

The Royal Society of Chemistry and the delivery of chemistry data repositories for the community

Antony Williams · Valery Tkachenko

Received: 16 April 2014 / Accepted: 25 July 2014 / Published online: 3 August 2014
© Springer International Publishing Switzerland 2014

Abstract Since 2009 the Royal Society of Chemistry (RSC) has been delivering access to chemistry data and cheminformatics tools via the ChemSpider database and has garnered a significant community following in terms of usage and contribution to the platform. ChemSpider has focused only on those chemical entities that can be represented as molecular connection tables or, to be more specific, the ability to generate an InChI from the input structure. As a structure centric hub ChemSpider is built around the molecular structure with other data and links being associated with this structure. As a result the platform has been limited in terms of the types of data that can be managed, and the flexibility of its searches, and it is constrained by the data model. New technologies and approaches, specifically taking into account a shift from relational to NoSQL databases, and the growing importance of the semantic web, has motivated RSC to rearchitect and create a more generic data repository utilizing these new technologies. This article will provide an overview of our activities in delivering data sharing platforms for the chemistry community including the development of the new data repository expanding into more extensive domains of chemistry data.

Keywords Data repository · ChemSpider · Crowdsourcing · InChI · NoSQL

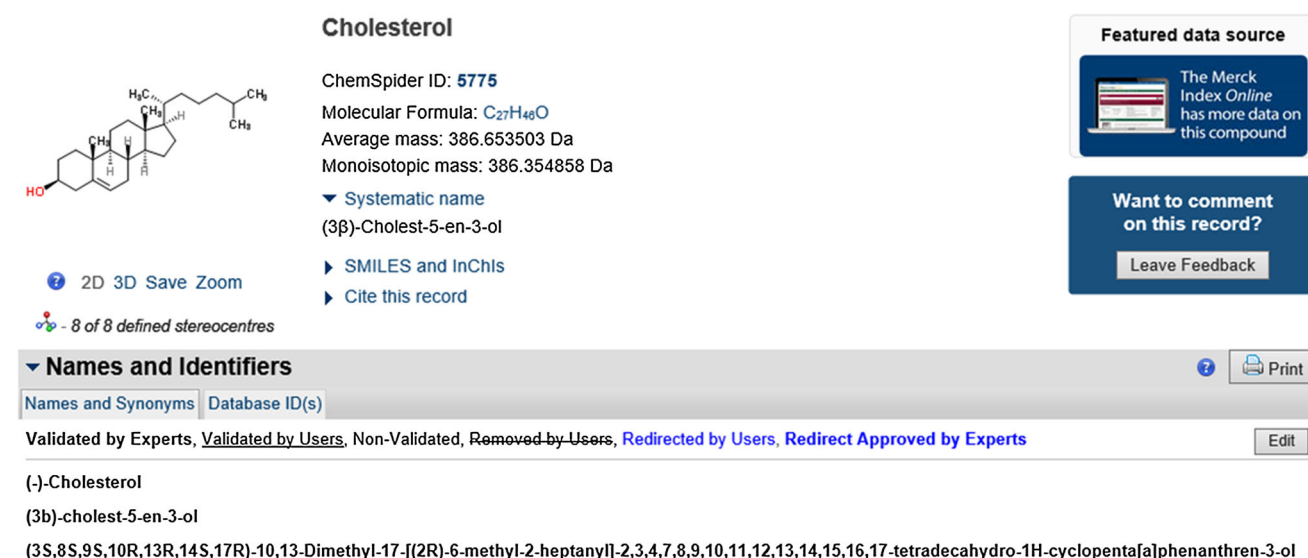
Introduction

The Royal Society of Chemistry (RSC) is the largest European organization with the specific mission of advancing the chemical sciences. With a worldwide network of almost 50,000 members and an international publishing business, activities span education, science policy and the promotion of chemistry to the public. In recent years, and as part of our mission, we provide access to chemistry related data and cheminformatics tools including ChemSpider [1–3], an online database and repository for chemistry which provides a platform for the community to source chemistry related data. However, in keeping with the growth of the so-called social web we ensured that users of the platform were also able to annotate and validate the data in the database as well as deposit and share their own data. This approach is consistent with the wiki approach, a so-called crowdsourcing activity. ChemSpider quickly became one of the chemistry community's primary platforms for sourcing data but was limited by the fact that only chemicals that could be represented by an InChI [4] could be hosted on the database. This significantly limited the handling of ambiguous materials, many organometallics, polymers and other classes of chemicals. Clearly a true data repository for chemistry should not be limited to only small organic molecules. In parallel new technologies are now available that offer greater flexibility for the handling of unstructured data and for making the data available to the growing semantic web.

