Environmental Health News -- Building an on-line news portal using Semantic Web technologies

Cormac Twomey
Siderean Software,
Los Angeles, California, U.S.A.
ctwomey@siderean.com

Todd Koym
Edgerton Foundation,
Washington, DC, U.S.A.
todd.koym@edgertonfoundation.org

John Peterson Myers
Environmental Health Sciences,
Charlottesville, Virginia, U.S.A.
jpmyers@ehsic.org

Abstract

Environmental Health News (EHN) is an on-line resource that covers issues concerning human health and the environment. Published daily by Environmental Health Sciences, its mission is to help increase public understanding of emerging scientific links between environmental exposures and human health. The EHN service provides a web site and delivers email newsletters to a growing audience of readers and subscribers, providing them with frequently updated summaries of breaking news, scientific research and reports by environmental organizations. All content is tagged with terms from the EHN subject taxonomy, which covers a broad spectrum of environmental health issues.

Keywords

Content management; Educational Web Applications; Faceted Navigation; Environmental Health; Taxonomy

1. Introduction

In July 2003, the Edgerton Foundation and Environmental Health Sciences contacted Siderean Software to investigate enhancing the EHN service by making use of Siderean's faceted navigation engine, Seamark. At the time, EHN was manually maintained, with about 30 to 40 article summaries added per day. This labor-intensive editorial process was an impediment to extending EHN's scope and audience. To achieve its ambition to be a primary resource for activists, scientists, media outlets and the public at large with respect to matters of environmental health, EHN needed an innovative content management system and information architecture that would allow it to grow. To address this, the Edgerton Foundation and Siderean Software jointly developed a new content management system built from the ground up using semantic web principles. The new EHN service encodes all content on the site in RDF, using standardized terms. Specifically, the new site was designed to:

- Aggregate content about health and the environment from distributed, heterogeneous sources.
- Support the creation and aggregation of value-added metadata about the content.
- Provide effective and flexible navigation and search over a large and growing repository of content metadata.
- Support the creation of customized syndicated feeds of the value-added metadata for delivery across diverse media platforms.

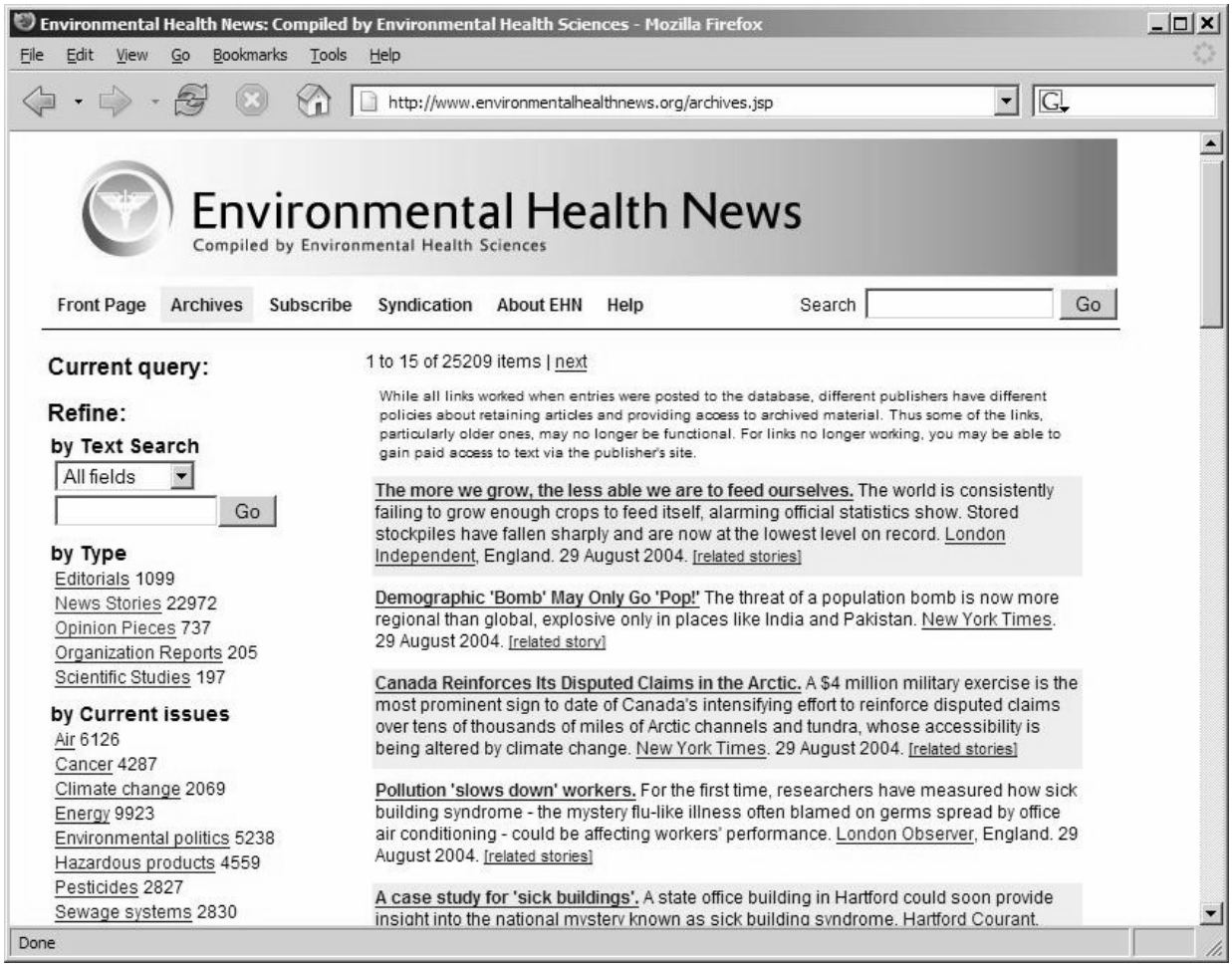
The new EHN service is a proprietary application designed by Siderean and EHN runs on a single RedHat Linux PC, hosted by Siderean. It is comprised of two main components:

- Apache Tomcat hosting Java Server Pages which provide the front-end user interface and editorial workflow.
- Siderean Seamark Server acting as the metadata store and navigation engine.

Both components run on the same machine, and communicate with each other via a SOAP interface. The current version of the EHN service can be visited at http://www.environmentalhealthnews.org/.

2. Navigational Features

2.1 Faceted Navigation



Archives Page

All articles on the site have been classified according to the EHN taxonomy. This categorization has been leveraged throughout the site, in particular on the site archives page, which allows the user to browse the results of their search using faceted navigation [1]. The top-level terms are shown as separate navigation facets on the archives page:

- Human Health condition
- Contamination agent
- Exposure pathway
- Ecological effects
- Infrastructure
- Solutions
- Emerging Science

The article type and publisher information are also exposed as navigation facets.

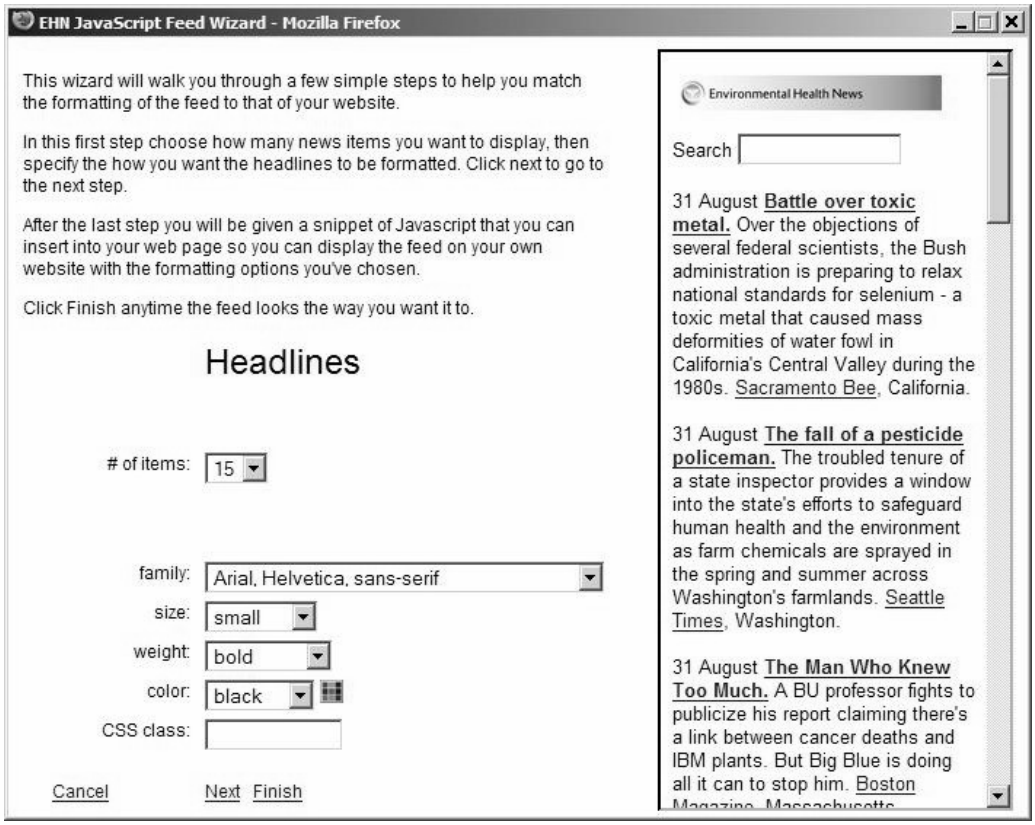
Example Query

2.2 Customized Syndication

One of the primary goals of the EHN project is not only to provide a centralized repository of environmental health content, but also to ensure that it can be leveraged and reused by third parties. To this end, all content on the EHN site is available through RSS 1.0 feeds. Indeed, each individual search results page can be obtained as RSS 1.0, providing site visitors with customized feeds tailored to their specific interests. These feeds can then be used by individuals to monitor any area of particular interest. For example, users of the site can create RSS feeds to monitor:

- new articles that relate to a topic or set of topics in the EHN taxonomy;
- new articles added to the site that are related to a particular article of interest;
- new articles that correspond to a text search, or
- any combination of the above

