

**Framework for Policy Aware Reuse of Content on  
the WWW**

by

Oshani Wasana Seneviratne

Submitted to the Department of Electrical Engineering and Computer  
Science

in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

© Massachusetts Institute of Technology 2009. All rights reserved.

Author .....  
Department of Electrical Engineering and Computer Science  
May 22, 2009

Certified by .....  
Tim Berners-Lee  
Professor  
Thesis Supervisor

Certified by .....  
Lalana Kagal  
Research Scientist  
Thesis Reader

Accepted by .....  
Terry P. Orlando  
Chairman, Department Committee on Graduate Theses



# Framework for Policy Aware Reuse of Content on the WWW

by

Oshani Wasana Seneviratne

Submitted to the Department of Electrical Engineering and Computer Science  
on May 22, 2009, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Computer Science and Engineering

## Abstract

This thesis focuses on methods for detecting and preventing license violations, in a step towards policy aware content reuse on the Web. This framework builds upon the Creative Commons (CC) Rights Expression Language, which provides a very clear and a widely accepted set of licenses grounded in Semantic Web technologies. These licenses are machine readable, and indicates to a person who wishes to reuse the content exactly how it should be used.

An experiment on CC attribution license violations on Flickr images revealed the attribution license violation rate on the Web to be around 70-90% from samples of Websites that had at least one embedded Flickr image. Therefore, it is evident that there should be robust mechanisms for detecting license violations on the Web and prevent those happening, if possible. The primary objective is to enable the user to do the right thing instead of preventing the user from doing the wrong thing.

As a solution, we have implemented (1) “Attribution License Violations Validator” for Flickr images and, (2) the more generic “Semantic Clipboard”. The “Attribution License Violations Validator” can be used to validate users’ work against any attribution license violation. The “Semantic Clipboard”, which is implemented as a component of the Tabulator Firefox extension, allows the user to copy an image with its license metadata expressed in *Resource Description Framework in annotations* (RDFa) in the original source document to any other document.

Thesis Supervisor: Tim Berners-Lee  
Title: Professor

Thesis Reader: Lalana Kagal  
Title: Research Scientist



## Acknowledgments

I would like to thank my advisor Tim Berners-Lee for his guidance during the past two years. He helped me learn the intricacies of Tabulator and many things about the Semantic Web.

Daniel Weitzner for all the advice he has given throughout, especially on certain legal aspects of Creative Commons licenses and the Copyright Law.

Lalana Kagal, for mentoring me, giving very good implementation suggestions and making sure that everything goes smoothly and timely.

Hal Abelson, for introducing me to the Creative Commons community to present this work and get lot of feedback from.

Gerry Sussman for advising me on the feasibility of applying data purpose algebra in the context of policy aware content reuse.

Nigel Shadbolt at University of Southampton in the UK, for helping me initiate this work during the summer of 2008 while I was undergoing the ‘Networks for Web Science Student Exchange’ under his supervision. Without his advice this work would not have materialized.

I am also very grateful to my lab mates and friends for providing help on bug busting by testing the software I have developed, proofreading my thesis and for all the encouragement they provided all throughout.

This work was carried out with generous funding from National Science Foundation Cybertrust Grant award number 04281, IARPA award number FA8750-07-2-0031, and UK Engineering and Physical Sciences Research Council (EPSRC) grant number EP/F013604/.



# Contents

<b>1</b>	<b>Introduction</b>	<b>15</b>
1.1	Problem Description . . . . .	15
1.2	Motivation . . . . .	15
1.3	The Need for Policy Awareness in Content Reuse . . . . .	16
1.3.1	Exposing Licenses . . . . .	17
1.4	Tools Developed . . . . .	17
1.5	Thesis Overview . . . . .	18
<b>2</b>	<b>Background</b>	<b>21</b>
2.1	Content Reuse Defined . . . . .	21
2.2	Policies for Rights Enforcement on the Web . . . . .	22
2.2.1	Digital Rights Management (DRM) . . . . .	22
2.2.2	Copyright through Licensing . . . . .	23
2.3	Data Purpose Algebra . . . . .	26
2.4	Inline Provenance using Metadata . . . . .	27
<b>3</b>	<b>License Violations Assessment on the WWW</b>	<b>31</b>
3.1	Experiment Setup . . . . .	31
3.1.1	Ensuring a Fair Sample . . . . .	31
3.1.2	Checking for Attribution . . . . .	33
3.2	Results . . . . .	33
3.2.1	Issues in this Experiment . . . . .	34
3.2.2	Refining the Results . . . . .	37

3.3	Extensions to this Experiment . . . . .	37
<b>4</b>	<b>Attribution License Violations Validator</b>	<b>41</b>
4.1	Checking for Attribution License Violations . . . . .	41
4.2	Design and Implementation of the System . . . . .	42
4.2.1	Site Crawler . . . . .	42
4.2.2	Flickr Query Evaluator . . . . .	43
4.2.3	License Checker . . . . .	44
4.2.4	Notification System . . . . .	45
<b>5</b>	<b>Semantic Clipboard</b>	<b>47</b>
5.1	Inspiration . . . . .	47
5.2	Enabling Awareness of Licenses . . . . .	48
5.3	Design and Implementation . . . . .	50
5.3.1	RDFa Extractor . . . . .	51
5.3.2	UI Enhancer . . . . .	53
5.3.3	Attribution XHTML Constructor . . . . .	54
5.3.4	User Interface . . . . .	55
<b>6</b>	<b>Related Work</b>	<b>59</b>
6.1	License Detection Tools . . . . .	59
6.2	License Embedding Tools . . . . .	60
6.3	Transferring Metadata with Content . . . . .	61
6.4	Commercial Applications for Detecting Violations . . . . .	62
<b>7</b>	<b>Summary</b>	<b>63</b>
7.1	Contributions . . . . .	63
7.2	Challenges . . . . .	64
7.2.1	Challenges for the Attributions License Violations Validator . . . . .	64
7.2.2	Challenges for the Semantic Clipboard . . . . .	66
7.3	Future Work . . . . .	67
7.3.1	Check for Other Types of License Violations . . . . .	67



7.3.2	Give Credit to the Original Content Creator as Requested . . .	68
7.3.3	Extend to Other Media Types . . . . .	68
7.3.4	License Granularity . . . . .	69
7.3.5	Persistent Data Storage . . . . .	69
7.3.6	User Study . . . . .	69
7.3.7	Widening Social Criteria . . . . .	70
7.4	Conclusion . . . . .	70
<b>A Model Using the AIR Policy Language</b>		<b>73</b>
A.1	Introduction . . . . .	73
A.2	Scenario . . . . .	74
A.2.1	CC-BY Policy Expressed in AIR . . . . .	75
A.2.2	Results . . . . .	75
<b>B Links to Code and Demos</b>		<b>77</b>
B.1	Experiment to determine Attribution License Violations on Flickr Images	77
B.2	Attribution License Violations Validator . . . . .	77
B.3	Semantic Clipboard . . . . .	78
B.4	Creative Commons License Scenarios in AIR . . . . .	78



# List of Figures

2-1	Illustration of Usage of CC Licensed Content . . . . .	25
2-2	License Expressed in RDFa . . . . .	25
2-3	License Composition Matrix for Detecting Share Alike Violations . . . . .	29
3-1	Results from the Experiment . . . . .	32
3-2	Example Result from the Experiment . . . . .	35
3-3	Different Ways of Attribution . . . . .	36
3-4	Attribution Violations and Precision . . . . .	38
4-1	The Design of the Validator . . . . .	43
4-2	Output from the Validator . . . . .	45
5-1	Venn Diagram for the Acceptable Use and Restricted Use of Content Depending on the CC License . . . . .	49
5-2	Design of the Semantic Clipboard . . . . .	51
5-3	Sample Image with Metadata about it Expressed in RDFa . . . . .	52
5-4	Firefox Menu to Specify User Purpose to See What is Available . . . . .	53
5-5	Image Overlaid With the Purpose Information . . . . .	54
5-6	Tooltip on an Image giving Visual Cues as to whether the Image can be Copied or Not . . . . .	54
5-7	Right Click Context Menu on an Image . . . . .	55
5-8	Metadata Description of Image given in Figure 5-3 . . . . .	56
5-9	RDF Metadata about the Image in Notation 3 Serialization . . . . .	57
5-10	Constructed Attribution XHTML Snippet . . . . .	57

6-1	CC Deed Page Displaying the Attribution XHTML . . . . .	60
A-1	Illustration of Inappropriate Usage of CC Licensed Content . . . . .	74
A-2	Creative Commons Attribution License Violations Scenario Output from the AIR Reasoner using the Justification UI in the Tabulator . .	75
A-3	Sample AIR Policy . . . . .	76

# List of Tables

3.1	Attribution License Violations Rates of the Experiment Samples . . .	34
3.2	Precision Values of the Experiment Samples . . . . .	37



# Chapter 1

## Introduction

### 1.1 Problem Description

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 Web 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 Web sites, have lots of increasingly complex information. Such information 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 and license terms violations, motivating research into effective methods to detect and prevent such violations.

### 1.2 Motivation

Creative Commons (CC) provides a very clear and a widely accepted rights expression language [21] using Semantic Web technologies that are used to compose a set of well-defined licenses. These licenses are both machine readable and human readable, and clearly indicates to a person who wishes to reuse content exactly how it should be used, by expressing the accepted use, permissions, and restrictions of the content.

However, even with these human-friendly licenses, we can expect license violations

to occur due to many factors: *a)* Users may be ignorant as to what each of the licenses mean. *b)* Users may forget or be too lazy to check and include the proper license terms. *c)* Users may give an incorrect license which violates the original content creator's intention. And last but not least, *d)* malicious users might 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, on which Web pages and by whom. But given the scale and the nature of the Web, the knowledge of such license violations is highly unlikely unless the original content creator comes across those by chance. An assessment on Creative Commons Attribution License Violations on Flickr images on the Web, as discussed in Chapter 3 revealed violation rates ranging from 70%-90%.

On the other hand, people who want to reuse content may be interested in knowing whether a particular content item can be reused or not, and if it can be reused, the conditions for reuse. Therefore, license aware tools for easy content reuse, and validators to verify works against any license violations will lead to the path of least resistance in generating creative works on the Web.

### **1.3 The Need for Policy Awareness in Content Reuse**

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 [5]. Information accountability provides another motivation to apply policies for data usage practices [49]. In this thesis we will limit the 'policy awareness' aspect to licenses that can be expressed semantically, that are widely deployed on a range of media, and that have a large community base. CC licenses fit this description perfectly. Therefore, we will be focussing our attention on CC licenses, keeping in mind the possibility of extending the system we develop to support other types of licensing mechanisms.



### 1.3.1 Exposing Licenses

Typically there are two ways in which metadata about licenses can be exposed:

1. Through APIs which expose the Licenses:

For example, Flickr allows users to specify the license associated with their images. This license information can then be queried through the Flickr API. This method is not very interoperable as API specific data wrappers have to be written for each service.

2. Through Resource Description Framework in Attributes (RDFa) [42]:

CC licenses can be expressed in machine readable form such as RDF [41] using RDFa. The content creator and consumer can use RDFa for rights expression and compliance respectively. RDFa allows machine understandable semantics to be embedded in the XHTML.

## 1.4 Tools Developed

Research undertaken in this thesis focuses on methods for detecting and helping users avoid license violations. We have developed several tools to achieve this.

1. **License Violations Validator** : *to verify your own work for attribution license violations.*

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, by misattributing or by not attributing the original content creator. In such cases, a *validator* which checks for CC license violations of content would be very useful. This will be analogous to the validators Web developers use to check whether their XHTML is valid by using the *W3C Markup Validation Service*, or semantic data producers checking to see if their data is in proper RDF syntax by using the *W3C RDF Validation Service*. Using such a tool,

content reusers can rectify the instances where they have inadvertently violated the CC licenses before they publish their work.

2. **Semantic Clipboard:** *for users to seamlessly reuse content on the Web while integrating the license metadata in a policy aware manner.*

This is designed to address the problem of users being lazy to check the license or inadvertently giving a wrong license or attribution information. Semantic Clipboard allows users to copy images, but adds the license metadata to the copied image so that the secondary work will be license compliant automatically.

Both these tools are built upon the Creative Commons Rights Expression Language (ccREL) [21]. There is no attempt to enforce the rights associated with content as in Digital Rights Management (DRM). The violator will not be automatically prevented if the license terms are violated. It merely guides the user as to how best the content should be reused, making sure that the policies governing content usage are properly adhered to.

## 1.5 Thesis Overview

Chapter 1 has given an introduction to the problem that is addressed in this thesis. The rest of the thesis is structured as follows:

Chapter 2 gives the background and an overview of the technologies used for policy aware content reuse.

