to these licenses from the more traditional, software-oriented open source community. For instance, the debian-legal mailing list, which is dedicated to analyzing intellectual-property licensing and other legal issues that may affect the Debian Linux community, has published an analysis of the Creative Commons licenses [Pro]. Interestingly, they conclude that they do not meet the definition of an "open-source" license as defined under Bruce Perens' *Debian Free Software Guidelines*,[23] which are also now better known as the *Open Source*

[23] `http://www.debian.org/social_contract`

Definition.[24] This means that CC-licensed content cannot be included in a Debian distribution, which could, in the future, potentially restrict both projects from a wealth of useful information. Specifically, they point to several general areas of Creative Commons which they believe are contradictory to the aims of Free Software.

The first contradiction is fairly obvious: any No Derivative Works license violates the freely modifiable spirit of open source [Pro]. Second, debian-legal also found that certain clauses in the "attribution" section, which is central to all Creative Commons licenses, could allow an author to effectively curtail derivative use of his work by forbidding the modifier from making direct or indirect reference to him [Pro]. In a similar vein, the attribution segment of the contract could also, if strictly interpreted, require attribution on equal footing to all contributors. For instance, "if Alice writes her autobiography, and includes lyrics from Bob's song in one chapter, she must give him credit for the entire work: "The Autobiography of Alice, by Alice and Bob", or even "The Autobiography of Alice and Bob" [Pro]. Finally, the non-commercial use variant violates two parts of the Debian Free Software Guidelines: there can be "no discrimination against fields of endeavor" and all licensees be allowed to sell copies of works [Pro].

This is not a surprising conclusion given the historic philosophy of the Free Software community. However, it is also quite possible to argue that, unlike software, audio, video, text, and images do not contain the functional component software does, and therefore should benefit from different types of copyright protection. One of the main arguments advanced for the cause of open-source software by several people, including most famously Eric Raymond [Ray], is that it is still possible to make large amounts of money and continue to fund the information-technology world both by building custom solutions on top of open-source software and by providing technical support for open-source software. Since the authors of this free software would also be the ones most knowledgeable about its inner workings, they stand to gain tremendously. In other words, open-source software moves the computer world from a property-based economic model to a service-based economic model.

This argument does not readily translate into the creative realm. Once a master painter creates a work of art, he can certainly, if it is a work in demand, recoup his expenses by selling prints of his work. Since he owns the copyright on the image, he can exclusively set the terms on which it

[24] `http://www.opensource.org/docs/definition_plain.php`

is redistributed. He cannot, however, sell service contracts on his painting. Giving away the image of his painting under the GNU GPL or, somewhat more appropriately, GNU Free Documentation License, would effectively be the same as releasing it into the public domain; there would be little to no way for him to earn back his money to continue creating work without reverting to an archaic system of elite patronage. Even Lessig has publicly stated that he does not want to take away the ability of creative people—musicians, artists, authors—to make a living at what they do [Les02].

The compromise solution is to try to introduce, as much as possible, a commons space where the kind of collaboration the Internet fosters can flourish, but the individual contributions that make up these great collaborative works are not lost. Since there is no way to build a perfect copyright-property control system like Xanadu that people will voluntarily use, Creative Commons seeks to do this through clear delineation of licensing terms, in both legal contracts and plain English, that do not place undue burdens on people who seek to engage in creative activity. The legalese behind CC might not yet be perfect, but the spirit of the movement is certainly moving in the right direction. Effectively, Creative Commons seeks to eliminate the producer-consumer model of creative production and replace it with the democratization of content already seen on the Web — blogs, Flash animations, home movies, garage bands, wikis, and everything else. This new licensing scheme, with its ability to make the world's intellectual wealth into a tremendous commons space, has obvious applications to the Semantic Web, which seeks to make the world's information into a single database.

