The Future of Chemical Information Is Now

by Antony J. Williams and Harry E. Pence

Search and retrieval of chemical information has been dramatically changed by the application of "Big Data" techniques. This development continues to be driven by the massive growth of chemical scientific literature and of online data and databases. Not only is there an expansion of the traditional avenues of publication, but many new contributing resources, such as open access journals, MOOCs (Massive Open Online Courses), Wikis, and blogs have arisen. Powerful tools, like APIs (application programming interfaces) and Big Data interrogation are providing innovative ways to retrieve and analyze data and connect different databases. Materials, pharmaceutical, and environmental research, to name just a few, are especially challenged by the need to organize and access vast amounts of data. What skill-sets will need to be developed in order to get the greatest value out of the available data? Will it be coding and information technology skills, or awareness and better delivery of the data by the available systems? We believe that, in the short term, efforts are needed to expand awareness and training.

Exponential Growth of the Amount of Chemical Information

One of the earliest efforts to quantitatively measure the growth of the scientific literature was made by Derek J. de Solla Price over five decades ago. [1] He determined that the number of scientific journals was increasing by about 5.6% per year, with a doubling time of 13 years, and that the number of abstracts in Chemical Abstracts was also growing exponentially. More recently, Larsen and von Ins have reported a similar rapid growth for scientific articles, with a slightly longer doubling time than Price, and they report that the Science Citation Index is covering a decreasing portion of the traditional scientific literature. [2] These authors also point out that publishing has expanded into new channels, such as open access archives and web pages, especially in the form of blog posts and an increasing number of alternative distribution channels available to any scientist (*e.g.* LinkedIn, GoogleDocs, Slideshare). In late 2014, there were about 28,100 active scholarly peer-reviewed English-language journals (plus a further 6450 non-English-language journals), collectively publishing about 2.5 million articles a year. [3] The number of publishers and journals continues to increase and is being boosted by the Open Access publishing movement. Faizul and Hilal surveyed

the number of chemistry journals listed in the Directory of Open Access Journals (DOAJ) and found 164 journals as of 2014. [4]

In parallel to the ongoing growth in the number of journals and articles and the availability of a number of for-fee databases, there has also been an enormous growth in the number of free online chemistry databases. The premier for-fee chemistry resources are CAS' SCIFinder, [5] with over 127 million chemical substances in its registry at the time of writing, [6] and Elsevier's Reaxys database, which contains over 100 million organic, inorganic, and organometallic compounds. [7] Both SCIFinder and Reaxys index tens of millions of chemical reactions and hundreds of millions of experimental facts. While there are a number of other commercial databases that can be acknowledged, it is the explosion in web-based data that continues to feed the Big Data revolution in access to chemistry. Wikipedia alone lists over forty chemical databases. [8] The majority of these are freely available resources on the Internet. They represent an increasing thrust in terms of public data dissemination and release as encouraged by funding agencies. [9]

Despite the name, Big Data in chemistry is not just the *amount* of chemical information. Often the size of the dataset is compounded by the complexity of the information. For example, chemical information is reported in a bewildering variety of structured and unstructured formats including closed, instrument-specific format data. Currently the majority of these data are not being made available in standardized open formats. For example, NMR spectra generated in research labs, if published, are normally reduced to an image, despite there being digital exchange standards for such data. [10] In our experience, the vast majority of scientists have never heard of the standard, do not know how to generate this form of the data and, in any case, would not know how to share it. Some publishers are starting to consider this need and efforts are underway to put a spectral database online. [11] However, this situation will not change until there is a general increase in awareness

of the problem and potential advantages that will result if the community collaborates in the delivery of data in a form that can be aggregated for better access. One approach to address this issue is an educational effort that includes professional chemists, as well as undergraduate majors.

Even though the data from one individual's sequenced DNA is only about 750 MB, analyzing the genome for a single person in order to find the best treatment for a disease represents a massive computing problem. It is estimated that just storing the genomic data for the entire U.S. population would represent 222 petabytes. [12] Approaches are being developed by the bioinformatics community to consume and digest omics data (*i.e.* genomics and metabolomics) to support personalized medicine. [13] Despite the increasing complexity of chemical information, it is likely only through analytical data dissemination that the masses of data will move towards the scale offered by the biomedical community. There is some movement towards sharing big data to support discovery, certainly in terms of file sizes, in the analytical sciences,.