Chapter 3 outlines the experiment conducted to assess the level of Creative Commons attribution license violations on the Web, and the results of that experiment.

Chapter 4 gives the implementation details of the “Attribution License Violations Detector and Validator” for Flickr images.

Chapter 5 gives the implementation details of the “Semantic Clipboard”.

Chapter 6 discusses related work in this area.

Chapter 7 gives a summary of the contributions, challenges and the future work.

Appendix A gives the scenario encoded in the AIR rule language and the output from the AIR reasoner.

Appendix B gives links to the source code, documentation and demos of the tools described in this thesis.



# Chapter 2

## Background

### 2.1 Content Reuse Defined

Content reuse, or in other words “mash-ups” have existed for as long as content has existed. Musicians routinely use other songs and tunes in their compositions. Collage art is considered to be creative, and even original although it is composed from many different sources. Scientists routinely utilize data from different sources to conduct their own experiments. However, mash-ups, as we know them now, are a peculiarly digital phenomenon of the Web age. They are entirely a product made possible by the portable, mixable and immediate nature of digital technology.

Reuse detection is important in domains such as plagiarism detection and even in biological sequence mining. Significant research has been carried out to detect reuse of text. This includes information retrieval techniques as mentioned in [32, 43], where the document is treated as a sequence of symbols and substring based fingerprints are extracted from the document to determine repetitive patterns.

A potential legal problem arises when more than one legally encumbered content or data streams are bound together in the form of a mash-up. The users of the original content should remain within the bounds of the permitted use of the components comprising the mash-up. They can choose to ignore these permissions, or follow them. Either way, this creates a burden on them. Ignoring the license terms puts them at the peril of breaking the law, and following them slows the creative process.

## 2.2 Policies for Rights Enforcement on the Web

Policies governing the reuse of digital content on the Web can take several forms. It can be the *Digital Rights Management* (DRM) approach exercised for example, by some commercial systems when distributing their media content. It could also be the *Copyrights* alternative, that requires anybody reusing the content, either to have explicit permission from the original content creator, or express the original content creators rights as specified by the CC.

### 2.2.1 Digital Rights Management (DRM)

Distribution and usage of copyrighted content is often controlled by up-front policy enforcement mechanisms such as DRM. These systems usually restrict access to the content, or prevent the content from being used within certain applications. The core concept in DRM is the use of digital licenses, which grant certain rights to the user. These rights are mainly usage rules which are defined by a range of criteria, such as frequency of access, expiration date, restriction to transfer to another playback device, etc. There are many applications that enforce DRM, such as Apple iTunes or Microsoft Windows Media Rights Manager. An example of a DRM enforcement would be a DRM software enabled playback device not playing a DRM controlled media transferred from another playback device, or not playing the media after the rental period for that media has ended.

The use of DRM to express and enforce rights on content on the Web 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 [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 requiring vendor lock-in.

## 2.2.2 Copyright through Licensing

Copyright is essentially the right to make copies. Prior to the Berne Convention, a multilateral treaty to which most countries now adhere [2], content creators were required to explicitly give the *copyright* (©) notice to indicate that all rights are reserved for their works. Otherwise, their creations will fall under the public domain. However, Article 5(2) of the convention provides that “the enjoyment and the exercise of [copyrights] shall not be subject to any formality” which means that any content that does not have any explicit license will be protected by copyright law, whether the author is aware of it or not. Many argue this to be over inclusive to inhibit creativity [12]. For example, when reusing content that are copyrighted, the reuser faces many obstacles such as locating the copyright holder, obtaining permission, and usually having to pay royalty.

In the context of digital content, a ‘license’ is describes the conditions of usage of copyrighted material. A license on a digital file can exist whether or not there are any corresponding users for it. A user should abide by the license that covers the usage, and if any of the conditions of usage described in that license are violated, then the user should cease using that content.

### Choosing a License

Choosing a license can be very confusing. In the open source regime alone there are as many as 90 different licenses [13]. If choosing a license is tough for the creator, imagine what understanding the license is like for the end user.

Creative Commons, a non-profit organization has been striving to provide a simple, uniform, and understandable set of 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. 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, used only for non-commercial use, distributed with

a similar license and will be allowed in other free culture media. To allow these use restrictions CC has composed four distinct license types: *BY* (attribution), *NC* (non-commercial), *ND* (no-derivatives) and *SA* (share-alike) that can be used in combinations that best reflects the content creator's rights.

In order to generate the license XHTML easily, CC offers a license chooser that is hosted at <http://creativecommons.org/license>. There, the users are given options to select for their works, and it will generate a snippet of XHTML that contains the RDFa [42] they need to embed when they are publishing the content on the Web.

### **Creative Commons Rights Expression Language (ccREL)**

*ccREL* [21] is the standard recommended by the CC for machine readable expression of the meaning of a particular license. Content creators have the flexibility to express their licensing requirements using this rights expression language and are not forced into choosing a pre-defined license for their works. Also, they are free to extend licenses to meet their own requirements. *ccREL* allows a publisher of a work to give additional permissions beyond those specified in the CC license with the use of the *cc:morePermissions* property to reference commercial licensing brokers or any other license deed, and a *dc:source* to reference parent works. Therefore, unlike in older CC recommendations, it is also possible to have content under a CC license that does not require attribution.

### **Anatomy of a CC License in RDFa**

Figure 2-1 illustrates a simple CC license applied to an image. Suppose Alice is an avid Flickr user and she uploads her photos 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. Flickr does not yet markup this license information in RDFa, and is still under the older CC 2.0 specification. However, the CC license badge hyperlinked to the CC *Deed Page*<sup>1</sup> of the corresponding license will be displayed in all of her photo album

---

<sup>1</sup>A Deed Page is a human readable page that describes the license type.



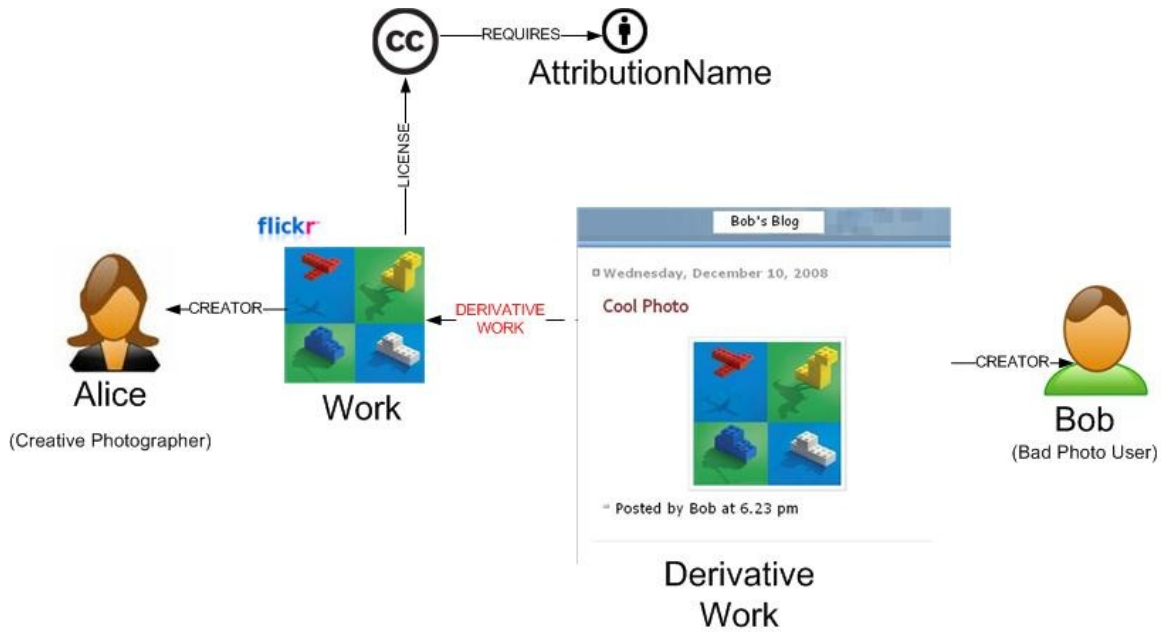


Figure 2-1: Illustration of Usage of CC Licensed Content

pages.

Supposing that Alice uses a service which marks up the license metadata in RDFa, the code snippet shown in Figure 2-2 shows how she would express her rights for one of her photos identified by the following URI: <http://flickr.com/photos/alice/somephoto.jpg>. The RDF data are given in red.

```
<div about="http://flickr.com/photos/alice/somephoto.jpg">
<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</span>
and link the credit to <a rel="cc:attributionURL"
  href="http://flickr.com/photos/alice">
  http://flickr.com/photos/alice
</a></div>
```

Figure 2-2: License Expressed in RDFa

The important things to note in this code snippet is that Alice has given the

subject (the URI of the photo) the following properties:

- *attributionName*: ‘Alice’
- *attributionURL*: ‘http://flickr.com/photos/alice’.

If we assume that Bob uses Alice’s photo in his blog or some other derivative work without embedding the XHTML that references Alice via the *attributionURL* or the *attributionName*, or does not at least mention the source of the photo - it is a clear violation of Alice’s CC license terms.

## 2.3 Data Purpose Algebra

One of the hazards of combining multiple data sources is that incompatible licenses can get mixed up creating a license that basically freezes the creative process. Take for example a Non-Commercial (NC) license that gets mixed with a Share-Alike (SA) license. A SA license requires that the resulting product be shared under exactly the same conditions as the component product under SA. The resulting license in our scenario becomes NC-SA. But while the result satisfies the first license by also being NC, it fails the second license by not being only SA. So, if the component product under SA was unencumbered by the NC clause, adding it to an NC restricted component violates the SA requirement. Also we cannot simply ignore the NC clause and give the resulting work only the SA license. This is because somebody else might use the resulting derivative work which does not have the NC clause in some commercial use violating the rights of the original creator who composed the NC component. Many of these license combination conflicts are given in Figure 2-3.

It has been shown that it is possible to model data usage policies programmatically by what is known as the *Data Purpose Algebra* [22], by describing each content item  $i$  in a data set, a source or agent that processes the data  $Q_d(i)$ , the category of data  $K_d(i)$ , and its purpose  $P_d(i)$ . When another agent combines two or more data sets, a new data set is created whose content, category and purpose are some function of the agent, content, category and purpose of each of the component data sets. Specifically,

if the agent or the source of the new data item is  $a'$ , the new category becomes the function  $\mathcal{K}(K_d(i))$  of the given category, and the allowed purposes of the new data item will be the more complex function  $\mathcal{P}(P_d(i), A_d(i), a', K_d(i))$  that may depend on the original purposes, the agents, and the category of the original data.

We believe that when reusing content on the Web, the same principle could be applied. For the content item  $i$  that is reused, the source function  $Q_d(i)$  would be represented by the URI of the content. The category of data  $K_d(i)$  will be represented by the content type (i.e. image, text, video, audio, etc.) that is being reused. The purpose  $P_d(i)$  will be determined by the CC license associated with it that specifies the allowed uses, restrictions and conditions. Then the function which composes the new set of purposes  $\mathcal{P}(P_d(i), A_d(i), a', K_d(i))$  should generate the XHTML that is required to embed the content with proper attribution.

## 2.4 Inline Provenance using Metadata

To be useful, metadata need to have three important characteristics: they have to be easy to produce, be embedded within the data they describe, and be easily readable. The easiest way to produce metadata is to have them be produced automatically. Any metadata that has to be produced manually by the user usually doesn't get produced at all. The easiest way to ensure that the link between metadata and the data they describe is not broken is by embedding the former inside the latter. This way, the two travel together inseparably as a package. Finally, metadata have to be accessible easily, readable both manually as well as programmatically. At best, the metadata should be readable by crawlers of various search engines. Since metadata and data are traveling together, if popular search engines such as Google and Yahoo can read the metadata, by default the data become available to anyone who searches for it. RDF [41] is the best known metadata format which satisfy all these criteria. It has lot of community support, adopted widely and is a W3C recommendation.

Extensible Metadata Platform (XMP) [52] is a technology that allows one to transfer metadata along with the content by embedding the metadata in machine

readable RDF. This technology is widely deployed in embedding licenses in free-floating multimedia content such as images, audio and video on the Web.

Another format which is nearly universal when it comes to images is the Exchangeable Image File format (EXIF) [14]. International Press Telecommunications Council (IPTC) photo metadata standard [25] is also another well known standard. The metadata tags defined in these standards cover a broad spectrum including date & time information, camera settings, thumbnail for previews and more importantly, the description of the photos including the copyright information. However, these latter two formats do not store these metadata in RDF. They both use key-value pairs.

