Framework for Policy Aware Content Reuse on the WWW

Oshani Wasana Seneviratne <oshani@csail.mit.edu>

February 10, 2009

1 Introduction

The World Wide Web (WWW) is a platform in which users can share their work very effectively. Due to the nature of the medium, content on the WWW including text, images, and videos can be reused and remixed rapidly. Scientific research data, social networks, blogs, photo sharing sites and other such applications known collectively as the social web, and even general purpose websites have lots of increasingly complex data. Data 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 or license terms violations, motivating research into effective methods to detect and prevent such violations.

2 Problem Description

The research work undertaken on this project will focus on methods for detecting and helping users avoid license violations, in order to create a framework for policy aware content reuse on the WWW. This framework will build on the Creative Commons Rights Expression Language (ccREL). However, it will differ from Digital Rights Management (DRM), because it will not attempt to enforce the rights embedded in media, but merely guide the user as to how best the content should be reused, making sure that the policies governing the content usage are properly adhered to.

Such a framework is very useful because, for typical web users wanting to share their content online, DRM techniques are often an overkill. The Creative Commons (CC) project, on the other hand, provides a very clear and a widely accepted rights expression language using semantic web technologies which is used to compose a set of well-defined licenses. These licenses are machine readable, and indicates to a person who wishes to reuse the content exactly how it should be used. However, unlike with DRM, if the license terms are violated, the violator will not be automatically penalized.

Our experiment on CC attribution license violations on Flickr images revealed a violation rate of 70%-90% on the web. We hope to extend this experiment to detect non-commercial and share-alike license violations as well. From the results of our initial experiments, it is evident that there should be robust mechanisms for detecting license violations, and prevent those happening if possible.

The framework we propose in this thesis will be implemented in two parts:

1. License Violations Validator : to verify your own work for any license violations (for e.g. this is in the same spirit as the W3C XHTML validator or the RDF validator, etc.)

2. Semantic Clipboard Application: for users to seamlessly reuse content on the web while integrating the license metadata in a policy aware manner. Note that the extent to which this application is feasible depends on the agreement of Operating System APIs with the core system we design.

3 Background

Policies governing the digital content on the web comes in two flavors: As already explained briefly, they can either use the DRM approach exercised for commercial entities or the Rights Expression alternative offered by the CC.

3.1 Digital Rights Management (DRM)

Distribution and usage of copyrighted content is often controlled by DRM systems. These systems usually restrict access to the content, or prevent the content from being used within certain applications, such as iTunes not playing a DRM controlled song or a movie not playing after the rental period has ended.

The use of DRM to express and enforce rights on content on the WWW raises several concerns. First, the consumer privacy and anonymity are compromised. The authentication process in the DRM system usually requires the user to reveal her identity to access the protected content. This could lead to profiling of user preferences, and monitoring of user activity at large [4]. The other concern is the ability to use content under the Fair Use doctrine [15]. Another huge criticism of DRM is the usability of the content, where the user is limited to using proprietary applications to view or play the digital content.

3.2 Creative Commons (CC)

CC is a non-profit organization that has been striving to provide a simple, uniform, and understandable licenses that content creators can use to issue their content under. These licenses provide a solution to the problem of copyright on the WWW, while ensuring that the culture of reusing existing works to foster creativity is not hindered. 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, or only for non-commercial use, distributed with a similar license etc.

'ccREL' [7] is the standard recommended by the CC for machine readable expression of copyright terms and licensing. Content creators have the flexibility to express their licensing requirements in ccREL and are not forced into choosing a pre-defined license for their work. Also, they are free to extend licenses defined by others to meet their own requirements.

CC has put much focus on coming up with ways to enable tool builders to use the CC licenses very effectively. There are currently several Mozilla Firefox extensions that are CC-License aware. MozCC [13] is one such tool. It provides a specialized interface for ccREL, and the user would receive 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. Operator [16] is another Firefox browser extension that detects microformats and RDFa in web pages that the user visits. Using Operator, it is possible to write a CC 'action script' that finds all CC licensed content inside a web page by looking at the RDFa syntax.

