EPSRC Networks for Web Science
Web Science Student Exchange Programme

STUDENT REPORT

Name

Oshani Wasana Seneviratne

Host institution of exchange

University of Southampton, UK

Dates of exchange (from/to)

July 29, 2008 – August 29, 2008 and January 1, 2009 – January 30, 2009

Host supervisor during exchange

Professor Nigel Shadbolt

Planned programme of research during exchange *(this should be taken from the submitted Web Science Student Exchange Application Form)*

My current research interest is in data provenance on the web; specifically on policy languages and their applications in order to provide explanations to end users in an intuitive manner. To that end, I have developed a graphical justification user interface in Tabulator, a Semantic Web browser which allows users to view the explanation provided by the AIR (Accountability in RDF) reasoner [1] in different views. These include:

(i) Default view in a simple rule language called N3

(ii) Justification view in a graphical layout that highlights the result of the reasoning and allows the explanation to be explored in a step-wise manner

(iii) Lawyer's view which provides an executive summary of the 5-step analysis comprising of the Issue, Rule, Facts, Analysis and Conclusion of the scenario being investigated

Justifications provided along the three dimensions mentioned above gives natural language explanations for questions about policy decisions. This includes explanations for failed results as well. The explanation generation process is separate from the regular query process, where it uses the proof tree generated by the reasoner, and the initial log and policy files in order to build up the explanation in natural language constructs.

While this works for static log and a policy files, a bigger research question as to how to apply this reasoning based on a policy file on some dynamically evolving log of data remains. It would be useful to take a proactive application of policy to user actions on the web, rather than the retrospective approach as mentioned in [1]. The interesting questions that arise would be; (i) how to gather the relevant bits of data on to a log file without causing any security and/or privacy implications and (ii) how to apply the relevant policy to such data collected within reasonable computational bounds. Once these questions are answered, we would have a system where web users are accountable for their actions, and preventive measures are taken when a policy violation is foreseen.

If I am selected for the exchange program I am interested in working in this general direction. However, I am always open to other interesting web science related research ideas as well.

--

[1] AIR reasoner - Integrated Policy Explanations Via Dependency Tracking, Lalana Kagal, Chris Hanson, Daniel Weitzner, http://dig.csail.mit.edu/2008/Papers/IEEE%20Policy/air-overview.pdf

The Report: This should include an account of all research undertaken during the exchange. Where this differs from the planned programme of research, you should include an explanation of why it was changed and an account of any extra activities carried out. Please also state how you benefited from the exchange (maximum 2000 words)

“Policy Aware Content Reuse: Detecting Creative Commons License Violations on Flickr images on the Web”

1. Introduction

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 [1]. Information accountability provides another motivation to apply policies for data usage practices [2]. On the Semantic Web, policy based systems may be implemented with reasoners on rule-based systems, where the rules represent laws, licenses, or policies that relate to the systems. For example, the AIR policy language [3] is designed to express and enforce policies to provide reliable assessments of compliance with rules and policies governing the use of information [4].

Policies governing digital content on the Web can take either the “Rights Enforcement” approach exercised by “Digital Rights Management” (DRM) or the “Rights Expression” alternative offered by the Creative Commons (CC). DRM will inhibit any inappropriate use, whereas CC will express the license information and expect the user to obey the terms expressed in those licenses. Also, for typical users wanting to share their content on-line, DRM techniques might seem overkill. On the other hand, Creative Commons (CC) provides a very clear and a widely accepted Rights Expression Language, ccREL [19] 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.

An experiment on CC attribution license violations on Flickr images revealed a violation rate of 70%-90% on the Web. Therefore, it is evident that there should be robust mechanisms for detecting license violations on the Web, and provide methods to prevent those if possible. The work carried out during this exchange program was focussed on designing a tool to accomplish this goal.

2. Assessment of CC-BY (Attribution) License Violations on the Web

This section describes an experiment conducted to detect how many web sites are using images from Flickr without properly attributing the original owner of the photo.