One major drawback of inline metadata formats such as XMP and EXIF is that, it is embedded in a binary file, completely opaque to nearly all users, whereas metadata expressed in RDFa will require colocation of metadata with human visible HTML. In addition to that, these metadata formats can only handle arbitrary number of properties using RDF predicates and these formats do not seem to take advantage of the rich expressivity offered by RDF.

<i>Resultant License</i>	<i>Component License</i>											
	Public Domain	CC-Zero	BY	BY-NC	BY-NC-ND	BY-NC-ND-SA	BY-NC-SA	BY-ND	BY-ND-SA	BY-SA	All Rights Reserved	No License
Public Domain	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
CC-Zero	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
BY	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
BY-NC	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
BY-NC-ND	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗
BY-NC-ND-SA	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗
BY-NC-SA	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗
BY-ND	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗
BY-ND-SA	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗
BY-SA	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗
All Rights Reserved	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
No License	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗

Figure 2-3: License Composition Matrix for Detecting Share Alike Violations

The rows represent the resultant license of the composite work and the columns represent the license of a component content item. The ✓ symbol is used when the corresponding resultant license *can* be given to the composite work if one of the components is under the license given by the column. However, if at least one of the cells is given the ✗ symbol, then the resultant license has a conflict with the corresponding component content item, and thus leads to a license violation.



# Chapter 3

## License Violations Assessment on the WWW

### 3.1 Experiment Setup

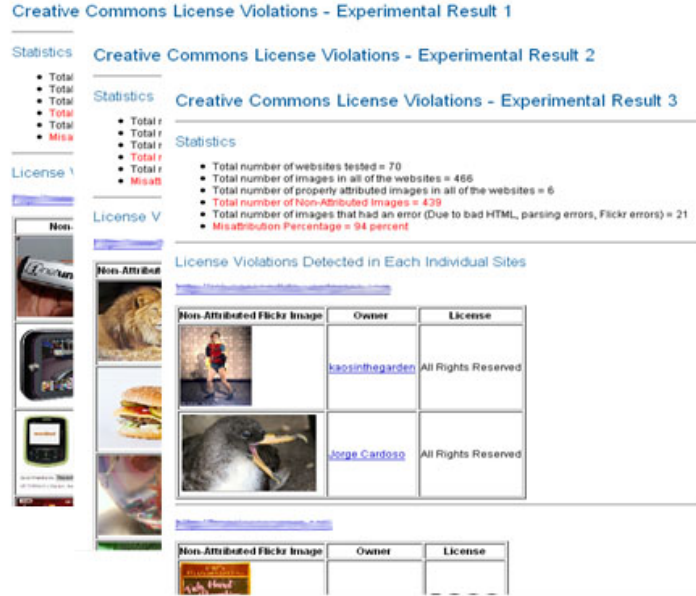
The goal of the experiment is to obtain an estimation for the level of CC attribution license violations on the Web. Since Flickr has over 100 million Creative Commons Licensed images (as of April 2009), detecting attribution license violations with Flickr images seem to be a good way of getting an approximate measurement of the level of license violations out there on the Web. Therefore, specifically, our task in the experiment is to gather quantitative evidence of attribution license violations for several samples of sites that embed Flickr images.

#### 3.1.1 Ensuring a Fair Sample

The Technorati blog indexer [45] 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 so on. Most importantly all the technorati data are time dependent, which means that the technorati *authority rank*<sup>1</sup> is based on most

---

<sup>1</sup>Authority Rank is a measurement that determines the top ‘n’ number of results from any query to the Technorati API.



**Screenshots of the results from the experiment**

Figure 3-1: Results from the Experiment

recent activity in a particular Web site.

Web sites used in this experiment were obtained through the *Technorati Cosmos* method . 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 that have the following general format [18] were used.

`http://farm<farm-id>.static.flickr.com/<server-id>/<id>_<secret>.  
(jpg|gif|png)`

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 the 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 rank goes down changing the result set returned.



Therefore, this factor was also used in generating a random sample of Web sites to check for attribution license violations.

### 3.1.2 Checking for Attribution

After a sample was collected, attribution for each of the images embedded in these sites were checked using few heuristics. Since Flickr is still using the older CC 2.0 recommendation, Flickr users do not have that much flexibility in specifying their own *attributionURL* or the *attributionName* values to state how they would like attribution to be given to them. However, it is considered general practice to give attribution by linking to the Flickr user profile or give the Flickr user name (which could be interpreted as the *attributionURL* and the *attributionName* respectively), or by the least, point to the original source of the image [23]. Therefore, the criteria for checking attribution consist of looking for the *attributionURL* or the *attributionName* or any *source citations* within a reasonable level of scoping from where the image is embedded in the Document Object Model (DOM).

## 3.2 Results

The results from 3 samples of websites gathered within 2 week intervals are shown in Table 3.1. These results have misattribution rates ranging from 78% to 94%. Figure 3-1 illustrates the results from few sample runs of the experiment. The overall summary for each of the samples include:

- Total number of Web sites tested
- Total number of images in all of the Web sites
- Total number of properly attributed images in all of the Web sites
- Total number of mis-attributed images
- Total number of instances that led an error (due to bad HTML, parsing errors, Flickr errors, etc)

Using these values, the misattribution percentage for each sample is calculated. Also, for each of the offending sites, the interface gives the non-attributed or the mis-attributed image, the original owner of the image and the license it is under.

These results indicate that there is a strong need to have awareness among reusers of content to check and honor the licenses associated with the content. However, although the misattribution rates in these samples seem to have a very high value, it should be kept in mind that the sample consists of only Web sites that have Flickr images embedded with a Flickr farm URI, and not a sample of general Web sites.

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

Table 3.1: Attribution License Violations Rates of the Experiment Samples

### 3.2.1 Issues in this Experiment

#### No Self-Attribution

The results from the experiment includes cases where users have not attributed themselves: i.e. user uploads her photos on Flickr, and uses those in the user’s own blog or Web site. Since those are user’s own photos, she is under the assumption that there is no need to attribute herself. This assumption is not entirely valid as the CC BY license deed [7] specifies: *“If You Distribute you must keep intact all copyright notices for the Work and provide (i) the name of the Original Author (or pseudonym, if applicable) ... ”*. This means that, if there is a license attached with the original content, the original user will become the reuser, and therefore will have to honor the license even though it is imposed by herself. This might seem absurd since it should not matter to her if she violates her own license terms. However if the user gives attribution to herself, it would in fact guide other people who want to reuse the content in that secondary work. Therefore, by not attributing herself, the user may be violating her own rights in the long run.

A solution to this issue is hard to realize, as it is difficult to infer the Web site owner from the data presented in the Web site. Even if that was possible, it is hard to make a correlation between the Flickr photo owner and the Web site owner. For example in Figure 3-2, the first attribution violation result 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 Web site where these particular photos are embedded may belong to the same person. Since these connections are not explicitly stated in a machine readable format such as RDF, it becomes very hard to determine the real owners of the image and the Web site programmatically.

## 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
	<a href="#">Tambako The Jaguar</a>	
	<a href="#">Arne List</a>	

Figure 3-2: Example Result from the Experiment

### Location of the Attribution

Majority of the Web sites crawled and examined in this experiment have not used ccREL in marking up attribution. Therefore, we used a heuristic to check for the existence of attribution. This heuristic includes the *attributionName* (constructed from the Flickr user name) or the *attributionURL* (constructed from the Flickr user

profile URI or the original source document’s URI). Visually this would correlate to including the attribution information immediately after the content that is being attributed as shown in Figure 3-3 Attribution Example 1. However, since there is no strict definition from CC as to how attribution should be scoped, someone could also attribute as shown in Figure 3-3 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. However, we discovered that different people use different levels of attribution scoping making it hard to detect the proper attribution in the HTML.

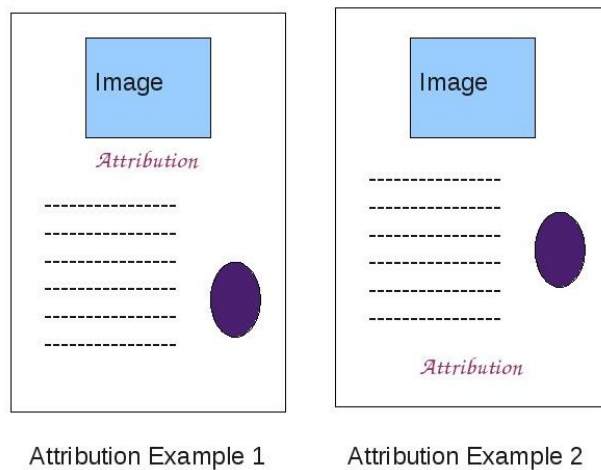


Figure 3-3: Different Ways of Attribution

### Blog Aggregators Ignoring Attribution Information

Tumble-logs such as tumblr.com cuts down the text and favors short form, mixed media posts over long editorial posts. Use of such blog aggregators is another related problem in getting an accurate assessment of attribution license violations. For example, in a blog post where a photo was reused, the original owner of the photograph may have been duly attributed. But when the tumble-log pulls in the feed from that post in the original Web site and presents the aggregated content, the attribution

details may be left out. This problem is difficult to circumvent, because there is no standard as to how aggregation should happen with the license and attribution details. Hence detecting such cases also becomes difficult.

### 3.2.2 Refining the Results

In order to validate the results against the issues mentioned in Section 3.2.1 the samples of Web sites used in the experiment were inspected manually to determine how many images were incorrectly identified as misattributed or non-attributed images and then the precision of the result from the sample was calculated by using the following formula:

$$Precision = \frac{\text{Correctly Identified Misattributed Images} \cap \text{Total Images Retrieved}}{\text{Total Images Retrieved}} \quad (3.1)$$

The results from the three samples are given in Table 3.2. This indicates a relatively low precision value, which means there are lot of false positives. These are mainly due to the fact that people do not attribute themselves when reusing their own content. Note that calculating the *recall* values for this experiment was not possible due to the manner in which the random samples were generated. The samples are obtained from the Web, where there is obviously no correct estimate of how many real attribution license violations are there.

Sample	Correctly Identified Images	Precision
1	183	55 %
2	113	42 %
3	268	39 %

Table 3.2: Precision Values of the Experiment Samples

## 3.3 Extensions to this Experiment

This experiment can be extended to check for Non Commercial use violations, Share Alike use violations and finding out license conflicts in the composite work.

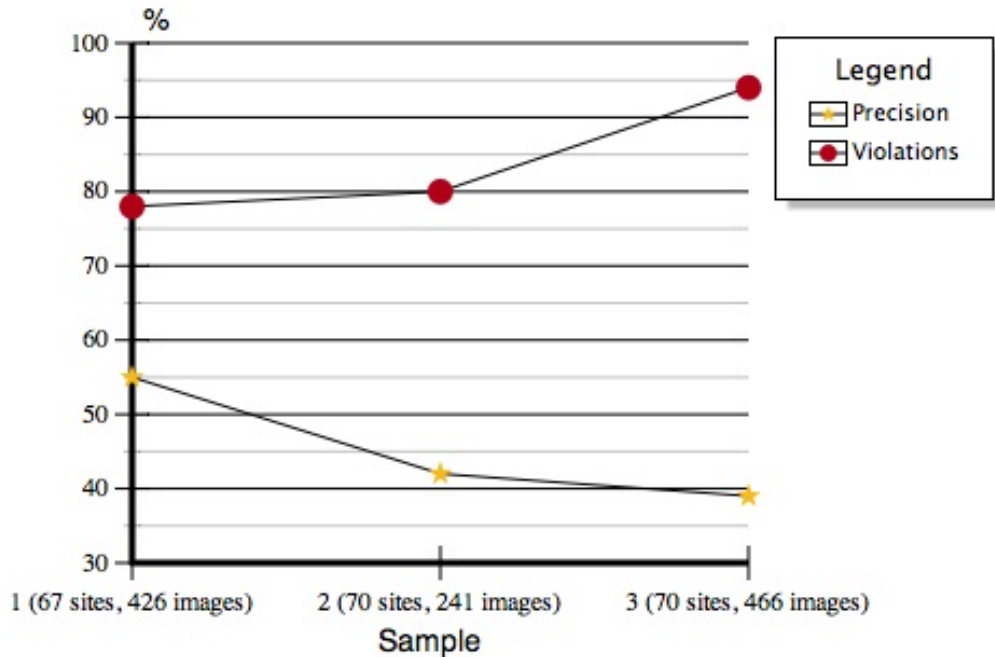


Figure 3-4: Attribution Violations and Precision

In terms of Non Commercial (NC) use, the CC deed specifies that a license including the NC term may be used by anyone for any purpose that is not “*primarily intended for or directed towards commercial advantage or private monetary compensation*”. However, this definition can be vague in certain circumstances. Take for example the case where someone uses a CC-BY-NC licensed image in her personal blog properly attributing the original content creator. The blog is presumably for non commercial use, and since she has given proper attribution it appears that no license violation has occurred. However there might be advertisements in the page that are generated as a direct result of the embedded image. Our user might or might not actually generate revenue out of these advertisements. But if she does, it could be interpreted as a ‘private monetary compensation’. Hence we believe that the perception as to what constitutes a Commercial Use is subjective. CC recently conducted an online user survey to gather general opinions as to what people perceive a commercial use is [11]. An important finding from this survey is that 37% of the creators who make money from their works do so indirectly through advertising. However there aren’t any clear cut definitions of a non commercial use yet to find out violations and

gather experimental results.