5.2 The Semantic Web and the Commons

Let us return to the ongoing case between Agence France Presse and Google. Imagine, as ludicrous as it sounds today, that AFP had published all of their content under a Creative Commons Attribution–ShareAlike–Noncommercial license. This would clearly mean that Google could use their news stories on their news aggregator in any way they wished, so long as they did not have advertising or subscriptions to the page. AFP would receive more hits to their members' webpages, and Google News users could continue to enjoy the benefits of algorithm-driven news aggregation.

Google, however, is a revenue-driven company, and it only makes sense that at some point they would like to make money off of their aggregation service. To do so, however, they would

have to contact AFP and arrange some sort of new licensing terms with them for commercial use, presumably paying them some sort of royalty much as current AFP for-profit organizations, like traditional newspapers and Yahoo! News do. This is where the ability of the true, W3C Semantic Web, in conjunction with Creative Commons, will really have the greatest potential to come together and shine.

When a Creative Commons license notice is placed on a webpage, there are several components visible to the Web browser and the person sitting at the computer. First, there is the Creative Commons "Some Rights Reserved" image, and a clickable link leading to the page describing the specific requirements of the license the original content producer has chosen for their work. Second, and much more interesting, there is an RDF descriptor of the license [CCF]. This places the requirements into an entirely machine-readable form, and makes them into simply another piece of data in the Semantic Web.

There is one piece missing that Creative Commons has not addressed. The foundation has specifically said that they are not in the business of helping content producers collect royalties [CCF]. Without this ability, the noncommercial clause of their license has no teeth. Leaving the collection of royalties and money for the exchange of content to the old-fashioned method of humans contacting each other and agreeing on terms does not make sense in the coming era of machine-readable metadata, and the clear solution to this is the creation of another RDF specification—one which can specify fees for commercial re-use.

In this scenario, Google News would make money by selling advertisements, or perhaps subscriptions, to their webpage, which acts as a content aggregator from multiple sources. These sources are supplied in the eventual RDF-based format of the Semantic Web, and employ various Creative Commons licenses. When it encounters a non-commercial restriction, it checks the royalty requirements and makes a payment based on use to the location specified in the Semantic Web document. Naturally, Google could also code-in spending restrictions to its robot, keeping it from being tricked into transferring large sums of money to less-than-scrupulous news providers. Critically, the onus would not on users making micropayments, but on the content producers to pay royalties. Such a system has a much higher chance of working, since it far more closely parallels existing models of economic compensation.

Fundamentally speaking, there seems to be little impetus here to honor the Creative Commons licenses, given the lack of hard-coded digital rights management in the system. However, at least in the case of relationships between content producers and content aggregators, it will not be hard for content producers to detect, through their web access logs, who exactly is retrieving information from their site. Moreover, even if a service provider should choose to strip out all CC licensing information, not pay royalties, and display this content as their own, it will be just as easy as it is today for the original content producer to find out: usually a Google search will do the trick. The difference will be that there is that a Creative Commons copyright holder will likely have a much more clear-cut court case.

5.3 Creative Commons and the Rest of the Internet

Content aggregators are, of course, only one part of the Internet. There are already websites in place other than Google News that are attempting to take advantage of the new methods allowed by Creative Commons both to profit and create new collaborative works. The BBC, reasoning that British taxpayers have already paid for the work of their cinematographers once, is placing large archives of their material online under a license similar to CC-Attribution-ShareAlike [Mer]. Lawrence Lessig, attempting to practice what he preaches, is revising his book *Code* using a CC-licensed wiki.[25] Former rap stars the Beastie Boys, along with David Byrne and several other artists, have released music under a CC license designed specifically to encourage sharing and sampling [CCB].

