Detecting Creative Commons License Violations on Images on the World Wide Web

Oshani Seneviratne
CSAIL, MIT
oshani@csail.mit.edu

Lalana Kagal
CSAIL, MIT
lkagal@csail.mit.edu

Daniel Weitzner
CSAIL, MIT
djweitzner@csail.mit.edu

Hal Abelson
CSAIL, MIT
hal@mit.edu

Tim Berners-Lee
CSAIL, MIT
timbl@w3.org

Nigel Shadbolt
ECS, University of Southampton, UK
nrs@ecs.soton.ac.uk

ABSTRACT

Creative Commons is a non-profit organization that has been striving to provide 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 Web, while ensuring that the culture of reusing existing works to foster creativity is not hindered. There are many online tools in photo sharing sites such as Flickr and EveryStockPhoto that generate Creative Commons license information associated with their content in machine readable form. This information is generally included in the metadata of the content. However, not much effort has been on actually detecting whether a license has been violated or not. This paper describes of a method for identifying whether a particular kind of Creative Commons license, namely attribution, has been violated, with a focus on Flickr images on the Web.

Categories and Subject Descriptors

1.2 [**Artificial Intelligence**]: Law; H.3.4 [**Information Systems**]: User Profiles and Alert Services; K.5 [**Computing Milieux**]: Legal Aspects of Computing; H.3.5 [**Information Systems**]: Web-Based Services

General Terms

Algorithms, Design, Human Factors, Languages, Legal Aspects, Verification

Keywords

World Wide Web, Semantic Web, License Violation Detection, Validation Service, Policy Compliance, Creative Commons Rights Expression Language (ccREL), RDFa, Accountability in RDF (AIR)

1. INTRODUCTION

The Web is a platform in which users can share their work very effectively. Due to the nature of the medium, content on the Web including text, images, and videos can be reused and remixed rapidly. Scientific research data, social networks, blogs, photo sharing sites and other applications known collectively as the Social Web, and even general purpose websites have lots of increasingly complex data. Data from several web pages could be very easily '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 such violations.

We start by reviewing terminology that will be used throughout this paper. A "work" is any material that may be licensed under Creative Commons (CC). "License Terms Violation" is the act of making use of a CC-licensed work in a way that violates the rights expressed by the original creator. "Creative Commons Rights Expression Language (ccREL)" [8] is the standard recommended by the CC for machine readable expression of copyright terms and licensing. A '"secondary content creator" is a person or an automatic tool which uses previous work in creating another work.

Information about the Creative Commons license of a particular work is usually specified as RDFa [17] on a web page. RDFa allows machine understandable semantics to be added to XHTML. There are many tools that extract RDF [16] from web pages marked up with RDFa, by using, for example, the W3C's RDFa Distiller [9] and provide license awareness when reusing works on the web as described in Section 2.3. However even with all the tools and licenses designed to warn users of their accepted use, permissions and restrictions, we can expect secondary content creators to fail to take the CC licenses into account while creating new content. This could be due to many factors: They may be ignorant as to what each of the licenses mean; It could also be that they somehow forget to check and include the proper license terms in their own work; Or they may simply give an incorrect license which violates the original content creators intention. We should also not forget that some secondary content creators will use previous works with malicious intent, intentionally ignoring the CC-license. Whatever the case may be, the original content creator would be interested in knowing when his/her licenses have been violated and on which web pages. But given the scale of the World Wide Web, the knowledge of such a license violation is highly unlikely, unless the original content creator comes across it by chance. Therefore, we expect original content creators to appreciate and welcome tools that will detect if their works have been used inappropriately.

There is another orthogonal case to this problem that can

WWW2009, April 20-24, 2009, Madrid, Spain.
.

