Harnessing Semantic Web technologies for solving the Dilemma of Content Providers

Claudia Wagner
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz
claudia.wagner@joanneum.at

Peter Scheir
Styria Media Group AG
Schönaugasse 64
8010 Graz
peter.scheir@styria.com

Alexander Stocker
Know-Center
Inffeldgasse 21a
8010 Graz
astocker@know-center.at

Wolfgang Halb
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz
wolfgang.halb@joanneum.at

ABSTRACT

The current business model of online newspapers is to create content and publish it on the Web. Content in turn attracts users to the online presence of a newspaper. This attention of users is monetized by presenting advertisements to them. The revenue of online content providers is mainly generated by advertisement and strongly depends on the number of users consuming their content.

Therefore - in general - content providers aim to attract as many users as possible with their content. However, to a large extent content providers make their content only available on one individual site in an unstructured form and consequently limit the accessibility and reusability of their content. Several reasons for this exist, among them are the way news have been published in the past, intellectual property rights and the way advertisements are currently sold.

In this paper we inspect the current situation of online newspapers and propose solutions for the current dilemma of content providers. We illustrate how Semantic Web technologies can help content providers to open their content, i.e. making their content reusable and integrable for third parties, without loss in revenue.

Categories and Subject Descriptors

H.4.4 [**Information Systems**]: Information Systems Application

General Terms

Semantic Web

Keywords

Linked Data, commercial aspects, licensing, usage restrictions

1. INTRODUCTION

In recent years the Web has been undergoing a dramatic change. The concept of Web 2.0 [5] summarizes several perceived changes on the Web, which have led, amongst others, to the emergence of innovative mash-up applications aggregating and combining content and data from different sources. As described in our previous work [8] the Content Republishing phenomenon, which can be observed on the Web 2.0, boosts diffusion and dissemination of content. Users may republish content by *reblogging* external content on their blogs, by *retweeting* microblog posts from other users on their own microblog or by *posting* external content to their Facebook[1] wall. Content becomes increasingly separated from individual sites and can be reused and recombined, providing new value for content consumers.

Professional online content providers, such as online newsportals, are very present on the Web since its early days and utilize it as one medium to publish their content. At present, to a large extent professional online content providers only publish their content on dedicated Web sites, most notably on their own ones. To successfully operate their business, they provide advertisements, acquired by themselves or taken from third parties (e.g. Google), embedded in their content. To determine the attractiveness for potential advertising customers online news providers measure the usage of their site by using Web analytics tools. These tools allow to record and analyze usage statistics on a very detailed level. As a rule of thumb the more users are attracted to an online portal, the higher the attractiveness for an advertiser, because more people can be reached by an advertising campaign. Therefore the amount of displayed advertisement directly influences the revenue of content providers, while usage statistics influence it indirectly.

Subsequently, content providers aim to reach and inform as many users as possible by the help of their content. But at present, they limit accessibility and reusability of their

[1] `http://facebook.com`

content, as content is usually only made available on one single site, most notably on their own. Unfortunately, by doing so, they are not able to benefit from potential cross-monetarization effects, as they are unwilling to disseminate their content. Currently, there is a lack of feasibility to measure content dissemination to make the increased reach transparent to advertising customers. Therein results the "dilemma of content providers".

In this paper we discuss Semantic Web technology oriented solutions to address the dilemma of content providers and to illustrate how professional online content providers can exploit new possibilities emerging through the recent paradigm shift of the Web. In Section 2 we introduce current models for online advertising and methods for the reuse of content on the Web. Section 3 provides a detailed description of the current dilemma of commercial content providers and Section 4 proposes approaches for solving the previously described problems. Finally, we relate our work with existing work in this area and draw conclusions for future work.

2. BACKGROUND

In this section we shed light on information that forms the basis of our work. We introduce current models for online advertising and methods for the reuse of content on the Web. The information introduced here sets the foundations for later sections.

2.1 Advertising based Business Models of online Content Providers

At present revenue of online content providers is mainly generated by advertisements and therefore is strongly dependent on the number of users consuming their content and the advertisements.