2.3 JavaScript Syndication



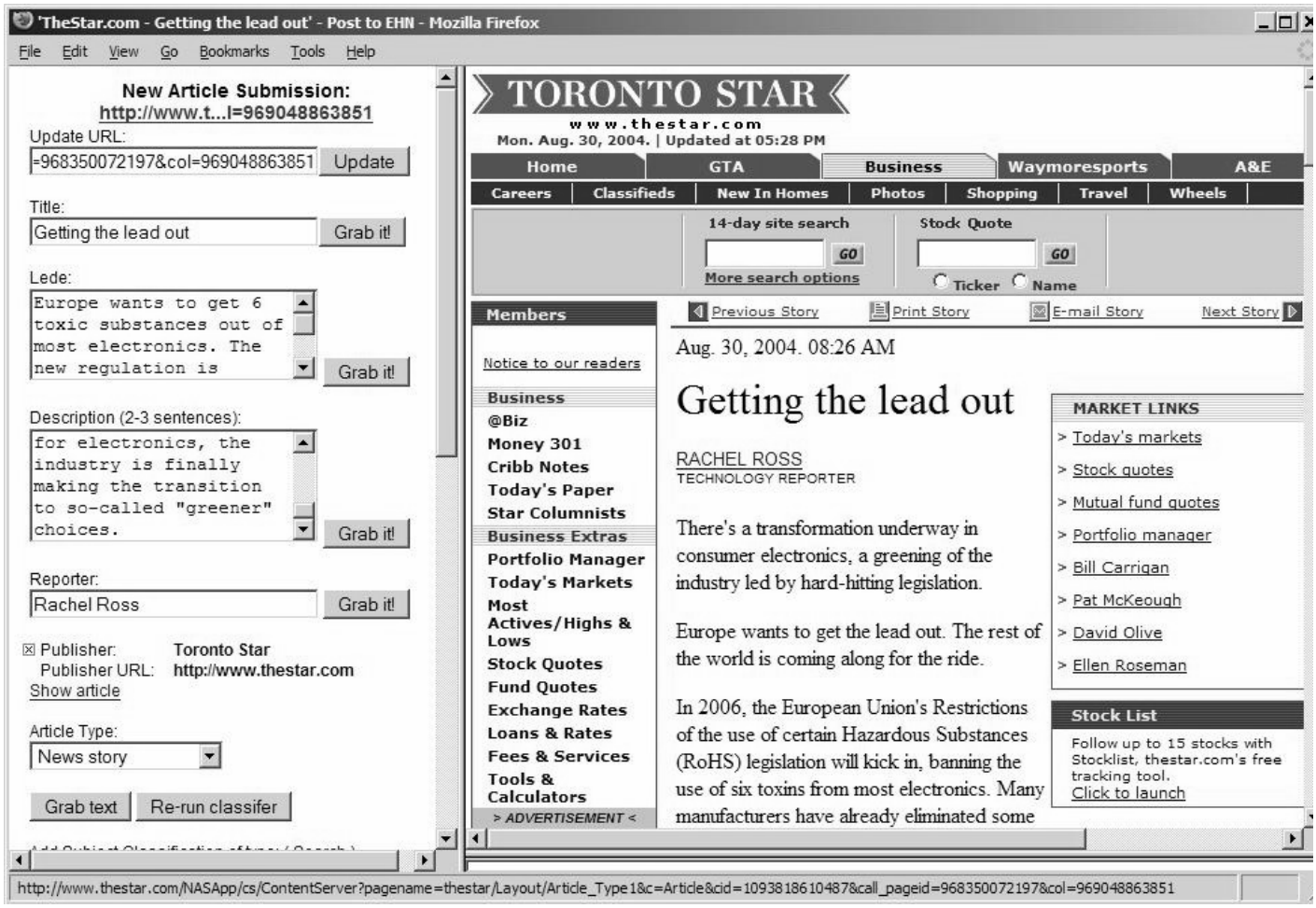
Javascript feed

In addition to RSS syndication, EHN also offers JavaScript-based syndication, allowing websites to drop a feed from EHN directly onto their website simply by adding an external JavaScript link to their web page. The site offers a JavaScript feed wizard which allows the user to customize the