The C6H6 NMR repository: an integral solution to control the flow of your data from the magnet to the public

Luc Patiny[a], *Michaël Zasso*[b], *Daniel Kostro*[b], *Andrés Bernal*[c], *Andrés M. Castillo*[d], *Alejandro Bolaños*[e], *Miguel A. Asencio*[a], *Norman Pellet*[a], *Matthew Todd*[f], *Nils Schloerer*[g], *Stefan Kuhn*[g], *Elaine Holmes*[h], *Sacha Javor*[i], *Julien Wist*[e]*

[a] Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[b] Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Batochime, CH-1015 Lausanne, Switzerland

[c] Departamento de Ciencias Básicas y Modelado, Universidad Jorge Tadeo Lozano, Bogotá, Colombia

[d] Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, A.A. 25360, Cali, Valle, Colombia

[e] Chemistry Department, Universidad del Valle, A.A. 25360, Cali, Valle, Colombia

[f] School of Chemistry, The University of Sydney, NSW 2006, Australia

[g] Department of Chemistry, University of Cologne, Köln, Germany

[h] Division of Computational and Systems Medicine, Imperial College, London, UK

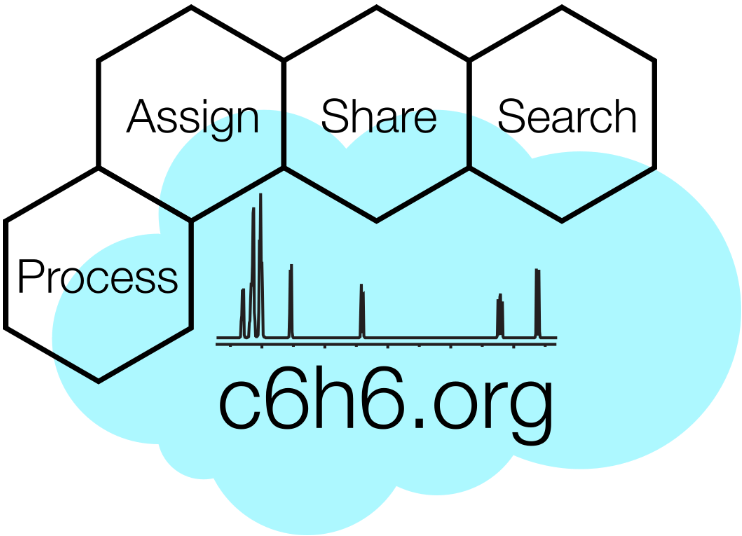
[i] Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland

*julien.wist@correounivalle.edu.co

Abstract

NMR is a mature technique that is well established and adopted in a wide range of research facilities from laboratories to hospitals. This accounts for large amounts of valuable experimental data that may be readily exported into a standard and open format. Yet the publication of these data faces an important issue: raw data are not made available; instead, the information is slimed down into a string of characters (the list of peaks). Although historical limitations of technology explain this practice, it is not acceptable in the era of internet. The idea of modernizing the strategy for sharing NMR data is not new and some repositories exist, but sharing raw data is still not an established practice. Here we present a powerful toolbox built on recent technologies that runs inside the browser and provides a means to store, share, analyse and interact with original NMR data. Stored spectra can be streamlined into the publication pipeline, to improve the revision process for instance. The set of tools is still basic, but is intended to be extended. The project is open source under the MIT license.

For the Table of Contents: C6H6 is a repository of raw NMR data coupled with a set of tools for spectra processing and analysis. It is open-source and runs on a standard web browser.



Keywords

Open science, Database, Online repository, Processing suit, NMR spectroscopy

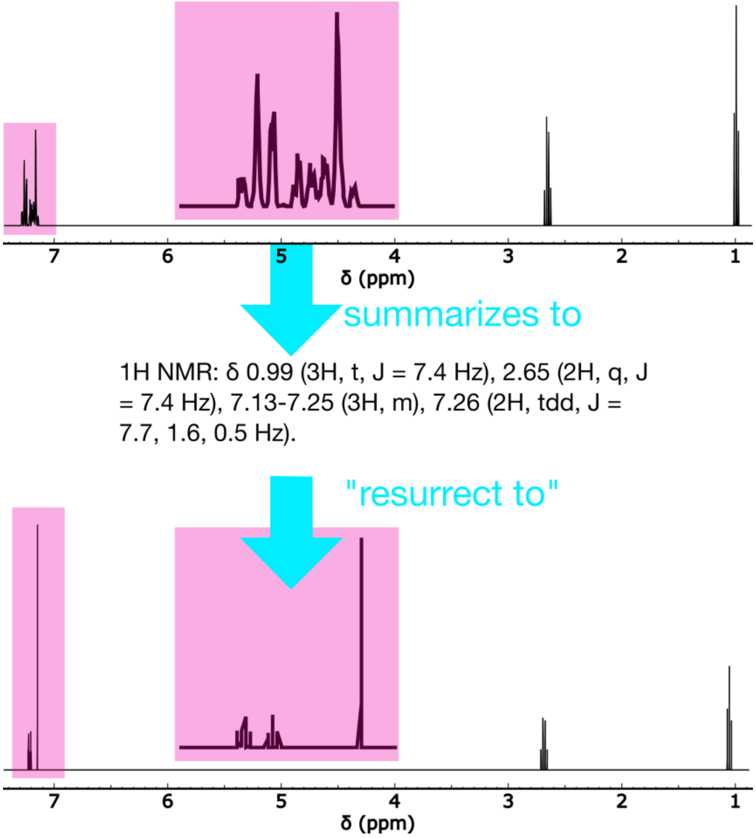
Introduction

Concern is growing, across different disciplines, over the current state of scholar data. Contemporary publication practices prevent us from taking advantage of research outcomes, from seamlessly integrating them in new knowledge discovery enterprises. To overcome this problem, the FAIR Data Principles[1] (proposed by a collective of 47 researchers around the world) attempt to set the standard for scientific data sharing. FAIR is an acronym for Findability (data are to be stored with rich metadata and registered or indexed in a searchable source), Accessibility (all data and metadata should be accessible through open, free and standard communication protocols), Interoperability (data and tools from non-cooperating resources should be able to integrate with minimum effort) and Reusability. The FAIR data initiative acknowledges the role of *computer agents* working for *human agents*, and the importance of making data FAIR for both.

