Policy Aware Content Reuse on the Web

Oshani Seneviratne, Lalana Kagal, and Tim Berners-Lee

MIT CSAIL, Cambridge
Massachusetts, USA
(oshani | lkagal | timbl)@csail.mit.edu

Abstract. The Web allows users to share their work very effectively leading to the rapid re-use and remixing of content on the Web including text, images, and videos. Scientific research data, social networks, blogs, photo sharing sites and other such applications known collectively as the Social Web, and even general purpose Web sites, have lots of increasingly complex information. Such information from several Web pages can be very easily aggregated, mashed up and presented in other Web pages. Content generation of this nature inevitably leads to many copyright and license terms violations, motivating research into effective methods to detect and prevent such violations.
An experiment on CC attribution license violations from samples of Web sites that had at least one embedded Flickr image revealed that the attribution license violation rate of Flickr images on the Web is around 70-90%. Therefore, it is evident that there should be robust mechanisms for detecting and preventing license violations on the Web. Our primary objective is to enable users to do the right thing and comply with CC licenses associated with Web media, instead of preventing them from doing the wrong thing or preventing violations of these licenses. As a solution, we have implemented two applications (1) Attribution License Violations Validator, which can be used to validate users' derived work against attribution licenses of reused media and, (2) Semantic Clipboard, which provides license awareness of Web media and enables users to copy them along with the appropriate license metadata.

1 Introduction

Content reuse, or in other words "mash-ups", have existed for as long as content has existed. Musicians routinely use other songs and tunes in their compositions. Collage art is considered to be creative, and even original although it is composed from many different sources. Scientists routinely utilize data from different sources to conduct their own experiments. However, mash-ups, as we know them now, are a peculiarly digital phenomenon of the Web age. They are entirely a product made possible by the portable, mixable and immediate nature of digital technology. A potential legal problem arises when more than one legally encumbered content or data streams are bound together in the form of a mash-up. The users of the original content should remain within the bounds of the permitted use of the components comprising the mash-up. They can choose to ignore these

permissions, or follow them. Either way, this creates a burden on them. Ignoring the license terms puts them at the peril of breaking the law, and following them slows the creative process.

Policies in general are pervasive in Web applications. They play a crucial role in enhancing security, privacy and usability of the services offered on the Web [3]. Information accountability provides another motivation to apply policies for data usage practices [28]. In this paper we limit the 'policy awareness' to scenarios that involve content reuse. Policies in this context comprise of licenses that can be expressed semantically, that are widely deployed on a range of media, and that have a large community base. Creative Commons (CC) licenses fit this description perfectly, as it provides a very clear and a widely accepted rights expression language implemented using Semantic Web technologies [11] . These licenses are both machine readable and human readable, and clearly indicates to a person who wishes to reuse content exactly how it should be used, by expressing the accepted use, permissions, and restrictions of the content.

Popular search engines including Google, Yahoo, and even sites such as Flickr, blip.tv, OWL Music Search and SpinXpress have advanced search options to find CC licensed content on the Web [5, 31, 9, 2, 20, 25]. However, even with these human-friendly licenses and the tools to support license discovery, license violations occur due to many reasons: Users may be ignorant as to what each of the licenses mean, or forget or be too lazy to check the license terms, or give an incorrect license which violates the original content creator's intention, or intentionally ignore the CC-license given to an original work in their own interests. Therefore, it is important that we have tools and techniques to make users aware of policies that they must follow while making the process of being license-compliant as painless as possible for the user, and make it difficult for someone to become license in-compliant either deliberately or by mistake.

This paper is organized as follows: Section 2 gives the background and an overview of the technologies used. Section 3 outlines an experiment conducted to assess the level of CC attribution license violations on the Web using Flickr images. This experiment provided the motivation to develop tools for policy aware content reuse as described later in the paper in section 4. Section 5 discusses the related work in this area and section 6 discusses some future work on the tools we have developed. Finally, we conclude the paper with a summary of the contributions in section 7.

2 Background

2.1 Policies for Rights Enforcement on the Web

Policies governing the reuse of digital content on the Web can take several forms. It can be upfront enforcement mechanisms such as *Digital Rights Management* (DRM) approach, or rights expression mechanisms such as *Creative Commons* licenses where users are given freedom to use content subject to several restrictions and conditions.

When it comes to DRM, distribution and usage of copyrighted content is often controlled by up-front policy enforcement. These systems usually restrict access to the content, or prevent the content from being used within certain applications. The core concept in DRM is the use of digital licenses, which grant certain rights to the user. These rights are mainly usage rules which are defined by a range of criteria, such as frequency of access, expiration date, restriction to transfer to another playback device, etc. An example of a DRM enforcement would be a DRM software enabled playback device not playing a DRM controlled media transferred from another playback device, or not playing the media after the rental period has ended. The use of DRM to express and enforce rights on content on the Web raises several concerns. First, the consumer privacy and anonymity are compromised. Second, the authentication process in the DRM system usually requires the user to reveal her identity to access the protected content leading to profiling of user preferences, and monitoring of user activity at large [8]. Third, the usability of the content is questionable, since the user is limited to using proprietary applications to view or play the digital content producing vendor lock-in. Similarly, 'copyright notice' or a 'license' describes the conditions of usage of copyrighted material. A user of that particular material should abide by the license that covers the usage, and if any of the conditions of usage described in that license are violated, then the original content creator has the right to take legal action against the violator.