appearance of the content to match the look & feel of their own site.

3. EDITORIAL FEATURES

3.1 Manual metadata markup

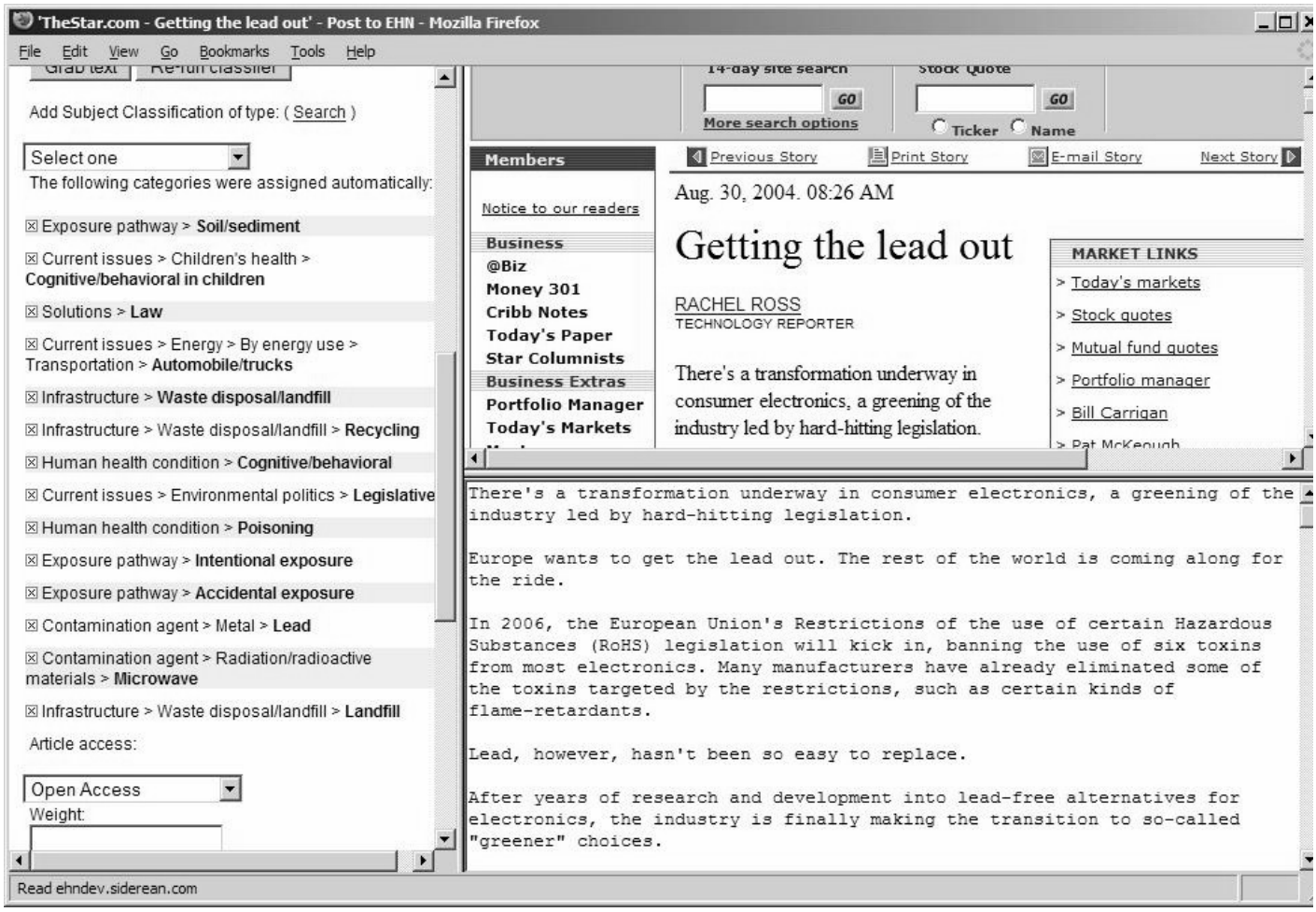
EHN editors submit new articles to the EHN site by using a simple HTTP bookmarklet when viewing the page to be submitted. This raises a window with a two pane view -- metadata on the left, article on the right:



Article Submission

3.2 Automatic metadata markup

The EHN taxonomy has also been tagged with matching rules, which allows the site to automatically tag the submission with dc:subject properties from the taxonomy. For the first version of the new EHN site, the rules are simply regular expressions which match against the article text. While it is planned in the future to use more sophisticated named-entity extraction techniques, this simple approach has proven effective – editors estimate that 90% of the effort involved in manual categorization is saved. The automatic categorization occurs when the contributor selects the text in the right-hand pane, and clicks "Grab Text." A third pane appears on the bottom left, displaying the article text for the editor to review. The results of the categorization appear on the left.

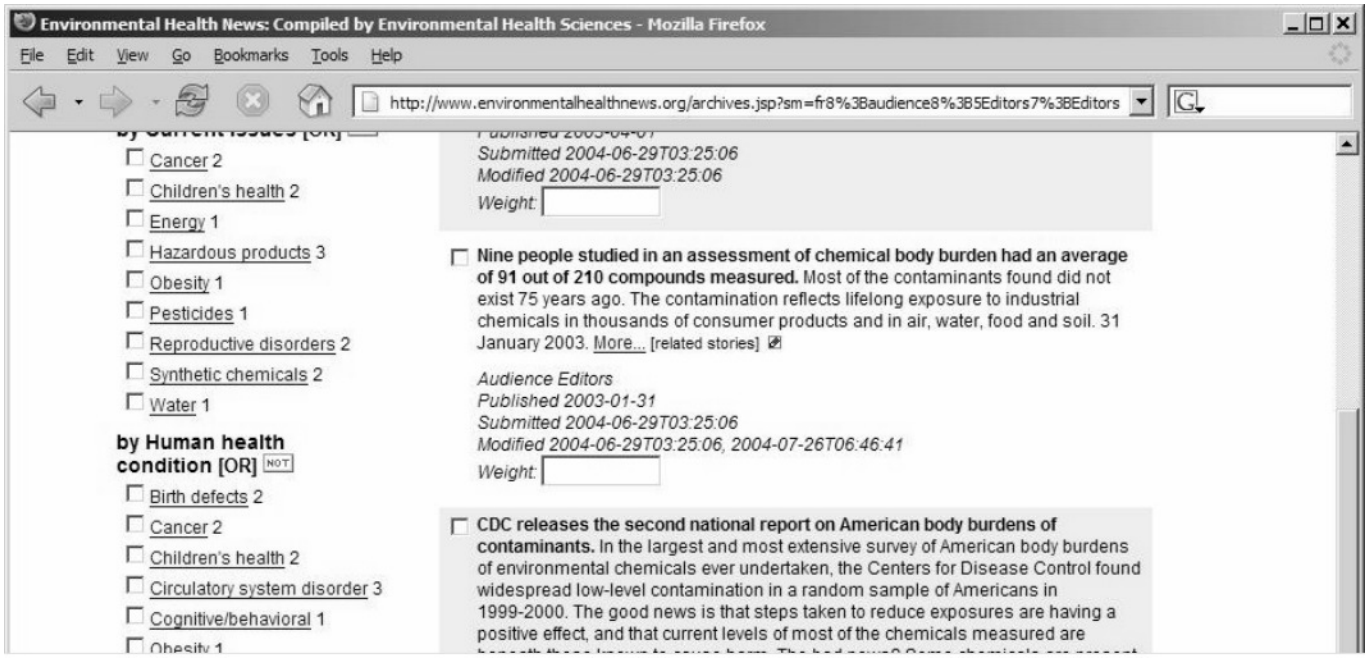


Automatic Markup

Other tasks, such as choosing related articles or assigning publisher information, are done by making use of faceted navigation pages. Typically however, the publisher is detected automatically as the article's URL is matched against a URL mask stored for the publisher.