The NMR community has recognized similar challenges and asks for completeness, accessibility, and both human- and machine-readability of published NMR data[2-4]. Surprisingly, while modern spectrometers already meet these requirements, generating full sets of experimental data that can be seamlessly shared and read both by machines and (through readily available software) by humans, our scientific publication practices break this achievement[4]. The full spectrum is reduced to a peak list, a time-consuming task that summarizes the findings while dismissing important parts of the experimental data (See Figure 1). This peak list may or may not be accompanied by an illustration of the spectra, rendered as an image that is a cryptogram to any algorithm attempting to interpret the underlying spectrum[5,6]. The final result is presented as a PDF, with data that are ultimately insufficient to seriously referee the publication, replicate it, or expand on it[7]. Peer-reviewers are thus doomed to vain attempts at checking the quality of the publication without access to the original data; readers have to "resurrect"[8] the spectra from the peak list to a more human-readable format; and developers have to parse and interpret these incomplete data to feed their algorithms. This absurd cycle engenders a workflow that is not only slow, but also prone to errors[7,9-11]. In this sense, one could even argue that only the (potentially wrong) results of the author's interpretation of the spectrum are being published, while the actual experimental data remains hidden. This is the opposite of what is expected of an empirical science.

The community has not been oblivious to this contradiction. Ten years ago, on the widely-read chemistry blog of Peter Murray, accompanying a low resolution image of an NMR spectrum taken from a scientific publication, one could read: "If an article costs USD 3000 then the scientific community deserves better. How many chemists have cursed the unreadability of numeric data mangled by graphics tools? There is no technical reason why the digital data shouldn't be deposited with the publisher, the institution, the department."[6] A decade ago, it was already clear that no representation could ever replace the original information and that technology was mature enough for that task.

As NMR techniques become part of the daily experimentalist's workflow and more and more structures are published, there is a growing concern about how to ensure the quality and veracity of the published assignments and structures[7,11]. In addition, it is not a simple task to define what is the minimum information that should be made available; indeed, this can keep scientists from different research fields talking for years. Therefore, it does seem very reasonable to, at least, ask for raw data to be stored and shared[4,10].



1H NMR: δ 0.99 (3H, t, J = 7.4 Hz), 2.65 (2H, q, J = 7.4 Hz), 7.13-7.25 (3H, m), 7.26 (2H, tdd, J = 7.7, 1.6, 0.5 Hz).

Figure 1: Illustration of the information lost during the current publication scheme. Raw data (up) are peak-picked, integrated, and signal multiplicities determined by the authors. This information is then summarized as a list of picked signals, sometimes referred to as "NMR text"[12], and published (mid). The "spectrum" at the bottom shows the information available to the reader after resurrection, *i.e.*, after a software is used to display the information contained in the NMR text. Although all relevant information is retained for the aliphatic region, the same does not happen in the aromatic region. Clearly, such a loss of information cannot be justified anymore by limitations of the technology available.

Note that publishing the raw spectra is a "less is more" kind of solution: since the spectrum already comes in an optimal shape right out of the spectrometer, we just need to publish it "as it is". This is not a new idea. Some authors endeavor to make their raw data available, either as supplementary material or via a website[13]. The first repositories of raw spectra appeared at least a decade ago[14]. Their importance is increasingly recognized for various applications and strategies, such as fingerprinting[15]; computer-assisted spectra analysis[16-26]; design of QSAR/QSPR descriptors to predict properties from spectra[27]; and identification of putative metabolites[28,29]. Nevertheless, it appears that availability of raw NMR data is still a significant issue.

Several dedicated databases exist[18,30] that provide spectroscopic assignments but they are just starting to include the original data[18]. ChemSpider[31] is a free *but not open* compound database that provides NMR raw files as subsidiary data. In a similar vein, metabolomics databases such as HMDB[32,33] and BMRB[34,35] include NMR data; some even provide access to the raw spectra[32,33]. Although they are not meant to become universal NMR repositories, it is worth noting that HMDB presents several features we consider desirable in a general-purpose repository (e.g. access to both raw data and assigned data, peak search capabilities).

NMRb[14], a repository of raw NMR datasets for the biosciences, seems to have disappeared. The SPECTRa project for sharing raw NMR data seems to have fallen into oblivion as well[36]. Lastly, in OSDB we found a recent effort that has not yet reached maturity[37].

The issue with these raw data NMR repositories has been that they have not given enough importance to interactivity. Offering little more than a download link to the jcamp files might not be enough. NMR spectroscopy produces complex data, requiring different forms of processing, visualization and searching. A repository needs to help different users interact with these data in the ways they want to; ways that may be as varied as the disciplines where NMR spectroscopy has earned a spot. For instance there is the view of the organic chemist, to whom NMR spectra are a means to elucidate molecular structures, and then there is the metabolomics view, in which the spectrum is a fingerprint. Organic chemists will browse the data looking for similar structure and assignment tables, while metabolomics researchers will browse the data looking for signals at a particular region of the spectra. Their different interests demand different capabilities from the repository. Other features such as peak-picking are equally crucial to both of them. We have not found a NMR repository that offers all these possibilities. Even more, doing so would not be enough, since there will always be somebody who wants to process NMR data in a previously unthought-of way. This means that extensibility is a must. What we are lacking, then, is a repository that provides extensible tools to extract useful information and to then convert that information into knowledge.

Here we present a repository that is intended to sit in between the spectrometer and the public, providing a set of efficient tools to manipulate spectra online and to browse the database. The whole system is open and its code is shared under the MIT license, as an invitation for others to join the project.

Data and Methods

The repository is composed of 3 parts, depicted in Figure 2a.