On the other hand, CC has been striving to provide a simple, uniform, and understandable set of licenses that content creators can use to issue their content under. Often, Web authors post their content with the understanding that it will be quoted, copied, and reused. Further, they may wish that their work only be used with attribution, used only for non-commercial use, distributed with a similar license and will be allowed in other free culture media. To allow these use restrictions CC has composed four distinct license types: *BY* (attribution), *NC* (non-commercial), *ND* (no-derivatives) and *SA* (share-alike) that can be used in combinations that best reflects the content creator's rights. In order to generate the license XHTML easily, CC offers a license chooser that is hosted at `http://creativecommons.org/license`. With some user input about the work that is being licensed, the license chooser generates a snippet of XHTML that contains the RDFa [23] to be included when the content is published on the Web. Creative Commons Rights Expression Language (*ccREL*) [11] is the standard recommended by the CC for machine readable expression of the meaning of a particular license. Content creators have the flexibility to express their licensing requirements using this rights expression language and are not forced into choosing a pre-defined license for their works. Also, they are free to extend licenses to meet their own requirements. ccREL allows a publisher of a work to give additional permissions beyond those specified in the CC license with the use of the *cc:morePermissions* property to reference commercial licensing brokers or any other license deed, and *dc:source* to reference parent works. Therefore, unlike in older CC recommendations, it is also possible to have content under a CC license that does not require attribution.

2.2 Inline Provenance using Metadata

To be useful, metadata need to have three important characteristics: they have to be easy to produce, be embedded within the data they describe, and be easily readable. The easiest way to produce metadata is to have them be produced automatically. Any metadata that has to be produced manually by the user usually doesn't get produced at all. The easiest way to ensure that the link between metadata and the data they describe is not broken is by embedding the former inside the latter. This way, the two travel together inseparably as a package. Finally, metadata have to be accessible easily, readable both manually as well as programmatically. At best, the metadata should be readable by crawlers of various search engines. Since metadata and data are traveling together, if popular search engines such as Google and Yahoo can read the metadata, by default the data become available to anyone who searches for it.

Extensible Metadata Platform (XMP) [30] is a technology that allows one to transfer metadata along with the content by embedding the metadata in machine readable RDF. This technology is widely deployed in embedding licenses in free-floating multimedia content such as images, audio and video on the Web. Another format which is nearly universal when it comes to images is the Exchangeable Image File format (EXIF) [7]. International Press Telecommunications Council (IPTC) photo metadata standard [15] is also another well known standard. The metadata tags defined in these standards cover a broad spectrum including date & time information, camera settings, thumbnail for previews and more importantly, the description of the photos including the copyright information. However, these latter two formats do not store metadata in RDF. One major drawback of inline metadata formats such as XMP and EXIF is that it is embedded in a binary file, completely opaque to nearly all users, whereas metadata expressed in RDFa will require colocation of metadata with human visible HTML. In addition to that, these metadata formats can only handle limited number of properties and lack the rich expressivity offered by RDF.

3 Motivation

Unless a particular piece of content on the Web has some strict access control policies, most users do not feel the need to check for the license it is under and be license compliant. To verify this hypothesis we conducted an experiment to assess the level of license violations on the Web. Specifically, the goal of the experiment was to obtain an estimation for the level of CC *attribution license violations* on the Web using Flickr images[1].

[1] As of April 2009, Flickr has over 100 million Creative Commons Licensed images. Thus it provided a large sample base for our experiment.

3.1 Experiment Setup

To ensure a fair sample of Web sites to check for attribution license violations, we used the Technorati blog indexer[2] without hand-picking Web pages to compose the sample. The Technorati blog indexer crawls and indexes weblog-style Web sites and keeps track of articles on the Web site, what links to it, what it links to, how popular it is, how popular the Web sites that link to it, and so on. Technorati data are time dependent, and therefore the technorati *authority rank*[3] is based on most recent activity in a particular Web site. To generate a fair sample of Web sites we used the *Technorati Cosmos* method[4], and the *authority rank* by retrieving results for Web sites linking to Flickr server farm URIs that have a specific format[5]. Since the Flickr site has several server farms, each time the experiment was run, the base URIs were randomly generated by altering the Flickr server *farm-id*s. In addition to that, we made sure that the samples were independent of each other by running the experiment after a two-week time gap for three times. This is because the *authority rank* given to a Web page by Technorati, and hence the results returned from the Cosmos method, dynamically changes as new content gets created[6].