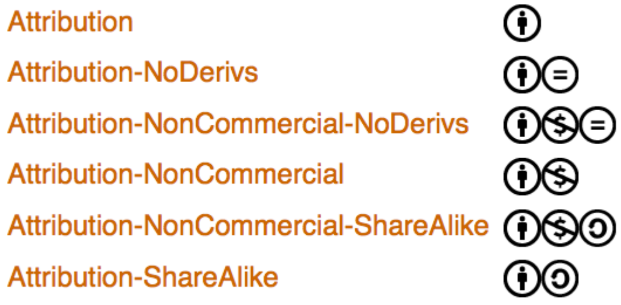
be solved by using the same method. Authors of works that may use several hundred other sources would be interested in knowing whether they have violated anybody else's CC license terms; For example, by failing to keep a proper citation list or by mis-attributing. In such a case, a 'validator' which checks for CC license violations of content on your website or work would be very useful. This is in the same spirit as Web developers checking whether their HTML is valid using the "W3C Markup Validation Service" or semantic data producers checking if their data is in proper RDF [16] by using the ”W3C RDF Validation Service". Using these tools secondary content creators can rectify the instances where they have inadvertently violated the CC licenses before they publish their work.

Although the intent of CC is not to implement an enforcement mechanism, we demonstrate that it is possible to detect violations for some CC license types on licensed content on the Web. This paper describes a tool that can be used to detect missing attribution details of Flickr images embedded in any given website.

2. BACKGROUND

2.1 Creative Commons (CC)

CC has defined licenses that permit restricted sharing and reuse that are clearly communicated in human readable form as well as in machine readable form. These set of licenses are simple, uniform and understandable. By reducing the size of the license space from a virtually infinite number to roughly a dozen licenses, CC has given end users a much clearer picture of what their rights are [8]. Figure 1 shows some of the human readable representations for few CC Licenses. For example, an ”Attribution-NonCommercial-ShareAlike” license indicates that the work should be attributed to the original creator, should not be used for any commercial use, and all the derivative works must be licensed under the same license.



Attribution

Attribution-NoDerivs

Attribution-NonCommercial-NoDerivs

Attribution-NonCommercial

Attribution-NonCommercial-ShareAlike

Attribution-ShareAlike

Figure 1: Some Common CC-Licenses

ccREL [8] is the specification for machine readable expression of copyright licensing terms and related information recommended by CC. This recommendation provides flexibility for content generators to express their licensing requirements beyond the pre-defined CC licenses. CC licenses are specified in RDFa [17], which is a method of embedding structured RDF in XHTML documents with minimal additional markup.

Our tool specifically looks at attribution details, as this is an important step in detecting CC license violations because all CC licenses contain this property by default, unless specified otherwise. We define the meaning of ”Attribution” in CC terminology as follows. Attribution: *”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 readable form of CC-BY (Attribution) license in Notation 3 [20], when applied to an image is as follows:

