JOURNAL OF CHEMICAL INFORMATION AND MODELING

Software Review

pubs.acs.org/jcim

chemicalize.org

chemicalize.org by ChemAxon Ltd.

chemicalize.org by ChemAxon is a free Web-based resource that adds extra contextual chemical information into Web pages and documents. When one "chemicalizes" a Web page or PDF document, all chemical names present are identified and highlighted in-place on the page. Hovering the cursor over one of the names produces a small image preview of the structure and links to a range of complementary services including structure-based predictions.

As a Web application, chemicalize.org has no complicated installation process or system requirements beyond the need for a Web browser and an Internet connection. In addition, all computationally intensive calculations are handled server-side, meaning performance is not dependent on the user's local processing power. It should function properly in any modern Web browser on a Mac, Windows, and Linux. Much of the core functionality is also available on mobile platforms such as iOS and Android.

Getting started is as simple as visiting http://chemicalize.org, where all the features are available to use immmediately. Access is free to the public, so no institutional subscription or login procedure is needed. Upon loading the home page, the user is greeted with a single text field and is invited to "Type a chemical name or URL to begin". The simplicity of this interface is reminiscent of Google's search page and is quite refreshing in comparison to some similar services, where it is common to be bombarded by long forms with dozens of options that many users rarely touch. In fact, across the chemicalize.org site, it is clear that there has been great attention to detail in terms of the user interface. This results in a Web application that is accessible to users of all abilities, while still providing a powerful tool to students, teachers, and researchers in many different fields.

As text is entered, chemicalize.org validates it on the fly, providing autocomplete functionality for chemical names and displaying which services are available for the given input. The five services that make up chemicalize.org are Webpage Viewer, Document Viewer, Properties Viewer, Chemical Search, and Web Search. These services are highly complementary to one another, and each is available from within the others. In addition to the text field, there are two small buttons that provide alternative input methods. Clicking the "Draw" button brings up a MarvinSketch Java applet, providing straightforward but fully featured structure drawing capabilities. Clicking "Upload" allows the uploading of PDF documents for the Document viewer as well as any chemical file formats that work with Marvin.[1] As with text input, the file is quickly validated, and the available services for that file format are displayed.

The service at the core of chemicalize.org is Webpage Viewer, which is accessed simply by entering a URL into the text field and clicking the Webpage Viewer button. Upon doing so, the page specified by the URL appears as shown in Figure 1, annotated with extra information added by chemicalize.org. A dotted line is added under every chemical name that has been recognized, and hovering the cursor over a name causes a small 2D structure image to appear, which in turn can be clicked to go to the Properties Viewer for that structure. A "chemical table of contents" is inserted along the top of the page, which displays every structure identified on the page and can be used to quickly jump to the location in the text where a structure is mentioned. It also contains a button for downloading all structures mentioned on the page in MRV or SDF format. Links within the page are modified such that pages that are subsequently navigated to are automatically "chemicalized" as well.

The quality of the name-to-structure matching is very good, albeit not perfect. On the basis of my usage, it seems to be broadly comparable to similar tools such as OPSIN.[2] The occasional false positives are not a major issue, but more annoyingly, it does also miss some obscure and newly named compounds. Of course, truly perfect matching would be impossible, as while algorithms work well for InChI, SMILES, and most IUPAC nomenclature, dictionaries are required for common names and would have to be constantly updated. Crowd-sourcing this task could be a future development avenue for chemicalize.org. Interestingly, there is already a "Report Error" link that provides a quick and easy way to report false positives, but there is no similar process for missed names.

The most significant shortcoming of Webpage Viewer is that it does not work with any page that requires a cookie-based login or IP address-based authorization. This makes it useless for all content that requires a subscription of any kind, including, perhaps most importantly, the HTML versions of many scientific publications. Because of cross-domain security restrictions, it also functions poorly with sites that make heavy use of JavaScript XMLHttpRequests (AJAX). AJAX is the technology used by Web pages to dynamically load extra content after the page has already loaded—one important example being many of the pages on the RSC Journals Web site.