The two most prominent advertising *pricing models* are:

1. Cost per impression (CPI) model:
   The cost per impression model ensures that the content provider gets a certain amount of money from the advertiser for having displayed an ad a certain amount of times on his Web site. The advertisement profit depends on how often an ad is displayed and the price the advertiser is willing to pay for an ad impression. The most popular form of this advertising model is classical banner advertising.

2. Cost per click (CPC) model:
   The cost per click model demands that a visitor clicks on an ad to visit the advertiser's site. That means the advertisement profit depends on the number of clicks on an ad and the price the advertiser is willing to pay for a click. The most popular and well known cost per click advertising program is Google Adsense[2].

In addition to the advertising pricing model also the advertising *serving model* influences the advertisement revenue:

1. Advertisements are served directly by (or on behalf of) the content provider:
   If advertisements are directly served by the content providers they can be either embedded statically into the page or be loaded dynamically. In the later case advertisements can be loaded from local or remote servers. Advertisement revenue is generated directly by the content providers. If an external server is used a (usually volume based) fee is paid for the usage of the ad server.

2. Advertisements are served by third party services:
   If third party services serve the advertisements they are loaded dynamically to the site. Content providers embed certain tags as placeholder into their Web pages. The appropriate advertisements are loaded dynamically from the ad server. The advertisement revenue is shared between advertiser and the content provider.

The two presented pricing models regulate the way money is paid from the advertiser to the publisher. The technical realization of ad serving does not depend on whether advertisements are billed by CPI or CPC models. In terms of technological realization the central difference is whether the ad is embedded statically into the page or loaded dynamically from a local or remote server.

2.2 Content Reuse Methods

If content are reused across application boundaries we must distinguish two different methods:

(1) Content Reuse by Value:
If content are copied by value two individual instances of the same content item are generated.

An advantage of this traditional copy and paste mechanism is that instances of the same content item can evolve independently from each other. However, the biggest disadvantage of copying content by value is that the content are not kept up to date. If the original content change the copied versions are not updated. Another disadvantage is that copyright statements often prohibit copying content by value.

(2) Content Reuse by Reference (Transclusions):
Transclusions allow including existing content into documents without duplicating it. Content are copied by reference and content values are fetched and reloaded on demand. The concept of transclusions has originally been introduced in the early 1960s by Ted Nelson.

Transclusions have several advantages compared with traditional cut-and-paste mechanisms: they allow for keeping copied content up to date, avoiding copyright problems and saving disk space [4]. As shown by [3] basic Web technologies such as plain HTML, JavaScript, CGI scripts and a specialized HTTP proxy application allow realizing transclusions on the Web.

The biggest disadvantage of transclusions is that for textual content they never had their breakthrough and never found wide adoption. Furthermore, if the document from where the transcluded content originates becomes unavailable, the compound document must handle the broken transclusion.

3. PROBLEM DEFINITION

In general content providers aim to reach and inform as many users with their content as possible. If they would open their content, interlink it with other content and describe it semantically, third parties could aggregate and use

[2] http://www.google.com/adsense

their content. Since content would become accessible via different sites, it could possibly attract more users and the value and reach of content could increase. Content providers thus could reach more users by allowing third parties (i.e. users, Web application developers, software agents) to display and use their content.

However, at present content provider are forced to limit the access to their content to their own sites because of the following reasons:

1. **Presenting advertisements:** Currently widely used advertising methods (see Section 2.1) force content providers to keep their content locked on their sites, because ads are bound to individual Web sites instead of being bound directly to the content. Consequently the revenue of content provider depends on users consuming content via their sites. Content providers cannot open their content to make them integrable and reusable for third parties without a loss in revenue.

2. **Licences:** Parts of the content published by content provider usually comes from third parties, such as press agencies or photographers. Original intellectual property can be hold by editors that create content exclusively for a newspaper or by news agencies which (pre-)produce content for several newsportals. This content falls under limited licenses. Only online newspapers paying for the content are allowed to publish it on their sites. Republishing by other sites is prohibited by the license of the content.