1) **A storage component** where raw NMR data and additional information can be stored from any webpage, from a third party software, or, if desired, directly from a spectrometer (See Figure 2b). This components consists of:
    a) **A data management system implemented in CouchDB**[38]. We chose CouchDB because it presents several pros for building the repository: it is document-based, it has been designed to work in a distributed manner, it is easily replicable, it scales horizontally, and it has a data-revisions functionality.
    b) **A data structure for chemical information**, described in the supplementary materials. The proposed data model has been defined using the JavaScript Object Notation[39] (JSON). The main advantages with this approach are that the natural text representation of the data is human readable and that it is supported by libraries in almost any modern programing language; in particular, it is natively supported by JavaScript in any web browser. Spectra are stored as attachments in JCAMP-DX format. Original data from

spectrometer manufacturers can be accepted in the future, whenever the format is well described. Molecules are stored separately as SDF files.

c) **A RESTful API (Application Programming Interface)** called rest-on-couch[40] that exposes the data to the web and allows the control of permissions on the documents. This API has been developed in JavaScript.

All the related sources are available in GitHub[41].

2) **A toolbox of JavaScript libraries** for data manipulation (processing and analysis)[42,43]. It is built over more than 60 libraries (some developed in-house, some borrowed from other open-source projects) that enable it to perform a gamut of operations on different kinds of information ranging from image analysis, fourier transform, multiplet analysis, spectra prediction and simulation to data mining and multivariate statistical analysis. Full details on the methods implemented in these libraries can be found in the projects' documentation; here we will just refer to those specifically concerning the NMR repository, which are available in GitHub from the cheminfo-js[44] and mljs[45] organizations:

a) Structure search uses the algorithms of DataWarrior[46].

b) 1D NMR peak picking uses the Global Spectra Deconvolution method described by Cobas *et al*[47]. 2D peak detection is performed by using the watershed algorithm for image segmentation[48] and by the identification of centroids of closed regions on the Laplacian of Gaussian of the 2D spectra[49]. If many spectra are available for the peak-picking process (e.g. both 1D and 2D ^{1}H), then a validation is performed that can identify fake or missing peaks by comparing their patterns.

c) ^{13}C-NMR chemical shifts are predicted using NMRshiftDB[17,18], while ^{1}H-NMR chemical shifts and coupling constants are predicted with Spinus[22,23]. 1D spectra are simulated from predicted shifts using the method of Castillo et al.[25]. Spin couplings in 2D spectra are predicted by calculating *n*-length paths between active nuclei in the corresponding molecule; for example, a COSY cross-peak pair is drawn for each pair of protons separated by up to 3 bonds. Cross-peaks coming from long-range couplings such as COSY couplings at >3 bonds, are included in the 2D spectrum when Spinus predicts a coupling constant > 2 Hz.

3) **A visualization tool** called *Visualizer*[50], developed in JavaScript and HTML5. Using an interface written in a programming language supported by all modern web browsers enables access to the application without having to install software on the client: everything runs in the browser and always uses the latest updates. It is built in a modular manner that allows to modify its behaviour directly from the browser, by executing code written inside the tool. This is similar to extending TopSpin's functionalities using AU programs[51] or jython[52] scripts.