CC defines attribution as “you let people copy, distribute, display, perform, and remix your copyrighted work, as long as they give you credit the way you request”. The machine processable representation of attribution in RDF is given in Listing 1. Here, we refer to the photograph at “http://flickr.com/photos/janedoe/photo.jpg”, and indicate that it is under CC-BY 3.0 license. Therefore, whenever that photo is reused, it is expected that the attribution includes a pointer to the value given by the 'attributionURL', i.e. “http://flickr.com/photos/janedoe” and specify the 'attributionName' as “Jane Doe”.

```
@prefix xhtml: <http://www.w3.org/1999/xhtml/vocab#>
@prefix cc: <http://creativecommons.org/ns#>

<http://flickr.com/photos/janedoe/photo.jpg>
   xhtml:license
   <http://creativecommons.org/licences/by/3.0> ;
   cc:attributionURL
   <http://flickr.com/photos/janedoe> ;
   cc:attributionName
   "Jane Doe" .
```

Listing 1: Machine readable form of CC-BY license

Web sites used in this experiment were obtained through the Technorati Cosmos [5]. The cosmos method can be used to retrieve results for web sites linking to a given base URI. Therefore, to obtain samples for the experiment, several of the Flickr server farm URIs which have this general format “http://farm{farm-id}.static.flickr.com/{server-id}/{id}_{secret}.(jpg|gif|png)” were used. Since Flickr has several server farms, to obtain a fair sample each time the experiment was run, the base URIs were randomly generated by altering the Flickr server farm-ids. In addition to that, randomness of samples was guaranteed by running the experiment after a small time gap (for e.g. a week or two). This is because the 'authority rank' given to a web site by Technorati, and hence the results returned from the Cosmos method dynamically changes as new content gets created. The links in the Technorati Cosmos are only valid for 180 days, and if there are no fresh links coming in to a site regularly, the authority rank goes down changing the result set returned. Therefore, this factor was also used in generating a random sample of sites to check for attribution license violations.

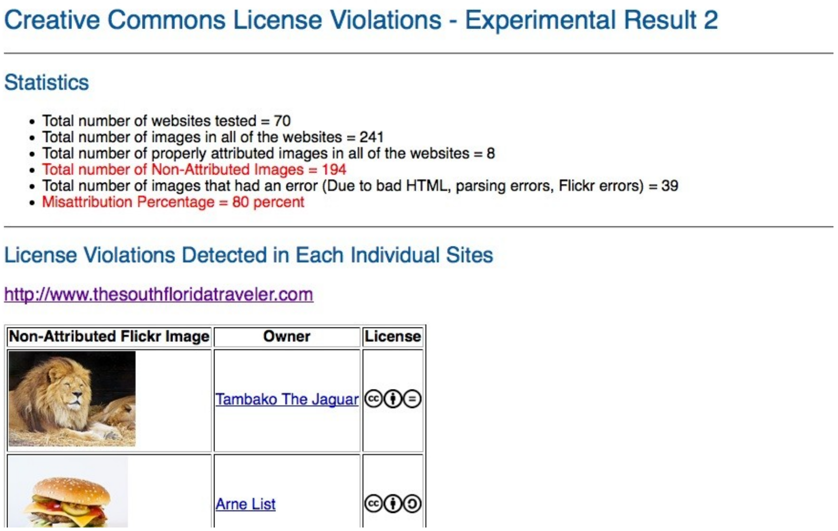
After a sample is collected, attribution for each of the images embedded in these sites were checked. Since Flickr is still using the older CC 2.5 recommendation, Flickr users do not have that much flexibility in specifying their own 'attributionURL' or the 'attributionName'. However, it is considered general practise to give attribution by linking to the Flickr user profile (attributionURL) or give the Flickr user name (attributionName) or by the least point to the original source of the image. Therefore, the criteria for checking attribution consist of looking for the 'attributionURL' or the 'attributionName' within a reasonable level of scoping from where the photograph is embedded in the Document Object Model (DOM) [6].