3.3 Editorial Review

All articles submitted to the site are marked as needing editorial review and are reviewed by the editors on the archives page, which displays additional controls to the logged-in editors:

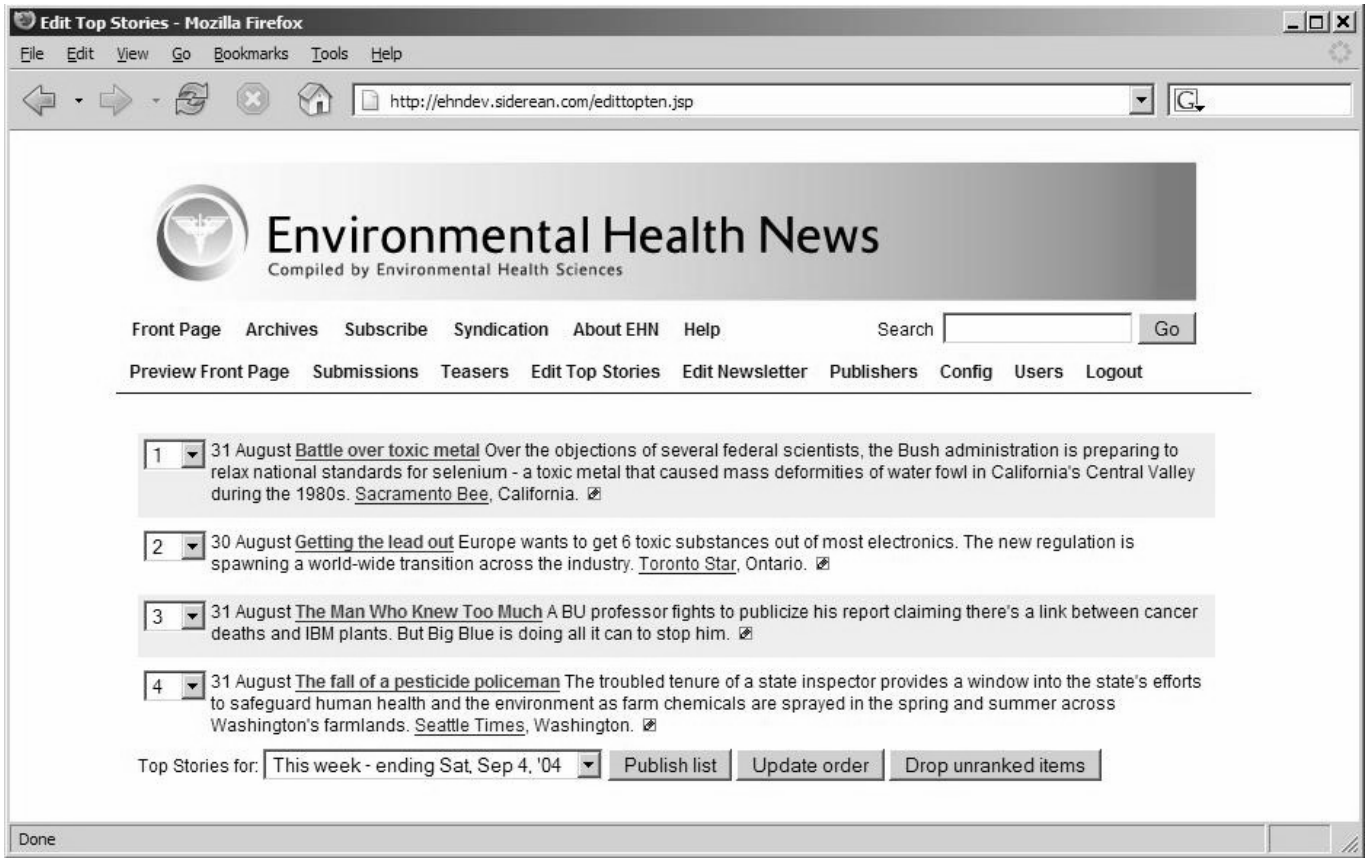


Editorial Review

This page also gives the editors the ability to mark articles for the week's top stories list or for the daily newsletter.