3. **Measuring attractiveness:** The attractiveness of online news sites for an advertiser is determined using performance indicators, such as page impressions, unique clients, visits or use time. The page impressions indicator shows how many pages where displayed in a distinct period of time. The unique client factor indicates the reach of portal in terms of unique users. The visits indicator summarizes how often users visited a site and subsequently retrieved a set of pages. The use time indicates how long an individual user has stayed on a site during one visit.

   All of these traditional performance indicators operate on a site level and not on a content bits or data level. That means these indictors assume that digital content bits are only published and consumed on one individual site.

These contradicting factors cause the current dilemma of content providers.

4. SOLUTION APPROACH

In this section we refer to the three problems which force content providers to limit the access to their content (see section 3) and show how Semantic Web technologies can be used to address these problems.

4.1 Presenting Advertisements

Since content provider need a way to open and interlink their content without loosing advertisement revenues, we suggest to semantically annotate content (e.g. articles and advertisement) published on a Web site to allow machines to interpret the semantics of content and the relations between them. Usage control policies allow controlling not only who may access which content, but also how the content may or may not be used afterwards [6]. Machine-interpretable usage control policies can be used to expose, for example, that an article can only be reused together with its advertisement.

Listing 1 shows an example of a semantically annotated news site. The different resources (advertisement, article, embedded video) that are displayed on the site and the relations between them are described. The article is an instance of the class `Post` from the SIOC[3] ontology and has certain properties such as `content`, `author` and `topic`. The fact that the article embeds a video is exposed by using the `embeds` property of the SIOC ontology which indicates that a resources embeds external content (which may be related with certain policies and licenses). The article is related with an advertisement which is an instance of the class `Ad` of our advertisement ontology by using the `has_ad` property.

An ad has the following properties:

- `has_advertiser/advertiser_of`: relates an ad with an advertisement customer (instance of class Person or Organization of the FoaF[4] ontology) who is paying for the ad

- `advertises`: relates an ad with a product (instance of class ProductOrService of the GoodRelations[5] ontology) for which an advertisement is made

- `code`: relates and ad with its HTML and/or JavaScript code

```
<div xmlns=”http://www.w3.org/1999/xhtml”
xmlns:content =
”http://purl.org/rss/1.0/modules/content/”
xmlns:sioc = ”http://rdfs.org/sioc/ns#
xmlns:ex=”http://example.com/terms/”>

<p about=”#article” typeof=”sioc:Post”
 property=”content:encoded”>
 <p>article text</p>
 <div property=”sioc:embeds”>
   <p about=”#video”
     property=”content:encoded”> video here </p>
 </div>
 <div property=”ex:has_ad”>
   <p about=”#ad” typeof=”ex:ad”>
     <span property=”ex:code”>
      <object><embed
      src=”http://mydomain.com/picturebutton.swf”
          type=”application/x-shockwave-flash”>
      </embed></object>
     </span>
     <span rel=”ex:has_advertiser” resource=
          ”http://www.gigasport.at/about/#company”/>
     <span rel=”ex:advertises” resource=
          ”http://www.gigasport.at/product/234/#this”/>
  </p>
 </div>
```

Listing 1: ”Semantic description of advertised content”

To describe how resources can be used and reused we are using the Open Digital Rights Language (ODRL[6]). ODRL is a vocabulary which allows expressing terms and conditions over assets (Web resources which include any physical

[3] `http://sioc-project.org/`

[4] `http://xmlns.com/foaf/spec`

[5] `http://www.heppnetz.de/ontologies/goodrelations/v1`

[6] http://odrl.net

or digital content). ODRL allows expressing permissions which indicate the actions that a certain party (e.g. user, role, group) is permitted to perform on a specific target asset manifestation (i.e. format) or a range of manifestations of the target asset. Constraints may optionally constrain permissions and duties may indicate requirements that must be fulfilled to receive a permission [1].