Perhaps more notable than these individual, scattered efforts is Bitzi, which purports to be a universal media catalog and a "metadata publishing company" [Bit]. Bitzi attempts to catalog the vast variety of files that cross peer-to-peer networks by distributing a cataloging utility to users. This cataloging utility computes a unique hash identifier for the file and checks to see if it has been entered into Bitzi's tracking database yet. It then allows the end-user either to create a new entry or to add on to previous information submitted by other users about the file. Eventually, at least in theory, it becomes users can authoritatively know the source, quality, and licensing information on a media file before downloading it. This effectively extends the Creative Commons, through

[25] `http://codebook.jot.com/WikiHome`

the Semantic Web and RDF, to P2P. If this technology becomes widespread in P2P clients, it will make users more aware, and perhaps more likely to respect, the intellectual commons. Bitzi has already begun a program to allow copyright-holders to identify their files in the database in a way such that the metadata they supply is considered authoritative.

Finally, Lulu.com is a publishing company that, unlike most mainstream groups, has embraced the Creative Commons. Under the Lulu system, authors retain complete control of their work; as such, they can license it however they choose. The publisher becomes only another means of media transmission. Moreover, Lulu has built an ebook distribution mechanism which allows open, unlocked PDFs to be sold or even downloaded for free as a promotional tool for books [Lul].

All of these efforts demonstrate that the transition from property to commons is already occurring; it remains to see how far it will go. The difficulty of letting go of the conception of intellectual property as absolute, as Lessig pointed out, will be tremendous, especially for the organizations that have built their businesses around it [Les02]. Even so, if there is enough pressure applied from the new businesses and collaborations of the Internet, it seems likely that a middle ground will, in time, be found.

6 Conclusion

Widespread reuse of content at this point seems inevitable. The increasing ability of computers to intelligently redact content customized for individual users, combined with the ease of acquiring unprotected data on the existing Internet, will cause the problem to grow to the breaking point. The only solution to this problem will be to recognize the reality that content will be reused and begin to make allowances for it. The Creative Commons' flexible licensing provides a mechanism for explicitly enumerating in what ways content may be reused, thereby at once encouraging certain forms of reuse and prohibiting others. Through this compromise, users' needs might be realistically fulfilled without hurting content producers, stifling innovation, or overthrowing the existing copyright system. Although hardly perfect, Creative Commons provides the model today for how reuse can be legal, commonplace, and profitable tomorrow.

References

[AFP] Complaint for preliminary and permanent injunction and copyright infringement. http://www.resourceshelf.com/legaldocs/afpvgoogle1.pdf. Accessed on April 25, 2005.

[Bit] About bitzi. http://bitzi.com/about/. Accessed on May 2, 2005.

[BLa] Tim Berners-Lee. Realising the full potential of the web. http://www.w3.org/1998/02/Potential.html. Accessed on April 22, 2005.

[BLb] Tim Berners-Lee. Semantic web roadmap. http://www.w3.org/DesignIssues/Semantic. Accessed on April 22, 2005.

[BLc] Tim Berners-Lee. What a semantic can represent. http://www.w3.org/DesignIssues/RDFnot.html. Accessed April 30, 2005.

[Car05] David Carr. The founder of salon is passing the mouse. *New York Times*, February 10 2005.

[CCB] The wired cd. http://creativecommons.org/wired/. Accessed on May 2, 2005.

[CCC] CCC Information Services v. Macclean Hunter Market Reports, 44 F.3d 61; 1994 U.S. App. LEXIS 34212; 33 U.S.P.Q.2D (BNA) 1183; Copy. L. Rep. (CCH) P27,328.

[CCF] Frequently asked questions — creative commons. http://creativecommons.org/faq. Accessed on May 2, 2005.

[CCT] Licenses explained. http://creativecommons.org/about/licenses/. Accessed on May 2, 2005.

[CCW] Wiki license — attribution-sharealike 0.5 (beta). http://creativecommons.org/drafts/wiki_0.5. Accessed on May 2, 2005.

[Cona] World Wide Web Consortium. Micropayments overview. http://www.w3.org/ECommerce/Micropayments/Overview.html. Accessed April 25, 2005.

[Conb] World Wide Web Consortium. Resource description framework (RDF). http://www.w3.org/People/Berners-Lee/1996/ppf.html. Accessed April 18, 2005.