The algorithm used for checking attribution is given in Listing: 2.

```
1 Collect a random sample of web sites which links to Flickr farm URIs
2 For each of the web sites:
3    Isolate the elements linking to a Flickr photo
4    For each photo found:
5       id = Extract the photo ID from the URI
6       Use the id to query Flickr
7       info = owner and license information
8       Check the containing element of the photo for attribution based on info
9       If not check the sibling and parent nodes
10     If 8 and 9 fail, attribution not given
11 Pretty print results
```

Listing 2: Algorithm for checking CC-BY license violations using the Technorati Cosmos

When this experiment was run for several times the mis-attribution rate was found to range between 70%-90%. This number includes the images for which no attribution has been given, or when the image was incorrectly attributed. Figure 1 gives the the results from sample # 2.



Creative Commons License Violations - Experimental Result 2

Statistics

- Total number of websites tested = 70
- Total number of images in all of the websites = 241
- Total number of properly attributed images in all of the websites = 8
- Total number of Non-Attributed Images = 194
- Total number of images that had an error (Due to bad HTML, parsing errors, Flickr errors) = 39
- Misattribution Percentage = 80 percent

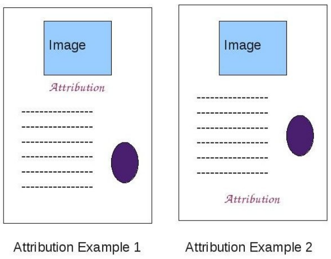
License Violations Detected in Each Individual Sites

http://www.thesouthfloridatraveler.com

Non-Attributed Flickr Image	Owner	License
	Tambako The Jaguar	🅭🅯⊜
	Arne List	🅭🅯🄎

Figure 1: Results from the experiment using sample #2

It should be noted however, that the mis-attribution results from the experiment includes cases where users have not self-attributed themselves (i.e. user uploads her photos on Flickr, and use those in her blog. Since it's her own photos, she is under the assumption that there's no need to attribute herself). A solution to this issue was hard to realize, as it is difficult to infer the web site owner from the data presented on the web site. And even if that was possible, it is hard to make a correlation between the Flickr photo owner and the web site owner. For example, the first attribution violation result shown in Figure 1 shows photos from the Flickr users “Tambako the Jaguar” and “Arne List”. People often assume pseudonyms on the web, and these two users might in fact be the same person, and the site where these particular photos are embedded “www.thesouthfloridatraveler.com” may belong to the same person. Since these connections are not explicitly stated in RDF or any other machine readable format, it becomes very hard to infer the connections.



Attribution Example 1

Attribution Example 2

Figure 2: Different ways attribution could be given

Another issue when checking for attribution is the scope in which the attribution is specified. The usual practise is to include the attribution information immediately after the content that is being attributed. For example, as shown in Figure 2, we would expect the attribution to be given as in 'Attribution Example 1' in the figure. However, since there is no strict definition of how attribution should be 'scoped', someone could also attribute as shown in 'Attribution Example 2' or it could be even buried within the text 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 work from the same original content creator, and by mistake failed to attribute some, but attributed some others.

Use of tumblelogs which cuts down the text and favours short form, mixed media posts over long editorial posts is another related problem in getting an accurate assessment of attribution license violations. For example, in a blog post where a photograph was reused, the original owner of the photograph may have been duly attributed. But when a tumblelog site such as tumblr.com pulls in the feed from that post and presents the aggregated content, the attribution details may be left out. This problem is also difficult to circumvent, because there is no standard as to how aggregation should happen or the scope with which attribution should be given in the DOM.

3. CC-BY License Violations Validator

The goal of the this tool is to check whether a particular site has any embedded Flickr images which are not properly attributed. This validator can be used in the same spirit as a web page creator would use the XHTML validator to validate the HTML mark-up.

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

1. ***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.

2. ***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 to, by querying the Flickr API using the photo id, and then constructs the Flickr user URI to check for attribution.

