Policy Aware Content Reuse on the Web

Oshani Seneviratne
Advisor: Tim Berners-Lee

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 | Properly attributed images = 28 | Misattributed images = 533 | Misattribution = 78% |
| Sample 2 | Properly attributed images = 8 | Misattributed images = 194 | Misattribution = 80% |
| Sample 3 | Properly attributed images = 6 | Misattributed images = 439 | Misattribution = 94% |

The results of the experiment summarized:

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

• Too lazy to check for the licenses hidden in the XHTML
• Ignorant as to what each of the licenses mean

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 via the Flickr API. Based on this information, the DOM of the page is checked for the proper attribution.

Find out about the license terms:

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.

Semantic Clipboard:

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

Notification System: This is used to make the data persistent across browser sessions.

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

RDF Store: Indexes and stores all the RDF data from the pages that the user has visited in a given browser session.

Semantic Clipboard: Acts as the control panel to co-ordinate the copy and paste operations.

More Information

http://creativecommons.org
http://rdfla.info

FlickrCC Attribution License Violations Validator

Semantic Clipboard

More Information

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

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

Contributions

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

Please send your comments to oshani@csail.mit.edu