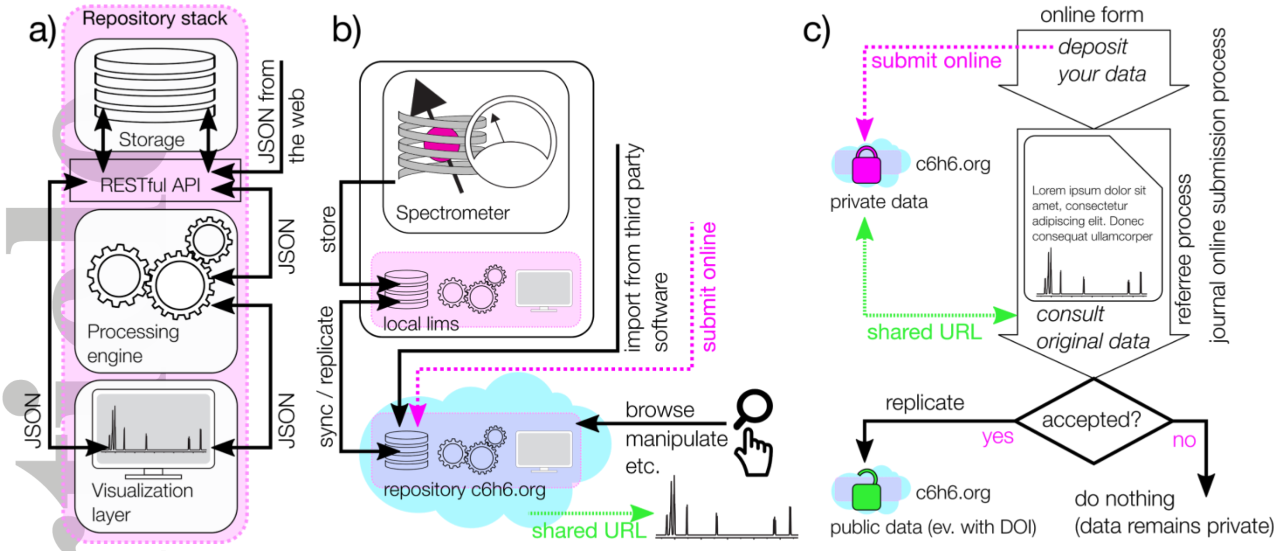


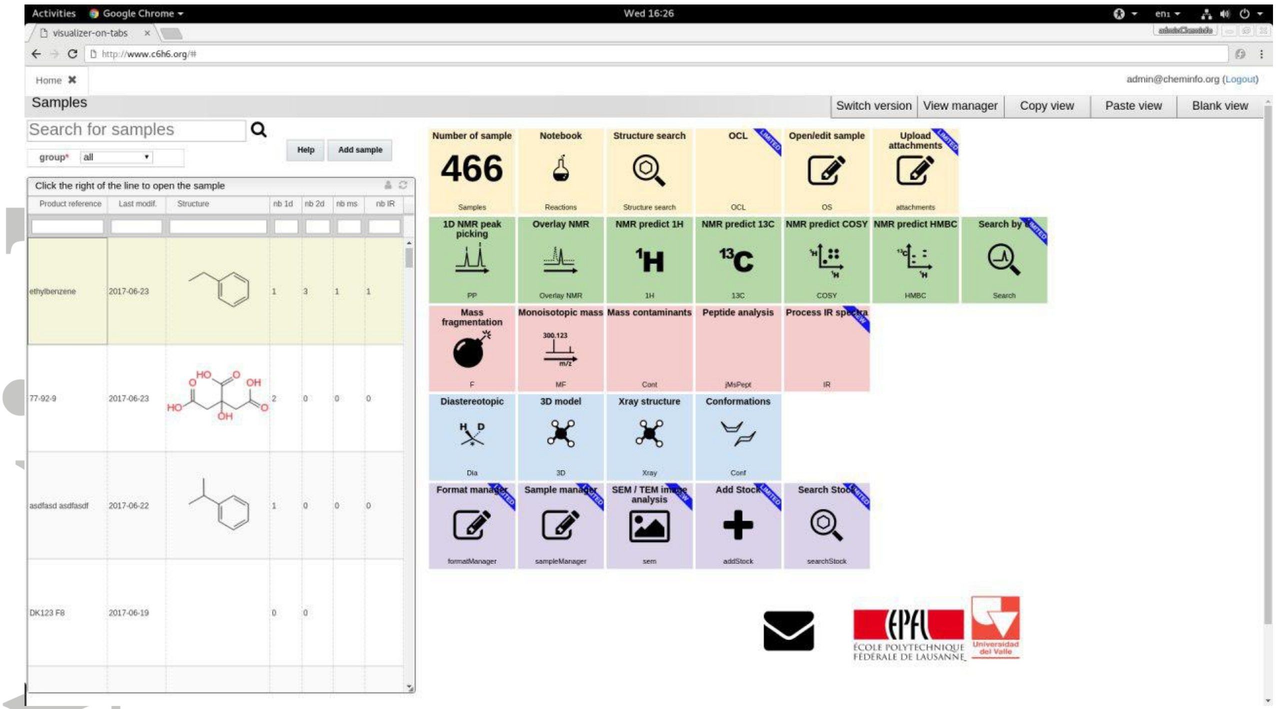
Figure 2: Schematics of the data flow of the proposed repository. a) The three components of the repository: the storage component is accessible to the web or to the other two components by a RESTful API. The processing engine and the visualization component allow to access the data and manipulate them. b) Synchronization with the laboratory: C6H6 may be directly linked to the spectrometer via synchronization/replication with a local Laboratory an Information Management System (LIMS); alternatively data may be imported from third-party software or submitted online. Efforts are underway to implement an experiment configuration and request queue, thus turning C6H6 into a full-feature LIMS. c) Insertion in the publication system: data is readily shared via an url; it is kept private during research and peer-review, then made public once the paper is accepted.

The whole system is available at github[53] and can be easily deployed using container technology[54], that permits to install and configure the service without having to install and configure all its components individually.

Results

Putting all the elements described in the previous section together allowed us to build C6H6, an NMR repository with several features to make it worth the effort consented by the users when sharing their data. A working implementation of our application can be accessed at www.c6h6.org. The concept behind C6H6 is as follows: a database of spectra (^{1}H, ^{13}C, COSY, HMBC and HSQC are currently supported) and other sample data (ID, origin, name, physical constants, see supplementary materials) is kept on secured servers. The webtool interacts with the database through the RESTful API, allowing the user to search, view, modify and download the data (See Figure 2b). Results of the query are processed on the client using JavaScript and the final result (e.g. a plot of the queried spectrum) is presented in the browser interface.

The user is asked to login before using the application; this is done in order to manage permissions and ensure safety and security of the user's data. The submitted data are private, unless the owner decides otherwise. Users may choose, for each register, whether to make it available to the public. Public data can be accessed without login, but it cannot be modified by anyone but the owner.



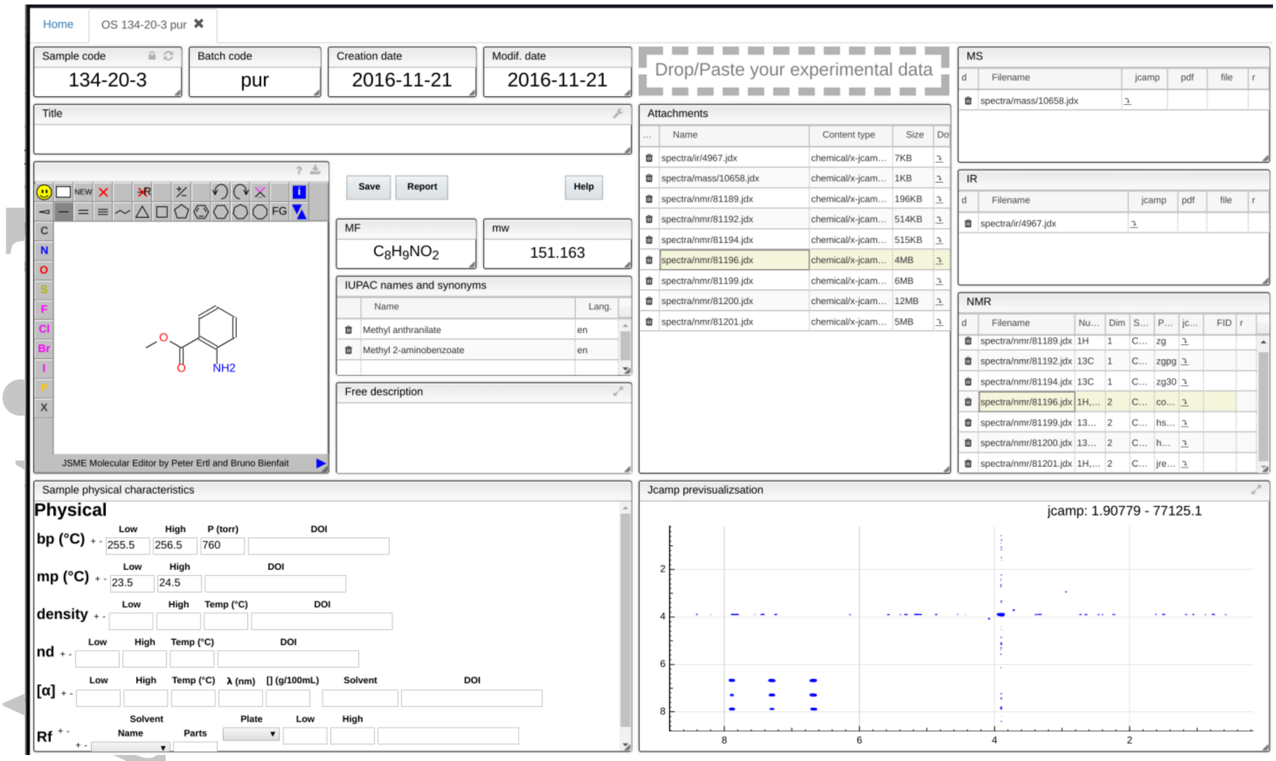
Product reference	Last modif.	Structure	nb 1d	nb 2d	nb ms	nb IR
ethylbenzene	2017-06-23		1	3	1	1
77-92-9	2017-06-23		2	0	0	0
asdfasd asdfasdf	2017-06-22		1	0	0	0
DK123 F8	2017-06-19		0	0		