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

<http://flickr.com/photos/alicesmith/random_number>
   xhtml:license
   <http://creativecommons.org/licences/by/3.0> ;
   cc:attributionURL
   <http://flickr.com/photos/alicesmith> ;
   cc:attributionName
   "Alice Smith" .
```

2.2 Flickr

Flickr.com is a popular photo sharing site. As of October 2008, Flickr hosts approximately 70 million CC-licensed images. The default license for any photo uploaded on Flickr is 'All Rights Reserved'. This means that the original creator does not allow any reuse of the photo and only allows others to view it on the Flickr site. However, users can change this default settings to attach CC licenses [5], as shown in Figure 2.

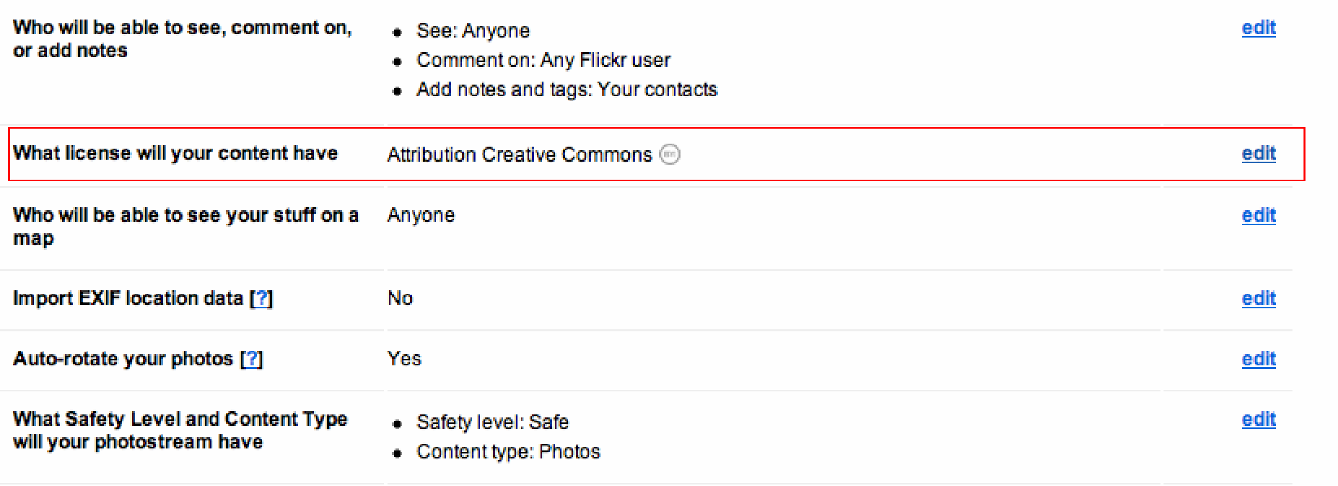
Currently Flickr denotes a license on each page containing images with a link to the relevant license qualified by rel=”license" as shown in Figure 3. However, this approach breaks down when multiple images are viewed on a single page, or when further information, such as the photographer's name, is required. Although ccREL provides a solution to this problem by recommending special properties called 'attributionName' and 'attributionURL', Flickr has not yet implemented a mechanism for it's users to customize these settings related to attribution.

Some rights reserved

Figure 3: Rights expressed in a Flickr Image. ***(Here the owner has specified the rights for the photos to indicate that they are licensed under CC Attribution license.)***

The Flickr API [4] provides a very convenient method to access the photos with several different public methods, as well as the information associated with the images such as tags, geo, EXIF, etc. The API also lets third party application developers to do almost everything that flickr.com can do. The most important functionality offered from the point of view of the tool described in this paper, is the ability to access the metadata for an image which includes license information.

Once a photo is uploaded using Flickr's upload web service, the image is given a URI. The URI takes one of the formats shown in Figure 4. It is possible to identify the photo id, as well as other identifying information related to



Who will be able to see, comment on, or add notes	• See: Anyone • Comment on: Any Flickr user • Add notes and tags: Your contacts	edit
What license will your content have	Attribution Creative Commons	edit
Who will be able to see your stuff on a map	Anyone	edit
Import EXIF location data [?]	No	edit
Auto-rotate your photos [?]	Yes	edit
What Safety Level and Content Type will your photostream have	• Safety level: Safe • Content type: Photos	edit

Figure 2: Changing the Default License Settings to Creative Commons License for Images on Flickr

the photo such as the Flickr's server id where it is hosted, secret number, size and the file format of the photo, using this photo URI alone, without querying Flickr.

