



Good Copyright Compliance Practices
for
Web Content Aggregations

Northern Light Whitepaper

Northern Light®

November 2013

TABLE OF CONTENTS

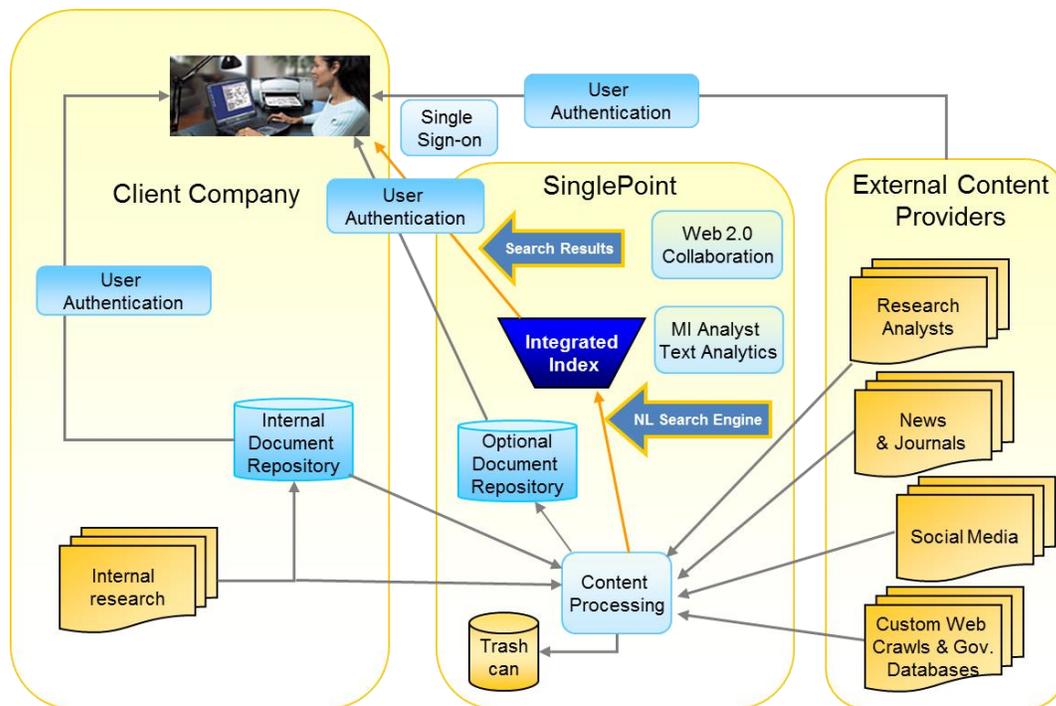
Background On Northern Light and SinglePoint	2
Northern Light Experience With Web Content Aggregations	2
Web Content and Copyright	3
Real World Example	3
Fair Use	4
The purpose and character of the defendant’s use	4
nature of the copyrighted work	4
amount and substantiality of the portion used in relation to the copyrighted work as a whole	5
effect of the use upon the potential market for, or value of, the copyrighted work.....	5
Web Search Engines and Copyright.....	5
AP and Meltwater.....	6
Lessons for Companies Buying Web-Based Aggregations.....	7
Conclusion.....	8
About the Author	9
Contact Information	9

BACKGROUND ON NORTHERN LIGHT AND SINGLEPOINT

Northern Light is in a single business: providing strategic research portals to global, new product, new technology, and innovation driven organizations in many industries. Our strategic research portals have been used for market research, competitive intelligence, product development, and technology research to global enterprises since 1999. Our current strategic research portal client list reads like a who's who in technology and industry, including leaders in many industries:

- Life sciences
- Financial services
- Telecom
- Information Technology
- Manufacturing
- Agribusiness
- Consumer products
- Logistics and distribution
- Corporate strategy consulting
- Energy

Altogether, there are over 200,000 individual users of our *SinglePoint™* strategic research portals at companies like the ones above. *SinglePoint* is a hosted, turnkey solution provided in its entirety by Northern Light. The client licenses, creates, or simply identifies the content to be included in the research portal, but after that Northern Light handles all other aspects of the portal including development, configuration, deployment, content aggregation, indexing and search, text analysis, collaboration, user management, document security, and reporting. Below is an overview of the *SinglePoint* solution.



A typical *SinglePoint* has 10-20 licensed external syndicated research sources with hundreds of thousands of research reports, several internal primary research repositories, a business news feed, and is used by 5,000 users within our client's organization. Northern Light also harvests content from five thousand websites accounting for 20 million news stories, and indexes 30 million scientific and journal articles.