Since Flickr is still using the older CC 2.0 recommendation, Flickr users do not have that much flexibility in specifying their own *attributionURL* or the *attributionName* values to state how they would like attribution to be given to them. However, it is considered general practice to give attribution by linking to the Flickr user profile or give the Flickr user name (which could be interpreted as the *attributionURL* and the *attributionName* respectively), or by the least, point to the original source of the image [13]. Therefore, the criteria for checking attribution consist of looking for the *attributionURL* or the *attributionName* or any *source citations* within a reasonable level of scoping from where the image is embedded in the Document Object Model (DOM).

3.2 Results

The results from 3 samples of Web sites gathered within 2 week intervals are given in Fig 1. These results have misattribution and non-attribution rates ranging from 78% to 94% signaling that there is a strong need to instigate license (or policy) awareness among reusers of content. The overall summary of the results includes the total number of Web sites tested, number of images in all of the

[2] Implemented using the `http://technorati.com/developers/api`

[3] Authority Rank is a measurement that determines the top 'n' number of results from any query to the Technorati API.

[4] Cosmos Query allows the retrieval of results for blogs linking to a given base URL

[5] This format is: `http://farm<farm-id>.static.flickr.com/<server-id>/<id>_<secret>.(jpg|gif|png)` (From `http://www.flickr.com/services/api/misc.urls.html`)

[6] However, this may have introduced a bias in our sample as the top sites from the Technorati blog indexer are probably well visited, hence more pressure for those sites to fix errors in attribution. We were not able to verify this in our experiment though.

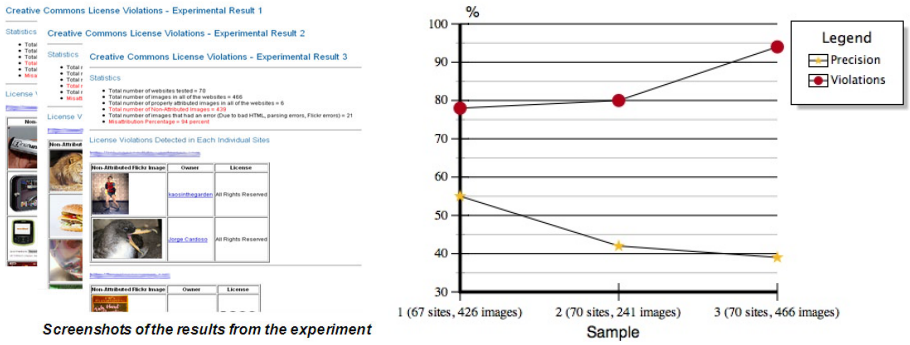


Fig. 1. Left: Screenshots of the Results from the Experiment (Refer to *http://dig.csail.mit.edu/2008/WSRI-Exchange/results* for more information). **Right:** Attribution Violations Rate and Precision With Correction for Self-Attribution.

Web sites, number of properly attributed images, number of misattributed or non-attributed images, and the number of instances that led to an error (due to bad HTML which led to parsing errors for example). Using these values, the percentage of misattribution and non-attribution for each sample was calculated.

3.3 Issues and Refining the Experiment

The results from the experiment includes cases where users have not attributed themselves: i.e. user uploads her photos on Flickr, and uses those in the user's own blog or Web site. Since those are user's own photos, she is under the assumption that there is no need to attribute herself. This assumption is not entirely valid as the CC BY license deed [4] specifies: *"If You Distribute you must keep intact all copyright notices for the Work and provide (i) the name of the Original Author (or pseudonym, if applicable) ... "*. This means that, if there is a license attached with the original content, the original user will become the reuser, and therefore will have to honor the license even though it is imposed by herself. This might seem absurd since it should not matter to the user if she violates her own license terms. However if the user gives attribution to herself, it would in fact guide other people who want to reuse the content in that secondary work. Therefore, by not attributing herself, the user may be setting a precedent for the violation of her own rights in the long run. A solution to this issue is hard to realize, as it is difficult to infer the Web site owner from the data presented in the Web site. Even if that was possible, it is hard to make a correlation between the Flickr photo owner and the Web site owner. However, we manually inspected the samples to see whether the misattributed images were actually from the user or not, and flagged those as *false positive* in the results set. After this correction, we found the precision rate of the experiment to be between 55% to 40%.

We also found out that a majority of the Web sites crawled and examined in this experiment have not used ccREL in marking up attribution. Therefore, we

used a heuristic to check for the existence of attribution. This heuristic includes the *attributionName*[7] or the *attributionURL*[8] or the original source document's URI. This would visually correlate to including the attribution information immediately after the content that is being attributed. However, since there is no strict definition from CC as to how attribution should be scoped, someone could also attribute the original content creator somewhere else in the document. This experiment only considers the types of attributions as given in the first category. The rationale behind this assumption is that, it is possible that the user intended to include more than one image from the same original content creator, and by mistake, failed to attribute some images only, while correctly attributing all the others.

