Detecting *Creative Commons License* Violations with *Flickr* Images on the Web

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

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

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

Hal Abelson
CSAIL, MIT
hal@mit.edu

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

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

ABSTRACT

Social networks, blogs, photo sharing sites and other applications known collectively as the social web has lots of increasingly complex data. There are many accountability issues associated with such over-exposed data on the web. This paper describes of a solution to an instance of such complex data usage relationships. The focus is on photo sharing sites such as Flickr, and blogging platforms such as Blogger or Live Journal (or simply any web site), with the aim of finding out if any Creative Commons Licences have been violated in reusing images.

Categories and Subject Descriptors

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

General Terms

World Wide Web, License Violation Detection, Policy Compliance, Search Algorithms

Keywords

Web Crawling, Semantic Web, Creative Commons Rights Expression Language, Accountability in RDF, Semantic Web Reasoning

1. INTRODUCTION

One could upload a photo on Flickr and attribute a Creative Commons (CC) license to it and make it publicly available. But there is no way of knowing if that photo was used for any other purpose by somebody else, unless she actually spots that. It would be nice to have an automatic way of checking whether any CC licensed photo was republished on a blog, but has not given attribution to the original creator when the license says "BY".

2. BACKGROUND ON THE CREATIVE COMMONS

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

Creative Commons (CC), has tried to answer questions pertaining to the promotion of reasonable and flexible copyright regime for the World Wide Web for several years now. One of the widely accepted solutions so far has been to create licenses that permit sharing and reuse with conditions, clearly communicated in human readable form. The other option is to leverage digital networks themselves to make licensed work more reusable and [1]

3. MOTIVATING SCENARIO

User Study - Harvest Some URIs and run the system over with all those

Focus this more as a validator

Science Commons work, where people have so many different sources, and they might want to run their own work on the validator to see any license violations of works.

Alice is an avid Flickr user and she uploads her photos to her account regularly. In her Flickr account settings she has applied "CC-BY-3.0" to all her photos by default. This means she allows anybody to Bob sees one of her photos which interest him, and he embeds the photo in his

4. RELATED WORK

DRM - Joan Feiganbaum's work

Much of the work that has been done in this area includes incorporating the CC license along with the meta data for an image. LibLicense /citeliblicense provides a low-level license metadata integration for applications.

XMP

5. SYSTEM DESIGN

Use of the AIR reasoner for more complicated license scenarios like original work licensed under CC-BY-SA being licensed with ACM license and CC-BY.

This system has 4 major components.

The crawler will look at a given site and determine if there are any embedded Flickr photos. If such photos are detected, License Checker will determine whether it is under a Creative Commons license. The User Checker will find out other identifying information related to the original creator of the photo. Since all Creative Commons licensed works should give attribution to the original creator, the crawler will again check whether the name or any other identifying information of the original creator appears on the page the

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

Figure 2: Format of Flickr Image URIs

photo is embedded on. The notifier will send a notification to the original creator about the data usage and license terms violation.

The crawler implements a basic BFS algorithm to check for Flickr image URIs in a given site.

A Flickr image URI takes one of the following formats:

From the photo URI, the id of the photo can be extracted. Using this id, all the information related to the photo could be obtained by calling several methods in the Flickr API. This information also includes the original creator's Flickr user account, name and CC license information pertaining to the photo. Again, using the crawler, the page is checked to see whether the original creator is attributed (only if the photo has a CC license attached). With the QDOS SPARQL endpoint, more of the photo owner's data (FOAF URI, etc) could be obtained, to perform a thorough search and notify in case of a license terms violation.

6. APPLICABILITY TO OTHER TYPES OF LICENSES

[2]

7. ISSUES

This will not work if the images are downloaded from Flickr and embedded in the site. Can write license information as EXIF data, but it could be easily overwritten BFS is only limited to all the links within the seed site Locality of the search for creator within the web page should be improved Notification to user for any license terms violation is not implemented. Needs user consent and worry about DPAs, etc.

8. FUTURE WORK

Extend to other CC licenses Extend to other data usage scenarios (for e.g. YouTube) Track provenance of images using metadata (instead of relying on the URIs) Social Verification: i.e. use the FOAF graph to control access for viewing, tagging and commenting on photo sharing sites Automatically inject the attribution details whenever a photo is linked

9. CONCLUSIONS

10. ACKNOWLEDGMENTS

This work was carried out for Web Science Research Initiative (WSRI) Exchange funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/F013604/1. We would also like to thank our colleague Harith Alani, Steve Harris and various devs on IRC channel 'cc' who helped us with numerous suggestions to explore. Last, but not least, we would also like to thank Mr. Puneet Kishore who contributed to us formulating the problem and coming up with a potential solution (he complained when one of us inadvertently published one of his CC licensed photos).