```
http://farm{farm-id}.static.flickr.com/{server-
id}/{id}_{secret}.jpg
or
http://farm{farm-id}.static.flickr.com/{server-
id}/{id}_{secret}_[mstb].jpg
or
http://farm{farm-id}.static.flickr.com/{server-
id}/{id}_{o-secret}_o.(jpg|gif|png)
```

Figure 4: Format of Flickr Image URIs

The tool we describe checks whether the images are properly attributed to the original content creators as specified in the license for the Flickr image. The key to query all these information is extracted from the Flickr image URI.

2.3 Related Work

Distribution and usage of copyrighted content is usually controlled by Digital Rights Management (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 rent period has ended. These schemes not only affect user privacy [10] but most DRM systems have been circumvented when deployed widely. On the World Wide Web, a more flexible approach to handling copyrighted content is required. 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. 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. Essentially the DRM approach to copyright control is very difficult and frustrating. We advocate expressing licensing rights as required for your works in CC, and demonstrate that violations of these rights can be identified making the use of CC licenses as a viable alternative.

CC has put much focus on coming up with ways to enable tool builders to use the CC licenses very effectively. Along these lines, there are currently several Mozilla Firefox extensions that are CC-License aware. MozCC [14] is one such tool. It provides a specialized interface for ccREL, and the user would receive visual cues when he/she encounters a page with RDFa metadata. 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 [15] 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. Both these methods will only assist a person intending to reuse the content on a web page, and not detect any violations of CC licenses.

Flickr also provides a search interface to search it's photos by the most of the common CC-license types [5]. This allows users to filter out photos that they can reuse. This is a nice method to make the user aware of the images which have different CC licenses. However this too does not prevent a user from using CC licensed images in a manner which violates the original content creators rights.

Several CC sponsored projects have also tried handling the reuse of CC-licensed images in applications. LibLicense [2] provides a low level license metadata integration for applications. There are several LibLicense based implementations such as License Tagger [13] - which is a cross-platform application to add license metadata files, and Flickr image reuse for openoffice.org [6], which automatically injects the CC license metadata whenever the image is embedded in an OpenOffice.org document.

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. We have not used the proprietary Attributor system, but we hope to provide a free and open source alternative, comparable, if not better than the solution they claim.

Most of the work in the literature seem to revolve around the task of providing tool and methods to reuse CC licensed images. Our approach is orthogonal to these solutions, because we aim to provide an effective method to enforce CC licenses by detecting certain kinds of license violations.

3. MOTIVATING SCENARIOS

We now illustrate few scenarios for which the tool described in this paper would prove to be useful.

3.1 Notify when somebody violates your CC-licenses

Suppose Alice is an avid Flickr user and she uploads her photos to her account regularly. In her Flickr account settings she has applied "CC-BY-2.0" to all her photos. This means she allows anybody to reuse her photos as long as they properly attribute her as given in her CC license terms. The following RDFa [17] snippet shows how anybody should attribute Alice if they are using Alice's photo.