4 Policy Awareness

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 WWW [2]. Information accountability provides another motivation to apply policies for data usage practices [23]. On the semantic web, policy based systems may be implemented with a reasoner on rule-based systems, where the rules represent laws, licenses, or policies that relate to the system. For example, REIN, which is named after the Rei policy specification language [20] and Notation 3 [22], describes a system offering policy management [11]. Stemming from that work, the AIR policy language is designed to express and enforce policies to provide reliable assessments of compliance with rules and policies governing the use of information [12].

4.1 The Need for Policy Awareness in Content Reuse

In this thesis we will limit the 'policy awareness' aspect to licenses which can be expressed semantically, widely deployed on a range of media, and has a large community base. CC licenses fit this description perfectly. Therefore, we will be focussing our attention on CC licenses, keeping in mind to extend the system we develop to other types of licensing mechanisms.

Information about the Creative Commons license of a particular work is usually specified as RDFa [19] on a web page, or in some cases, it can be obtained via a simple query to the service endpoint which hosts the content. RDFa allows machine understandable semantics to be added to XHTML. There are many tools that extract RDF [18] from web pages marked up with RDFa, by using, for example, the RDFa Distiller [8].

CC too provides the RDFa which should be included to attribute the original content creator. This could be obtained on the corresponding CC deed page when followed from the CC license link appearing on the original content page.

However even with all the tools and licenses designed to warn users of their accepted use, permissions and restrictions, we can expect occurrence of license violations due to many factors. Users may be ignorant as to what each of the licenses mean. Users may forget to check and include the proper license terms in their own work. Or the users may simply give an incorrect license which violates the original content creators' intention. We should also not forget malicious users who may intentionally ignore the CC-license given to an original work in their own interests. Whatever the case may be, the original content creator would be interested in knowing when her licenses have been violated and on which web pages. But given the scale of the WWW, the knowledge of such a license violation is highly unlikely, unless the original content creator comes across it by chance.

Also, people who create works that may use several hundred or so other sources would be interested in knowing whether they have violated anybody else's CC license terms: For example, by failing to keep a proper citation list or by misattributing. In such cases, a 'validator' which checks for CC license violations of content would be very useful. This will be similar to web developers checking whether their XHTML is valid using the 'W3C Markup Validation Service' or semantic data producers checking if their data is in proper RDF [18] by using the 'W3C RDF Validation Service'. Using these tools content re-users can rectify the instances where they have inadvertently violated the CC licenses before they publish their work.

5 Related Work

5.1 Commercial Image Trackers for Detecting Violations

Attributor [1], a commercial application, claims to continuously monitor the web for its customers' photos, videos, documents and let them know when those have been used elsewhere on the web. Then it offers to send notices to the offending websites notifying link request, offers for license or request for removal.

Another commercial application called PicScout [17] claims that it is currently responsible for detecting over 90% of all online image infringements detections.

5.2 Transferring Metadata with Content

There is tool called News Credit [14] developed by the Media Standards Trust with the aim of making online news transparent. Although the main purpose of this tool is not to honor the license information, by embedding microformats with some specific enhancements allow journalists to embed basic information

# of Websites	Total # of Images	Misattribution
67	426	78 %
70	241	80 %
70	466	94 %

Figure 1: Results from an experiment to see how many CC attribution license violations are there in a sample of websites

to their news articles online which helps establishing an article's authorship and provenance.

There is also some work on annotating XHTML documents with provenance metadata using RDFa [9], which presents a method for performing copy and paste operations on XHTML documents that preserves the metadata. This tool also incorporates a Creative Commons reasoning engine that reads document metadata and makes licensing decisions for annotated documents.

6 Design Proposal

6.1 Experiment to Assess the Level of Creative Commons License Violations

Since it is interesting to have an estimation of how many license violations are out there on the WWW, we hope to conduct experiments by using samples of blogs generated using the Technorati Cosmos.

6.1.1 Experiment Setup

Technorati blog indexer [21] crawls and indexes weblog-style web sites gathering lots of information. It 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 are, how popular the people that read it are, and etc. Most importantly all the technorati data are time dependent, which means that the technorati rank is based on most recent activity in a particular website.