[Conc] World Wide Web Consortium. The World Wide Web: Past, present, and future. http://www.w3.org/People/Berners-Lee/1996/ppf.html. Accessed April 18, 2005.

[Fei] Feist publications, inc., v. rural telephone service co., 499 u.s. 340 (1991). http://www.law.cornell.edu/copyright/cases/499_US_340.htm. Accessed on April 12, 2005.

[Fou] Electronic Frontiers Foundation. Trusted computing: Promise and risk. http://www.eff.org/Infrastructure/trusted_computing/20031001_tc.php. Accessed April 18, 2005.

[Gar05] John Gartner. The news for feed(s). April 7 2005.

[Ger05] Michele Gershberg. 2994 ad revenue tops dot-com boom levels. *Computerworld*, April 2005.

[Gooa] Google corporate earnings information. http://www.google.com/intl/en/corporate/business.html. Accessed April 30, 2005.

[Goob] Google corporate information: Company overview. http://www.google.com/corporate/index.html. Accessed on April 24, 2005.

[Gooc] Google news (beta). http://news.google.com/intl/en_us/about_google_news.html. Accessed on April 25, 2005.

[Good] Google web apis - faq. http://www.google.com/apis/api_faq.html. Accessed on April 24, 2005.

[Har] Harper and Row v. Nation Enterprises, 471 U.S. 539.

[INS] Associated press v. international news service. http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=us&vol=248&invol%=215. Accessed on April 25, 2005.

[Kes03] Michelle Kessler. More shoppers proceed to checkout online. *USA Today*, 2003.

[Les02] Lawrence Lessig. The architecture of innovation. *Duke Law Journal*, 51(6), 2002.

[Lul] Lulu basics. http://www.lulu.com/help/node/view/1713. Accessed on May 2, 2005.

[Mer] Philip Merrill. Bbc launches creative archive license enabling uk public to reuse broadcast programs. http://www.grammy.com/news/artswatch/2005/0418creativearchive.aspx. Accessed on May 2, 2005.

[Nel92] Theodore Holm Nelson. *Literary Machines*. Mindful Press, 93.1 edition, 1992.

[Per] Juan Carlos Perez. Google removing agence france presse from google news. http://www.pcworld.com/news/article/0,aid,120125,00.asp. Accessed on April 25, 2005.

[Pro] Evan Prodromou. debian-legal summary of creative commons 2.0 licenses. http://people.debian.org/~evan/ccsummary. Accessed on May 2, 2005.

[Ray] Eric S. Raymond. *The Cathedral and the Bazaar*. Accessed on May 2, 2005.

[Roh02] Jason Rohrer. token_word: a xanalogical transclusion and micropayment system. http://hypertext.sourceforge.net/token_word/rohrer__token_word.pdf, 2002. Accessed on March 28, 2005.

[Tec] Technorati: Tags. http://www.technorati.com/help/tags.html. Accessed May 1, 2005.

[Tod] Andrew Todd. Hardware is law?

[W3Ca] Common markup for micropayment per-fee-links. http://www.w3.org/TR/Micropayment-Markup/. Access April 30, 2005.

[W3Cb] Semantic web activity statement. http://www.w3.org/2001/sw/Activity. Accessed on April 22, 2005.

[Wik] Google. http://en.wikipedia.org/wiki/Google. Accessed on April 24, 2005.

[Wol95] Gary Wolf. The curse of xanadu. http://www.wired.com/wired/archive/3.06/xanadu_pr.html, June 1995. Accessed on March 4, 2005.

[WSJ05] Wsj's online subscriptions outperform print. http://slashdot.org/article.pl?sid=05/04/15/139207, April 15 2005. Accessed April 15, 2005.

[Yah] Yahoo! reports first quarter 2005 earnings results. http://yhoo.client.shareholder.com/news/Q105/YHOO0419-123456.pdf. Accessed May 2, 2005.