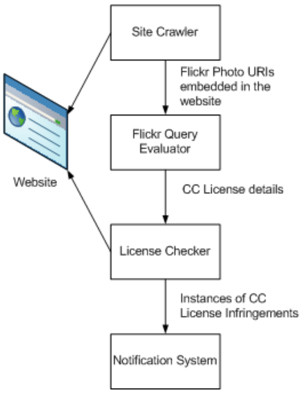
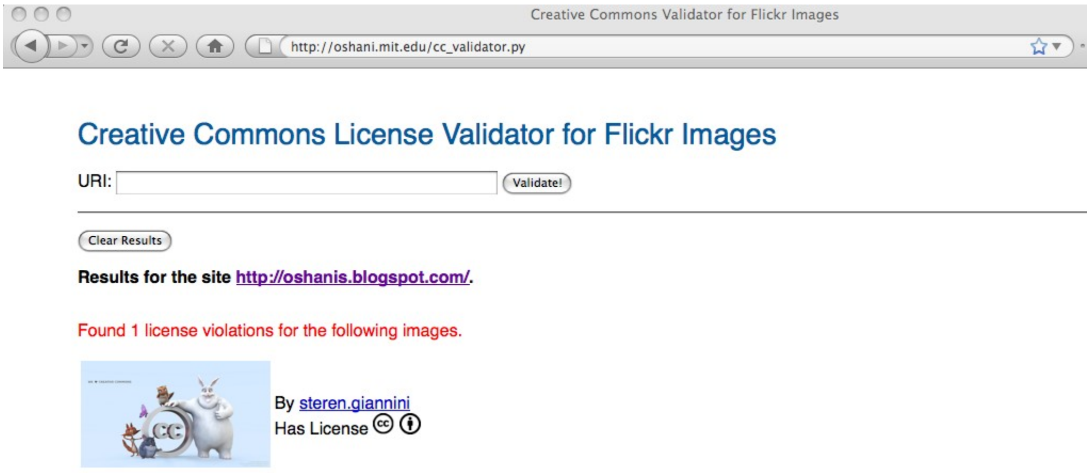


Figure 3: Design of the CC Validator

3. ***License Checker:*** If a photo has a CC license attached, according to the CC 2.5 specification, 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.

4. ***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 to Flickr or a third party service provider such as QDOS (http://qdos.com).

Figure 4 shows the result from the validator when a web site with a mis-attributed image is input to the tool. It gives the image, the name of the person who owns the image as given in the Flickr site hyper-linked to that user's Flickr account, and the CC-license attached with that particular image. The user is able to go back and correct the mis-attributions by using the information given from the validator.



Creative Commons Validator for Flickr Images
http://oshani.mit.edu/cc_validator.py
Validate!
Clear Results

Creative Commons License Validator for Flickr Images

URI:

Results for the site http://oshanis.blogspot.com/.

Found 1 license violations for the following images.

By steren.giannini
Has License

Figure 4: Output from the validator from a site with a non-attributed image

4. Model Using the AIR Policy Language

The CC license violations validator service mentioned earlier will help an "honest user to remain honest" by validating her work. But imagine a situation in which the content transfers from one individual to another individual, and then on to another individual. During the whole transfer process, the content may be altered in some way, and a new set of policies could be added. In such a scenario, the original content creator may be interested in following the trail of transactions involving the work that she created, and figure out if her work is used in any way that she did not intended it to be used. Also, it would be useful for her to see if any policy violations have taken place to take any corrective action if necessary. To realize such a system, we need to use more expressive policy languages. The new CC recommendation, ccREL [19], has an expanded set of license options. Accountability In RDF (AIR) policy language [3], supports explanations for policy decisions and provides efficient and expressive reasoning. By combining ccREL and AIR, we are able to express the policies that represent rules stating how it should be reused. Given that we have a transaction log indicating how the content was used, we can reason over that using the AIR policy which describes the CC license.