Blog Aggregators such as Tumble-logs (for example tumblr.com) cuts down the text and favors short form, mixed media posts over long editorial posts. Use of such blog aggregators is another related problem in getting an accurate assessment of attribution license violations. For example, in a blog post where a photo was reused, the original owner of the photograph may have been duly attributed. But when the tumble-log pulls in the feed from that post in the original Web site and presents the aggregated content, the attribution details may be left out. This problem is difficult to circumvent because there is no standard as to how aggregation should happen with the license and attribution details.

4 Tools to Enable Policy Awareness

As a proof of concept we have developed couple of tools that can be used to enable policy awareness when reusing *images* on the Web. Both the tools are currently limited to image reuse at the moment, but can be easily extended to support other types of media.

4.1 Attribution License Violations Validator for Flickr Images

When someone aggregates content from many different sources, it is inevitable that some attribution details may be accidentally forgotten. The Attribution License Violations Validator is designed to check whether the user has properly cited the source by giving the due attribution to the original content creator. In order to make sure that no CC license terms of the user are violated, the author can run the CC License Violations Validator and see if some sources have been left out or whether some have been misattributed.

Design and Implementation: This tool has four major components as shown in the left half of Fig 2. Once the user gives the URI where the composite work can be found, the site crawler will search for all the links embedded in the given

[7] *attributionName* is constructed from the Flickr user name
[8] *attributionURL* is constructed from the Flickr user profile URI

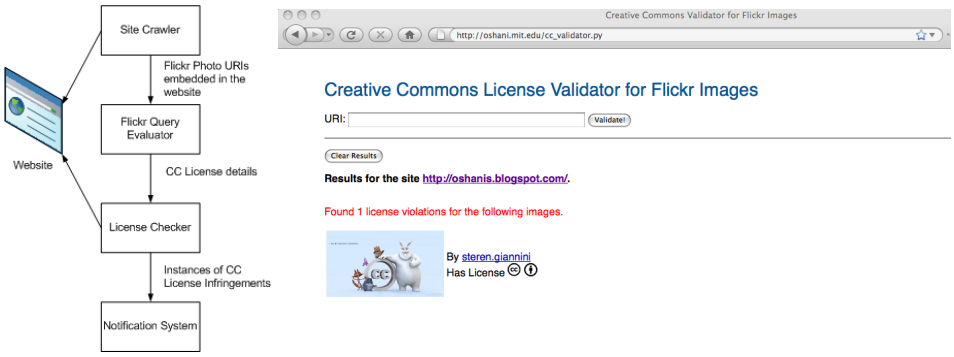


Fig. 2. Left: The Design of the Validator. **Right:** Output from the Validator showing the image that was not attributed properly, who the image belongs to and what license it is under.

Web page and filter out any embedded Flickr photos. From each of these Flickr photo URIs, it is possible to glean the Flickr photo id. Using this photo id, all the information related to the photo is obtained by calling several methods in the Flickr API. This information includes the original creator's Flickr user account id, name and CC license information pertaining to the photo, etc. Based on the license information of the Flickr photo, the tool checks for the attribution information that can be either the *attributionName*, *attributionURL*, source URI or any combination of those within a reasonable scoping in the containing DOM element in which the image was embedded. The 'reasonable scoping' in this case, is taken to be to be within the parent or the sibling nodes in the DOM. If such information is missing, the user is presented with the details of the original content creator's name, the image along with its URI, and the license it is under, enabling the user to compose the XHTML required to properly attribute the sources used.

Challenges and Limitations: The license violations detection can only work if the the image URI is actually linked from the Flickr site. Therefore if a user wants to cheat, she can easily do so by changing the image URI by uploading to another Web space for example. Another complication is that a Flickr user can upload and assign CC licenses regardless of that user having the actual rights to do so. In other words, if someone uploads a copyrighted photo from *Getty Images* and assigns a CC license on Flickr, and an innocent user downloads and uses this photo, then that user will be violating copyright law without the user knowing it. Therefore, we need to have some capability to track provenance of image data, and be able to identify whether a particular image has been used elsewhere in a manner that violates the original license terms.

One of the major assumptions we have made in developing this tool is that attribution to be specified within the parent node or the sibling nodes of the