Any analytical chemistry laboratory running spectral analyses currently produces multi-megabyte file sizes. Multidimensional NMR spectroscopy data files can measure from 10s to 100s of megabytes, but tandem mass spectrometry files, especially in proteomics, can consume terabytes of space. The online database, the Center for Computational Mass Spectrometry at UC San Diego,

already has datasets over 1 terabyte in size. The largest of these is almost 13 terabytes in size. [14] Despite the sheer size of these datasets, and their applications in proteomics, can such data be used for discovery purposes? A recent report regarding how such approaches can be used for the identification of new antibiotics suggests that there really are needles in the haystack that can be extracted from mass spectrometry data. [15,16]

The Global Natural Product Social (GNPS) molecular network, launched in 2015, [17] is also pursuing a Big Data approach to the discovery of novel drug candidates from enormous quantities of mass spectral data assembled from the work of over a hundred laboratories. As has been the case for the analytical sciences for many years, data generation is rarely the main problem. The analysis algorithms associated with data interpretation remain the bottleneck, accompanied by a lack of data interchange standards to migrate data into alternative data processing platforms. Analysis, however, can be assisted by the availability of online data resources, providing access to tens of millions of chemical structures and tens of thousands of analytical spectra. These resources represent tens of millions of dollars of investment in informatics architecture, but also in the collection, curation, and annotation of the data.

Similar to the work that GNPS is doing, scientists can today access online "big data collections" that can be used in the identification of chemical compounds. With nothing more than a monoisotopic mass or molecular formula derived by mass spectrometry, and access to an online database, structure candidates can be identified using a search for "known unknowns", searching for well-known chemicals that are held in public databases, though the scientists themselves do not know the candidate structure. This approach has been demonstrated previously. [18,19] Chemists can combine this information with an online search of NMR data for tens of thousands of chemicals [20] and even use Robien's Spectral Robot Referee [21] to help confirm structural hypotheses. Mobile applications on phones and tablets, hosting over 700 000 chemicals (with masses and predicted

Table 1: Data Storage Units

Unit	Symbol	Size in bytes*
kilobyte	KB	1 000
megabyte	MB	1 000 000
gigabyte	GB	1 000 000 000
terabyte	TB	1 000 000 000 000
petabyte	PB	1 000 000 000 000 000
exabyte	EB	1 000 000 000 000 000 000

* a byte is a group of binary digits or bits (usually eight))
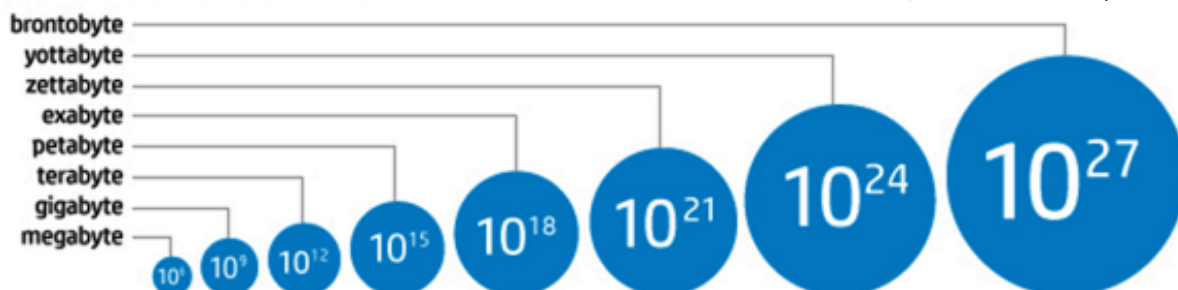
13C NMR spectra), can also assist in the identification of known unknowns—so rather sizeable data collections can now be held in the hand! [22]

Big Data Tools for Chemical Information

Most chemists have been using Big Data tools, even though they might not have been aware of it, since search engines, like Google or Bing, use a combination of MapReduce and Hadoop to distribute a search among multiple servers and then analyze the huge amounts of information that result. [23] Hadoop facilitates the distributed processing of massive unstructured data sets across large computer clusters, while MapReduce distributes work to various nodes within the cluster (or 'Map'), organizes it, and then 'Reduces' the results into a coherent answer for a particular query. [24] Big Data tools are also being used for data analytics in many areas of chemistry. Large arrays of inexpensive sensors connected through a computer cloud may generate very large datasets in environmental chemistry. [25] In addition to size, environmental data may be complicated because it consists of records in differing formats. Data collected in recent years may be distributed across a large number of databases, each with a different data model, potentially using different data ontologies (if any), with the potential for enormous migration and data integration efforts to mesh together valuable data. Historical data, while valuable, may not even be available in digital form, other than as scanned documents where optical character recognition software will produce only some limited level of retrieval, so these are only available for consumption into Big Data analysis tools using the limited metadata associated with the documents.