```
This photo is licensed under a
< a rel="license"
   href="http://creativecommons.org/licenses
   /by-nc-nd/2.0/">
   Creative Commons license
</a>.
If you use this photo within the terms of the
license or make special arrangements to use
 the photo, please list the photo credit as
<span property="cc:attributionName">
Alice  Smith</span>
and link the credit to
<a rel="cc:attributionURL"
   href="http://flickr.com/photos/alicesmith">
   http://flickr.com/photos/alicesmith
</a>
```

(Note that currently Flickr does not allow users to add much finer grain control to change the 'attributionURL'. It simply displays the CC license type and implicitly assumes that the 'attributionURL' is in fact the Flickr user's account URI.)

Now, suppose that Bob, an avid blogger, sees one of Alice's Flickr photos that interests him. He embeds that photo in his blog in a posting, but forgets to check the license attached to that photo and attribute Alice by giving the 'attributionName' and the 'attributionURL'. In this scenario, unless Alice sees Bob's blog post, there is no automatic way to detect the license violation. Although not implemented, it is relatively straightforward to provide an 'alert system' to notify users of such a violation, provided that they are subscribed to a notification service.

There could be larger consequences if Alice's CC license was in fact "CC-BY-NC-2.0", which prohibits the photo to be used for commercial use, and has to be properly attributed to Alice. Suppose Bob uses this particular photo in an online advertisement. Then in addition to checking for the missing attribution details, it should also be possible to detect that the depiction of the image is actually within an advertisement and is used for a commercial use. It should be mentioned that violations of this nature are not handled using the tool mentioned in this paper. As of now, we only provide a solution for detecting missing attributions.

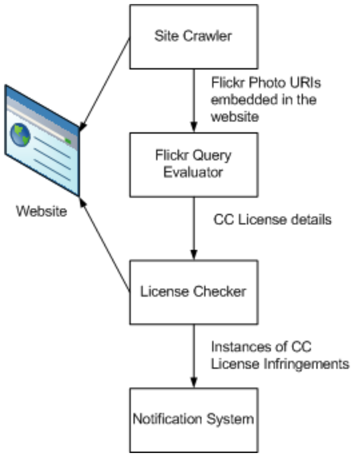


Figure 5: Creative Commons License Validator

3.2 Verify the your own work before publishing

When an author aggregates content from many different sources, it is inevitable that some attribution details are accidentally forgotten. Science Commons [18], which facilitates an open protocol for promoting the reuse of scientific data will benefit greatly from such a validator, which can establish that the material used in a derivative work is with appropriate reference to the originating licensing conditions.

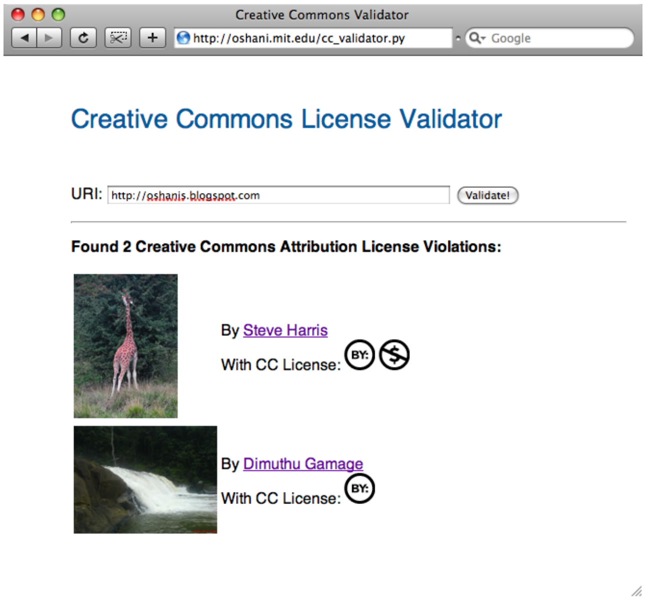
In order to make sure that no CC license terms are violated, the author can run the CC License Validator and see if some sources have been left out or whether some have been mis-attributed. The same principle could be applied to any compilation of a work using various different sources on the web. We can expect some people to check for CC license violations as they would use the W3C markup validation service to verify their HTML to be valid.

4. FLICKR-CC LICENSE VALIDATOR

The goal of the Flickr CC License Validator, is to check whether a particular site has Flickr image instances that are not properly attributed.

As shown in Figure 5 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 and determine if there are any embedded Flickr photos. This crawler avoids straying outside of the site for safety reasons as well as for efficiency reasons, but instead simply dig down into a single web page looking for embedded Flickr images. In order to follow the links that are provided within the web page, the crawler first parses the HTML pages and identify links to other resources. Then it queues them to a 'to-visit' queue, and then repeats this process using the first item from the 'to-visit' queue. As a link is checked, any new links that are found are loaded onto the same queue. An 'already-viewed' queue is also maintained to avoid digging into any link the crawler has seen in


Creative Commons Validator