In constructing this experiment, we obtained a random sample of websites which embedded Flickr images by using the Technorati 'cosmos' method. This method gives the blogs which link to a particular URI. Therefore, the sample is extracted from Technorati by giving the set of Flickr farm URIs [5]. Since the cosmos is dynamic, it generates a reasonably fair sample of websites linking to Flickr images. Out of this sample of websites, attribution is checked as follows:

- For each Flickr image found in the site, obtain the photo id from the URI.
- Query the Flickr API using the photo id, and get all the information related to that photo.

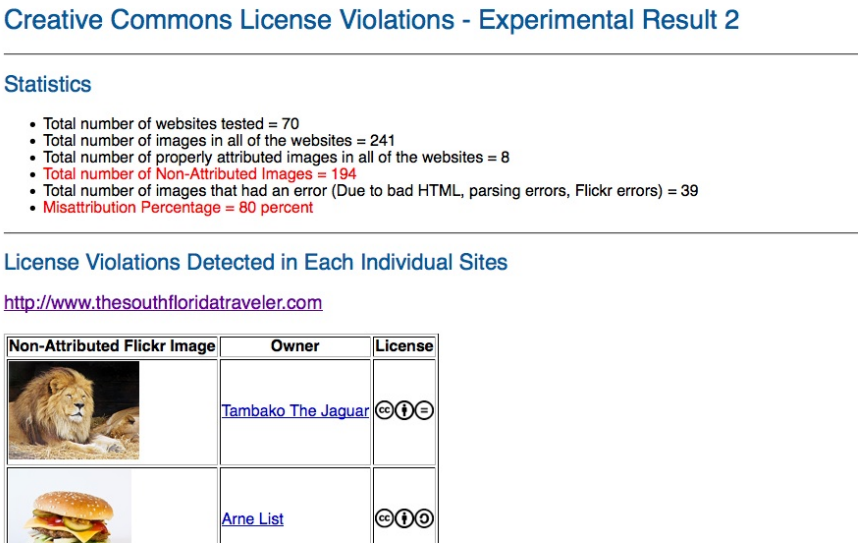

Figure 2: Results from the experiment using sample 2

- Check for the name of the original owner (as they've given to Flickr - which could be either the username or the real name) in the visible text within a reasonable scope surrounding the image node in the HTML DOM. Note the scope of checking is highly customizable, and for most cases it seems that checking within the parent node and the sibling nodes of the containing element seems to work.
- If it does not register a hit, it is taken as a License terms violation.

6.1.2 Results

The results from 3 samples of websites gathered within 2 weeks is shown in Figure 1. These results indicate that there is a strong need to have awareness among users of content on the web to check the licenses associated. Figure 2 illustrates the results from a sample run on the experiment. Since there is a high probability to have false negatives, which could be due to attribution given some where else in the DOM tree or the original creator and the owner and the re-user of the content being the same person (then not feeling obliged to attribute), etc, we propose to handpick a small sample and then finetune the results obtained from the automated experiment.

We hope to extend the same experiment as used to check for attribution license violations to investigate the extent of non-commercial use and share-alike license violations. We believe checking for non-commercial use to be very

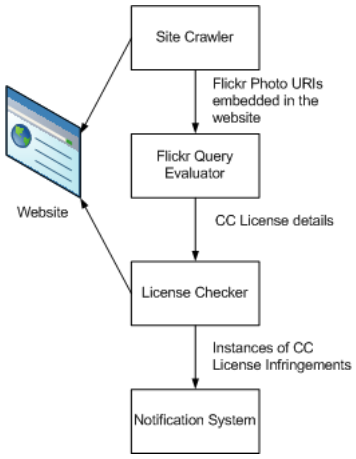

Figure 3: The Design of the Validator

tricky as the term and the question as to what comprises a commercial use is not very well defined.

6.2 CC Attributions License Violations Detector

The goal of the this tool is to check whether a particular site has embedded images (preferably from Flickr) which are not properly attributed.

As shown in Figure 3 this system has four major components.