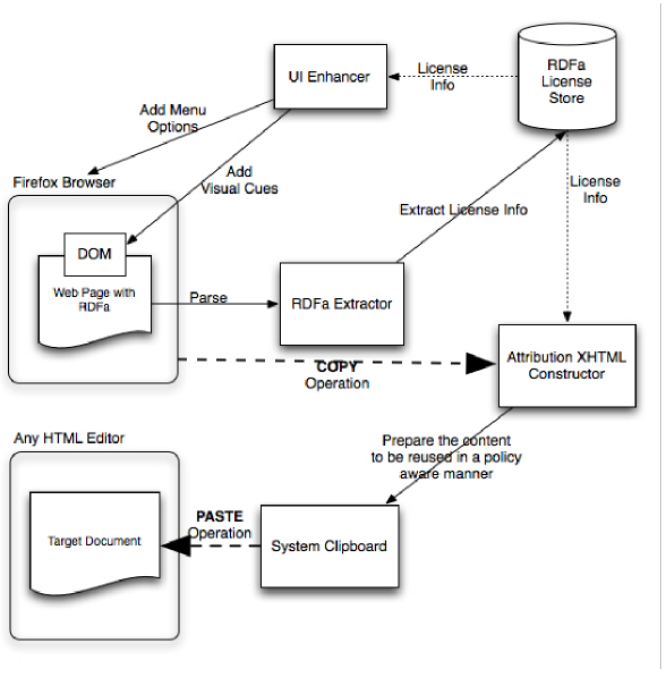
containing image element. Otherwise we classify it an instance of non-attribution. This assumption works practically and seems to be the most logical thing to do. However, since there is no standard agreement as to what the correct scoping for attribution is, this assumption can lead to a wrong validation result. The solution to this problem can be in two folds. (1) CC should give a guideline as to what the correct scoping of attribution should be relative to the content that is attributed. (2) Flickr (or any other such service) should expose the license metadata as RDF, instead of providing an API to query with. Exposing license metadata as RDF is preferred as it enables data interoperability and relieves the tool authors from having to write data wrappers for each service.

4.2 Semantic Clipboard

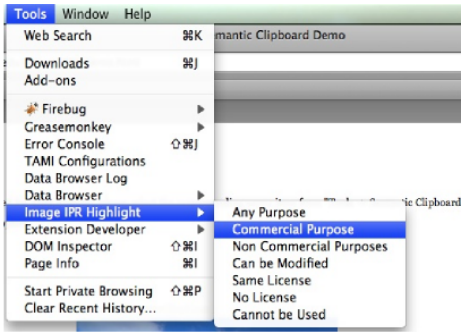
This is a Firefox Web browser based tool integrated with the Tabulator, a linked data browser that can be installed as a Firefox extension [27]. The primary goal of this tool is to let users reuse content with minimal effort.

Design and Implementation: The design of the Semantic Clipboard is given in Fig 3. The tool uses the 'RDFa Extractor' to glean the Creative Commons license information expressed in RDFa from the HTML page the user browses. The 'UI Enhancer' implements several menu options in the Firefox browser to select licensed images with the proper intention. The available options are given in Fig 3. For example, if a user want to see images that can be used for 'Commercial Purposes', she can select the corresponding menu item. Then the images that do not have the CC-NC clause (Creative Commons Non Commercial use) will be highlighted with an overlay on the image. The 'Attribution XHTML Constructor' is called when the user issues a copy instruction on a particular image by right-clicking on the image and selecting the context menu option 'Copy Image with License' as shown in Fig 3. Based on the license information for that particular image, the attribution XHTML snippet is constructed as specified by Creative Commons, and copied to the system clipboard. Currently two data flavors are supported: ASCII text and HTML. Therefore if the target application accepts HTML such as in a rich text editor, the source text (with the angle brackets) will not be displayed.

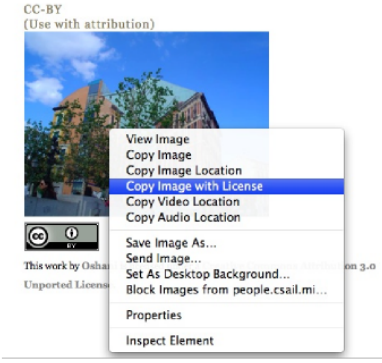
Data Purpose Algebra Analogy: It has been shown that it is possible to model data usage policies programmatically by what is known as the *Data Purpose Algebra* [12] by describing each content item i in a data set, a source or agent that processes the data $Q_d(i)$, the category of data $K_d(i)$, and its purpose $P_d(i)$. When another agent combines two or more data sets, a new data set is created whose content, category and purpose are some function of the agent, content, category and purpose of each of the component data sets. Specifically, if the agent or the source of the new data item is a', the new category becomes the function $\mathcal{K}(K_d(i))$ of the given category, and the allowed purposes of the new data item will be the more complex function $\mathcal{P}(P_d(i), A_d(i), a', K_d(i))$ that may



Design of the Semantic Clipboard



Firefox Menu



Context Menu on an Image

Fig. 3. Semantic Clipboard Architecture and the User Interface

depend on the original purposes, the agents, and the category of the original data.

When reusing content on the Web the same principle could be applied. For the content item i that is reused, the source function $Q_d(i)$ would be represented by the URI of the content. The category of data $K_d(i)$ will be represented by the content type (i.e. image, text, video, audio, etc.) that is being reused. The purpose $P_d(i)$ will be determined by the CC license associated with it that specifies the allowed uses, restrictions and conditions. Then the function which composes the new set of purposes $\mathcal{P}(P_d(i), A_d(i), a', K_d(i))$ should generate the XHTML that is required to embed the content with the proper attribution.