Researchers are being overwhelmed by so much data that experimental results can be overlooked or repeated unnecessarily. This is increasingly important at a time where reproducibility overall is being called into question. [26] Mullin reports, according to one estimate, "... 40 % of all R&D experiments are repeat runs necessitated by inefficient experimental design or inadequate IT." [27] Another arena where Big Data tools are useful is predicting protein structures. Ovchinnikov and co-workers have used Big Data techniques to better predict 3D protein structures. [28] Big Data tools are also valuable in chemical toxicology, where the use of high-throughput screening produces both structured and unstructured information that is so large and complex that it is difficult to analyze using traditional methods. [29] Just managing the chemistry data (in terms of chemical compounds and challenges of chemical structure detail *vs.* the myriad associated identifiers) is enough of a problem. [30] Working to blend the chemistry data with the associated bioassay screening data into a form consumable by scientists as openly accessible data, and for consumption by Big Data algorithmic approaches, is a significant and often underestimated difficulty.

Big Data, artificial intelligence, and machine learning are today commonly aggregated into the same conversation. This is certainly true when it comes to the promise of these approaches combined through the IBM Watson project. [31] In 2013, MD Anderson partnered with IBM to pursue a cure for cancer, starting with leukemia, [32] and medical centers are already implementing Watson to help oncologists make data-driven decisions. [33] While promise remains, a little of the luster has rubbed off recently with the announcement that the Anderson-IBM collaboration has been halted after a scathing report from auditors at the University of Texas says the project cost MD Anderson more than $62 million and yet did not meet its goals. [34] This example indicates how these approaches might not solve all of our current challenges, but it does not mean we should not attempt these efforts as, after all, this is research.

Big Data Tools to Search the Web for Chemical Information

Text-mining of chemistry data in patents and documents has been underway for many years [35] and IBM text-analytics, and Watson specifically, have been applied to problems in the life sciences. By analyzing millions of pages of text in the medical literature, patents, genomics, and chemical and pharmacological data, Watson made novel connections. Early results suggest that Watson can indeed accelerate the identification of novel drug candidates and targets by harnessing Big Data. [36] Similar technology could be applied to the mass extraction of chemical reactions from literature articles and patents, as demonstrated by Lowe, [37] and then used as the basis of retrosynthetic reaction synthesis algorithms such as those underpinning Wiley's ChemPlanner. [38]

While there are certainly some naysayers regarding the potential of these supercomputers, *i.e.*, "neither could compete with a toddler at some of the most basic forms of human cognition," [39] this level of negativity has been pointed at various technologies at some point in their development, whether it be the potential contributions of solar power, self-driving cars, or even putting a man on the moon. All of these, clearly, have proven to be possible.

While the breakthrough technologies are on the bleeding edge and newsworthy, there are many capabilities already available for every Internet consumer. More and more information is available in online databases, mined from the literature in ever increasing large scale

data and then made available as downloadable datasets (*e.g., ca.* 300 000 melting points extracted from patent literature, [40] connected by appropriate Application Program Interfaces (APIs) and increasingly available via the semantic web). Wikipedia both delivers data for consumption and, increasingly, is being served by the developing Wikidata project. [41] It is getting easier to access and integrate data with components, add-ins, and widgets.