- *Site Crawler:* This will search for all the links embedded in the given site using a Breadth-First-Search algorithm to determine embedded images. This crawler avoids straying outside of the site , but instead simply dig down into a single web page.
- *Flickr Query Evaluator:* If the Site Crawler detects any embedded Flickr images, this will extract the photo id from the Flickr URI. Using this photo id, all the information related to the photo could be obtained through the Flickr API. Typically, the license attached with an image should either be 'All Rights Reserved' or should include a CC license (which may have a combination of Attribution, NonCommercial and ShareAlike CC license terms). This module also checks to which Flickr user this photo belongs, by querying the Flickr API using the photo id, and then constructs the Flickr user URI to check for attribution.
- *License Checker:* If a photo has a CC license attached, regardless of the purpose for what it is used for, the photo should be given proper attribution. Therefore, if the Flickr Query Evaluator determines that a Flickr photo on a particular page has a CC License, it checks for the Flickr User URI constructed in the value for 'attributionURL' property, and the Flickr User Name in the 'attributionName' property.

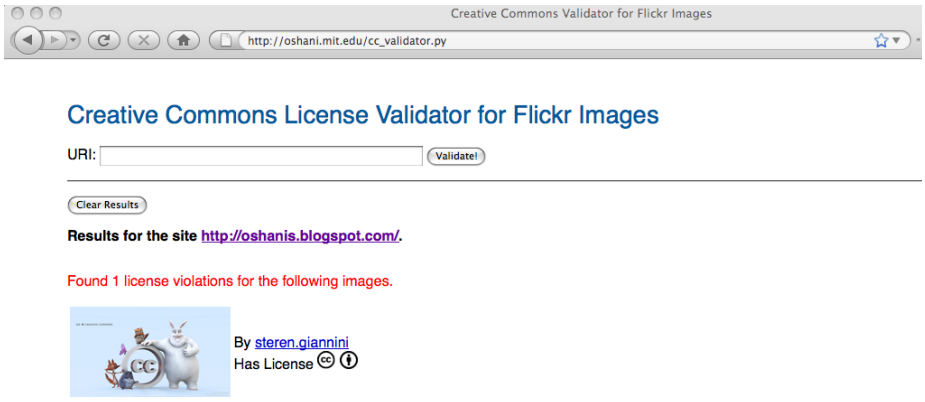

Figure 4: Output seen when an offending blog is validated using the tool

- *Notification System:* This will pretty-print and report the images which are missing attributions in a web interface. This module could be extended to provide actual notifications to the original content creator by integrating with a service provider such as Flickr.

6.3 Semantic Clipboard Application

We also hope to develop a web browser based application which lets users to reuse content with minimal effort. This will be an extension of the work done on "XHTML Documents with Inline, Policy-Aware Provenance" [9]. This "Clipboard" application will register the parts of the DOM tree which include the content that are intended to be reused along with their provenance and license information. Such addition of policy metadata will enable the transfer of content between applications very easily, and make people more aware of policies associated with content reuse.

It will be designed keeping in mind that users will want to reuse content with point and click operations which can be used to notify the application to copy the content with the appropriate license metadata as appearing on the web page. Then, these references should be available within an editor application. When the content is reused an inbuilt reasoner will determine whether it is an appropriate use, and if it is it will allow the copying of content and the license information will be automatically injected into the target document.

6.4 Project Contributions

We hope to provide an accurate assessment as to how much license violations are there on the web, and thus highlight the need to make web users more aware of

the licensing options available. The systems we propose will implement policy assurance using existing semantic web technologies. The Creative Commons License Validator will be implemented in the same spirit as the W3C XHTML validator and the like. The Clipboard application will provide users a point and click type of interaction to register the content they wish to reuse in their browser itself and then later integrate those in to their documents along with the license metadata seamlessly. This will also notify if there is an incompatibility in the type of licenses being attached. Furthermore, we also hope the work on this project will spur some ideas as to preserve data provenance and use the same concept in similar applications.