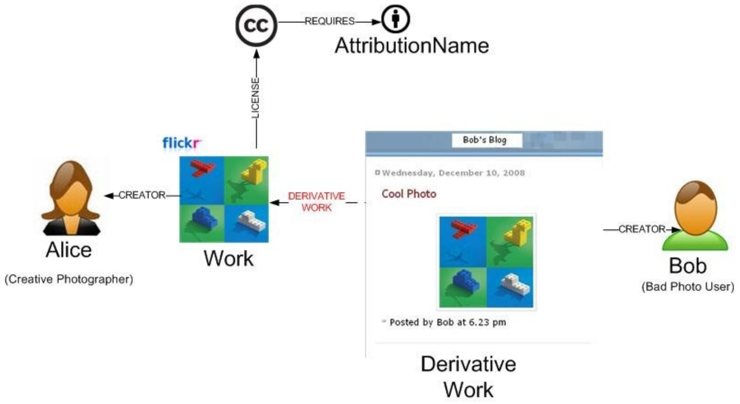


Figure 5: Content Use Scenario

As an example, consider the scenario given in Figure 5, where Alice, a content creator, takes a photograph and uploads to Flickr under the CC-BY license. Then Bob comes along and uses that photo in a derivative work without attributing Alice. This is clearly a violation of Alice's original CC-BY attribution license. Assuming that all the transactions composing this event gets logged, we can use the following AIR policy against the transaction log to derive a conclusion and a set of justifications for that conclusion.

The AIR policy for CC-BY given below first defines several prefixes such as 'cc', 'xhtml', 'dc' and 'air'. These prefixes are merely notational conveniences to refer to concepts defined in other documents. Several variables such as ':EVENT', ':P1' are introduced using the @forAll directive. Then the policy description specifies a rule which indicates that upon matching a pattern where the 'cc:attributionName' is not present in an instance of reuse of the content, then it should be a CC-BY license violation.

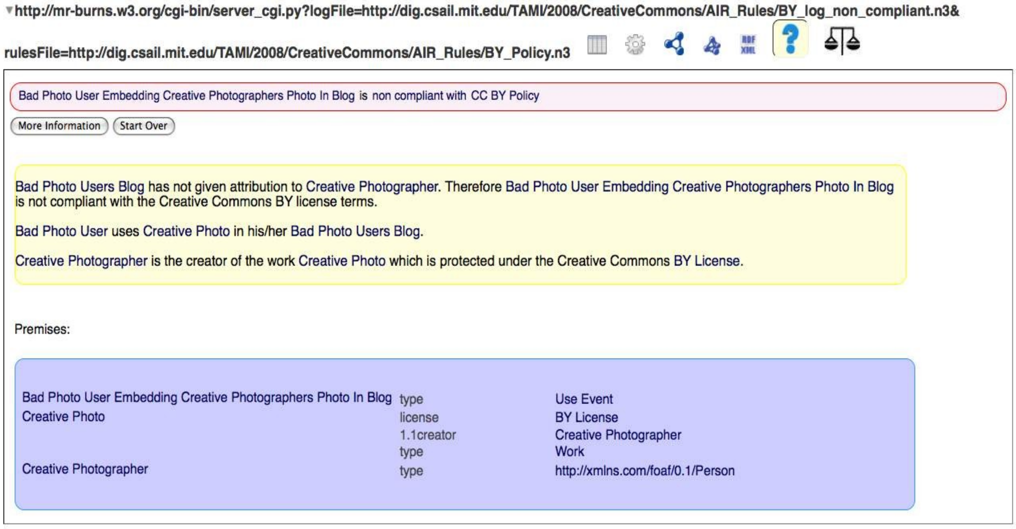
```
@prefix cc: <http://creativecommons.org/licenses/by/3.0>.
@prefix xhtml: <http://www.w3.org/1999/xhtml/vocab#>.
@prefix dc: <http://purl.org/dc/elements/1.1>.
@prefix air: <http://dig.csail.mit.edu/TAMI/2007/amord/air#>.

@forAll :EVENT, :P1, :P2, :WORK, :LICENSE, :DERIVATIVE.

:CC_BY_Policy a air:Policy;
   air:rule [
      air:pattern {
         :EVENT a air:UseEvent;
             cc:work :WORK.
         :P1 a foaf:Person.
         :WORK dc:creator :P1;
             xhtml:license :LICENSE.
         :LICENSE cc:requires cc:AttributionName. };
    air:rule [
      air:pattern {
                  :DERIVATIVE cc:derivativework
                  :WORK;
                  dc:creator :P2 . };
          air:rule [
              air:pattern {
                :DERIVATIVE cc:attributionName :P1 . };
             air:assert {
                :EVENT air:compliant-with
                :CC_BY_Policy. };
             air:alt [
                air:assert {
                :EVENT air:non-compliant-with
                :CC_BY_Policy. };
            ];
         ];
      ];
  ].
```