Listing 2 extends the example of listing 1 with usage policies which indicate that the video can only be reused i.e. be aggregated, modified and extracted (permission) together with the ad (duty). The meaning of permissions and duties is defined by the ODRL data dictionary[7].

```
<div xmlns="http://www.w3.org/1999/xhtml"
 xmlns:ox="http://odrl.net/1.1/ODRL-EX#"
 xmlns:od="http://odrl.net/1.1/ODRL-DD#">

 <p typeof="ox:Permission">
   <span rel="ox:has_asset" resource="#video"/>
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
        content="excerpt" />
   </span>
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
        content="aggregate" />
   </span>
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
        content="modified" />
   </span>
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
        content="annotated" />
   </span>
   <span rel="ox:has_duty">
      <p typeof="ox:Duty">
        <span rel="ox:has_object" resource="#ad" />
        <span rel="ox:has_action">
           <span typeof="ox:Action" property="ox:name"
                      content="display" />
        </span>
      </p>
  </span>
</p>
</div>
```

Listing 2: "Usage Policies binding ad to video"

Specifying that a certain asset can only be reused together with a certain ad allows content providers to separate their content from their sites without loosing advertisement revenues. Exposing this information in a machine-interpretable way by using well defined semantics and agreed-upon ontologies is a crucial factor, because it allows third party applications (e.g. end-user tools such as reblogging tools or mash-up creation tools such as Yahoo Pipes[8] and Semantic Web Pipes[9]) to interpret these policies.

If content are reused by copying content values, the content copying or aggregation tools must be able to interpret policies and ensure that the copied version of the content is also published under certain terms and conditions. If content are reused by reference (as transclusion) the proxy service which resolves the reference must be able to interpret policies.

4.2 Licenses and Usage Restrictions

Since parts of the content published by content provider usually comes from third parties, such as press agencies or photographers, original intellectual property can be hold by editors that created the content. Therefore content providers need a way to expose under which conditions what kind of content usage is allowed. To address this problem we again suggest the use of Semantic Web policies. With the help of policies machines will be able to interpret if the permission to reuse a piece of content in a certain context is given or not.

Listing 3 extends the example of listing 1 with usage policies which indicate that the article can only be displayed. Content reuse actions, such as extractions, aggregations and modifications, are explicitly prohibited by machine-interpretable polices.

```
<div xmlns="http://www.w3.org/1999/xhtml"
 xmlns:ox="http://odrl.net/1.1/ODRL-EX#"
 xmlns:od="http://odrl.net/1.1/ODRL-DD#">

 <p typeof="ox:Prohibition">
   <span rel="ox:has_asset" resource="#article" />
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
      content="extract" />
   </span>
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
      content="aggregate" />
   </span>
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
      content="modify" />
   </span>
</p>

<p typeof="ox:Permission">
   <span rel="ox:has_asset" resource="#article" />
   <span rel="ox:has_action">
      <span typeof="ox:Action" property="ox:name"
      content="display" />
   </span>
</p>
</div>
```

Listing 3: "Copy restricted content"

4.3 Measuring Attractiveness

Content providers need a way to measure and analyze usage statistics of their content which may be distributed on the Web. Performance indicators for Web sites, such as page impressions, unique clients, visits and use time, need to be extended to be able to take usage of off-site content - i.e. content not presented on the original site - into account. We see three options for addressing this issue described as problem 3 in Section 3:

(1) One solution would be to force third party applications to attribute the source of content with detailed provenance metadata which can be used to find all third party applications using a certain resource and to create a provenance index. Third party applications must expose their usage statistics in a machine-interpretable way to allow content provider to take them into account when analyzing the usage statistics of their content.

    This approach however has several limitations: First the content provider or third party tracking services must crawl the Web to find distributed content and build a provenance index. Second the measurement and exposure of usage statistics must be standardized in order to allow different applications combine their usage statistics. Third content providers need ways to esti-

[7] http://www.w3.org/TR/odrl/#24552

[8] http://pipes.yahoo.com/pipes/ (20.06.2009)

[9] http://pipes.deri.org/ (20.06.2009)

mate trustworthiness of usage statistics exposed by third party applications.