http://oshani.mit.edu/cc_validator.py

Creative Commons License Validator

URI: http://oshanis.blogspot.com Validate!

Found 2 Creative Commons Attribution License Violations:

By Steve Harris
With CC License: BY:

By Dimuthu Gamage
With CC License: BY:

Figure 6: Creative Commons License Validator Interface showing few of the License Violations it has detected in a Blog.

In this instance, the validator has detected two CC license violations in this particular blog. It reports by giving the photo, the name of original creator hyperlinked with the Flickr User URI, and the CC license that was attached to the photo in a human readable form (using icons that are hyperlinked to the CC Deed).

the past. This results in breadth-first traversal and the crawler carefully avoids moving to another site by not following non-local links.

- *Flickr Query Evaluator:* If the Site Crawler detects any embedded Flickr Photos, this module will extract the photo id from the Flickr URI based on the URI format given in Figure 4. Using this photo id, all the information related to the photo is obtained by calling several methods in the Flickr API. This information also includes the original creator's Flickr user account id, name and CC license information pertaining to the photo. It then queries the Flickr API for the CC License details using the FlickrLib python library [7]. The response from Flickr is obtained in JSON format [11], and after parsing that for the relevant license we can determine the license attached to the photo. The license should be either 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. There is also a hook to override checking the attribution details inside RDFa in the HTML. This is because the user can specify the attribution name and the URL in the web page without restricting it to be strictly in the RDFa. Whatever the approach taken, only the content within the containing DOM [3] node containing the image is checked for the attribution details. This is to avoid accounting for 'phantom' attribution. For example, if there are two images with the same attribution details located in different levels of scope in the HTML, and only one image is properly attributed, the License Checker will not say that both images are properly attributed.

- *Notification System:* Currently this tool will only report the photos which are missing attributions in a web interface as shown in Figure 6. We hope to integrate actual 'notification' capabilities so that it will notify the original content creator of the violation. It is then up to them to take remedial action (for e.g. ask the violator to properly attribute or ask them to take down the photo, etc.).

The validator described in this paper is hosted at:

`http://oshani.mit.edu/cc_validator.py`

Any offending website or blog could be given in the URI box to run the validator. If it finds any CC violations, it will report as shown in Figure 6.

5. CC LICENSE MODELS IN AIR

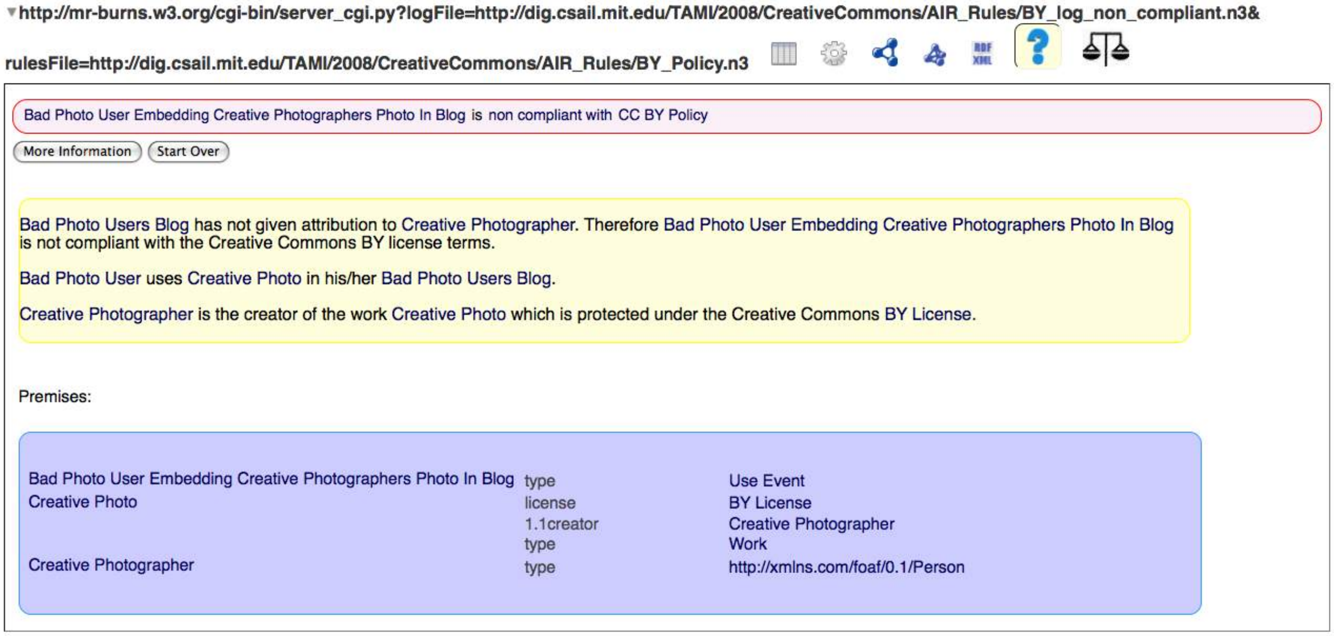
The Flickr-CC validator outlined in Section 4 is limited to checking for CC license violations by finding out missing attribution details. However in order to model other license violations greater expressivity is needed. Accountability In RDF (AIR) policy language [12] provides a robust framework for expressing policy compliance in a variety of environments. Thus the scenario for detecting CC licenses can be easily modeled in AIR. When given an appropriate transaction log which details the events that lead to the license violation, along with the AIR policy, the AIR reasoner (implemented within the bounds of the AIR policy language) will provide a final outcome with the justification behind the reasoning.

Since the CC licenses are expressed in RDFa, which is machine readable, it is relatively straightforward to convert the CC license and express it in an AIR policy written in N3 [20].

The AIR policy for CC-BY license as expressed in ccREL is as follows:

```
@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.
```



▾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 7: CC License Violation using the AIR reasoner as viewed on the Justification User Interface

```
: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.
      };
    air:rule [
      air:pattern {
        :DERIVATIVE cc:derivativework
        :WORK;
        dc:creator :P2 . };
          air:rule [
              air:pattern {
                :DERIVATIVE cc:attributionURL :P1 .
                };
             air:assert {
                :EVENT air:compliant-with
                :CC_BY_Policy. };
             air:alt [
                air:assert {
                :EVENT air:non-compliant-with
                :CC_BY_Policy. };
            ];
         ];
      ];
  ].
```

(This AIR policy defines ’CC_BY_Policy’. This has a rule that tries to match the pattern for any event ’EVENT’, which is of type air:UseEvent, and has some cc:work ’WORK’, by some foaf:person ’P1’, licensed under the xhtml:license ’LICENSE’. Provided that this pattern is matched, it fires another rule which checks whether this work is used in another derivative work ’DERIVATIVE’ by some other person ’P2’. If this derivative work has given the proper cc:attributionURL of the person ’P1’, then we assert that the ’EVENT’ is compliant with ’CC_BY_Policy’, else we assert that it is non-compliant with the ’CC_BY_Policy’.)

However, this generic conversion from ccREL to AIR policies might not work all the time because ccREL also allows a work to be licensed with ’cc:additionalPermissions’.

This property can specify any other licensing terms including commercial licenses or it could even specify that no attribution is needed at all. In such a case, it is important to have a customized license generating mechanism in addition to the generic type of license generation mentioned above. Therefore, as illustrated in Figure 8 we use a Policy Generator for CC Licenses to generate AIR policies. An RDFa scraper looks at the terms with which a work is licensed under, and combine it with the CC license deed to generate an AIR policy. The user is then able to customize the policy.

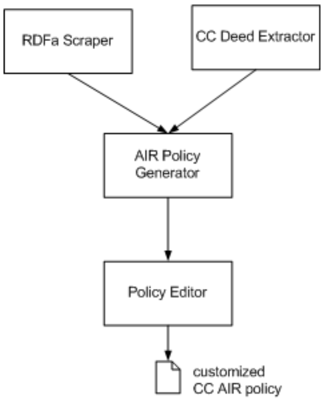


Figure 8: Policy Generator for CC Licenses

In order to 'detect' a violation, and to provide a justification behind the reasoning, the AIR reasoner requires the AIR representation of the CC License as well as the the facts about the work and its usage on the scraped Web page. It will reason over this information to deduce whether or not a violation occurred and will generate an explanation behind the reasoning. This 'proof' of violation can be viewed in the Tabulator [19], as shown in Figure 7 using a custom application called the Justification User Interface, which was built using the modular *'Pane system'* in the Tabulator.There is a Firefox sidebar extension which will invoke the reasoner given the URIs of the relevant log file and the policy file. One could also invoke the reasoner by appending the proper parameters to the URI of the reasoner. The explanations are produced in the form of proof trees in RDF/N3, and indicates whether the given facts comply with the policy specified. It also gives the premises on which this decision was based on.

Though the AIR representation for the CC-BY license is currently straightforward, we can represent other more complex licenses in similar manner and reason over their violations.

6. LIMITATIONS OF FLICKR-CC

ccREL allows a publisher of a work to give additional permission beyond those specified in the CC license with the use of the 'cc:additionalPermissions' property to reference commercial licensing brokers, and a 'dc:source' to reference parent works. Therefore a document with a CC license that requires attribution may be usable without attribution [8]. Flickr has not yet implemented this feature to let the users control their rights on the individual photos, photo sets, or their entire collection of photos. Thus the tool described in this paper is not capable of exploring the extra license information to determine whether attribution should be verified using either the 'cc:attributionName' or the 'cc:attributionURL' or both or some other license or whether attribution should be checked at all.

Another limitation is that this tool is only able to verify embedded images with a specific URI format as given in Figure 4. It will not work if the images are downloaded from Flickr, uploaded to some other server and and embedded in the site. Flickr does not yet provide XMP embeddings [21] in images themselves, otherwise an XMP parser would be able to recognize the image's CC License data, even if the URI will not be from Flickr.

7. FUTURE WORK

7.1 Other License Types

