Policy Aware Content Reuse on the Web

Oshani Seneviratne, Tim Berners-Lee
Decentralized Information Group, CSAIL, MIT
{oshani, timbl}@csail.mit.edu

The Problem

Reusing content saves resources and fosters creativity. However, reusing a particular piece of content without honoring the license expressed with it may violate the original content creator's rights. There are several reasons this situation might happen. The person reusing the content may be:

- too lazy to check for the licenses hidden in the XHTML
- weary of the multi-step operations required to embed the license metadata
- ignorant as to what each of the licenses mean

At the same time, the original content creator would also be interested in knowing whether someone has violated his or her license terms.

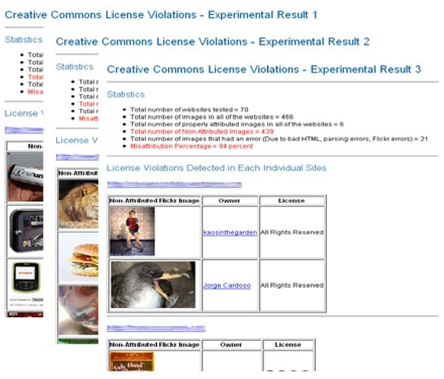
How much of a problem is this?

Flickr has over 100 million Creative Commons Licensed images. Given a sample of web pages which embed such images, how many of these are properly attributed as specified in their licenses?

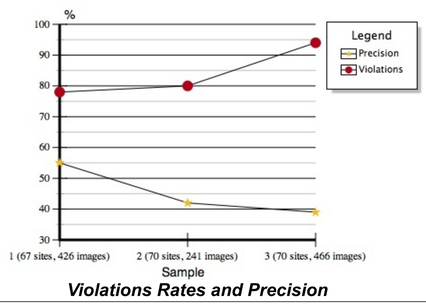
A simple experiment was conducted to get an assessment on this, and the results are as follows:

Sample 1 (67 sites, 426 images)	Properly attributed images = 28 Misattributed images = 333 Misattribution = 78 %
Sample 2 (70 sites, 241 images)	Properly attributed images = 8 Misattributed images = 194 Misattribution = 80 %
Sample 3 (70 sites, 466 images)	Properly attributed images = 6 Misattributed images = 439 Misattribution = 94 %

Results of the experiment summarized



Screenshots of the results from the experiment



Violations Rates and Precision

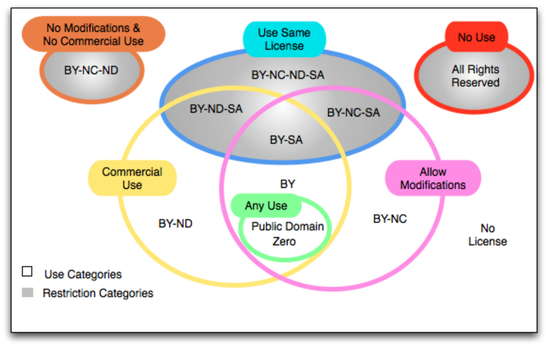
The Solution

Build **Policy Aware Systems**, such as:

- Validators to tell users what information is missing or inaccurate
- Seamlessly integrate metadata by detecting and assisting in embedding the licenses
- Notify users if their content is used in an inappropriate manner

Background

Policies are pervasive in web applications as they play a crucial role in enhancing security, privacy and usability of services offered on the Web. Use of Creative Commons licenses is the widely accepted method of expressing rights of the original content creators when it comes to digital multimedia content on the Web



CC Licenses

More Information
http://creativecommons.org
http://rdfa.info

How can you Extract License Metadata?

1. Through APIs which expose the licenses. E.g. Flickr
2. Through RDFa (Resource Description Framework in Attributes)

A simple scenario which illustrates a rights violation of a content creator:

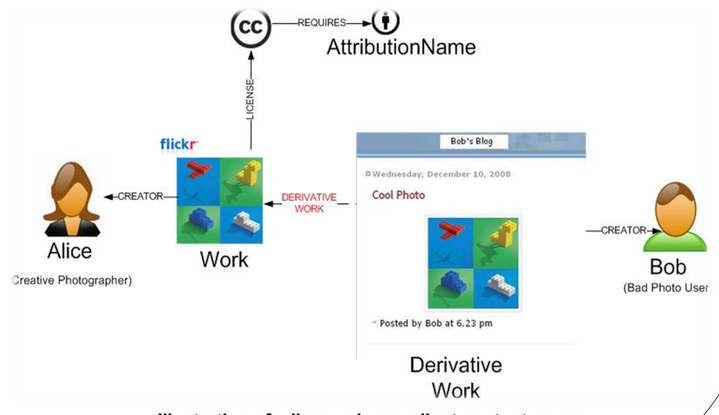
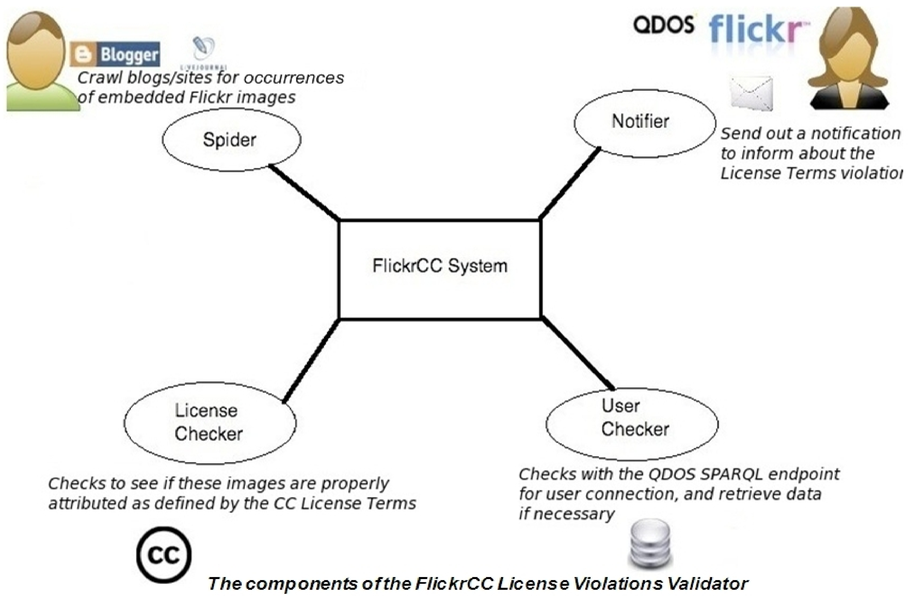


Illustration of a license incompliant content reuse

CC Attribution License Violations Validator



The components of the FlickrCC License Violations Validator

Goal

Check whether a particular site has any embedded Flickr images which are not properly attributed as specified in the Creative Commons license.

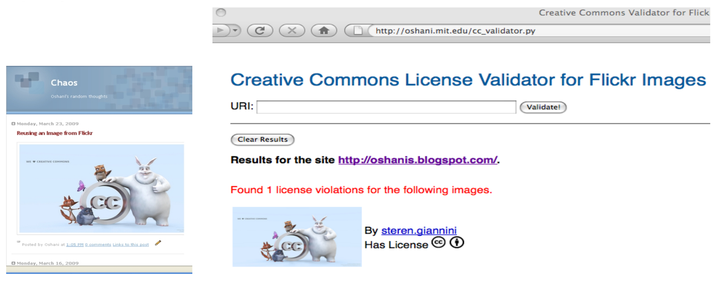
Components

Spider: This is a site crawler which searches for all the links in a given seed site using a Breadth First search algorithm to determine any embedded Flickr images.

License Checker: This extracts the photo id from the Flickr image URI. Then all the information related to the photo is obtained through the Flickr API. Based on this information, the DOM of the page is checked for the proper attribution.

Notification System: This will pretty-print and report the images with missing attributions in a Web interface. The user can then use the missing information in his or her own work to be license compliant.

User Checker (optional): This module can be used to send actual notifications to the original content creators for any violations if the system is linked to some user base.