11. ADDITIONAL AUTHORS

12. REFERENCES

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

[2] Tim Berners-Lee and Dan Connolly and Lalana Kagal and Jim Hendler and Yosi Scharf. N3Logic: A Logical Framework for the World Wide Web. *Journal of Theory and Practice of Logic Programming (TPLP), Special Issue on Logic Programming and the Web*, 2008.

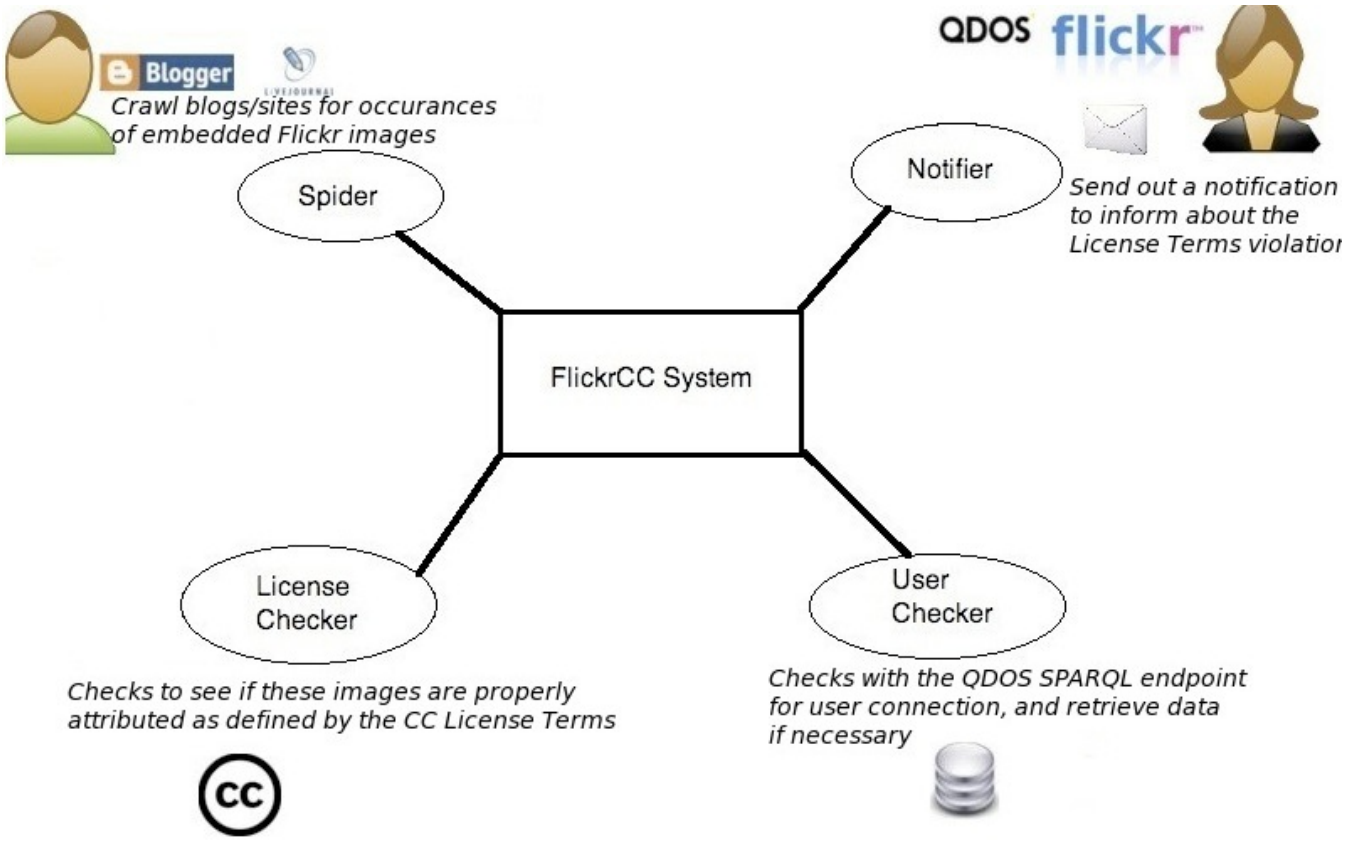

Figure 1: System Design