Each *SinglePoint* client has a custom search index built from the specific content providers and content services that the client licenses and/or wishes to have in the client's strategic research portal. If a client needs a third-party research provider that is not in our universe of existing relationships, Northern Light has an established process for bringing new sources into the *SinglePoint* system. Clients will never outgrow Northern Light's ability to integrate content on the clients' behalf.

Northern Light *SinglePoint* portals enjoy a high ROI, paying for themselves ten to twenty times per year. Northern Light publishes a whitepaper on this topic titled *SinglePoint Strategic Research Portal ROI Analysis* and it is available in the Knowledge Center on our corporate website at <http://northernlight.com/knowledgecenter>.

NORTHERN LIGHT EXPERIENCE WITH WEB CONTENT AGGREGATIONS

As part of the *SinglePoint* portal product line, Northern Light provides search and access to several aggregations of Web content to clients. These include:

- *Northern Light Business News* – 40,000 news stories per day from 5,000 business and technology-focused sources, with 15 million articles in the archive.
- *IT Analyst Blogs* and *IT Analyst Tweets* – over 1.5 million posts from 2,300 IT analyst tracked by name through the social Web.
- *Life Sciences Conference Abstracts and Posters* – over 1.0 million abstracts and posters from over 1,500 research conferences and meetings held around the world since 2010.
- *Government News and Publications* – 30,000 documents a day representing every news story, announcement, and official document published by over 200 branches, agencies, and departments of the U.S. Government.

WEB CONTENT AND COPYRIGHT

Web content is unruly, misbehaved, messy, scattered, badly formatted, when formatted at all. It's also informative, insightful, timely, and crucial for business and technical research and analysis.

Web content that is "free" and "openly-accessible" to the public is a fast growing content category that often has the best and most current strategic information. Web news aggregations are now more comprehensive and useful than most licensed newswires. This trend is spreading like an out-of-control wildfire to other content categories including competitive intelligence, social media, scientific research, and regulatory and compliance information.

Copyright infringement can carry substantial financial penalties for companies that distribute or use copyrighted content improperly. However, this issue is of special concern with Web content because, despite being easily accessible without payment of a subscription fee, the content found on webpages is almost always copyrighted. Organizations, users, and even some news aggregators frequently confuse "free" and "openly-accessible" with "copyright-free."

REAL WORLD EXAMPLE

To illustrate the importance of getting ahead of the copyright issue, consider this example. A global company solicited bids from Web news aggregators to establish a monitoring system of news coverage of their brands. The requirement included making and storing actual images of the news articles for later retrieval. Two web content aggregators, let's call them A and B, competed for the business.

Aggregator A declined to provide the requirement to store and reproduce the full-text of the Web news articles on the basis of copyright considerations and was dropped from the bidding process. Aggregator B agreed to the requirement regarding storing images of the webpages. The company spent hundreds of thousands of dollars on the solution with Aggregator B.

The *night* before the solution was to deploy to internal users, someone in the marketing department, perhaps remembering Aggregator A's original objection to the requirement to store and reproduce the full-text of the aggregated Web news articles, decided to ask the company's attorneys to sign off on the solution. The legal depart took one look at the new brand tracking system and vetoed its deployment on the spot.

Why? Because, despite the fact that only "free" and "openly-accessible" content was in the solution, the storing and reproduction of the full-text of the aggregated content potentially violated the copyrights of the content owners. The whole solution was abandoned and the substantially-into-six-figures budget for the project completely lost.

How can your company avoid a situation like this one?

FAIR USE

As anyone knows who has used a card catalog in a library or read a book review on Amazon, indexing, excerpting, and summarization of other people's copyrighted content are not new practices. Indexing services are as old as the publishing business. Book reviewers often reveal substantial portions of a plot and reveal the nature of the protagonists' personalities. Also, general practice in a wide variety of fields is to quote with attribution from copyrighted works without much, if any, fear of getting sued for it.

From a legal viewpoint, the principal of "fair use" governs the legality of indexing, excerpting, and summarization products. Regarding fair use doctrine, courts apply a four-part test to determine if the copied material is fair use or not.

THE PURPOSE AND CHARACTER OF THE DEFENDANT'S USE

Copying for nonprofit educational purposes is more likely to be deemed fair use than copying of a commercial nature. For example, copying material for classroom use may be permissible even if that same material may not be copied for commercial purposes. So a teacher might be able to copy a news article and give it out to students in a classroom to facilitate a lesson while it would be unacceptable for a marketing department to include the full-text of the same news article in a newsletter to employees.