The tool described in the paper is currently able to check CC-BY (attribution) license violations. It is relatively straightforward to check for violations of this type as the terms of attribution are specified in the license, i.e. what attribution name to give, and what attribution URL to point to, etc. A violation would simply imply that the attribution details are either missing or incorrect when the image is reused. However, checking for other types of license violations could be a bit more complex. In addition, situations where an image (or any other work) is used for a reasonably justifiable Fair Use could complicate the actual detection and reporting of a violation by the tool.

In Sections 7.1.1 and 7.1.2, since any CC licensed work should by default be given attribution (unless specified in 'additionalPermissions'), we assume that the images are properly attributed. We now envision the problem of detecting the following types of CC license violations.

7.1.1 Non commercial Use

Detecting whether a photo 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. However, given the nature of images, it is easy to alter the image data, and the metadata. Therefore, there should be proper image data provenance methods to detect violations of this nature.

Another interesting parallel problem would be when a photo is embedded in a site, which has dynamic advertisements generated based on the content on the page. Some of these advertisements may be direct consequence of the embedded image itself. Would it then be considered a non-commercial use violation? How will a tool decide which advertisements on the site actually correspond to the image embedded? These questions are hard to answer even with human judgement. Therefore, it is difficult to imagine a tool will automatically detect these kinds of license violations.

7.1.2 Share Alike

Violations of this nature can be such that either the original image is unchanged or changed, and a conflicting license is given. If the image is unchanged, the solution would be check the RDFa in both the original page and in the page where the image was embedded, and see if the latter complies with the original CC license. However, if the image has been changed when creating the derivative work, or as discussed in Section 6, if it was uploaded to some other server and then linked from, a method to capture the image data provenance should be incorporated in to the tool as well.

7.2 Other Media Types

We have only explored one domain of creative works on the web - namely Flickr images. There are billions of videos uploaded on YouTube, and potentially countless number of documents on the web, which have CC licenses applied.

While MPAA, RIAA and other such big organizations are working towards preserving the rights of the works of their artists on YouTube, other video / audio sharing sites and peer-to-peer file sharing networks, there are no viable alternatives for normal 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 free-floating content will be highly appreciated.

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.

9. ACKNOWLEDGMENTS

This work was carried out for Web Science Research Initiative (WSRI) Exchange funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/F013604/1, National Science Foundation Cybertrust Grant award number 04281, and IARPA award number FA8750-07-2-0031.

We would also like to thank our colleagues Harith Alani, Steve Harris and various developers on IRC channel #cc who helped us with their feedback on the project idea and numerous suggestions to explore.

10. REFERENCES

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

[2] Creative Commons LibLicense. http://wiki.creativecommons.org/Liblicense.

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

[4] Flickr API. http://www.flickr.com/services/api.

[5] Flickr Creative Commons Images. http://www.flickr.com/creativecommons.

[6] Flickr image reuse for openoffice.org. http://labs.creativecommons.org/2008/07/12/flickr-image-re-use-for-openofficeorg-demo-availlable/.

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

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

[9] Ivan Herman. RDFa Distiller, 2008.

[10] Joan Feigenbaum and Michael J. Freedman and Tomas Sander and Adam Shostack. Privacy Engineering for Digital Rights Management Systems. *Digital Rights Management Workshop*, 2006.

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

[12] Lalana Kagal and Chris Hanson and Daniel Weitzner. Using Dependency Tracking to Provide Explanations for Policy Management. *IEEE Policy*, 2008.

[13] License Tagger. http://wiki.creativecommons.org/Licensetagger.

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

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

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

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

[18] Science Commons. http://sciencecommons.org/.

[19] Tim Berners-Lee. Tabulator Redux: Browing and Writing Linked Data . In *Linked Data on the Web Workshop at WWW08*, 2008.

[20] 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.

[21] XMP - Extensible Metadata Platform. http://www.adobe.com/products/xmp/index.html.