Share Alike license violations occur when some content is reused with a conflicting license or remixed with content that have conflicting licenses. The CC survey [11] states that the most popular CC license of the respondents is BY-NC-SA. Therefore, it would be interesting to have an estimation of SA license violations as well. Figure 2-3 illustrates possible cases when a license conflict might occur. For example, consider the case where the resultant license given to some composite work is BY-NC-SA, and that the two components it is composed of are licensed under BY-NC and BY-ND-SA respectively. The BY-NC component could be used in the work as the resultant license also has the BY-NC term. However the other component which has the BY-ND-SA specifies that if it is to be used in some other work, the same license has to be given to the composite work. BY-ND-SA is not the same as BY-NC-SA, therefore this leads to a license violation.





# Chapter 4

## Attribution License Violations Validator

### 4.1 Checking for Attribution License Violations

When someone aggregates content from many different sources, it is inevitable that some attribution details may be accidentally forgotten. The Attribution License Violations Validator will figure out whether the user has properly cited the source by giving the due attribution to the original content creator. In other words, this is essentially a tool to help an honest person remain honest when reusing content on the Web.

According to the CC user survey [11], out of content types such as photos, text, blogs, online journals, videos, songs, games, mash-ups, podcasts and other such media types, photos seem to be the most common type of work created on the Web. In terms of redistribution and reuse sending via email, posting on a social networking site and posting on a blog or Web site run by someone else seems to be very popular. Therefore as a proof of concept, we have implemented this tool to pinpoint any attribution license violations on Flickr photos used in composite works that are in the form of HTML pages on the Web. In order to make sure that no CC license terms of the user are violated, the author can run the CC License Violations Validator and see if some sources have been left out or whether some have been misattributed.

Once the user gives the URI where the composite work can be found, the site crawler will search for all the links embedded in the given site and filter out any embedded Flickr photos. From each of these Flickr photo URIs, it is possible to glean the Flickr photo id. Using this photo id, all the information related to the photo is obtained by calling several methods in the Flickr API. This information includes the original creator's Flickr user account id, name and CC license information pertaining to the photo.

If a Flickr photo has a CC license attached, regardless of the purpose for which it is used, the photo should be given proper attribution as Flickr is still using the older CC 2.0 recommendation (as of April 2009). Therefore, if it was determined that a Flickr photo on a particular page has a CC License, the tool checks for the attribution information that can be either the *attributionName*, *attributionURL*, source URI or any combination of those within a reasonable scoping in the containing DOM element in which the image was embedded. The 'reasonable scoping' is defined to be some where within the parent or the sibling nodes in the DOM. If such information is missing, the user is presented with the details of the original content creator's name, the URI and the license it is under, enabling the user to compose the XHTML required to properly attribute the sources used.

## **4.2 Design and Implementation of the System**

This tool has four major components as shown in Figure 4-1.

### **4.2.1 Site Crawler**

This will search for all the links embedded in the given Web site starting from the document at the URI input by the user. The crawler uses a Breadth-First-Search algorithm and determines if there are any embedded Flickr photos. It avoids straying outside of the Web site for safety reasons as well as for efficiency reasons, but instead simply digs down into the Web page looking for embedded Flickr images. Once it is done looking for Flickr images embedded in that page, it follows links to other pages

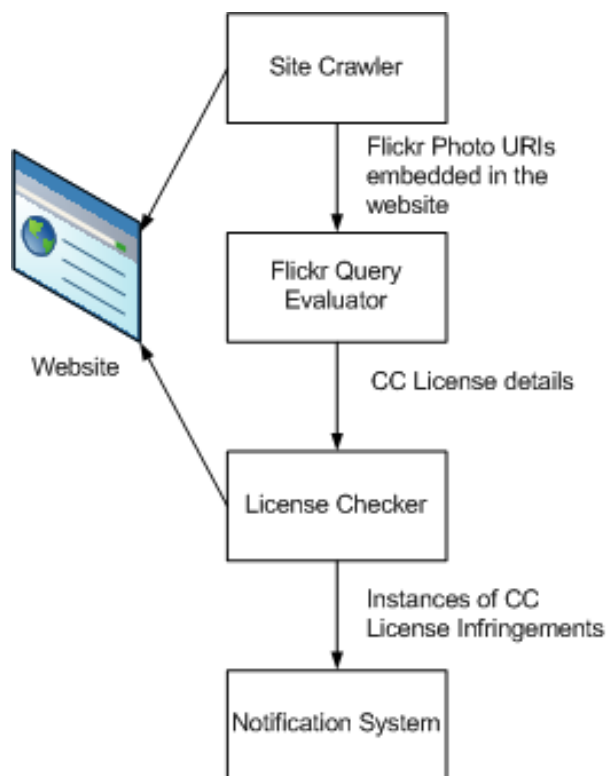


Figure 4-1: The Design of the Validator

within the same site. In order to follow these links, the crawler first parses the HTML 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 the past. This results in breadth-first traversal. The crawler avoids moving to another site by not following non-local links.

## 4.2.2 Flickr Query Evaluator

If the Site Crawler detects any embedded Flickr photos, this module will extract the photo id from the Flickr URI assuming that the URI is in one of the three formats given below:

`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)
```

Using this extracted photo id, all the information related to the photo is obtained by calling several methods in the Flickr API. This information includes the original creator's Flickr user account id, name and CC license information of the photo. It then queries the Flickr API for the CC License details <sup>1</sup>. The response from Flickr is obtained in the JSON data format [29], and after parsing that for the relevant license, we can determine the license attached to the photo. The license given by this query would be either *All Rights Reserved* or it would include a *CC license* which may have a combination of Attribution, Non-Commercial, No-Derivative and Share-Alike CC license terms. This module will also query the original photo owner's Flickr id, and then construct the Flickr user profile URI to check for attribution. This usually takes the following format:

```
http://www.flickr.com/photos/{Flickr User ID}
```

This module has very high reliance on the URI structure used for Flickr photos and the Flickr user profiles, as it performs several string operations on these URIs to obtain the photo id, and to construct the *attributionURL* to check for attribution.

### 4.2.3 License Checker

If a photo has a CC license attached with it, according to the older CC 2.0 recommendation that Flickr photos are still under (as of April 2009), the photo should be given proper attribution regardless of the purpose for which it was used. 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 or the Flickr User Name within the

---

<sup>1</sup>We used the FlickrLib python wrapper API [20] to query Flickr and obtain the license information.

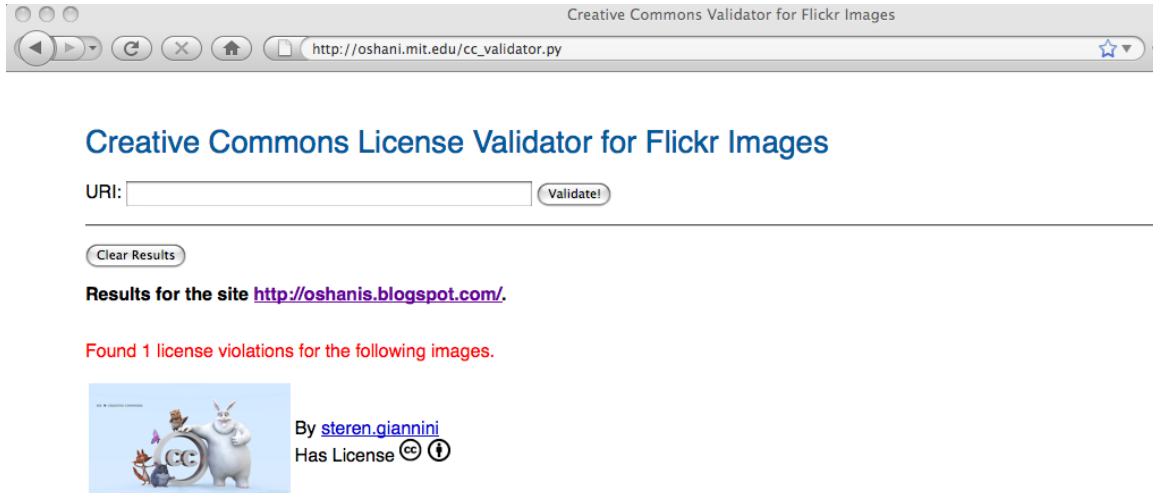


Figure 4-2: Output from the Validator

containing parent DOM node and the sibling nodes of the image for the attribution details (the same criteria used in the experiment described in Chapter 3). The reason for not doing a page level attribution check is because when two or more Flickr images are embedded in a page and if only one of those is properly attributed, this will result in an incorrect license violation detection.

#### 4.2.4 Notification System

In the current implementation of the validator, the “Notification System” is a Web interface which accepts the URI of a Web site to validate via a REST interface. If the validator finds any CC license violations, it will report those as shown in Figure 4-2. It will output the problematic image, who the original owner of the image is, and the license it is under. It is then expected that the user will go back to her compilation and correctly attribute the image in question using the information that appears on the user interface. This module is named the “Notification System” because when integrated with a closed world system such as QDOS [40] that has Flickr user account information integrated, it can be used to send notifications to the original content creator regarding the license violation.



# Chapter 5

## Semantic Clipboard

### 5.1 Inspiration

This draws inspiration from the work done on ‘XHTML Documents with Inline, Policy-Aware Provenance’ [27] by Harvey Jones. Jones developed a document format that can represent information about the sources of its content, a method of excerpting from these documents that allow programs to trace the excerpt back to the source, a Creative Commons reasoning engine which calculates the appropriate license for the composite document, and a bookmarklet that uses all these components to recommend permissible licenses. However, there are several issues with this tool:

1. All the source documents that the user needs to copy from, have to be annotated with a special document fragment ontology.
2. It can only be used for text copying and pasting.
3. The implementation of the *Paste Operation* is limited to inserting copied XHTML in the top level of the document, and does not allow copying inside existing document fragments.

The Semantic Clipboard described in this chapter tries to address these issues. Specifically:

1. The reliance of an external document fragment ontology is completely eliminated.
2. Image reuse is supported. But this tool this currently does not support text copying.
3. Since the Semantic Clipboard utilizes the operating system's clipboard, the image with the associated license metadata in XHTML can be copied to any editor application at any location as source text.

The only requirement for the Semantic Clipboard to work is that the license information about the work to be expressed in RDFa in the source documents.

## 5.2 Enabling Awareness of Licenses

License properties describe the permissions, prohibitions, and requirements of a license. Permissions declare a permission given by the license, above and beyond what default copyright law allows. Prohibitions prohibit a particular use of the work, specifically affecting the scope of the permissions provided by a permission. Requirements describe actions required by the user when making use of the permissions [10].

CC offers a handful of licenses that are applicable to most use and restriction scenarios. However, if these licenses are buried in the HTML or if the end user does not understand the CC license icons that signal the accepted use of the content, the user may not be aware of the licensing options available to her.

Also it would be useful to know which of the images that the user sees on a given Web page can be used for commercial purposes, which images allow modifications, which images have to be reused under the same license, etc. Figure 5-1 shows all the different CC license types grouped in to the most common *Use* and *Restriction* categories. Based on these categories, a menu is implemented to enable the user to easily select the images based on the user's intentions.

The *Use* categories available in the Semantic Clipboard menu are as follows:



1. Any Use: This category includes images licensed under ‘CC0’ or images that are under the ‘Public Domain Certification’. These are images where no rights are reserved, and reuse is allowed for any purpose without restriction under copyright.
2. Commercial Use: This category includes images that essentially have no ‘Non-Commercial’ use clause in the license. We imagine this option to be useful to individuals or organizations who would like to reuse existing images on the Web without having to worry about any legal ramifications.
3. Allow Modifications: This category includes images which do not have the ‘No-Derivatives’ clause in the CC license. These images can be reused in other free culture media.

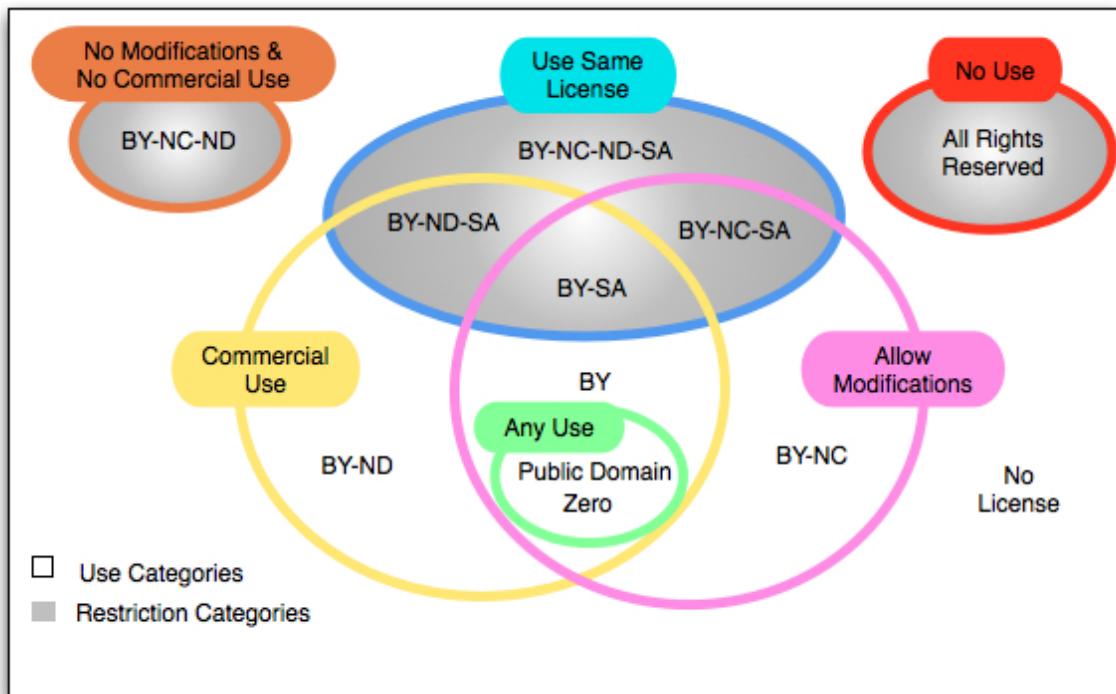


Figure 5-1: Venn Diagram for the Acceptable Use and Restricted Use of Content Depending on the CC License

The Venn diagram given in Figure 5-1 also includes few *Restriction* categories represented by the areas colored in gray. These are also incorporated in the Semantic

Clipboard menu to give users more awareness about which images cannot be used for a specific purpose based on their intention:

1. No Use: These are images that have explicitly given a copyright notice indicating that all rights are reserved.
2. Use Same License: This category includes images which are under a CC ‘Share-Alike’ license. Users should take care when combining such images with some other licensed works.
3. No Modifications and Commercial Use: These restrictions specify that no derivatives are allowed and the image can only be copied for commercial use. However, this allows the user to redistribute the composite work under a different license.

In the case of when there is no license, we make no decision. The original content creator may have had the intention to share the work with others. But since there is no indication of it expressed explicitly, copyright law specifies that those images are copyrighted and are not to be used. But we do not make this assertion, as the original content creator’s intention might have been otherwise.

To make these allowed uses and restrictions clearer, the Semantic Clipboard gives visual cues to the user to signal what can be copied based on the intention of the user as described in Section 5.3.2.

## 5.3 Design and Implementation

This is a Firefox Web browser based tool integrated with the Tabulator, a linked data browser that can be installed as a Firefox extension [47]. The primary goal of this tool is to let users reuse content with minimal effort.

When a Web page loads, the Semantic Clipboard registers the parts of the DOM that include the content intended to be reused along with their provenance and license metadata expressed in RDFa. Then it creates an XHTML snippet according to the Creative Commons specification for proper attribution. The design overview of the

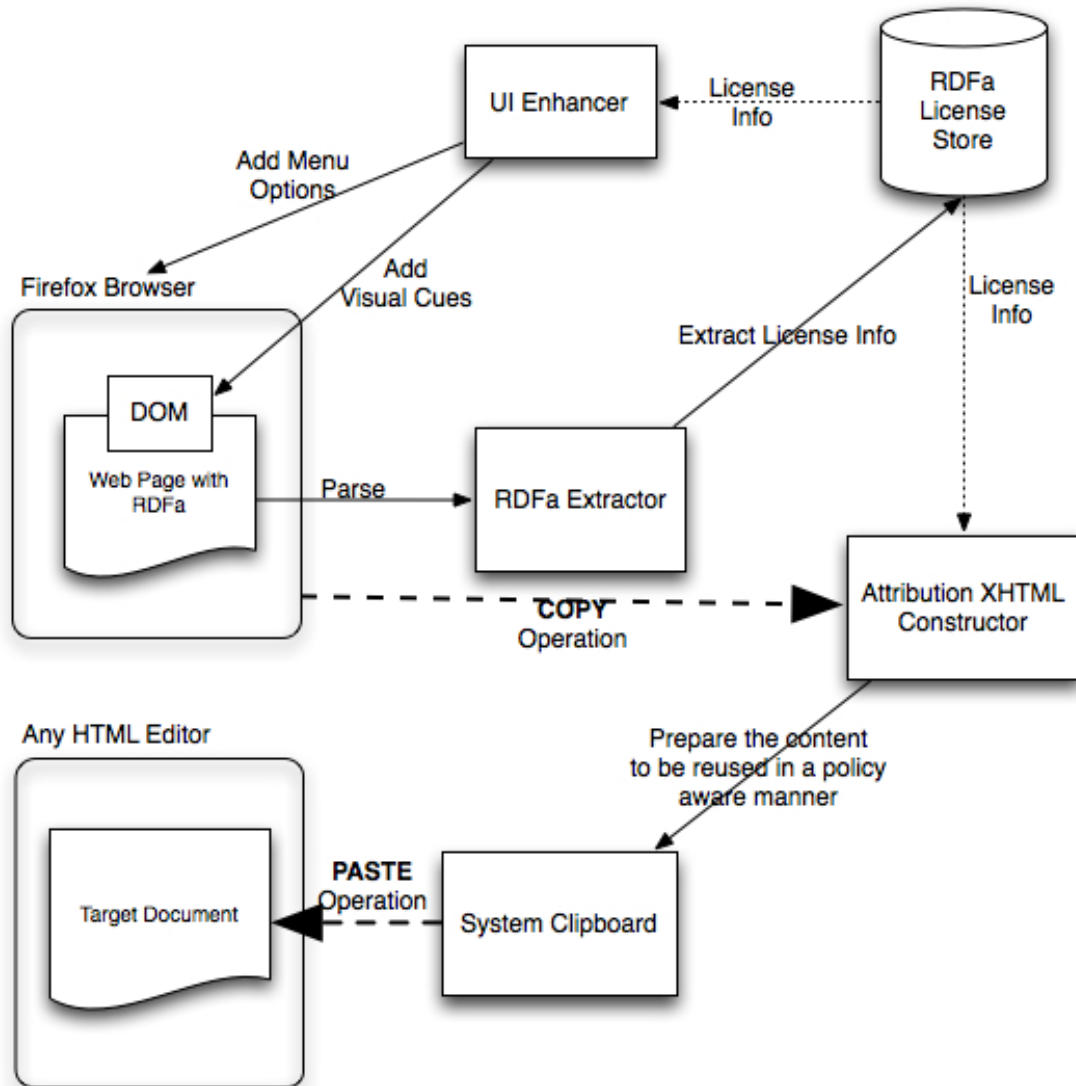


Figure 5-2: Design of the Semantic Clipboard

Semantic Clipboard is given in Figure 5-2. The following sections describe what each of the components are designed to do.

### 5.3.1 RDFa Extractor

This module will extract all the RDFa embedded in the HTML page the user browses. For example, consider the image shown in Figure 5-3. The creator of the image has embedded the image, and given some information regarding it in both human



*Plaza Mayor, Madrid.*

Tags: Spain, Madrid, Plaza Mayor, King Philips III, Statue

Taken on April. 25th by **Oshani**. (Lat: 40.4156/ Long:-3.7074)

This work by **Oshani** is licensed under a **Creative Commons Attribution-Share Alike 3.0 Unported License**.

Figure 5-3: Sample Image with Metadata about it Expressed in RDFa

readable and machine readable formats. If you look more carefully at the source of this embedded image in the document, you would be able to see the XHTML snippet adorned with RDFa as shown in Figure 5-8. This describes the image in terms of the creator, some tags, the date the photo was created, latitude, longitude, and the license it is under. All the *license metadata* about the image are given in red and the *other metadata* are given in blue.

The RDFa Extractor implemented in the Semantic Clipboard uses the JQuery Javascript Library [28] to parse the embedded RDFa in Web pages. JQuery simplifies HTML DOM traversing, event handling, and Ajax interactions for rapid web development. By using a JQuery specific ‘selector’ for the attributes, the RDFa from web pages are extracted. For example, the extracted RDF triples in the above XHTML snippet in Notation 3 serialization [3] is shown in Figure 5-9.

For this particular application we are only concerned about the license information associated with the images. Therefore, out of all the extracted metadata, only the

license information will be kept in an in-memory ‘RDFa License Store’ to be used in generating the attribution XHTML snippet.

### 5.3.2 UI Enhancer

We have implemented several menu options in the Firefox browser to select licensed images with the proper intention. These are based on the allowed *Use* and *Restriction* categories as discussed in Section 5.2. The options available within the tool are as given in Figure 5-4.

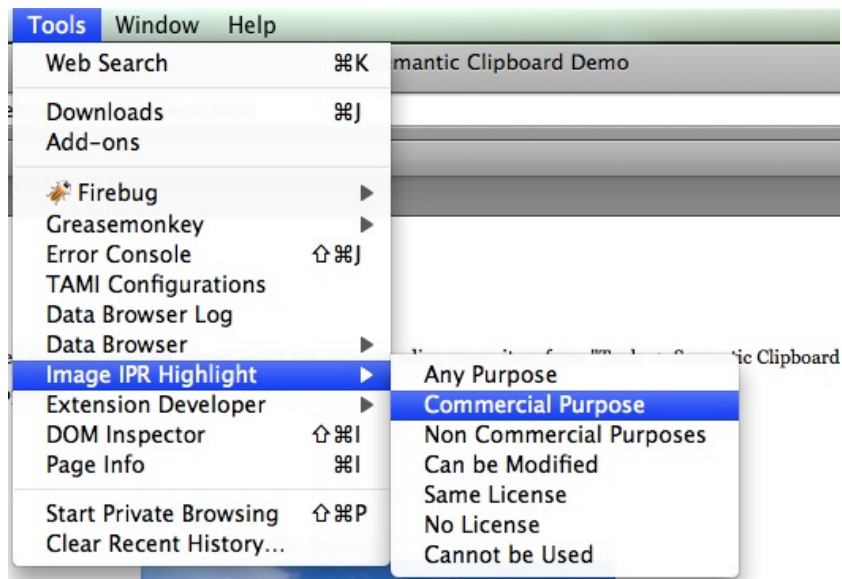


Figure 5-4: Firefox Menu to Specify User Purpose to See What is Available

Once the user selects the corresponding menu option, if there are any images that satisfy the criteria, those images will be overlaid with a small ‘text notice’. For example, in Figure 5-4, it shows that the user wants to see all the images on the Web page she’s browsing which can be used for commercial purposes. Once this option is selected, all the images on the page which have CC-Zero, CC-BY, CC-BY-SA, CC-BY-ND, or CC-BY-ND-SA licenses (essentially all the licenses which does not have a NC clause) will be overlaid with the ‘text notice’. An example is shown in Figure 5-5.

In addition, more subtle visual cues such as tool tip texts will be displayed if the

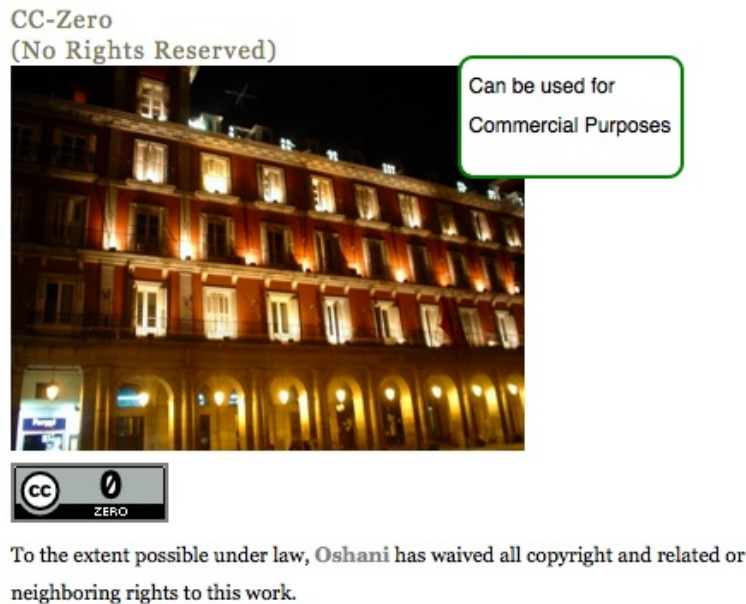


Figure 5-5: Image Overlaid With the Purpose Information

image has a CC license that allows it to be reused. An example of such a tooltip text is shown in Figure 5-6. Note that this tooltip will not be displayed if the image already has a ‘title’<sup>1</sup>. We hope that these visual cues will enable the end-user to select the images for copying in a more policy aware manner.



Figure 5-6: Tooltip on an Image giving Visual Cues as to whether the Image can be Copied or Not

### 5.3.3 Attribution XHTML Constructor

The user can issue a copy instruction on a particular image by right-clicking on the image and selecting the context menu option ‘Copy Image with License’ as shown

---

<sup>1</sup>The tooltip displays whatever the string that is given to the ‘title’ property of the image.

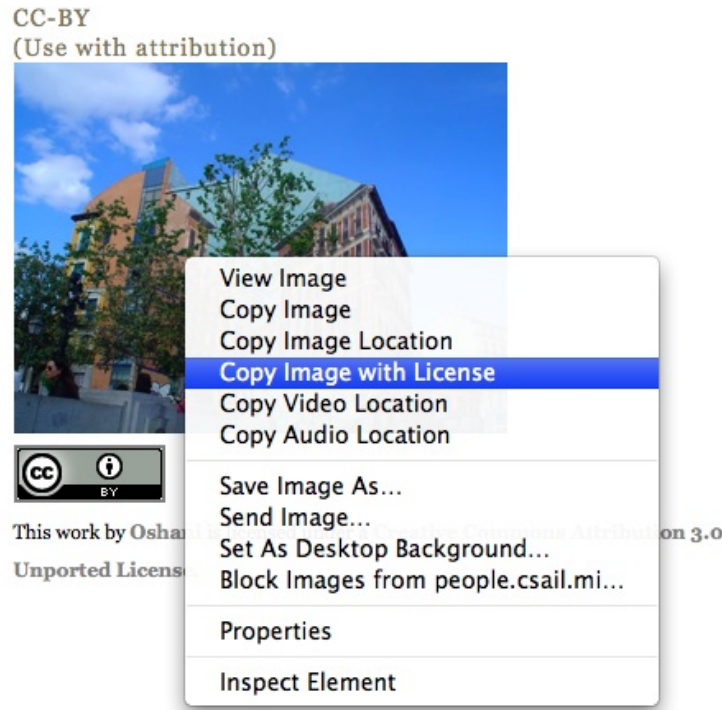


Figure 5-7: Right Click Context Menu on an Image

in Figure 5-7. Then, based on the image URI that acts as an index to all the subjects gleaned from the page, the RDFa license triples corresponding to this subject (i.e. the URI of the image that is being copied) are retrieved from the RDFa License Store. This constitutes the triples with the predicates *attributionName* and the *attributionURL*. Based on these triples, the attribution XHTML snippet is constructed. An example attribution XHTML snippet with the image embedded is given in Figure 5-10. This XHTML snippet will be copied to the system clipboard using a Mozilla specific component (namely ‘@mozilla.org/widget/clipboardhelper;1’ that implements nsIClipboardHelper) [48]. This allows the user to paste it in to any application that accepts ASCII text *data flavor* as input.

### 5.3.4 User Interface

The Firefox browser XUL [16] overlay is over-written such that the context menu that appears when images are right clicked will have the extra menu option ‘Copy Image with License’ as shown in Figure 5-7. Once this option is selected, the license data in

the RDFa associated with the image will be extracted as described in Section 5.3.1, the attribution XHTML will be constructed as described in Section 5.3.3, and this XHTML snippet will be made available in the system clipboard. Also, as mentioned in Section 5.3.2, the user interface in the browser can be overlaid with the necessary information about the allowed uses and restrictions of the images based on the licenses that are attached with those. This will assist the user in making an informed decision about what image(s) the user can copy for a specific purpose.