Listing 3: AIR Policy for CC-BY license

When this policy is run against the log of events collected from Alice's and Bob's transactions using the AIR reasoner, we would obtain a result which indicates if there were any policy violations and the justifications behind those. When you view the outcome on the Tabulator's [20] 'Justifications' pane, the final conclusion will be given first, followed by the reasons behind the conclusion which ensures that the user can view the annotated transaction log in a meaningful manner. This is shown in Figure 6.



▾http://mr-burns.w3.org/cgi-bin/server_cgi.py?logFile=http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/AIR_Rules/BY_log_non_compliant.n3&rulesFile=http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/AIR_Rules/BY_Policy.n3

Bad Photo User Embedding Creative Photographers Photo In Blog is non compliant with CC BY Policy

Bad Photo Users Blog has not given attribution to Creative Photographer. Therefore Bad Photo User Embedding Creative Photographers Photo In Blog is not compliant with the Creative Commons BY license terms.

Bad Photo User uses Creative Photo in his/her Bad Photo Users Blog.

Creative Photographer is the creator of the work Creative Photo which is protected under the Creative Commons BY License.

Premises:

Bad Photo User Embedding Creative Photographers Photo In Blog	type	Use Event
Creative Photo	license	BY License
	1.1creator	Creative Photographer
	type	Work
Creative Photographer	type	http://xmlns.com/foaf/0.1/Person

Figure 6: Output from the reasoner using the Justification UI in the Tabulator

5. Related Work

5.1 Commercial Image Trackers for Detecting Violations

Attributor[7], 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[6] claims that it is currently responsible for detecting over 90% of all on-line image infringements detections.

5.2 License Detection Tools

There are several CC license detection tools including the MozCC[9] and Operator[10] which could be installed as Firefox extensions. Once installed these applications will notify if a given web page has any CC-licensed works. This enables a user to be aware of the license terms, and use the CC licensed works appropriately. In addition, most popular search engines including Google, Yahoo and even sites such as Flickr[11], blip.tv[12], OWL Music Search[13] and SpinXpress[14] have advanced search options to find any CC-licensed content on the web.

5.3 License Embedding Tools

There are several tools which can be used to automatically embed the license meta-data in certain applications such as: ThinkFree[15] - a web based commercial office suite, License Tagger[16] – a cross platform application to add license meta-data in many content files. There is also an implementation on Flickr image re-use for OpenOffice.org[17] which allows a user to directly pick an image from the Flickr website and automatically inject the license meta-data along with it. News Credit[18] developed by the Media Standards Trust with the aim of making on-line news transparent, uses micro-formats to allow journalists to embed basic information to preserve the provenance of the information including the license meta-data.

6. Choice of this Project

I was motivated to explore this particular topic because of an incident that happened during the 'Summer Doctoral Programme' (SDP) held at the Oxford Internet Institute just few weeks before I started the exchange programme at University of Southampton. I participated in the SDP along with 28 other doctoral students, and after the programme ended, some of the students started sharing the photographs taken during the SDP by posting links to those in the mailing list. Few students, including myself, started uploading these photographs on social networking sites such as Facebook. This annoyed a particular student, who had given the Creative Commons 3.0 BY-NC license to his photographs, which meant those cannot be used for any commercial use, and whenever republished, attribution has to be given to him. Since all the students were connected through the social networking site, he was able to see who has uploaded his photographs without attributing him. Of course, the 'license violators' had no intention of actually violating the original license. It was merely an oversight on their part where they ignored the license or was not aware of what it meant and how it should be used. This made me think that there should be a tool in between the DRM and the ccREL continuum, to help users to reuse content in a policy aware manner.

As was stated in the planned programme of work, I was interested in investigating real time application of policies on user actions on the Web. However, this idea seemed too broad to cover within the period of the exchange programme. Therefore, with the help of Prof. Shadbolt, I narrowed the original idea to conduct research on detecting license violations on image reuse, explore the social awareness on policies and thus evaluate the effectiveness of the methods that enable rights expression on the Web.