For example, the PubChem project provides access to chemistry data for about 94 million chemical substances and about 1.2 million bioassay measurements. [42] The data are not only available by browsing the PubChem website data, but their widgets can be directly integrated with other websites so that they are accessible to different audiences interested in the data in different contexts. [43] For example, environmental chemists surfing the U.S. Environmental Protection Agency (EPA) CompTox Chemistry Dashboard have direct access to the PubChem bioassay data via the use of PubChem widgets. [44]

Increasingly, chemical information is stored on the Internet in the form of videos. The Internet hosts many chemistry videos (for example the Periodic Table of Videos [45] in addition to the *Journal of Video Experiments*), [46] and chemistry videos are also be found in many chemistry-related Massive Open Online Courses (MOOCs). [47] If the use of virtual and augmented reality environments develops as expected, even more valuable data will be made available. At present, the ability to search these resources is limited by the associated metadata. The challenge becomes how to search and get the greatest benefit out of these complex datasets and environments.

How should Chemists be Trained to Use Big Data?
Does the increased importance of Big Data mean that every chemist needs to learn how to use Big Data tools? In-depth knowledge of how such tools and algorithms work is unlikely to be useful to all chemists, since it is more probable that most chemists will simply use the results of Big Data algorithms and searches. It is, and will be, more important for chemists to recognize the strengths and weaknesses of a Big Data approach, rather than to be able perform a direct analysis themselves. A complex computer algorithm, such as that used to analyze Big Data, is for many simply a black box, and for most it is human nature to assume that the algorithm is performing its function in an ideal manner rather than questioning the results. Consider the parallels with how many individuals interact with a general online Internet search, accepting partial results rather than insisting

upon completeness.

However, we think training is required at two different levels. First, there is a need for some chemists to have in-depth training in data analysis. These individuals, who would have a combined background in Computer Science and in Chemistry, will have skills needed to ensure that chemical-specific information is used appropriately in Big Data analyses. Second, there is a need for practicing chemists to have a background in Big Data analytics sufficient to recognize the potential uses of these techniques, as well as some of the potential pitfalls. Some training is already available: a Google search will turn up a number of courses that may be appropriate. For example, there is an Online Learning Cheminformatics Course sponsored by the Committee on Computers in Chemical Education of the ACS Division of Chemical Education. [48] David Wild, director of the Indiana University Cheminformatics Program, maintains a *Cheminformatics Education Portal* (ICEP), a repository of freely accessible cheminformatics educational materials, [49] as well as an online course, Introducing Cheminformatics: Navigating the World of Chemical Data. [50]

What does the Future Hold for Searching Chemical Information?
The Internet has catalyzed a shift in expectations for many stakeholders in terms of chemical information. The primary consumers of chemical information today are those driving the search at their desktop, on their tablet, or on their phone using one of the common browsers. The majority of chemists likely think they know enough in terms of Internet searching that they can find what they are looking for themselves and do not need training. They are satisfied with a simple search box. While such results can be found with the majority of Internet searches, it is a fallacy to consider that such approaches are not without issues. Chemical information professionals, and librarians specifically, commonly bring significant experience to the array of commercial platforms available and are able to answer questions, supply training in "search strategies," and teach users how to best utilize Internet resources. In academia especially, basic training in cheminformatics and chemical information resources is increasingly encouraged. It is hoped that this will expand both in depth and in general availability across institutions. Certainly there are no signs that the rate of publishing is decreasing, that the amount of data coming online is slowing, or that the expectations for high productivity and more innovation are waning. To the contrary, it is likely that chemists will contribute directly to the growth in Big Data, using both personal publishing platforms (*e.g.* blogs), community collaboration tools for

information dissemination (*e.g.* wikis), and data sharing platforms. In fact, many funding agencies are now demanding the release of data associated with scientific research: the relevant skills to do so need to be developed and the supporting tools to make it happen need to be continually improved.

Conclusion

It has been two decades since Carla Hesse predicted that the future of the book might consist of "paths of inquiry, modes of integration, and moments of encounter." [51] That may yet serve as a good description of the future of chemical information, since data, information, and knowledge is hardly static—it is changing moment by moment and accessible via a web search. Chemical researchers may integrate the results of multiple search modes using a variety of paths to the data (*i.e.*, literature, online data, analytical tools). The resulting vast amount of information may allow for less than the optimal time for careful examination of all but the most obviously essential resources. Despite the vast amounts of information to be surveyed, both online and print, it will still be necessary to insure that as little as possible of the most important sources are not lost. This will require new levels of sophistication and ingenuity from the researchers of the future.