Both of these issues are due to the fact that when viewing the "chemicalized" version of a Web page, what is actually being shown is a page served from ChemAxon's servers, which have just downloaded the original page, analyzed, and modified it before sending it on. As it is ChemAxon's servers downloading the original page rather than the user, none of the login credentials are present, and the IP address is different. One possible solution to this problem would be to create additional browser extension versions of chemicalize.org, which would not encounter the same restrictions.

Document viewer is a recent addition to chemicalize.org. It works in much the same was as Web site Viewer but instead "chemicalizes" PDF documents, making it useful for journal articles. The document can be specified either by typing a URL into the text input field or by uploading a PDF file. Similarly to Web site Viewer, the original document is displayed with chemical names underlined and a "chemical table of contents" is displayed across the top.

Published: February 7, 2012

dx.doi.org/10.1021/ci300046g | *J. Chem. Inf. Model.* 2012, 52, 613–615

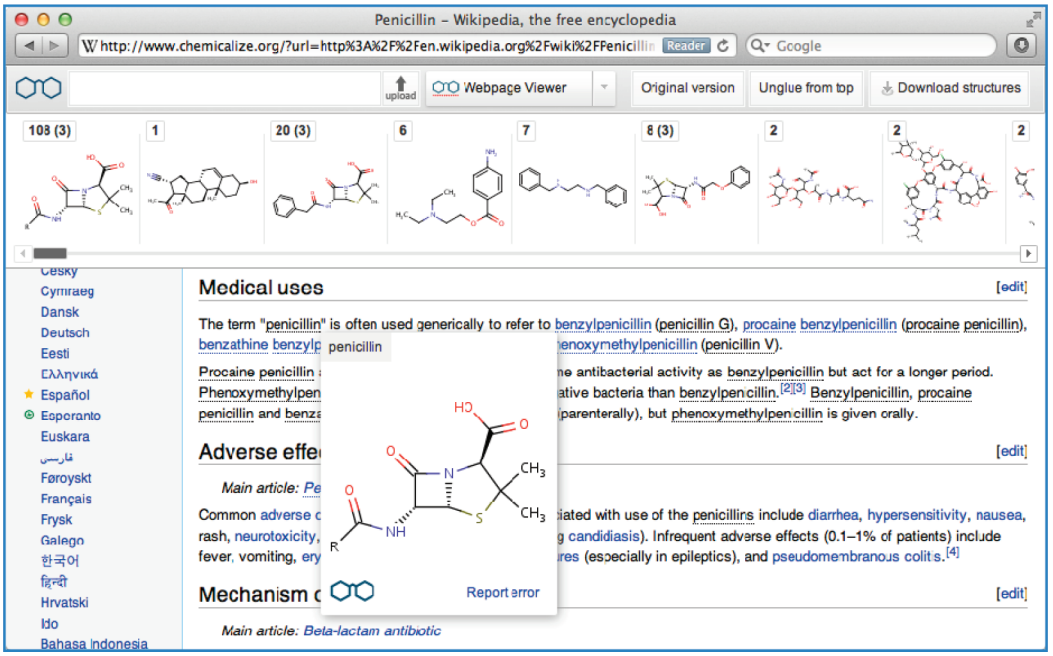

Figure 1. A "chemicalized" Web page, showing the chemical table of contents, underlined chemical names, and a structure image preview window.

Creating an application like this that works within the confines of a Web browser is an impressive feat. First, support for viewing PDFs is inconsistent across different browsers and often requires third-party plugins that cannot be guaranteed to be present. chemicalize.org solves this problem by first converting each page of the PDF into a PNG image before displaying it. The increased file sizes that result from this are elegantly dealt with using an AJAX solution that only loads pages when the user scrolls down to view them. A significant drawback to this approach, however, is that the document text is no longer selectable, making copying and pasting impossible.

Second, Document Viewer is impressive because PDF files are notoriously difficult to parse in a logical manner as the format is purely designed around how the document will look when printed on paper. Despite this, quick testing indicates that chemical name identification works at a similar level to Web site Viewer, apart from in PDF files that were created by scanning in paper documents. With scanned PDFs, Document Viewer fails completely because the text information is not available; however, ChemAxon say they are working on a solution involving optical character recognition (OCR).