Challenges and Limitations: One of the hazards of combining multiple data sources is that incompatible licenses can get mixed up creating a license that basically freezes the creative process. Take for example a Non-Commercial (NC) license that gets mixed with a Share-Alike (SA) license. An SA license requires that the resulting product be shared under exactly the same conditions as the

component product under SA. The resulting license in our scenario becomes NC-SA. But while the result satisfies the first license by also being NC, it fails the second license by not being *only* SA. We cannot simply ignore the NC clause and give the resulting work only the SA license because somebody else might use the resulting derivative work which does not have the NC clause for some commercial use violating the rights of the original creator who composed the NC component. The Semantic Clipboard does not handle such license conflicts.

5 Related Work

Reuse detection is important in domains such as plagiarism detection and even in biological sequence mining. Significant research has been carried out to detect reuse of text. This includes information retrieval techniques as mentioned in [18, 24], where the document is treated as a sequence of symbols and substring based fingerprints are extracted from the document to determine repetitive patterns.

The CC License ***Syntax*** Validation Service [14] can be used to parse documents for embedded licenses in RDFa. After parsing the document, this service gives a list of licensed objects and each of their license authorship, version, jurisdiction, whether the license has been superseded or deprecated and whether the work is allowed in free cultural works, etc. However, it does not give the information as to whom the attribution should be given when reusing these license objects like in the attribution license violations validator we have developed. In addition to that, CC has put much focus on coming up with ways to enable tool builders to use the CC licenses very effectively. For example, the *live box* on the *License Deed Page* as shown in Fig 4 suggests how to attribute a particular work. This is created when a CC license hyperlink that has the *attributionName* and the *attributionURL* properties to the *License Deed Page* is dereferenced. There are also several license aware Mozilla Firefox extensions developed by the CC. MozCC [19] is one such tool. It provides a specialized interface for displaying CC licenses, where the user receives visual cues when a page with RDFa metadata is encountered. This includes the display of specific CC branded icons in the browser status bar when the metadata indicates the presence of a CC license. However, this software does not offer the capability to copy the license attribution XHTML as in the Semantic Clipboard that we have developed.

Fig. 4. CC Deed Page Displaying the Attribution XHTML

The Semantic Clipboard was actually inspired from the work done on 'XHTML Documents with Inline, Policy-Aware Provenance' [16] by Harvey Jones. Jones

developed a document format that can represent information about the sources of its content, a method of excerpting from these documents that allow programs to trace the excerpt back to the source, a CC reasoning engine which calculates the appropriate license for the composite document, and a bookmarklet that uses all these components to recommend permissible licenses. But this tool requires all the source documents that the user needs to copy from, to be annotated with a special document fragment ontology, and the *Paste Operation* is limited to inserting copied XHTML in the top level of the document only, i.e. it does not allow copying inside existing document fragments. The Semantic Clipboard address these issues by eliminating the reliance of an external document fragment ontology and utilizing the operating system's clipboard to paste the image with the associated license metadata in XHTML. The only requirement for the Semantic Clipboard to work is that the license information about the work to be expressed in RDFa in the source documents.

There are several tools which can be used to automatically embed the license metadata from Flickr. Applications such as ThinkFree, a Web based commercial office suite [26], and the open source counterpart of it: the "Flickr image reuse for OpenOffice.org" [10] are few examples of such applications. These applications allow the user to directly pick an image from the Flickr Web site and automatically inject the license metadata with it into a document in the corresponding *office suite*. A severe limitation of this approach is that they only support Flickr images. The Semantic Clipboard can be used to copy any image in to any target document with the license as long as the license metadata is expressed in RDFa.

Attributor [1], a commercial application, claims to continuously monitor the Web for its customers' photos, videos, documents and to let them know when they have been used elsewhere on the Web. Then it offers to send notices to the offending Web sites notifying link request, offers for license, request for removal or a share of the advertisement revenue of that page. Another commercial application called PicScout [21] claims that it is currently responsible for detecting over 90% of all online image infringements detections. They also claim to provide the subscribers of their services with a view into where and how their images are being used online. The problem with these services is that it penalizes the infringers after-the-fact, rather than encouraging them to do the right thing upfront [17]. Since their implementations are based on bots that crawl the Web in search of infringes, these services take up valuable Internet bandwidth [29]. Also, these services are not free, which bars many content creators who wish to use such services to find license violations of their content from using the service.