```

<div about="http://example.com/images/someImage.jpg">
  
  <a rel="license"
    href="http://creativecommons.org/licenses/by-sa/3.0">
    </a>
  <em property="dc:description">Plaza Mayor, Madrid.</em> <br />
  Tags:
    <span property="dc:subject">Spain</span>,
    <span property="dc:subject">Madrid</span>,
    <span property="dc:subject">Plaza Mayor</span>,
    <span property="dc:subject">King Philips III</span>,
    <span property="dc:subject">Statue</span>
  Taken on
    <span property="dc:date" datatype="xsd:date" content="2009-04-26">
  April. 25th</span> by
    <a rel="dc:creator" href="http://people.apache.org/~oshani/foaf.rdf#me">
      <span property="foaf:name">Oshani</span></a>.
    (Lat: <span property="geo:latitude">40.4156</span>/
    Long:<span property="geo:longitude">-3.7074</span>)
    This work by <a xmlns:cc="http://creativecommons.org/ns#"
href="http://people.apache.org/~oshani/foaf.rdf#me"
property="cc:attributionName"
rel="cc:attributionURL">Oshani</a>
    is licensed under a <a rel="license"
href="http://creativecommons.org/licenses/by-sa/3.0">
    Creative Commons Attribution-Share Alike 3.0 Unported License</a>.
</div>

```

Figure 5-8: Metadata Description of Image given in Figure 5-3



```
<http://example.com/images/someImage.jpg> cc:attributionName "Oshani" ;
cc:attributionURL <http://people.apache.org/~oshani/foaf.rdf#me> ;
dc:creator <http://people.apache.org/~oshani/foaf.rdf#me> ;
dc:date "2009-04-26"^^xsd:date ;
dc:description "Plaza Mayor, Madrid." ;
dc:subject "King Philips III", "Madrid", "Plaza Mayor", "Spain", "Statue" ;
dc:license <http://creativecommons.org/licenses/by-sa/3.0/> ;
geo:latitude "40.4156" ;
geo:longitude "-3.7074" .
```

Figure 5-9: RDF Metadata about the Image in Notation 3 Serialization