Figure 3: C6H6 landing page. On the left the list of available samples with simple search features. The tiles on the right give access to a set of tools. Each will open in a new tab (in the application, not in the browser).

Figure 3 shows the main interface. On the left we have the list of samples available to the user, a basic search utility, and the *Add sample* button. On the right we can access the tool set. Upon entering or opening a sample the user is taken to a new tab[1] (Figure 4) where sample attributes (structure, name, physical constants, etc.) can be edited and spectra can be uploaded. Similarly, clicking on a tool takes the user to a new tab where an interface to perform the corresponding task is presented (Figure 5).

As it can be seen in these snapshots, support for other techniques such as Mass Spectrometry and Infrared spectroscopy is being developed. Therefore, the necessary non-proprietary formats, such as NetCDF[55,56], will be ported to our application. In addition, work is in progress to fully support and test a recently proposed standard for the inclusion of NMR assignments as associated data items in SDF files[3]. This new format may help the process of importation of data analysed using a third-party software (See Figure 2b).

[1] These are tabs inside the application, like those used by modern web browser and other software.



Sample code: 134-20-3
Batch code: pur
Creation date: 2016-11-21
Modif. date: 2016-11-21
Title:
MF: $C_8H_9NO_2$
mw: 151.163

IUPAC names and synonyms

Name	Lang.
Methyl anthranilate	en
Methyl 2-aminobenzoate	en

Free description:

Attachments

...	Name	Content type	Size	Do
🗑	spectra/ir/4967.jdx	chemical/x-jcam...	7KB	⤵
🗑	spectra/mass/10658.jdx	chemical/x-jcam...	1KB	⤵
🗑	spectra/nmr/81189.jdx	chemical/x-jcam...	196KB	⤵
🗑	spectra/nmr/81192.jdx	chemical/x-jcam...	514KB	⤵
🗑	spectra/nmr/81194.jdx	chemical/x-jcam...	515KB	⤵
🗑	spectra/nmr/81196.jdx	chemical/x-jcam...	4MB	⤵
🗑	spectra/nmr/81199.jdx	chemical/x-jcam...	6MB	⤵
🗑	spectra/nmr/81200.jdx	chemical/x-jcam...	12MB	⤵
🗑	spectra/nmr/81201.jdx	chemical/x-jcam...	5MB	⤵

MS

d	Filename	jcamp	pdf	file	r
🗑	spectra/mass/10658.jdx	⤵			

IR

d	Filename	jcamp	pdf	file	r
🗑	spectra/ir/4967.jdx	⤵			

NMR

d	Filename	Nu...	Dim	S...	P...	jc...	FID	r
🗑	spectra/nmr/81189.jdx	1H	1	C...	zg	⤵		
🗑	spectra/nmr/81192.jdx	13C	1	C...	zgpg	⤵		
🗑	spectra/nmr/81194.jdx	13C	1	C...	zg30	⤵		
🗑	spectra/nmr/81196.jdx	1H,...	2	C...	co...	⤵		
🗑	spectra/nmr/81199.jdx	13...	2	C...	hs...	⤵		
🗑	spectra/nmr/81200.jdx	13...	2	C...	h...	⤵		
🗑	spectra/nmr/81201.jdx	1H,...	2	C...	jre...	⤵		

Sample physical characteristics

Physical

	Low	High	P (torr)	DOI
bp (°C) + -	255.5	256.5	760	

	Low	High	DOI
mp (°C) + -	23.5	24.5	

	Low	High	Temp (°C)	DOI
density + -				
nd + -				

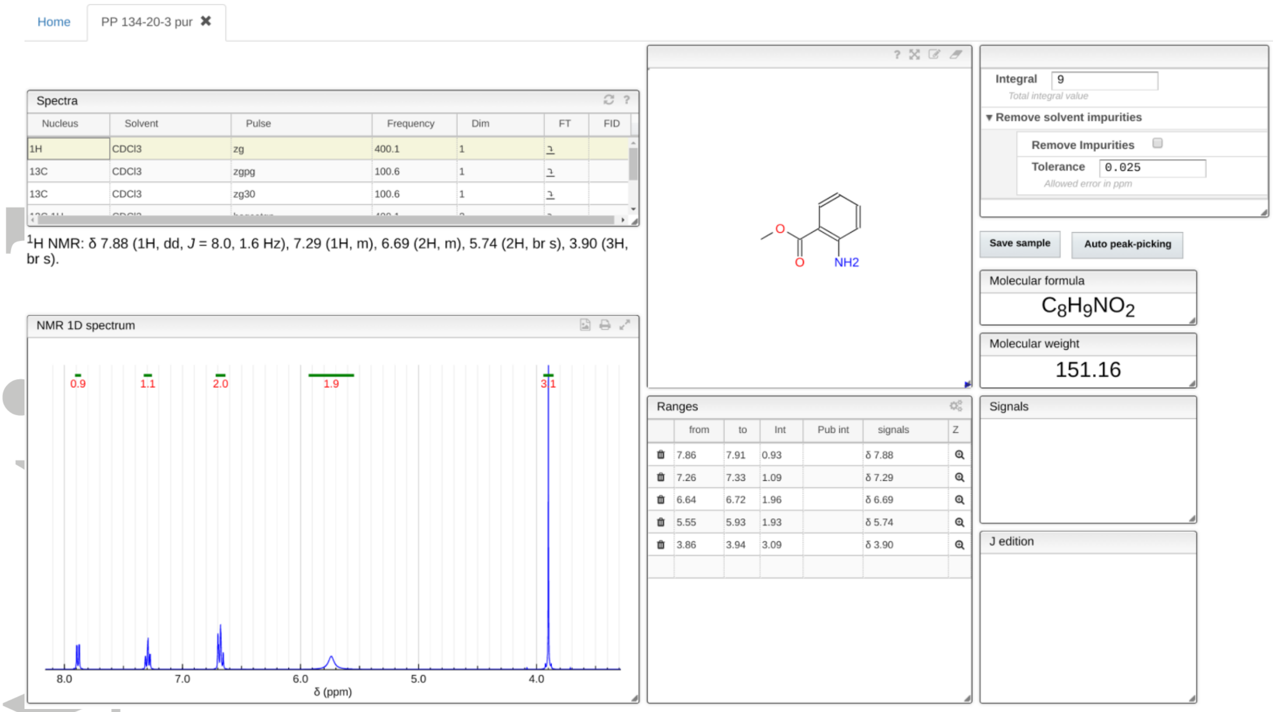
	Low	High	Temp (°C)	λ (nm)	[] (g/100mL)	Solvent	DOI
[α] + -							