6.5 Project Milestones and Deliverables

- **Phase I:** Assessment on how much license violations are there on the web through an experiment with Flickr images
- **Phase II:** Creative Commons License Validator to verify your own work
- **Phase III:** Semantic Clipboard application to manage sources along with license data

7 Challenges

Although we believe that our solution will sufficiently address the problem of inappropriate content reuse on the web, we are not making the promise that it will handle all possible cases. In the case of checking license violations with images, a malicious user could very easily change the image URI by uploading the same image on a different server. However, in the realm of semantic web, we assume that every resource in the web is made persistent with a unique URI and therefore we overlook this problem.

8 Conclusion

We live in an era of increasing user generated content. We need tools, techniques and standards that strike an appropriate balance between the rights of the originator and the power of reuse. Building systems to support this balance would seem to be an important element in building a transparent and accountable Web.

We have demonstrated that it is possible to detect CC Attribution license violations of Flickr images on the Web. This will allow original content creators to control who is using their works and whether their licenses have been honored. The tool described in this paper is best used as a CC License validator, although it would not be impossible to build a notifier which will alert users of CC license violations on their photos. Possible extension of the tool is to incorporate complicated CC licenses to be represented in the AIR policy language and reason

out the violations. Although the work described in the paper is limited to Flickr images, it is possible to apply the same concept to other works on the web which have licenses expressed in ccREL.

References

[1] Attributor. http://www.attributor.com.

[2] Bonatti, P. A., Duma, C., Fuchs, N. E., Nejdl, W., Olmedilla, D., Peer, J., and Shahmehri, N. Semantic web policies - a discussion of requirements and research issues. In *ESWC* (2006), pp. 712–724.

[3] DOM - Document Object Model. http://www.w3.org/DOM/.

[4] Feigenbaum, J., Freedman, M. J., Sander, T., and Shostack, A. Privacy engineering for digital rights management systems. In *Digital Rights Management Workshop* (2001), T. Sander, Ed., vol. 2320 of *Lecture Notes in Computer Science*, Springer, pp. 76–105.

[5] Flickr Farm URL. http://www.flickr.com/services/api/misc.urls.html.

[6] FlickrLib. http://monotonous.org/2005/11/26/flickrlib-05/.

[7] Hal Abelson, Ben Adida, Mike Linksvayer, Nathan Yergler. ccREL: The Creative Commons Rights Expression Language. *Creative Commons Wiki* (2008).

[8] Ivan Herman. RDFa Distiller, 2008.

[9] Jones, H. C. Xhtml documents with inline, policy-aware provenance. Master's thesis, Massachusetts Institute of Technology, May 2007.

[10] JSON - JavaScript Object Notation. http://www.json.org/.

[11] Kagal, L., Berners-Lee, T., Connolly, D., and Weitzner, D. Using semantinc web technologies for policy management on the web. In *21st National Conference on Artificial Intelligence (AAAI)* (July 2006).

[12] Kagal, L., Hanson, C., and Weitzner, D. J. Using dependency tracking to provide explanations for policy management. In *POLICY* (2008), pp. 54–61.

[13] MozCC. http://wiki.creativecommons.org/MozCC.

[14] News Credit. http://newscredit.org.

[15] on exclusive rights, F. U. L. http://www4.law.cornell.edu/uscode/17/107.html.

[16] Operator. https://addons.mozilla.org/en-US/firefox/addon/4106.

[17] picScout. http://www.picscout.com.

[18] RDF - Resource Description Framework. http://www.w3.org/TR/rdf-syntax-grammar/.

[19] RDFa. http://www.w3.org/2006/07/SWD/RDFa/syntax/.

[20] Rei : A Policy Specification Language. http://rei.umbc.edu//.

[21] Technorati API. http://technorati.com/developers/api/.

[22] Tim Berners-Lee and Dan Connolly and Lalana Kagal and Jim Hendler and Yosi Scharf. N3Logic: A Logical Framework for the World Wide Web. *Journal of Theory and Practice of Logic Programming (TPLP), Special Issue on Logic Programming and the Web* (2008).

[23] Weitzner, D. J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., and Sussman, G. J. Information accountability. *Communications of the ACM* (June 2008).