3.4 Weekly Top Stories

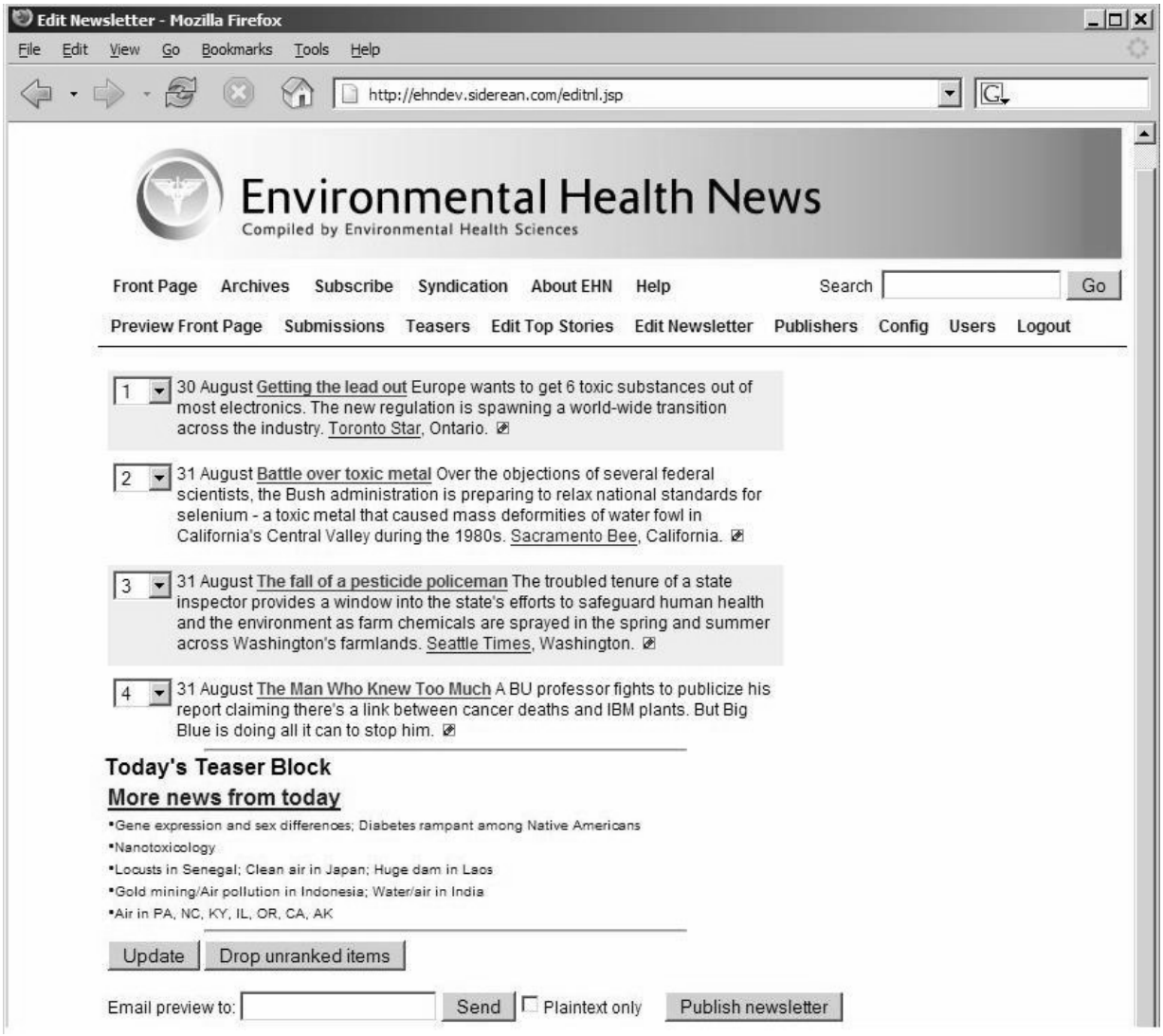
Once a number of articles have been selected for inclusion in the week's top stories list, they can be reviewed, reordered and published on the Top Stories page:



Top Stories

3.5 Daily Newsletter

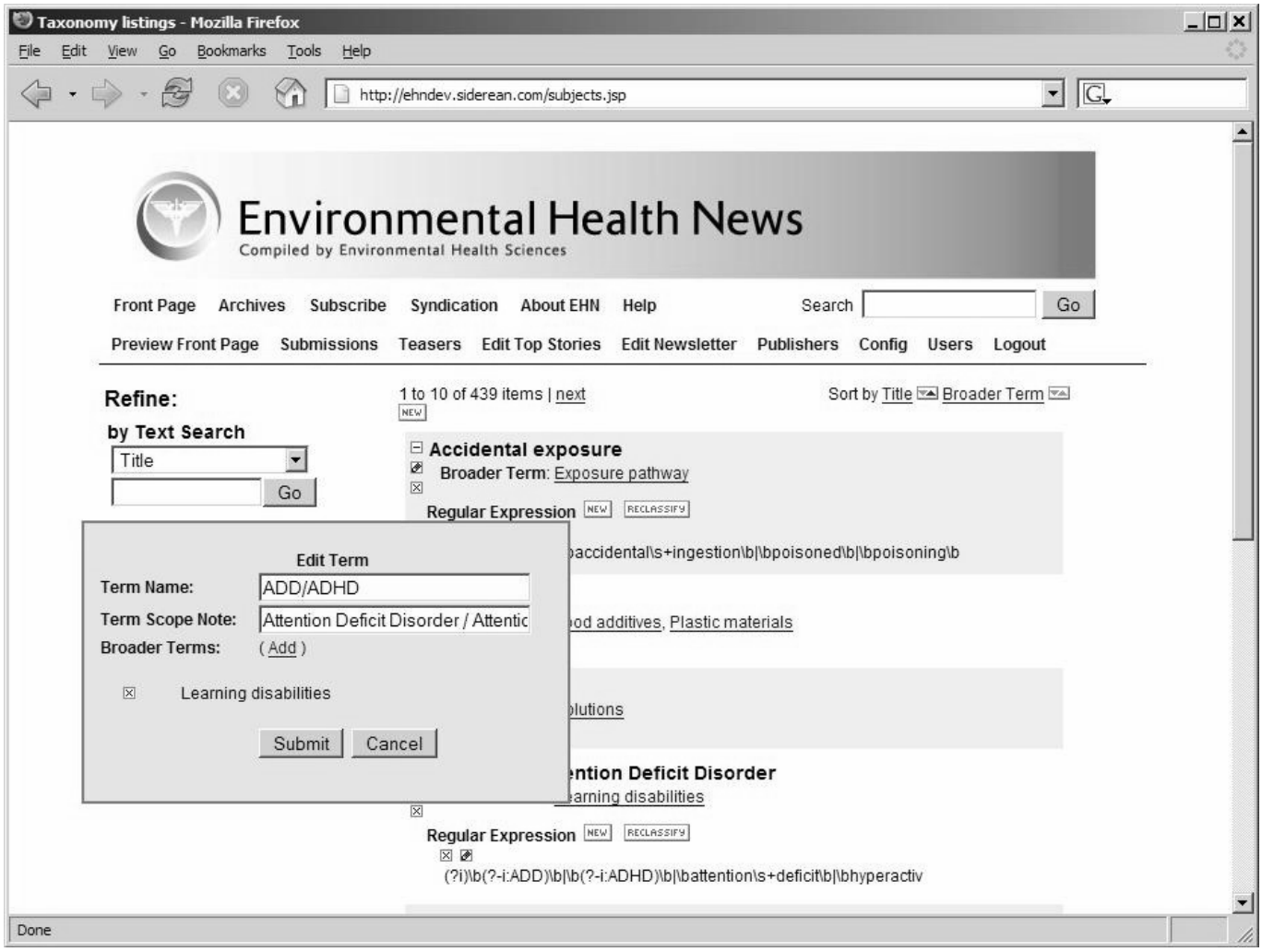
Similar to the Top Stories, the selection of articles marked for inclusion in the day's newsletter can be reviewed and reordered on the newsletter page, prior to publication. The "Teaser List," additional information summarizing the rest of the day's content on the site, is also displayed as it will be included in the newsletter. Other static newsletter text used to frame and format the newsletter is specified on a separate page.



Newsletter Page

3.6 Taxonomy Editing

The latest version of the EHN site also provides the site editors with the ability to manipulate and edit the EHN taxonomy directly, rather than by external means. Using the taxonomy editing page, terms can be added, removed, renamed and re-parented. Article text-matching rules associated with the terms (which allow for automatic classification of articles as they are submitted) can also be added, edited or deleted.



Taxonomy Editing

4. Metadata

4.1 RDF Vocabulary

The metadata representing the content on the EHN site is represented using standardized namespaces where appropriate. This includes the DCMI's Element Set [2] and Metadata Terms [3] to describe basic information about each article described on the site, SWAD-Europe's TIF [4] for representing the EHN Taxonomy and the Prism element set [5], to cover properties particular to publishing, which Dublin Core does not cover. Additionally, the EHN site uses the following RDF Classes to describe its resources:

- dcmitype:Service
  used to describe article publishers
- ehn:Post
  used to describe individual submissions. Subclasses include ehn:Story for News items, ehn:Editorial & ehn:Opinion for editorials & opinion articles, ehn:Study for Scientific Studies and ehn:Report for Organization Reports
- ehn:PostCollection
  represents a collection of posts. Subclasses include ehn:Newsletter which represents the Newsletters published daily by the editors, and ehn:ToptenList which is used to represent each week's list of top stories

5. References

[1] Jennifer English, Marti Hearst, Rashmi Sinha, Kirsten Swearingen, Ka-Ping Yee: *Flexible Search and Navigation using Faceted Metadata*, Jannuary 2002

[2] *Dublin Core Element Set, Version 1.1*, Dublin Core Metadata Initiative, http://dublincore.org/documents/dces/

[3] *DCMI Metadata Terms*, Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmi-terms/

[4] *Thesaurus Interchange Format (TIF)*, SWAD-Europe, http://www.w3c.rl.ac.uk/SWAD/thesaurus/tif/tif.html

[5] *Prism -- Publishing Requirements for Industry Standard Metadata*, http://www.prismstandard.org/about/