	Solvent Name	Parts	Plate	Low	High	
Rf + -	▾		▾			

Figure 4: Sample edition tab. Any kind of NMR data can be uploaded using the drag & drop area. 1D and 2D spectra can be displayed and zoomed in (lower right). Accompanying structures may be drawn in the JSME editor[57] (left) or input directly as an SDF file. Additional information such as molecular formula and molecular weight are calculated on the fly, while drawing. Similarly, more complex processes can be triggered automatically when data is uploaded, such as automatic peak-picking and assignment.

Discussion

The issue we want to address with C6H6 is one of data availability and usability. NMR spectroscopy is a powerful technique that has rightfully attracted the attention of people with different backgrounds and interests. Making NMR data available means that the data published copes with the interests of all these parties. As discussed above, no optimal solution in this sense has yet been achieved by existing repositories. With C6H6 we thus intended to appeal to all disciplines where NMR is of relevance.



Spectra

Nucleus	Solvent	Pulse	Frequency	Dim	FT	FID
1H	CDCl3	zg	400.1	1		
13C	CDCl3	zgpg	100.6	1		
13C	CDCl3	zg30	100.6	1		

^{1}H NMR: δ 7.88 (1H, dd, *J* = 8.0, 1.6 Hz), 7.29 (1H, m), 6.69 (2H, m), 5.74 (2H, br s), 3.90 (3H, br s).

Integral: 9
Total integral value

▾ Remove solvent impurities

Remove Impurities □

Tolerance: 0.025
Allowed error in ppm

Molecular formula: $C_8H_9NO_2$

Molecular weight: 151.16

Ranges

	from	to	Int	Pub int	signals	Z
	7.86	7.91	0.93		δ 7.88	
	7.26	7.33	1.09		δ 7.29	
	6.64	6.72	1.96		δ 6.69	
	5.55	5.93	1.93		δ 5.74	
	3.86	3.94	3.09		δ 3.90	

Figure 5: Peak-picking interface. Peak-picking can be performed either manually (by right-clicking on the spectra) or automatically. Assignment can be performed by selecting a signal (green rectangle on top of each peak) and then selecting a proton in the molecule.

For the organic chemist, an NMR spectrum is, above anything else, a means to determine or confirm the identity of a compound of interest. This community is probably the most familiar with the data generated by the technique, but are also probably the least interested in directly reading the whole output of the spectrometer (though there will be times when they will absolutely want to!). Yet they probably will not be satisfied with just reading the peak list, either. First, because this is not the representation of the spectrum they are most adept at reading; that would be the standard 2D plot. But, most importantly, because peak lists are not “true” experimental data. Indeed, they have already gone through a peak-picking and assignment process with which they may not agree and that may well hide the presence of impurities in the sample. In fact, checking and confronting the author’s choices on this regard is a key objective of peer-reviewing in this area. Overall, researchers in organic chemistry and related disciplines seek for the ability to *graphically* navigate the spectroscopic data on different stages of processing, from full-processed peak-picked and assigned spectra to, in extraordinary cases, the rawest data provided by the spectrometer. For these users, C6H6 packages a basic set of computer-assisted spectra processing tools and plotting capabilities along with the raw data repository.

In many other fields an NMR spectrum is first and foremost a fingerprint. Metabolomics is a key example. In this community NMR is used to characterize the composition profile of complex samples of biological origin (e.g. a blood sample). The ultimate goal is to measure metabolic responses to different stimuli by detecting statistically significant changes in the NMR profiles of intervened and control samples. In this setting, assignment of signals is often not a priority and only required to confirm already identified biomarkers. Instead, other necessities arise. Access to full-resolution spectra is an obvious must: even if it may be trimmed later during data processing, the researcher wants to start with the full fingerprint to make sure that no relevant features are missed. Knowledge of the precise conditions under which the spectrum was taken is equally important, since ^{1}H-NMR spectra vary significantly (for fingerprinting purposes) with parameters such as static magnetic field intensity, buffer or solvent composition, and temperature. C6H6 allows to store such information, and to search e.g. for spectra recorded at a particular field intensity. Tools such as spectra superposition and signal search may also be useful to the metabolomics community.

For the developer data availability means access to raw, full-resolution data that is key to the development of new methods and algorithms used to process and analyze spectra. Indeed, reliable data sets are needed to test and validate new methods. This is true not only for NMR, and one can look into other fields to realize what could happen to computer-assisted NMR analysis once data availability is given the importance it deserves. For example, in the field of visual media analysis, ImageCLEF[58] provides developers with curated datasets of images that developers use for training and testing their algorithms. The impact of this initiative is not to be underestimated, as ImageCLEF has become a major driving force behind a rapidly developing field. We believe a similar initiative could be equally valuable to the development of automatic NMR analysis; though surprisingly, previous attempts[14] seem to have faced limited success.