Clicking any structure image in Webpage Viewer or Document Viewer will take you to that structure's page in Properties Viewer. Properties Viewer can also be accessed directly from the chemicalize.org home page by typing in a chemical name, by drawing a structure, or by uploading a file containing chemical information. The calculated and predicted properties available include elemental analysis, geometry, polarizability, logP, logD, and pK_a. The properties are organized into boxes that can be hidden, resized, and rearranged. When changes are made to the layout, chemicalize.org saves the customized arrangement so it can be automatically used in future. There are also three preset layouts, "Synthetic Chemist", "Medicinal Chemist", and "Basic".

The calculations are performed on the fly, meaning there is no chance that any properties will be missing from the database. Once they have been calculated once, properties are stored in a cache, and subsequent viewings load much faster. Integration with other online services such as ChemSpider[3] would be very welcome. ChemSpider already links to the chemicalize.org page for each structure, but a more tightly integrated system where each service could display data from the other would be useful.

All the properties (along with everything else on chemicalize.org) are licensed under a Creative Commons Attribution–NonCommercial–ShareAlike license. This is generally quite a permissive license, although some care has to be taken with regards to the implications of the "NonCommercial" part.

The final two services that make up chemicalize.org are Chemical Search and Web Search. Chemical Search provides a way to search the chemicalize.org structure database that has been built up using the structures in Web pages and documents that users have "chemicalized". Currently, there is a reported total of 185,000 structures linked to 355,000 Web pages, which will automatically grow over time as users "chemicalize" more content. As with Properties Viewer, queries can be made by chemical name, structure drawing, or file upload. Exact match, substructure, and similarity (using Tanimoto coefficient) search types are available. From the results, it is possible to view each structure in Properties Viewer, as well as view a list of all the previously "chemicalized" Web pages that mention each structure. Web search provides another way of finding Web pages that mention a structure, using Google as a backend instead to provide a much more comprehensive search.

There is no official documentation for chemicalize.org beyond a quick video demonstration and a brief paragraph summarizing each of the five core services on the main page. It is not really missed, however, as the intuitive interface lends itself well to "learning by trying", and there is a user forum where support staff and other users are available to help should any problems arise. The downside to this, combined with the emphasis on a clean, accessible interface, is that niche or advanced features can be slightly hidden. This leads to a few "Easter egg" features that are not obvious enough to be discovered through normal use. For example, dragging any structure up into the text field uses it as a new input, and double-clicking

dx.doi.org/10.1021/ci300046g | J. Chem. Inf. Model. 2012, 52, 613–615

on a structure in Properties Viewer displays it in a MarvinView applet that allows 3D viewing.

In conclusion, chemicalize.org is a highly useful tool for adding chemical structure information to Web pages and documents, while also providing an easy to use Web interface to ChemAxon's structure-based predictions. From a technical point of view, the application makes use of a number of novel concepts and technologies such as crowdsourcing and on the fly calculations and is a pioneer in terms of providing cheminformatics solutions for free via a Web browser. As the world is slowly moving toward more semantic, data-driven methods of sharing scientific results, chemicalize.org helps bring the existing body published of work up to this level. The interface excels at being intuitive and accessible, and the service fits in extremely well alongside complementary online resources such as OPSIN and ChemSpider. As free services offered via a Web interface they are revolutionizing the way structure and property data are accessed.

Matthew Swain*

Cavendish Laboratory, University of Cambridge,
J J Thomson Avenue, Cambridge CB3 0HE,
United Kingdom

AUTHOR INFORMATION

Corresponding Author

*E-mail: mcs81@cam.ac.uk.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) File Formats in Marvin. https://www.chemaxon.com/marvin/help/formats/formats.html (accessed January 13, 2012).

(2) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Comput. Sci.* **2011**, *51*, 739−753.

(3) Pence, H. E.; Williams, A. J. ChemSpider: An Online Chemical Information Resource. *Chem. Educ.* **2010**, *87*, 1123−1124.