```
<div>
By
<a rel="cc:attributionURL" property="cc:attributionName"
  href="http://people.apache.org/~oshani/foaf.rdf#me" > Oshani</a> /
  <a rel="license" href="http://creativecommons.org/licenses/by-nc/3.0/">
    Creative Commons Attribution-Noncommercial 3.0 Unported</a></div>
</div>
```

Figure 5-10: Constructed Attribution XHTML Snippet



# Chapter 6

## Related Work

### 6.1 License Detection Tools

Popular search engines including Google, Yahoo and even sites such as Flickr, blip.tv, OWL Music Search and SpinXpress have advanced search options to find CC licensed content on the Web [8, 53, 17, 4, 37, 44]. In addition to these search engines, there are many tools that extract RDF from Web pages marked up with RDFa. RDFa Distiller [26] is one such RDFa parser. Even the CC License *Syntax* Validation Service [24] can be used to parse documents for embedded licenses in RDFa. After parsing the document, this service gives a list of licensed objects and each of their license authorship, version, jurisdiction, whether the license has been superseded or deprecated and whether the work is allowed in free cultural works. However, it does not give to whom the attribution should be given when reusing these license objects like the attribution license violations validator discussed in Chapter 4.

CC has put much focus on coming up with ways to enable tool builders to use the CC licenses very effectively. For example, the “*live box*” on the *License Deed Page* suggests how to attribute a particular work. This “*live box*” is created when a CC license hyperlink that has the *attributionName* and the *attributionURL* properties to the *License Deed Page* is dereferenced. Javascript code in the that page will scrape RDFa metadata from the referring page and construct the attribution XHTML as shown in Figure 6-1. There are also several license aware Mozilla Firefox extensions

developed by the CC. MozCC [34] is one such tool. It provides a specialized interface for displaying CC licenses, where the user receives 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. This extension is not supported in the latest versions of Firefox, and does not offer the capability to copy the license attribution XHTML as in the Semantic Clipboard that we have developed.

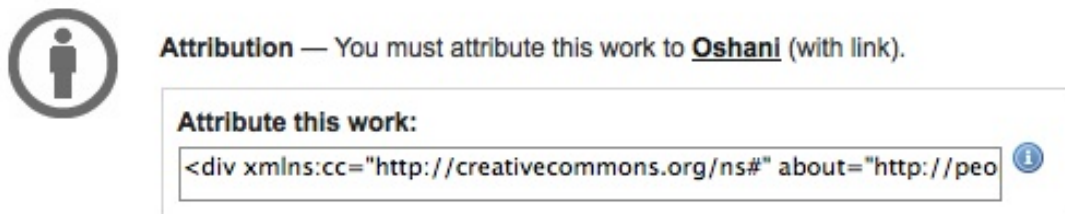


Figure 6-1: CC Deed Page Displaying the Attribution XHTML

Operator [36] is another Firefox browser extension that detects micro-formats and RDFa in Web pages that the user visits. Using Operator, it is possible to write an ‘action script’ that finds all CC licensed content inside a Web page by looking at the RDFa syntax. However, similar to the MozCC Firefox Extension, this cannot also be used to copy license information with the content as in the Semantic Clipboard.

## 6.2 License Embedding Tools

There are several tools which can be used to automatically embed the license metadata from Flickr. Applications such as ThinkFree, a Web based commercial office suite [46], and the open source counterpart of it, the “Flickr image reuse for OpenOffice.org” [19] are few examples of such applications. These applications allow the user to directly pick an image from the Flickr Web site and automatically inject the license metadata with it into a document in the corresponding *office suite*. A severe limitation of this approach is that they only support Flickr images. The Semantic Clipboard can be used to copy any image in to any target document with the license as long as the

license metadata is expressed in RDFa.

The python library ‘liblicense’ provides a straight forward way for developers to build license-aware applications by utilizing a pluggable module system for reading and writing metadata from specific file types [9]. This allows for the extraction and writing license information for files in GTK <sup>1</sup> applications. License Tagger [33] is built using this library and allows a user to add license metadata to audio, video, text and images. This application works for desktop applications and requires another application to interpret the embedded license metadata. However with the Semantic Clipboard, as long as the license metadata is expressed in RDFa, the content can be reused and embedded in another file in a license aware manner.

### 6.3 Transferring Metadata with Content

There is a tool called ‘News Credit’ [35] developed by the Media Standards Trust that was developed with the aim of making online news transparent. This tool embeds micro-formats with some specific enhancements to allow journalists to specify basic information to their news articles online. This helps establish an article’s authorship and provenance.

As mentioned in Chapter 5, there has been some work on annotating XHTML documents with provenance metadata using RDFa [27]. This work presents a method for performing copy and paste operations on XHTML documents in a way that preserves the metadata. This tool also incorporates a Creative Commons reasoning engine that reads document metadata and makes licensing decisions for annotated documents.

We also find inspiration in digital photos and their Exchangeable Image File Format (EXIF) [14] and the Extensible Metadata Platform (XMP) standard [52]. This information describes the photo, is embedded inside the photo itself, and is readable using simple tools. Therefore, it would be possible to embed license information inside the photo as well along with the other metadata. But this information can be easily overwritten should by a malevolent reuser.

---

<sup>1</sup>GTK is a tool kit for building graphical user interfaces.

## 6.4 Commercial Applications for Detecting Violations

Attributor [1], a commercial application, claims to continuously monitor the Web for its customers' photos, videos, documents and to let them know when they have been used elsewhere on the Web. Then it offers to send notices to the offending Web sites notifying link request, offers for license, request for removal or a share of the advertisement revenue of that page. Another commercial application called PicScout [38] claims that it is currently responsible for detecting over 90% of all online image infringements detections. They also claim to provide the subscribers of their services with a view into where and how their images are being used online.

The problem with these services is that it penalizes the infringers after-the-fact, rather than encouraging the them to do the right thing upfront [31]. Since their implementations are based on bots that crawl the Web in search of infringes, these services take up valuable Internet bandwidth [51]. Also, these services are not free, which bars many content creators who wish to use such services to find license violations of their content from using the service.

# Chapter 7

## Summary

### 7.1 Contributions

We have provided an assessment on the number of license violations on the Web, and thus highlighted the need to make Web users more aware of the licensing options available to them. The tools we have implemented provide policy awareness using existing Semantic Web technologies. These tools also provides a platform to use the data exposed on the Semantic Web.

The tools are:

1. *Creative Commons License Violations Validator for Flickr Images:*

This validator can be used to check if a certain Web site has used content inappropriately and thus violated anybody else's license terms when reusing Flickr images on the Web. If there are any violations, the validator provides the necessary information to include to make it license-compliant.

2. *Semantic Clipboard:*

This detects reusable images while a user is browsing and enables them to seamlessly integrate those images along with their metadata. Such addition of license metadata enables the transparent transfer of content between applications, and makes people more aware of the policies associated with content reuse.

## 7.2 Challenges

There are several limitations in the software we have developed for this project. Some of them are due to the lack of clear-cut definitions of license terms that can be directly transcribed to programmable verification methods. Some are due to portability problems. Some are due to user practices, while some are due to wide adoption of certain technologies. These challenges, and the possible future directions to circumvent those are given in the following sections.

### 7.2.1 Challenges for the Attributions License Violations Validator

#### Tracking Provenance

The validator sufficiently addresses the problem of inappropriate content reuse, specifically Flickr image reuse, on the Web. However we are unable to guarantee that it will handle all the cases. A malicious user could very easily change the image URI by uploading the same image on a different server. The license violations detection can only work if the the image URI is actually linked from the Flickr site.

Another interesting scenario is that any uploader can assign CC licenses to images on Flickr regardless of that user having the actual rights to do so. In other words, if someone uploads a copyrighted photo from *Getty Images* and assigns a CC license on Flickr, and an innocent user downloads and uses this photo, then that user will be violating copyright law without the user knowing it.

Therefore, we need to have some capability to track provenance of image data, and be able to identify whether a particular image has been used elsewhere in a manner that violates the original license terms. There are two hurdles in achieving that. *First*, image transformations are very easy to perform with image manipulation programs available today. Hence, even if there is a giant database of images (perhaps implemented by indexing the fingerprints of all the images on the Web), if one chooses to cheat, it can be easily done by changing a small unnoticeable pixel in the image and



changing the fingerprint as it is non-invariant to image transformations. Therefore, we are assuming that this tool will be used in a non-adversarial manner by honest people who wish to validate their documents against attributions license violations. *Second*, provenance might mean different things to different people. For example, one might think of provenance preservation as linking to the original URI, without embedding it in some other location. However, one might also argue that provenance would be preserved if the content is embedded in a new location, but mentions the original URI from where it came from. In the validator we have implemented, we assume that provenance is preserved by linking to the original image URI from Flickr.

### **Subsequent Changes in the License**

There is no law preventing a user from changing the license given to an image at any point in time. This can even be from a CC license to copyright protected. This makes the reuser vulnerable to copyright infringement claims by the original owner, putting the burden of proof on the reuser's shoulders. Therefore, a mechanism to record temporal changes in the licensing information will be very useful.

### **Scope of the Human Readable Attribution Notice**

One of the major assumptions we have made in developing this tool is that attribution to be specified within the parent node or the sibling nodes of the containing image element. Otherwise we classify it an instance of misattribution. This assumption works practically and seems to be the most logical thing to do. However, since there is no standard agreement as to what the correct scoping for attribution is, this assumption can give a wrong validation result. The solution to this problem can be in two folds. (1) CC should give a guideline as to what the correct scoping of attribution should be relative to the content that is attributed. (2) Flickr (any other such service) should expose the license metadata as RDF, instead of providing an API to query with. The latter method is preferred as it enables data interoperability and relieves the tool authors from having to write data wrappers for each service.

## Limited Flickr Support for License Expression

Flickr is still using the older CC 2.0 recommendation which specifies that the license metadata be included in HTML comments. This makes it impossible to process the metadata programmatically. This recommendation has been superseded by ccREL [21] that was published in March 2009. But Flickr has not yet implemented this recommendation, and as a result, users are not able to express *CC-Zero* (which means no rights are reserved with the image), or specify additional use restrictions via the *cc:morePermissions*, or use many of the other numerous additions in the new recommendation that supports richer semantics including the *cc:attributionName* and *cc:attributionURL*.

## 7.2.2 Challenges for the Semantic Clipboard

### License Granularity in the HTML DOM

Content creators have the freedom to assign whatever the license they see fit to their works. However, there is no standard agreement as to the granularity to which the license is applied. For example the decision about whether to apply the license to the entire document where the content is found, or whether to apply it to the individual content item is not specifically stated in the ccREL. Semantic Clipboard assumes that a license is applied to each image, and extracts the RDF metadata for that image in order to ascertain what the license for that particular image is. This assumption holds if the original content creator uses a tool when publishing her work that will automatically embed the license metadata for each content item rather than giving an all-encompassing license at the page level.

### Browser Dependence

One of the major drawbacks of the Semantic Clipboard is that it is Firefox browser-dependent. Firefox has only 22.48% of the browser market share according to [6]. Therefore, it seems that this application will not be widely adopted if it is not made available in all of the other major browsers. Developing an Opera Widget, a Chrome

Extension, a Safari Plugin, an Internet Explorer Content Extension or completely making this tool browser independent seem to be a viable future direction of the project in terms of making the tool available to the masses.

### **Usability vs. Operating System Independence**

It would be convenient to copy the attribution XHTML as *rich text* into a text editor which accepts hyperlinked text and displays it without the source text. Including this feature would be useful from a usability perspective, but it would raise concerns on the portability of the Semantic Clipboard across different operating systems. The Semantic Clipboard needs to have knowledge of the allowed data formats in the target applications. But this is a very operating system dependent parameter. The *data flavor* used when copying an image in the current implementation is ‘ASCII text’. Thus the application is portable to any Operating System. We have only tested the application on Mac OS X 10.5.6 and Ubuntu 9.04, but we assume that it should work in other platforms.

## **7.3 Future Work**

### **7.3.1 Check for Other Types of License Violations**

Detecting whether an image 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. But as was discussed in Chapter 2, from the point of view of the content consumers, it is hard to give a precise definition as to what constitutes a *Commercial Use*. For example, if the images are used in a Web site that has subscribed to a dynamic advertisement generation service, it can be easily argued that the advertisements that appear on the page and the image content that is embedded in the page have no direct correlation. However, as discovered by the CC survey [11], large number of content creators actually generate revenue indirectly through advertisements. Therefore, if the definition of *non commercial use* becomes

clearer and much more objective, the validator can be used to check for such violations as well.

It would also be interesting to check for share-alike license violations. These violations happen when a conflicting license is given when the content is reused. The solution, therefore, is to check the RDFa in both the original page and in the page where the image was embedded to see if the latter is the same as the original CC license.

### **7.3.2 Give Credit to the Original Content Creator as Requested**

The requirement for attribution is a form of social compensation for reusing one's work. While the mentioning of one's name or the URI when attributing trades attention to an individual, other forms of *attention mechanisms* can also be implemented. For example, a content creator can obligate the users of her works to give monetary compensation or require that they include certain ad links in the attribution XHTML or give attribution in an entirely arbitrary manner. These extra license conditions can be specified using the *cc:morePermissions*. Tools could be built to interpret these conditions and give credit to the original creator as requested.

### **7.3.3 Extend to Other Media Types**

We have only explored one domain of content, specifically image reuse on the Web. However, there are billions of videos uploaded on YouTube, and potentially countless number of documents on the Web, which have various types of licenses applied. While organizations such as Mobile Picture Association of America (MPAA), Recording Industry Association of America (RIAA) and other such big organizations are working towards preserving the rights of the works of their artists on YouTube, other video and audio sharing sites and peer-to-peer file sharing networks, there are no viable alternatives for ordinary 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 very useful.

### **7.3.4 License Granularity**

The tool we have developed only works if every image found on the page has its own license. Possible extensions of this tool would be to determine the license of an image when it does not have a license of its own, but is contained within a page that has a license or is a member of a set of images (e.g. a photo album) that has a license. Protocol for Web Description Resources (POWDER) [39], a mechanism that allows the provision of descriptions for groups of online resources, seems like a viable method to making the license descriptions about the resources explicit. Tool builders can then rely on the POWDER descriptions to help users to make appropriate content reuse decisions.

### **7.3.5 Persistent Data Storage**

Currently, images that are copied with their metadata to the Semantic Clipboard are overwritten when new content is copied to the clipboard. In other words, the tool only supports copying of one image at a time. It would be useful to have a persistent data storage to register content along with their license metadata, index them, make them persistent across browser sessions and use it whenever the user needs it.

### **7.3.6 User Study**

It would be interesting to measure how user behavior changes with the introduction of tools such as the License Violations Validator and the Semantic Clipboard. A measurement of the increased (or decreased or unchanging) level of license awareness would be an important metric in determining the success of these tools. Therefore, we would like to perform a controlled user study to determine how many violations a set of users make before and after the introduction of these tools during a limited period of time.

### 7.3.7 Widening Social Criteria

The central idea presented in this thesis is not limited to Creative Commons licenses. One could also combine it with a reasoning engine that reasons over security clearances to produce policy-aware tools that can calculate the appropriate level of data classification. Similarly, one could combine it with a tool that is aware of assertions regarding data reuse and create a tool that checks for violations of citizens' privacy [50]. Therefore, we hope the work on this thesis will spur research on how best to preserve data provenance in other domains.

## 7.4 Conclusion

As the license violations experiment indicated, there is a strong lack of awareness of licensing terms among content reusers. This raises the question as to whether the machine readable licenses are actually working. Perhaps more effort is needed to bring these technologies to the masses, and more tools are needed to bridge the gap between the license-aware and the license-unaware. An important research question that stems from this work is the method of provenance preservation of content on the Web. We have trivially assumed URIs as the provenance preservation mechanism when developing the tools described in this thesis. However, it would be an interesting challenge to track provenance based on the content itself, without having to rely on a unique identifier. This would enable us to find out license violators easily, in addition to validating one's own work for any violations. Also, programmatically determining whether a particular reuse of material is allowable or not is subjective, especially since some of the laws and standards have been quite ambiguous in defining these terms.

In general, social constraints are functions of any part of the blossoming Social Web we are experiencing today. As we are living in an era of increasing user generated content, these constraints can be used to communicate the acceptable uses of such content. We need tools, techniques and standards that strike an appropriate balance between the rights of the originator and the power of reuse. The rights of the originator can be preserved by expressing what constitutes appropriate uses and

restrictions using a rights expression language. These rights will be both machine and human readable. Reuse can be simplified by providing the necessary tools by leveraging these machine readable rights to make the users more aware of the license options available and ensure that the user be license or policy compliant. Such techniques can be incorporated in existing content publishing platforms or validators or even Web servers to make the process seamless. This thesis has demonstrated several tools that lay the foundation for such policy aware systems. We hope these will stimulate research in this area in the future.





# Appendix A

## Model Using the AIR Policy Language

### A.1 Introduction

This appendix describes expression of the CC-BY policy using the Accountability In RDF (AIR) policy language [30], and demonstrates a scenario in which this policy is used to generate a proof which indicates a person has violated somebody else's CC-BY policy.

The CC license violations validator service mentioned in Chapter 4 will help an honest user to remain honest by validating the user's 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 intend 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, has an expanded set of license options. AIR policy language 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 content should be reused. Given that we have a transaction log indicating how the content was used, we can reason over it using the AIR policy which describes the associated CC license.

## A.2 Scenario

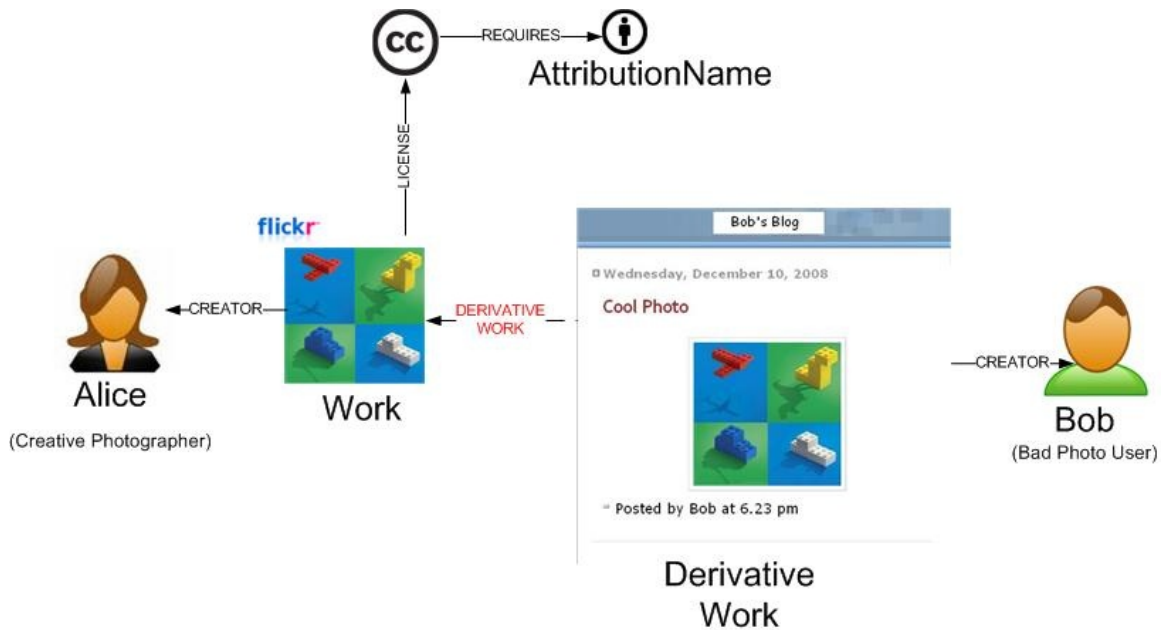


Figure A-1: Illustration of Inappropriate Usage of CC Licensed Content

As an example, consider the scenario given in Figure A-1, where Alice, a content creator, takes a photograph and uploads it 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 get logged, we can use the AIR policy given in Figure A-3 against the transaction log to derive a conclusion and a set of justifications for that conclusion.

The AIR policy for CC-BY given in Figure A-3 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 the log of an instance of content reuse, then there is a CC-BY license violation.

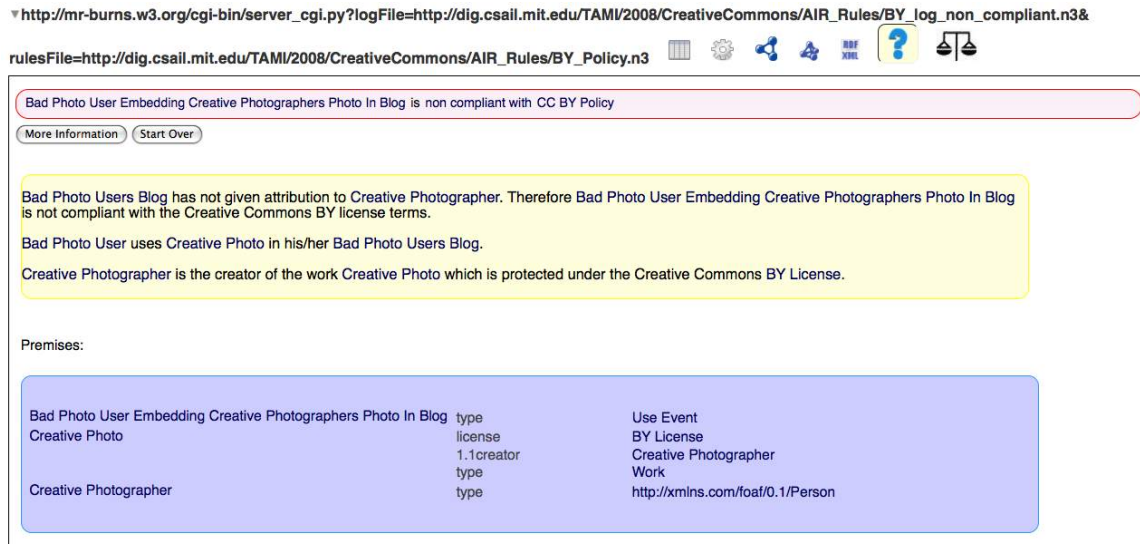


Figure A-2: Creative Commons Attribution License Violations Scenario Output from the AIR Reasoner using the Justification UI in the Tabulator

## A.2.1 CC-BY Policy Expressed in AIR

## A.2.2 Results

When this policy is run against the log of events collected from Alice’s and Bob’s transactions using the AIR reasoner, it would indicate if there were any policy violations and provide the justifications behind them. When viewed in the Tabulator’s ‘Justifications’ pane, the final conclusion of the outcome 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 A-2.