Another consideration more relevant to commercial settings is whether the use of the copied material is to produce a "transformative" work that is something new and different than the original. Works that are transformative might be able to incorporate copyrighted material. A transformative work adds substantial new creative expression (e.g., a parody) or new functionality not found in the copied work by itself (e.g., sifting large amounts of information or text analytics across many news stories).

NATURE OF THE COPYRIGHTED WORK

Generally, facts are not copyrightable, but creative expression of facts can be. So it could be acceptable to copy facts from a news article and republish them, but republishing the copyright holder's creative expression of those facts in addition to the facts themselves would be less acceptable.

For example, hypothetically copying "Google acquired Motorola today and announced it would join the Android patent lawsuit brought by Apple against Samsung" might be okay because it is totally factual. But one would be on thinner ice copying and repeating "Trying to be the dog instead of the tail in the Android patent lawsuit by Apple against Samsung, Google acquired Motorola today." The "dog" and "tail" language moves in the direction of creative expression.

AMOUNT AND SUBSTANTIALITY OF THE PORTION USED IN RELATION TO THE COPYRIGHTED WORK AS A WHOLE

A key consideration is whether the copied version is so complete as to substitute for the original. The items provided by the indexing and excerpting products have to be sufficiently brief or short so as to not substitute for the original. There is no bright line for how much of a work can be copied before it is too complete, and courts look at whether the “heart” of the copied work has been reproduced and whether there is objective evidence on whether the users of the copied work do not then use the original work. Click-through rates, for example, can illuminate the issue of how complete an excerpt is, with low click-through rates suggesting the excerpts are so complete the need for the original is diminished.

EFFECT OF THE USE UPON THE POTENTIAL MARKET FOR, OR VALUE OF, THE COPYRIGHTED WORK

This factor considers whether the copied work is marketed in competition with the original work in existing and potential markets and whether or not existing licensing programs are being circumvented by the alleged infringer. For example, if established licensing programs for the copyrighted material are not being utilized in favor copying the content from websites that have paid a licensing fee to the copyright owners, then the copying is less likely to be deemed fair use.

WEB SEARCH ENGINES AND COPYRIGHT

The largest aggregation and indexing of Web content is done by Web search engines such as Google. A number of courts in different cases have ruled in favor of Web search engines that have been sued by website content owners for copyright infringement. In those cases, they have found that fair use permits Web search engines to provide indexes, citation metadata, tags, and excerpts of copyrighted content.

Supporting this finding, courts have ruled that the provision of a Web-content index supports a transformative purpose which is the ability of users to sift through a large amount of information which practically speaking would be impossible without the search engine indexes. This sifting process is different than any purpose intended for the original content, and hence transformative.

An additional factor supporting the success of the fair use defense by Web search engines has been that providing links to original material on the Web supports the notion that the indexing, excerpting, and summarization of Web content as practiced by Web search engines does not negatively impact the audience for the copyrighted material, and perhaps actually improves it. The question is whether the copied work, in this case the search excerpts reproducing text from the webpages on search results, substitutes for the original content on the webpages. Google News reports a 56% click-through rate which has been cited as evidence the excerpts of copyrighted webpages in Google’s search results are not in fact substitutes.

AP AND MELTWATER

The Associated Press (AP) recently sued Web news aggregator Meltwater for copyright infringement of freely-accessible Web news content. This case is illustrative because the court reviewed the entire history of Web-content copyright litigation in its decision. Meltwater defended itself by claiming to be a Web search engine and citing the long string of favorable decisions for Web search engines in prior copyright litigation. In a sweeping victory, the court granted summary judgment to AP on its claim that Meltwater had committed copyright infringement, and that its copying was not protected by fair use.

Trying to briefly summarize the high points of the 94-page court decision, the court ruled that Meltwater was not protected by fair use doctrine in part because:

- Meltwater’s claim that their use of the copyrighted material from news websites was transformative in the sense that Web search engines have been found to provide a transformative use was rejected by the court. The click-through rate in Meltwater’s application was reported to be .08% which is about one-thousandth of the 56% click-through rate that Web search engines like Google News report. The court cited this fact in concluding that the Meltwater application is not transformative.
- Meltwater’s excerpts were so complete that the need to click through to the original article was effectively eliminated making them substitutes for the original articles. Meltwater’s excerpts were between 5% and 60% of the original article. The court discussed court cases that considered a wide range of reproduction of the text (e.g., 8%, 33%) and in the end did not establish a “bright-line” standard. The court said one test of whether the excerpts were too complete was whether the amount of copied text was necessary to accomplish a transformative purpose. Having rejected the transformative purpose of the Meltwater application, the excerpt length could not be used to justify the copying of any given percentage of text from news articles.
- Search engines do not deliver the full-text or complete excerpts that replace the need for the original documents, which is one reason they have high click-through rates. The court observed that another type of business, clipping services, make exact copies of article text and deliver the copies to clients. Obviously, having a copy eliminates the need for their clients to consume the original article in its original location. The court held that Meltwater was a clipping service instead of a search engine. Making a copy of an article (clipping it) is not transformative because the purpose of the copy is the same as the purpose of the original. The court observed that legitimate clipping services pay fees to the publishers for the rights to deliver the full-text to their clients.
- Meltwater always reproduced the entire “lede” paragraph in its excerpts. The lede is the first paragraph of the news story and often contains original creative expression by the article’s author. While facts cannot be copyrighted, original expression can be and Meltwater’s own court filings implied the lede was creative expression. Also, the court observed that in a news story, the lede is the “heart” of the article. (As every high school English composition student knows, the first paragraph of a news story contains the “who, what, when, and where” of the article.) With the lede reproduced in a summary or excerpt, one has both the author’s creative expression, and the “heart” of the article which might well substitute for the original document.

- The court observed that Meltwater’s application included a feature that facilitates the copying, pasting, saving, reuse, and distribution of the full-text of articles inside the users’ organizations for uses such as newsletters. The court decision describes a Meltwater feature called the “Article Editor” that launches a display of a news article in one browser window when a link is clicked and a paste-to box in the Article Editor window open at the same time. Meltwater’s collateral encouraged a copy and paste process between the open windows.
- AP has a licensing program that Meltwater could have used. The court held that not enforcing AP’s copyright would give Meltwater an unfair advantage over its license-paying competitors, while depriving AP of licensing revenue on its property.
- The court observed that Meltwater continued to incorporate AP content into its news aggregation despite it being common industry knowledge that AP did not want to be included in such services.

LESSONS FOR COMPANIES BUYING WEB-BASED AGGREGATIONS

So, what lessons should companies wanting to include Web content in their business research activities take from this court decision? There are some key points that will reduce the risk of copyright violation in news and research applications that use Web content:

- Do not copy and distribute the full-text of copyrighted news articles or other content from websites - even internally.
- Make sure your applications do not encourage copying and redistribution of the full-text.
- Lastly, actively discourage your employees from working around copyrights by use of techniques such as copying webpages or article text and attaching the content to emails, posting it on internal collaboration sites, or using it in newsletters.

And for Web content aggregated by an external vendor, make sure your content aggregator:

- Provides only indexes, citation metadata, tags, and short excerpts. These activities are traditional fair use by indexing services.
- Provides excerpts that are a small fraction of the original text. These pieces should not be so complete as to substitute for the original document.
- Does not serve copies of copyrighted material from the aggregator’s own repositories rather than connecting you to the original content. Does not specifically target the lede paragraph in news stories for excerpts.
- Demonstrates a click-through rate on the solution above 50%. The aggregator’s reporting system should help you measure that for your organization’s use of the service as well. (For example, Northern Light’s click-through rates are 76% as measured by our portal reporting systems.)

- Does not have search portal application features or documentation that facilitate copying, saving, or distributing the full-text from websites.
- Uses links to the original documents in every use case, promoting the use of the original material. Portal applications should save *links* when documents are bookmarked, share *links* when the articles are shared, and post *links* when articles are posted to collaboration pages. Working with links to the original webpages instead of copies of the content from the webpages is a copyright-compliant means of saving, sharing, and posting.
- Has a policy and practice to exclude content from any source that objects to being included.
- Licenses full-text content with redistribution rights from sources that are deemed to be essential.

Lastly, have your legal department review the solutions that you are considering licensing for copyright compliance. Do this *before* you license the solution!

CONCLUSION

Use of content published originally on the Web is an unavoidable activity for any market intelligence, product marketing, technology research, or strategic planning professional. But you have to keep your eye on the copyright-ball when using Web content.

The bottom line is this: it is not hard to stay compliant if you pay just a little attention to good practices.

ABOUT THE AUTHOR

Despite his habit of late-night reading of court decisions for entertainment, author C. David Seuss is not an attorney. (So don't take anything in this whitepaper as legal advice!) Rather, he is the CEO of Boston-based Northern Light, an aggregator of Web and licensed news and research content, and a provider of strategic research portals for global leaders in many industries.

CONTACT INFORMATION

C. David Seuss
CEO
Northern Light
One Constitution Center
Boston Massachusetts, USA, 02129
david@northernlight.com
Mobile: 1-617-515-5771