References

1. Price, D. J. d. S., *Science Since Babylon.* Yale University Press: New Haven, 1975.
2. Larsen, P., O., von Ins, M., The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index. *Scientometrics* **84**(3):575–603, 2010. https://dx.doi.org/10.1007/s11192-010-0202-z.
3. Ware, M., Mabe, M., *The STM Report, 4th Ed*. www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf (accessed 2 March 2017).
4. Faizul, N., Hilal, A., *Analysis of Open Access Scholarly Journals in Chemistry*. http://digitalcommons.unl.edu/libphilprac/1100/ (accessed 2 March 2017).
5. SCIFinder. https://scifinder.cas.org/ (accessed 3 March 2017).
6. CAS REGISTRY. www.cas.org/content/chemical-substances (accessed 2 March 2017).
7. ReaxysR Fact Sheet. www.elsevier.com/__data/assets/pdf_file/0005/91616/RDS_FactSheet_Reaxys_Oct_2016-WEB.PDF (accessed 2 March 2017).
8. https://en.wikipedia.org/wiki/Category:Chemical_databases (accessed 2 March 2017).
9. NIH Request for Information (RFI). https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html (accessed 2 March 2017).
10. Published JCAMP-DX Protocols. www.jcamp-dx.org/protocols.html (accessed 2 March 2017).
11. Chalk, S. J., The Open Spectral Database: an Open Platform for Sharing and Searching Spectral Data. *J Cheminform* **14**(8):55, 2016.
12. Gualtieri, M., *Is 750MB Big Data?* http://blogs.forrester.com/mike_gualtieri/12-12-05-is_750mb_big_data (accessed 4 June 2014).
13. Alyass, A., Turcotte, M., Meyre, D., From Big Data Analysis to Personalized Medicine for all: Challenges and Opportunities. *BMC Medical Genomics* **8**:33 https://dx.doi.org/10.1186/s12920-015-0108-y
14. Submitted MassIVE Datasets. https://massive.ucsd.edu/ProteoSAFe/datasets.jsp#%7B%22table_sort_history%22%3A%22main.fileSizeKB_dsc%22%7D (accessed 2 March 2017).
15. Patringenaru, I., Big Data for Chemistry. http://ucsdnews.ucsd.edu/pressrelease/big_data_for_chemistry (accessed 3 March 2017).
16. Mohimani, H., *et. al.,* Dereplication of Peptidic Natural Products Through Database Search of Mass Spectra. *Nature Chem Bio* **13**:30-37, 2017.
17. The Future of Natural Products Research and Mass Spectrometry. https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp (accessed 3 March 2017).
18. Little J.L., *et. al*., Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider. *J Am Soc Mass Spectr.* **23**(1):179-85, 2012.
19. McEachran, A. D., Sobus, J.R., Williams, A. J., Identifying Known Unknowns Using the US EPA's CompTox Chemistry Dashboard. *Anal Bioanal Chem* **409**(7):1729–1735, 2017.
20. Kuhn, S., NMRShiftDB. http://nmrshiftdb.nmr.uni-koeln.de/ (accessed 3 March 2017).
21. CSEARCH Robot Referee. http://nmrpredict.orc.univie.ac.at/c13robot/robot.php (accessed 3 March 2017).
22. Blinov, K., CompTox Mobile. https://itunes.apple.com/us/app/comptox-mobile/id1179517689?ls=1&mt=8 (accessed 3 March 2017).
23. Langit, L., Hadoop MapReduce Fundamentals 1 of 5. www.youtube.com/watch?v=7FcMhTTG1Cs (accessed 21 June 2015).
24. Pusala, M.K., Salehi, M.A., Katukuri, J.R., Xie, Y., Raghavan, V., Massive Data Analysis: Tasks, Tools, Applications, and Challenges. in *Big Data Analytics: Methods and Applications* Springer, 2016.
25. Vitolo, C., *et.al.,* Web Technologies for

Environmental Big Data. *Environ. Modell. Softw.* **63**:185-198, 2015.

26. Baker, M., 1,500 Scientists Lift the Lid on Reproducibility. *Nature* **533**(7604):452–454, 2016.

27. Mullin, R., Breaking Big. *Chem. Eng. News* **91**(42):19-21, 2013.

28. Ovchinnikov, S., *et. al,,* Protein Structure Determination Using Metagenome Sequence Data. *Science* **355**(6322):294-298, 2017.