(2) Another option would be to only allow reusing content by reference and prohibit copying content values. Transclusions allow content provider to continue controlling their content and handling requests from target applications which transclude their content. Consequently the usage statistic can still be measured by the content providers on their sites, although the content is consumed and accessed on other sites.

In this case the target application is however also not allowed to cache content on their platform. Every time a user wants to see the transcluded content on the target application it must be reloaded from the source application.

(3) Finally a promising option would be the use of an already existing (and often even in place) measuring methods: counting pixels or so called tags. Here a tag is embedded into the source code of the page that should be counted. Usually, a HTML `img` tag located on a remote server is used for this purpose.

Recently also remote JavaScript files which are embedded in HTML pages are used for this purpose. Counting methods are invoked on each page load. A well known tool that analyzes usage statistics of web sites via remote JavaScript files is Google Analytics[10]. Also national organizations such as the OEWA[11] in Austria are using this technique and publish performance content collected via this method.

Semantic annotations can be used to describe measuring resources (i.e., the counting script or image) and relate them with the content resources for which usage statistics should be measured. Formal policies can be used to expose that certain resources can only be reused together with certain measuring resources. Thus the usage of reused content will always be trackable. Based on the URL of a site where a resource is displayed the Web analytics tools will be able to distinguish between usage statistics of the original content and the reused content.

5. RELATED WORK

Usage control policies allow controlling not only who may access which content, but also how the content may or may not be used afterwards [6]. Usage control policies of digital content on the Web can take several forms:

The Copyright approach usually 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 license. Several experiments have been conducted in [7] and show the lack of license-awareness on the Web in general. Especially, the level of CC attribution license violations on the Web has been explored in this work.

The Digital Rights Management (DRM) approach usually restrict access to the content, or prevent the content from being used within certain application. DRM approach allows content owner to specify how their content can be used and reused. However, classic DRM methods cannot be used by content providers because first they want to open and interlink their content and not lock them and second it is not possible to completely control the dissemination of digital assets. Companies such as Apple had to learn this lesson and stopped implementing DRM methods [2]. The newly unlocked tracks now include the purchaser's name and other personally identifying information so that if users violate the usage policies they can be held accountable for that.

6. CONCLUSION

In this paper we have described the dilemma of professional content providers who on the one hand want to reach as many users as possible with their content but on the other hand must lock their content on their sites for various reasons such as advertisement revenue, copyright issues and the way usage statistics are generated.

To overcome these problems we have described how Semantic Web technologies can be used to semantically describe content bits and related usage restrictions in a machine-interpretable way. Finally we have proposed solutions for tracing content dissemination on the Web and generating cross-platform usage statistics.

7. REFERENCES

[1] S. Guth and R. Ianella. Odrl v2.0 - core model. Draft Specification, January 2009.

[2] S. Jobs. Thoughts on music. http://www.apple.com/hotnews/thoughtsonmusic/, February 2007.

[3] J. Kolbitsch and H. Maurer. Transclusions in an html-based environment. *Journal of Computing and Information Technology*, 14:161 – 174, 2006.

[4] H. Krottmaier and D. Helic. Issues of transclusions. In *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (E-Learn 2002)*, 2002.

[5] T. O'Reilly. What is web2.0? design patterns and business models for the next generation of software. web, 2005.

[6] Pretschner, Hilty, and Basin. Distributed usage control. In *Communications of the ACM*, 2006.

[7] O. W. Seneviratne. Framework for policy aware reuse of content on the www. Master's thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2009.

[8] C. Wagner and E. Motta. Data republishing on the social semantic web. In *in Proceedings of Workshop "Trust and Privacy on the Social Semantic Web" co-located with the ESWC 2009*, 2009.

[10] `http://www.google.com/analytics/`

[11] Austria Web Analysis `http://oewa.at/` (20.06.2009)