Website to Validate

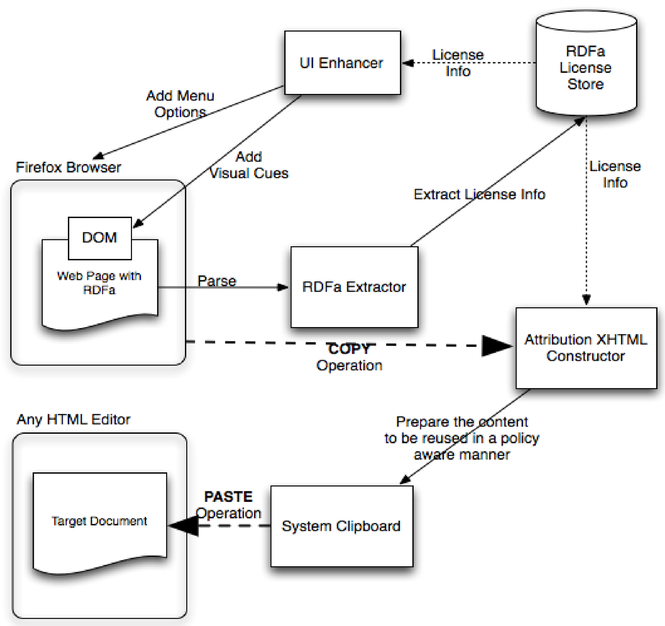
Validator

License Violations Validator

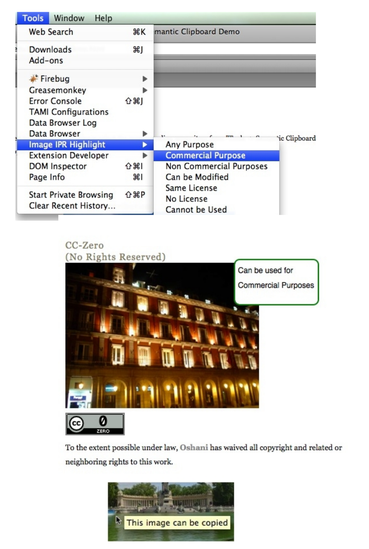
Try it out!
http://oshani.mit.edu/cc_validator.py

More Information
http://dig.csail.mit.edu/2008/WSRI-Exchange

Semantic Clipboard



Architecture of the Semantic Clipboard



Semantic Clipboard User Interface is integrated in the Browser

Goal

Enable transfer of content between Web applications with minimal effort in a policy aware manner, i.e. when content is copied, license metadata is also copied and pasted appropriately in the target application.

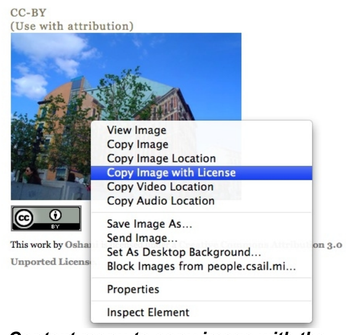
Components

RDFa Extractor: Extracts all the semantic information in the form of RDF attributes embedded in the HTML page the user browses.

UI Enhancer: Adds visual cues to the page for easy identification of images that can be copied based on the user's intended use.

RDFa License Store: Indexes the License data of images in a given browser session.

Attribution XHTML Constructor: Creates the attribution XHTML snippet as stated in the CC specification upon a copy instruction. Then it places this snippet in the system clipboard.



Context menu to copy image with the license metadata

All of these components are implemented in the **Tabulator**, a Semantic Web Browser which can be installed as a Firefox Extension.

Try it Out!
http://dig.csail.mit.edu/2007/tab

More Information
http://dig.csail.mit.edu/2009/Clipboard

Contributions

- Assessment on the level of policy-awareness on the Web
- Provide a platform to use the data exposed on the Semantic Web
- A License Violations Validator for Flickr images:
  - to check for license violations
  - use the information given by the validator to be policy-compliant
- Semantic Clipboard:
  - to detect reusable content while browsing
  - seamlessly integrate such content along with their metadata

Challenges

- Tracking provenance of content on the Web is hard
- Subsequent changes to a CC license cannot be prevented
- Lack of proper definitions from CC for the scoping of the human readable attribution in the DOM, and the license granularity
- Limited Flickr support for license expression
- Usability vs. Operating System independence

Future Work

- Assess the level of violations with regards to other types of licenses such as 'no commercial use', 'share alike' and 'no derivatives'
- Assess the level of license violations on other types of media
- Extend to licenses embedded in free-floating content
- Explore new and efficient ways of license violations detection
- Improve the User Interfaces of the CC license violations validator and the Semantic Clipboard