But beyond the importance of testing sets to validate new methods for NMR analysis is the requirement of larger datasets *allowing* the method itself to *function*. Chemical shift prediction is the clearest example: all state-of-the-art chemical shift predictors *need* a database of assigned spectra to work[17-24,26]. For all we know, this is probably the way it will always be[26]. Chemists often complain about the high cost of commercial suites for computer-assigned NMR analysis, but this cost acknowledges the enormous value of the spectra databases that actors in the private sector have managed to amass. If public science wants to compete with these corporations and produce its own, freely available applications, it needs its own, freely available NMR databases.

It must be emphasized that C6H6 is not just a webpage or web service: it is a full-fledged application designed to run in the browser. The “webpage” is intended to perform as traditional software; all the code necessary to perform the task is downloaded and executed within the browser. This solves the issue of operating system compatibility that often makes downloading and installing new software a troublesome task. Furthermore, the application is designed to be readily extensible: the JavaScript code necessary to perform new tasks can be stored inside the tool or implemented as an external library and called directly by the *visualizer*. In this manner, using the JavaScript language allows us to benefit from the biggest and fastest growing development community over the world. Finally, the same code can be executed either on the server side or directly in the browser (client side), a very powerful argument in favor of javascript programing language that makes it very suitable for mobile devices.

Ongoing efforts are currently focused on converting C6H6 into a full LIMS (See figure 2b) that provides a request and queueing system to encourage users to provide at least a minimal set of data to describe their experiment. This queue can be configured to automatically set up and trigger experiments in order to ensure that most experiments are performed with optimal parameters. Once the experiment is finished the data are automatically sent back to the server, thereby ensuring that all the data are correctly stored, thus improving the traceability of the data. Validated data can then be shared via an url and streamlined into the publication pipeline (see Figure 2c).

From a decade of running NMR facilities, we understand that sharing data is not a spontaneous act. It should be encouraged either by lowering the effort required to share and by providing added value, which are goals that we attempted to achieve, or by including this as a requirement within the publication pipeline (See Figure 2c). We appeal to the NMR community and publishers to include depositing raw data as a necessary condition for publication. We insist that sharing of raw data may be key to the future of public science for fundamental reasons. Over the past years, reproducibility of scientific studies has emerged as a major concern. Providing open access to the raw data, experimental designs and source code has been proposed as one avenue to mitigate this problem. Therefore, sharing electronic laboratory notebooks represents a true step towards incremental and reproducible science[59]. LIMS and repositories are key elements in the construction of such new ways to share and publish research outcomes. While the technology is there, the main challenge remains

Conclusions

We believe the repository presented here enables a sharing and publication model that can warrant the quality, traceability, transferability and agility required by contemporary NMR-related research; an ideal that the current publication system has not properly achieved.

But the sole existence of a repository does not guarantee that these ideal will be achieved. In order to ensure the comprehensiveness and correctness of the data stored, the repository needs to be inserted into the peer-review and publication pipeline, which in turns demands the collaboration of researchers and publishers. We invite the NMR community to participate in this joint effort.

References

[1] M.D. Wilkinson *et al*, *Sci. Data*. **2016**, 3:160018. DOI:10.1038/sdata.2016.18

[2] A.M. Clark, A.J. Williams, S. Ekins, *J. Cheminform.* **2015**, *7*:9. DOI:10.1186/s13321-015-0057-7

[3] D. Jeannerat, *Magn. Reson. Chem.* **2016**, *55*, 7-14. DOI:10.1002/mrc.4527.

[4] J. Wist, *Magn. Reson. Chem.* **2016**, *55*, 22-28. DOI:10.1002/mrc.4533.

[5] P. Murray-Rust, R. Smith-Unna, *D-Lib Magazine* **2014**, *20*. DOI:10.1045/november14-murray-rust

[6] P. Murray-Rust, "Save our spectra". https://blogs.ch.cam.ac.uk/pmr/2007/08/12/save-our-spectra/ [9 june 2017]

[7] G.F. Pauli, M. Niemitz, J. Bisson, M.W. Lodewyk, C. Soldi, J.T. Shaw, D.J. Tantillo, J.M. Saya, K. Vos, R.A. Kleinnijenhuis, H. Hiemstra, S.-N. Chen, J.B. McAlpine, D.C. Lankin, J.B. Friesen, *The J. Org. Chem* **2016**, *81*, 878–889. DOI:10.1021/acs.joc.5b02456.

[8] D. Banfi, L. Patiny, *Chimia* **2008**, *62*, 280-281. DOI:10.2533/chimia.2008.280

[9] A. Eklund, T. E. Nichols, H. Knutsson, *Proc. Nat. Acad. Sci.* **2016**, *113*, 7900-7905. DOI:10.1073/pnas.1602413113.

[10] J. Bisson, C. Simmler, S.-N. Chen, J. Brent Friesen, D.C. Lankin, J.B. McAlpine, G.F. Pauli, *Nat. Prod. Rep.* **2016**, *33*, 1028. DOI:10.1039/c6np00022c

[11] W. Robien, *Trac-Trend. Anal. Chem.* **2009**, *28*, 914-922. DOI:10.1016/j.trac.2009.03.012

[12] NMR guidelines for ACS journals. http://pubs.acs.org/paragonplus/submission/acs_nmr_guidelines.pdf [September 21, 2017]

[13] K.A. Badiola, D.H. Quan, J.A. Triccas, M.H. Todd, *PLoS ONE* **2014**, *9*, e111782. DOI:http://dx.doi.org/10.1371/journal.pone.0111782

[14] J. L. Pons, T. E. Malliavin, D. Tramesel, M. A. Delsuc, *Bioinformatics* **2004**, *20,* 3707-3709. DOI:10.1093/bioinformatics/bth450.

[15] J.G. Napolitano, D.C. Lankin, T.N. Graf, J. Brent Friesen, S.N. Chen, J.B. McAlpine, N.H. Oberlies, G.F. Pauli, *J. Org. Chem.* **2013**, *78*, 2827–2839. DOI:10.1021/jo302720h

[16] C. Steinbeck, S. Krause, S. Kuhn, *J. Chem. Inf. Comp. Sci.* **2003** *43*, 1733-1739.