```

@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. ;
      ];
    ];
  ];
].

```

Figure A-3: Sample AIR Policy

# Appendix B

## Links to Code and Demos

### B.1 Experiment to determine Attribution License Violations on Flickr Images

Results from the Experiment:

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

The Source Code Used to run the Experiment:

<http://dig.csail.mit.edu/2008/WSRI-Exchange/src/experiment.py>

### B.2 Attribution License Violations Validator

Project Home Page:

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

The CC-license Violations Validator Service:

[http://dig.csail.mit.edu/2008/WSRI-Exchange/src/cc\\_validator.cgi](http://dig.csail.mit.edu/2008/WSRI-Exchange/src/cc_validator.cgi)

Source Code:

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

The presentation given at the Creative Commons Tech Summit:

<http://dig.csail.mit.edu/2008/Talks/1212-CCTechSummit-os>

## **B.3 Semantic Clipboard**

Project Home Page:

<http://dig.csail.mit.edu/2009/Clipboard>

Source Code:

<http://dig.csail.mit.edu/2007/tab/ext/chrome/content/semclip.js>

Tabulator Firefox Extension (to install the Semantic Clipboard):

<http://dig.csail.mit.edu/2007/tab/>

Demo Page:

<http://people.csail.mit.edu/oshani/SM/demo.html>

## **B.4 Creative Commons License Scenarios in AIR**

Project Home Page:

<http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/>

CC-BY Policy:

[http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/AIR\\_Rules/BY\\_Policy.n3](http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/AIR_Rules/BY_Policy.n3)

Log File:

[http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/AIR\\_Rules/BY\\_log.n3](http://dig.csail.mit.edu/TAMI/2008/CreativeCommons/AIR_Rules/BY_log.n3)

# Bibliography

- [1] Attributor - Subscription based web monitoring platform for content reuse detection. <http://www.attributor.com>.
- [2] Berne Convention for the Protection of Literary and Artistic Works. <http://law-ref.org/BERN/index.html>.
- [3] Tim Berners-Lee, Dan Connolly, Lalana Kagal, Yosi Scharf, and Jim Hendler. N3logic: A logical framework for the world wide web. *Journal of Theory and Practice of Logic Programming*, 2007.
- [4] blip.tv - Hosting, distribution and advertising platform for creators of web shows. <http://blip.tv/>.
- [5] Piero A. Bonatti, Claudiu Duma, Norbert E. Fuchs, Wolfgang Nejdl, Daniel Olmedilla, Joachim Peer, and Nahid Shahmehri. Semantic web policies - a discussion of requirements and research issues. In *ESWC*, pages 712–724, 2006.
- [6] Browser Market Share - 2009. <http://marketshare.hitslink.com/browser-market-share.aspx?qprid=0>.
- [7] Creative Commons BY 3.0 Unported Legal Code. <http://creativecommons.org/licenses/by/3.0/legalcode>.
- [8] Creative Commons Customized Search in Google. Google Advanced Search Enables CC-Customized Searching, <http://creativecommons.org/press-releases/entry/5692>.
- [9] Creative Commons LibLicense library which integrates license metadata with content. <http://wiki.creativecommons.org/Liblicense>.
- [10] Creative Commons License Properties. License Properties, <http://wiki.creativecommons.org/License-Properties>.
- [11] Creative Commons Noncommercial study interim report, <http://mirrors.creativecommons.org/nc-study/NC-Use-Study-Interim-Report-20090501.pdf>. some title.
- [12] Debate on copyright and wrongs: "Existing copyright laws do more harm than good". <http://www.economist.com/debate/days/view/310>.

- [13] Different Open source Licenses. <http://www.opensource.org/licenses/alphabetical>.
- [14] Exchangeable Image File Format . <http://www.exif.org/specifications.html>.
- [15] Joan Feigenbaum, Michael J. Freedman, Tomas Sander, and Adam Shostack. Privacy engineering for digital rights management systems. In Tomas Sander, editor, *Digital Rights Management Workshop*, volume 2320 of *Lecture Notes in Computer Science*, pages 76–105. Springer, 2001.
- [16] Firefox XUL Documentation. <https://developer.mozilla.org/En/XUL>.
- [17] Flickr API. <http://www.flickr.com/services/api>.
- [18] Flickr Farm URL. <http://www.flickr.com/services/api/misc.urls.html>.
- [19] Flickr image reuse for openoffice.org. <http://wiki.creativecommons.org/Flickr-Image-Re-Use-for-OpenOffice.org>.
- [20] FlickrLib - Python wrapper for the Flickr API. <http://monotonous.org/2005/11/26/flickrlib-05/>.
- [21] Hal Abelson, Ben Adida, Mike Linksvayer, Nathan Yergler. ccREL: The Creative Commons Rights Expression Language. *Creative Commons Wiki*, 2008.
- [22] Chris Hanson, Tim Berners-Lee, Lalana Kagal, Gerald J. Sussman, and Daniel J. Weitzner. Data-purpose algebra: Modeling data usage policies. In *POLICY*, pages 173–177. IEEE Computer Society, 2007.
- [23] How to attribute Flickr images. <http://www.squidoo.com/cc-flickr#module12311035>.
- [24] Hugo Dworak, Creative Commons License Validation Service. <http://validator.creativecommons.org/>.
- [25] International Press Telecommunications Council Photo Metadata Format. <http://www.iptc.org/IPTC4XMP/>.
- [26] Ivan Herman. RDFa Distiller, 2008.
- [27] Harvey C Jones. Xhtml documents with inline, policy-aware provenance. Master’s thesis, Massachusetts Institute of Technology, May 2007.
- [28] JQuery Javascript Library by John Resig. <http://jquery.com>.
- [29] JSON - JavaScript Object Notation. <http://www.json.org/>.
- [30] Lalana Kagal, Chris Hanson, and Daniel Weitzner. Using dependency tracking to provide explanations for policy management. *IEEE Policy 2008*, 2008.



- [31] Ken Doctor, Blog Entry on "Attributor Fair Syndication Consortium Completes Newspaper Trifecta" . <http://www.contentbridges.com/2009/04/attributor-ad-push-on-piracy-completes-newspaper-trifecta.html>.
- [32] Jong Wook Kim, K. Seluk Candan, and Junichi Tatemura. Efficient overlap and content reuse detection in blogs and online news articles. In *18th International World Wide Web Conference (WWW2009)*, April 2009.
- [33] License Tagger. <http://wiki.creativecommons.org/Licensetagger>.
- [34] MozCC - Firefox extension to discover Creative Commons licenses. <http://wiki.creativecommons.org/MozCC>.
- [35] News Credit - WordPress plugin. <http://newscredit.org>.
- [36] Operator Firefox Addon which discovers Microformats and RDFa in web pages. <https://addons.mozilla.org/en-US/firefox/addon/4106>.
- [37] OWL Music Search. <http://www.owlmusicsearch.com/>.
- [38] picScout - Image tracker for stock photography agencies and professional photographers. <http://www.picscout.com>.
- [39] Protocol for Web Description Resources (POWDER). <http://www.w3.org/2007/powder/>.
- [40] QDOS, Product developed by Garlik to measure the online status of an individual. <http://qdos.com/>.
- [41] RDF - Resource Description Framework. <http://www.w3.org/TR/rdf-syntax-grammar/>.
- [42] RDFa - Resource Description Framework in Attributes. <http://www.w3.org/2006/07/SWD/RDFa/syntax/>.
- [43] N. Shivakumar and H. Garcia-Molina. Scam: A copy detection mechanism for digital documents. In *Second Annual Conference on the Theory and Practice of Digital Libraries*, 1995.
- [44] SpinXpress - Collaborative media production platform. <http://spinxpress.com>.
- [45] Technorati API. <http://technorati.com/developers/api/>.
- [46] Think Free - Java based web office suite. <http://www.thinkfree.com/>.
- [47] Tim Berners-Lee. Tabulator Redux: Browing and Writing Linked Data . In *Linked Data on the Web Workshop at WWW08*, 2008.
- [48] Using the Mozilla Clipboard, Mozilla Developer Center. <https://developer.mozilla.org/en/Using-the-Clipboard>.

- [49] Daniel J. Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Gerald Jay Sussman. Information accountability. *Communications of the ACM*, June 2008.
- [50] Daniel J. Weitzner, Harold Abelson, Tim Berners-Lee, Chris Hanson, James Hendler, Lalana Kagal, Deborah L. McGuinness, Gerald Jay Sussman, and K. Krasnow Waterman. Transparent accountable data mining: New strategies for privacy protection. Technical Report MIT-CSAIL-TR-2006-007, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, January 2006.
- [51] William Faulkner, Tales From The IT Side: "PicScout, Getty Images and Goodbye iStockPhoto..!". <http://williamfaulkner.co.uk/wordpress/2007/09/picscout-getty-images-and-goodbye-istockphoto/>.
- [52] XMP - Extensible Metadata Platform. <http://www.adobe.com/products/xmp/index.html>.
- [53] Yahoo Creative Commons Search. <http://search.yahoo.com/cc>.