6 Future Work

We envision that we could apply the same principle as checking for *attribution* license violations to check for other types of license violations. Detecting whether an image has been used for any commercial use would be of much interest to content creators, especially if the second use of the image decreases the monetary value of the original image. The CC deed for Non Commercial (NC) use specifies

that a license including the NC term may be used by anyone for any purpose that is not *"primarily intended for or directed towards commercial advantage or private monetary compensation"*. However, this definition can be vague in certain circumstances. Take for example the case where someone uses a CC-BY-NC licensed image in her personal blog properly attributing the original content creator. The blog is presumably for non commercial use, and since she has given proper attribution, it appears that no license violation has occurred. However there might be advertisements in the page that are generated as a direct result of the embedded image. Our user might or might not actually generate revenue out of these advertisements. But if she does, it could be interpreted as a 'private monetary compensation'. Hence we believe that the perception as to what constitutes a 'Commercial Use' is very subjective. CC recently conducted an online user survey to gather general opinions as to what people perceive a 'Commercial Use' is [6]. An important finding from this survey is that 37% of the creators who make money from their works do so indirectly through advertisements on their Web sites. Therefore, it seems that there aren't any clear cut definitions of a 'Non Commercial Use' yet to find out violations and gather experimental results. But, if the definition of *non commercial use* becomes clearer and much more objective, a validator can be implemented to check for such violations as well.

It would also be interesting to check for share-alike license violations. These violations happen when a conflicting license is given when the content is reused. The solution, therefore, is to check the RDFa in both the original page and in the page where the image was embedded to see if the latter is the same as the original CC license.

The requirement for attribution is a form of social compensation for reusing one's work. While mentioning one's name or the WebID when attributing draws attention to an individual, other forms of *attention mechanisms* can also be implemented. For example, a content creator can obligate the users of her works to give monetary compensation or require that they include certain ad links in the attribution XHTML or give attribution in an entirely arbitrary manner. These extra license conditions can be specified using the *cc:morePermissions* property. Tools could be built to interpret these conditions and give credit to the original creator as requested.

The tool we have developed only works if every image found on the page has it's own license. Possible extensions of this tool would be to have higher level of granularity to determine the license of an image when it does not have a license of it's own, but is contained within a page that has a license or is a member of a set of images (e.g. a photo album) that has a license. Protocol for Web Description Resources (POWDER) [22], a mechanism that allows the provision of descriptions for groups of online resources, seems like a viable method to making the license descriptions about the resources explicit. Tool builders can then rely on the POWDER descriptions to help users to make appropriate content reuse decisions.

Currently, images that are copied with their metadata to the Semantic Clipboard are overwritten when some new content is copied to the clipboard. In other words, the tool only supports copying of one image at a time. But it would be useful to have a persistent data storage to register images or any other Web media along with their license metadata, index, make those persistent across browser sessions and use whenever the user needs it in a license-compliant manner.

One of the major drawbacks of the Semantic Clipboard is that it is Firefox browser-dependent. Developing an Opera Widget, a Chrome Extension, a Safari Plugin, an Internet Explorer Content Extension or completely making this tool browser independent seem to be a viable future direction of the project.

It would be interesting to measure how user behavior changes with the introduction of tools such as the License Violations Validator and the Semantic Clipboard. A measurement of the increased (or decreased or unchanging) level of license awareness would be an important metric in determining the success of these tools. Therefore, we plan to perform a controlled user study in the future.

We have only explored one domain of content, specifically image reuse on the Web. However, there are billions of videos uploaded on YouTube, and potentially countless number of documents on the Web, which have various types of licenses applied. While organizations such as Mobile Picture Association of America (MPAA), Recording Industry Association of America (RIAA) and other such big organizations are working towards preserving the rights of the works of their artists on YouTube, other video and audio sharing sites and peer-to-peer file sharing networks, there are no viable alternatives for ordinary users who intend to protect their rights using CC. Thus a solution of this nature which detects CC license violations based on the metadata of other types of free-floating Web media will be very useful.

7 Conclusion

As the license violations experiment indicated, there is a strong lack of awareness of licensing terms among content reusers. This raises the question as to whether the machine readable licenses are actually working. Perhaps more effort is needed to bring these technologies to the masses, and more tools are needed to bridge the gap between the license-aware and the license-unaware. An important research question that stems from this work is the method of provenance preservation of content on the Web. We have trivially assumed URIs as the provenance preservation mechanism when developing the tools described. However, it would be an interesting challenge to track provenance based on the content itself, without having to rely on a unique identifier. This would enable us to find out license violators, in addition to validating one's own work for any violations. Also, programmatically determining whether a particular reuse of material is allowable or not is subjective, especially since some of the laws and standards have been quite ambiguous in defining these terms.

In general, social constraints are functions of any part of the blossoming Social Web we are experiencing today. As we are living in an era of increasing user generated content, these constraints can be used to communicate the acceptable uses of such content. We need tools, techniques and standards that strike an appropriate balance between the rights of the originator and the power of reuse. The rights of the originator can be preserved by expressing what constitutes appropriate uses and restrictions using a rights expression language. These rights will be both machine and human readable. Reuse can be simplified by providing the necessary tools by leveraging these machine readable rights to make the users more aware of the license options available and ensure that the user be license or policy compliant. Such techniques can be incorporated in existing content publishing platforms or validators or even Web servers to make the process seamless. This paper has demonstrated several tools that lay the foundation for such policy aware systems. We hope these will stimulate research in this area in the future.