[17] S. Kuhn, N.E. Schlörer, *Magn. Reson. Chem.* **2015**, *53*, 582-589. DOI:10.1002/mrc.4263.

[18] NMRShiftDB website. http://nmrshiftdb.org [June 21 2017]

[19] N. Haider, W. Robien, *Nachr. Chem.* **2016**, *64*, 196–198. DOI:10.1002/nadc.20164047147

[20] CSEARCH / NMRPREDICT-Server. http://nmrpredict.orc.univie.ac.at/ [21 June 2017]

[21] Wiley SSR. https://www.wsslabs.com [July 4 2017]

[22] Y. Binev, M.M. Marques, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2007**, *47*, 2089-2097. DOI:10.1021/ci700172n

[23] SPINUS website. http://www2.chemie.uni-erlangen.de/services/spinus/ [1 June 2017]

[24] ACD/LABS NMR shift predict webpage. http://www.acdlabs.com/products/adh/nmr/nmr_pred/ [23 June 2017]

[25] A.M. Castillo, L. Patiny, J. Wist. *J. Magn. Reson.* **2011**, *209*, 123-130. DOI:10.1016/j.jmr.2010.12.008.

[26] A. M. Castillo, A. Bernal, R. Dieden, L. Patiny, J. Wist, *J. Cheminform.* **2016**, *8*:26. DOI:10.1186/s13321-016-0134-6

[27] R.P. Verma, C. Hansch, *Chem. Rev.* **2011**, *111*, 2865–2899 DOI:10.1021/cr100125d

[28] D. Tulpan, S. Léger, L. Belliveau, A. Culf, M. Cuperlović-Culf, *BMC Bioinformatics* **2011**, *12*, 400. DOI:10.1186/1471-2105-12-400

[29] K. Bingol, L. Bruschweiler-Li, D. Li, B. Zhang, M. Xie, R. Brüschweiler, *Bioanalysis*, *8*, 557-573. DOI:10.4155/bio-2015-0004

[30] SDBSWeb. http://sdbs.db.aist.go.jp (National Institute of Advanced Industrial Science and Technology) [June 21 2017]

[31] Chemspider website. www.chemspider.com [June 21 2017]

[32] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, *Nucleic Acids Res.* **2013**, *41*(D1):D801-7.

[33] The Human Metabolome Database http://www.hmdb.ca/ [June 21 2017]

[34] E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R.K. Wenger, H. Yao, J.L. Markley, *Nucleic Acids Research* **2008**, *36*, D402-D408 DOI:10.1093/nar/gkm957

[35] Biological Magnetic Resonance Data Bank website. http://www.bmrb.wisc.edu/ [21 June 2017]

[36] SPECTRa website. https://spectradspace.lib.imperial.ac.uk:8443/handle/10042/25 [27 June 2017]

[37] S.J. Chalk, *J. Cheminform.* **2016**, *8*:55. DOI:10.1186/s13321-016-0170-2

[38] CouchDB website. http://couchdb.apache.org/ [29 June 2017]

[39] JavaScript Object Notation Description. https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/JSON [29 June 2017]

[40] rest-on-couch project source. https://github.com/cheminfo/rest-on-couch [29 June 2017]

[41] Cheminfo sources. https://github.com/cheminfo [29 June 2017]

[42] Lactame.com libraries. www.lactame.com [29 June 2017]

[43] https://www.npmjs.com/search?q=maintainer:cheminfo-bot [29 June 2017]

[44] Cheminfo-js sources.https://github.com/cheminfo-js [29 June 2017]

[45] mljs sources. https://github.com/mljs [29 June 2017]

[46] DataWarrior website. http://www.openmolecules.org/datawarrior [1 June 2017]

[47] C. Cobas, F. Seoane, S. Domínguez, S. Sykora, A.N. Davies, *Spectrosc. Eur.* **2011**, *23*, 26-30.

[48] L. Belaid, W. Mourou, *Image Anal. Stereol.* **2011**, *28*, 93-102. DOI:10.5566/ias.v28.p93-102

[49] J. Canny, *IEEE T. Pattern Anal.* **1986** *6*, 679-698.

[50] Visualizer project source. https://github.com/npellet/visualizer [29 June 2017]

[51] NMR Suite AU Programs Reference Manual. Bruker Analytik GmbH, **1999**

[52] Jython website. http://www.jython.org/ [July 4 2017]

[53] roc-eln-docker project source. https://github.com/cheminfo/roc-eln-docker [29 June 2017]

[54] Docker website. https://www.docker.com/ [29 June 2017]

[55] R.K. Rew, G. P. Davis, *IEEE Comput. Grap.* **1990**, *10*, 76-82.

[56] NetCDF website. https://www.unidata.ucar.edu/software/netcdf/ [4 July 2017]

[57] B. Bienfait, P. Ertl, *J. Cheminform.* **2013**, *5*:24.

[58] ImageCLEF@ICPR website. http://www.imageclef.org/2010/ICPR [7 June 2017]

[59] A.E. Williamson, P.M. Ylioja, M.N. Robertson, Y. Antonova-Koch, V. Avery, J.B. Baell, H. Batchu, S. Batra, J.N. Burrows, S. Bhattacharyya, F. Calderon, S.A. Charman, J. Clark, B. Crespo, M. Dean, S.L. Debbert, M. Delves, A.S.M. Dennis, F. Deroose, S. Duffy, S. Fletcher, G. Giaever, I. Hallyburton, F.-J. Gamo, M. Gebbia, R. Kiplin Guy, Z. Hungerford, K. Kirk, M.J. Lafuente-Monasterio, A. Lee, S. Meister, C. Nislow, J.P. Overington, G. Papadatos, L. Patiny, J. Pham, S.A. Ralph, A. Ruecker, E. Ryan, C. Southan, K. Srivastava, C. Swain, M.J. Tarnowski, P. Thomson, P. Turner, I.M. Wallace, T.N.C. Wells, K. White, L. White, P. Willis, E.A. Winzeler, S. Wittlin, M.H. Todd, *ACS Cent. Sci.*, **2016**, *2*, 687–701. DOI:10.1021/acscentsci.6b00086