7. Benefits from the Exchange

I undertook the exchange program in two phases. During the first phase, the idea was fleshed out by implementing a prototype system as a proof of concept. During the period in between the two phases, I was able to get feedback from my advisers and peers back at my home institution, MIT, and refine the idea a bit further. Finally in the second phase, I did an assessment of how much of a problem there is on the Web when it comes to CC license violations, and how effective the tool I developed would be.

It was a good opportunity to do research in a different environment but still be connected to my core research at MIT. I was also very fortunate to meet exceptionally bright people at Southampton, whom I was able to discuss my ideas with, and get very valuable feedback from. I presented part of this work at the Creative Commons Technology Summit 2009 [21] held at Cambridge, USA on 12th of December 2008, which gave me an opportunity to showcase the tool to the CC community to get their opinions as well.

In retrospect, I firmly believe that this exchange has helped me immensely in my research, allowed me to network with peers from another institution, and last but not least, enabled cross fertilization of a very nice idea into fruition.

8. Conclusion

The work done on this exchange program includes a study on the level of policy awareness among users on the Web when it comes to reusing content. The study and the tool focussed on photographs from Flickr as the 'content', and Creative Commons licenses as the 'policy' governing the reuse. From the experimental results we can conclude that there is a considerable problem when it comes to attributing content as stated in the Creative Commons license the original work is under. This could be due to the ignorance of the users about what each of the license terms means, or it could be because there are not many useful tools to automatically embed the license meta-data along with the content when reusing. As an intermediate solution to the problem, a Creative Commons License Violations Validator was implemented to aid a user to validate any work that includes photos from Flickr.

9. Tidbits

The source code which was used to run the experiment can be found at:
http://dig.csail.mit.edu/2008/WSRI-Exchange/src/experiment.py

The entire set of results from the experiment is available at:
http://dig.csail.mit.edu/2008/WSRI-Exchange/results/

The CC-license violations validator service mentioned in the paper can be found at:
http://dig.csail.mit.edu/2008/WSRI-Exchange/src/cc_validator.cgi

The video of that talk I gave at the CC Tech Summit is available at:
http://www.archive.org/details/cc-techsummit-200812-video

The presentation used for that talk is available at:
http://dig.csail.mit.edu/2008/Talks/1212-CCTechSummit-os

10. Acknowledgements

I would like to thank Prof. Nigel Shadbolt for guiding me all throughout the project, Harith Alani for supervising me during the first phase of the project, and my friends at Southampton including Albert Au-Yeung, Ilaria Liccardi and Asma Ounnas for their comments and suggestions. Many thanks to Susan Davies for helping me in numerous ways; ranging from finding accommodation to scheduling meetings with Prof. Nigel, and overall making sure that my exchange was going smoothly. I would also like to thank my advisors at MIT: Prof Sir Tim Berners-Lee, Daniel Weitzner and Prof. Hal Abelson for encouraging me to go on this exchange program, and giving their feedback on the project.

11. References

[1]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.

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

[3] AIR Policy Language. http://tw.rpi.edu/proj/tami/AIR_Policy_Tutorial

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

[5] Technorati Cosmos. http://technorati.com/developers/api/cosmos.html

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

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

[8] PicScout. http://www.picscout.com.

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

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

[11] Flickr. http://www.flickr.com

[12] blip.tv. http://blip.tv/

[13] OWL Music Search. http://www.owlmusicsearch.com/

[14] SpinXpress. http://spinxpress.com/

[15] ThinkFree. http://www.thinkfree.com/

[16] License Tagger. http://wiki.creativecommons.org/License_tagger

[17] Flickr Image Re-Use for OpenOffice.org
http://wiki.creativecommons.org/Flickr_Image_Re-Use_for_OpenOffice.org

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

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

[20] Tabulator. http://dig.csail.mit.edu/2007/tab/

[21] Creative Commons Technology Summit, MIT, USA held on December 12th, 2002. http://wiki.creativecommons.org/Creative_Commons_Technology_Summit_2008-12-12