29. Zhu, H,. Zhang., J., Kim, M.T., Boison, A., Sedykh, A., Moran, K., Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants. *Chem.Res. Toxicol.* **27**:1643-1651, 2014.

30. Richard, A. M., *et. al.,* ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **29**(8):1225–1251, 2016.

31. Go Beyond Artificial Intelligence with Watson. www.ibm.com/watson (accessed 3 March 2017).

32. MD Anderson Taps IBM Watson to Power "Moon Shots" Mission Aimed at Ending Cancer, Starting with Leukemia. www-03.ibm.com/press/us/en/pressrelease/42214.wss (accessed 3 March 2017).

33. Jupiter Medical Center Implements Revolutionary Watson for Oncology to Help Oncologists Make Data-Driven Cancer Treatment Decisions. www-03.ibm.com/press/us/en/pressrelease/51517.wss (accessed 3 March 2017).

34. Herper, M., MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine. www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/#315e86543776 (accessed 3 March 2017).

35. Trippe, A. Hunting for Hidden Treasures: Chemistry Text Mining in Patents and Other Documents. www.patinformatics.com/hunting-for-hidden-treasures-chemistry-text-mining-in-patents-and-other-documents (accessed 3 March 2017).

36. Chen, Y., Argentinis, JD. E., Griff, W., IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin. Ther.* **38**(4):688–701, 2016.

37. Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Doctoral Thesis, www.repository.cam.ac.uk/handle/1810/244727, Cambridge University: Cambridge, UK, 2012.

38. Wiley ChemPlanner. www.chemplanner.com/what-is-chemplanner/ (accessed 3 March 2017).

39. Grunewald, W., FYI: *Which Computer Is Smarter, Watson Or Deep Blue?* www.popsci.com/science/article/2012-12/fyi-which-computer-smarter-watson-or-deep-blue (accessed 3 March 2017).

40. Tetko, I. V., Lowe, D.M., Williams, A..J., The Development of Models to Predict Melting and Pyrolysis Point Data Associated with Several Hundred Thousand Compounds Mined from PATENTS. *J Cheminform* **8**(2), 2016. https://dx.doi.org/10.1186/s13321-016-0113-y.

41. Perez. S., Wikipedia's Next Big Thing: Wikidata, A Machine-Readable, User-Editable Database Funded By Google, Paul Allen And Others. https://techcrunch.com/2012/03/30/wikipedias-next-big-thing-wikidata-a-machine-readable-user-editable-database-funded-by-google-paul-allen-and-others/ (accessed 3 March 2017).

42. PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed 15 May 2014).

43. PubChem Widgets v2.0f. https://pubchem.ncbi.nlm.nih.gov/widget/docs/widget_help.html (accessed 3 March 2017).

44. EPA Chemistry Dashboard. https://comptox.epa.gov/dashboard/dsstoxdb/results?utf8=%E2%9C%93&search=atrazine#bio-activity (accessed 3 March 2017).

45. Haran, B,. Periodic Videos. www.periodicvideos.com/ (accessed 3 March 2017).

46. JoVE: Journal of Video Experiments. www.jove.com/ (accessed 3 March 2017).

47. MOOC List: Chemistry. www.mooc-list.com/tags/chemistry (accessed 3 March 2017).

48. Belford, R., E, DivCHED CCCE: Cheminformatics OLCC. http://olcc.ccce.divched.org/ (accessed 3 March 2017).

49. Wild, D., Indiana Cheminformatics Information Portal. https://icep.wikispaces.com/ (accessed 3 March 2017).

50. Wild, D., Introducing Cheminformatics: Navigating the world of chemical data. http://i571.wikispaces.com/ (accessed 3 March 2017).

51. Nunberg, G., (.ed.), *The Future of the Book*, 31. University of California Press,: Berkeley, CA, USA, 1996.

Antony J. Williams <tony27587@gmail.com> is a cheminformatician at the National Center for Computational Toxicology, U.S. Environmental Protection Agency in Research Triangle Park, USA. ORCID.org/0000-0002-2668-4821

Harry E. Pence <Harry.Pence@oneonta.edu> is SUNY Distinguished Teaching Professor Emeritus in the Department of Chemistry and Biochemistry, SUNY at Oneonta, USA. ORCID.org/0000-0002-0412-9018