Acknowledgements

Some parts of this work was done while the first author was undergoing the 'Networks for Web Science Research Exchange' program under Nigel Shadbolt at University of Southampton, UK. Special thanks for his guidance throughout the project. In addition, the authors wish to thank their colleagues, Danny Weitzner, Hal Abelson, Gerry Sussman, and other members at DIG for their contribution to the ideas expressed in this paper.

This work was carried out with generous funding from National Science Foundation Cybertrust Grant award number 04281, IARPA award number FA8750-07-2-0031, and UK Engineering and Physical Sciences Research Council (EPSRC) grant number EP/F013604/.

References

1. Attributor - Subscription based web monitoring platform for content reuse detection. `http://www.attributor.com`.
2. blip.tv - Hosting, distribution and advertising platform for creators of web shows. `http://blip.tv`.
3. P. A. Bonatti, C. Duma, N. E. Fuchs, W. Nejdl, D. Olmedilla, J. Peer, and N. Shahmehri. Semantic web policies - a discussion of requirements and research issues. In *ESWC*, pages 712–724, 2006.
4. Creative Commons BY 3.0 Unported Legal Code. `http://creativecommons.org/licenses/by/3.0/legalcode`.
5. Creative Commons Customized Search in Google. `http://creativecommons.org/press-releases/entry/5692`.
6. Creative Commons Noncommercial study interim report. `http://mirrors.creativecommons.org/nc-study/NC-Use-Study-Interim-Report-20090501.pdf`.
7. Exchangeable Image File Format . `http://www.exif.org/specifications.html`.

8. J. Feigenbaum, M. J. Freedman, T. Sander, and A. Shostack. Privacy engineering for digital rights management systems. In T. Sander, editor, *Digital Rights Management Workshop*, volume 2320 of *Lecture Notes in Computer Science*, pages 76–105. Springer, 2001.
9. Flickr API. `http://www.flickr.com/services/api`.
10. Flickr image reuse for openoffice.org. `http://wiki.creativecommons.org/Flickr-Image-Re-Use-for-OpenOffice.org`.
11. Hal Abelson, Ben Adida, Mike Linksvayer, Nathan Yergler. ccREL: The Creative Commons Rights Expression Language. *Creative Commons Wiki*, 2008.
12. C. Hanson, T. Berners-Lee, L. Kagal, G. J. Sussman, and D. J. Weitzner. Data-purpose algebra: Modeling data usage policies. In *POLICY*, pages 173–177. IEEE Computer Society, 2007.
13. How to attribute Flickr images. `http://www.squidoo.com/cc-flickr\#module12311035`.
14. Hugo Dworak, Creative Commons License Validation Service. `http://validator.creativecommons.org`.
15. International Press Telecommunications Council Photo Metadata Format. `http://www.iptc.org/IPTC4XMP`.
16. H. C. Jones. Xhtml documents with inline, policy-aware provenance. Master's thesis, Massachusetts Institute of Technology, May 2007.
17. Ken Doctor, Blog Entry on "Attributor Fair Syndication Consortium Completes Newspaper Trifecta". `http://www.contentbridges.com/2009/04/attributor-ad-push-on-piracy-completes-newspaper-trifecta.html`.
18. J. W. Kim, K. S. Candan, and J. Tatemura. Efficient overlap and content reuse detection in blogs and online news articles. In *18th International World Wide Web Conference (WWW2009)*, April 2009.
19. MozCC - Firefox extension to discover Creative Commons licenses. `http://wiki.creativecommons.org/MozCC`.
20. OWL Music Search. `http://www.owlmusicsearch.com`.
21. picScout - Image tracker for stock photography agencies and professional photographers. `http://www.picscout.com`.
22. Protocol for Web Description Resources (POWDER). `http://www.w3.org/2007/powder`.
23. RDFa, Resource Description Framework in Attributes. `http://www.w3.org/2006/07/SWD/RDFa/syntax`.
24. N. Shivakumar and H. Garcia-Molina. Scam: A copy detection mechanism for digital documents. In *Second Annual Conference on the Theory and Practice of Digital Libraries*, 1995.
25. SpinXpress - Collaborative media production platform. `http://spinxpress.com`.
26. Think Free - Java based web office suite. `http://www.thinkfree.com`.
27. Tim Berners-Lee and James Hollenbach and Kanghao Lu and Joe Presbrey and Eric Prud'ommeaux and mc schraefel. Tabulator Redux: Browing and Writing Linked Data . In *Linked Data on the Web Workshop at WWW08*, 2008.
28. D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman. Information accountability. *Communications of the ACM*, June 2008.
29. William Faulkner, Tales From The IT Side: "PicScout, Getty Images and Goodbye iStockPhoto..!". `http://williamfaulkner.co.uk/wordpress/2007/09/picscout-getty-images-and-goodbye-istockphoto`.
30. XMP - Extensible Metadata Platform. `http://www.adobe.com/products/xmp/index.html`.
31. Yahoo Creative Commons Search. `http://search.yahoo.com/cc`.