ChemSpider

ChemSpider is a free, online chemical database providing access to chemicals and chemistry related data for over

A. Williams (✉) · V. Tkachenko
Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest,
NC 27587, USA
e-mail: tony27587@gmail.com



Cholesterol

ChemSpider ID: **5775**

Molecular Formula: **C₂₇H₄₆O**

Average mass: 386.653503 Da

Monoisotopic mass: 386.354858 Da

▼ Systematic name
(3β)-cholest-5-en-3-ol

► SMILES and InChI

► Cite this record

2D 3D Save Zoom

8 of 8 defined stereocentres

Names and Identifiers

Names and Synonyms Database ID(s)

Validated by Experts, Validated by Users, Non-Validated, Removed by Users, Redirected by Users, Redirect Approved by Experts

(-)-Cholesterol

(3b)-cholest-5-en-3-ol

(3S,8S,9S,10R,13R,14S,17R)-10,13-Dimethyl-17-[(2R)-6-methyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopenta[a]phenanthren-3-ol

Print

Edit

Featured data source

The Merck Index Online has more data on this compound

Want to comment on this record?

Leave Feedback

Fig. 1 The header of the chemical record for cholesterol (<http://www.chemspider.com/5775>) in ChemSpider. The entire record spans multiple pages including links to patents and publications, pre-

calculated and experimental properties and links to many data external data sources and informational websites

thirty million unique chemical compounds, sourced and linked out to over five hundred separate data sources on the Web, and multiple services providing access to patents, publications and cheminformatics-based property prediction and services. ChemSpider is not just a search engine layered on enormous quantities of chemistry data, it is also a crowdsourcing platform for chemists who contribute their data, skills, and knowledge to the enhancement and curation of the database.

ChemSpider was initially developed as a hobby project to contribute a free resource to the chemistry community. Released in March 2007 with a seed collection of just over 10 million chemicals it expanded to over 20 million chemicals over the next 2 years and new functionality was added to facilitate content curation and depositions of data. ChemSpider was acquired by the RSC in May 2009 and the original vision of providing a structure-centric database for chemistry was expanded to focus on becoming the world's foremost free access chemistry database.

ChemSpider can be described as the “Google for Chemistry” and a “Wikipedia for chemists”. By aggregating data from over 500 different data sources and connecting them by means of chemical structure as the primary record in the database, ChemSpider has been able to link many popular online databases and resources including Wikipedia, chemical vendor websites, Google Patents and Books, and both open- and closed-access chemistry journals. Where possible, each chemical record retains the links out to the original source of the data. These links let a ChemSpider user source information of particular interest,

including where to purchase a chemical, chemical toxicity, metabolism data, and so on.

The database content in ChemSpider (exemplified in Fig. 1) has been developed as a result of contributions and depositions from chemical vendors, commercial database vendors, government databases, publishers and individual scientists. The database can be queried using structure/substructure searching and alphanumeric text searching of chemical names and both intrinsic, as well as predicted, molecular properties. Various searches have been added to the system to cater to various types of users including, for example, mass spectrometrists and medicinal chemists. ChemSpider is very flexible in its applications and the diverse nature of the available searches.

Curation involves ensuring the accuracy of the data in a digital database and registered users can enter information and annotate and curate the records. The chemical community has been forthcoming in adding information, including new chemical structures, associations between structures and publications, analytical data such as spectra, and in the curation of chemical identifiers and property data. Accessing scientific publications is performed by integrating to open internet services on websites such as PubMed [5], Google Scholar [6] and Google Patents [7] (see Fig. 2). Validated chemical names, either approved by users of the database or using dictionaries developed within the RSC, are used as the basis of a search against the PubMed database that is restricted to only title and abstract. All of these application programming interfaces (APIs) are called in a similar way: a list of approved synonyms associated with a particular ChemSpider record is listed,

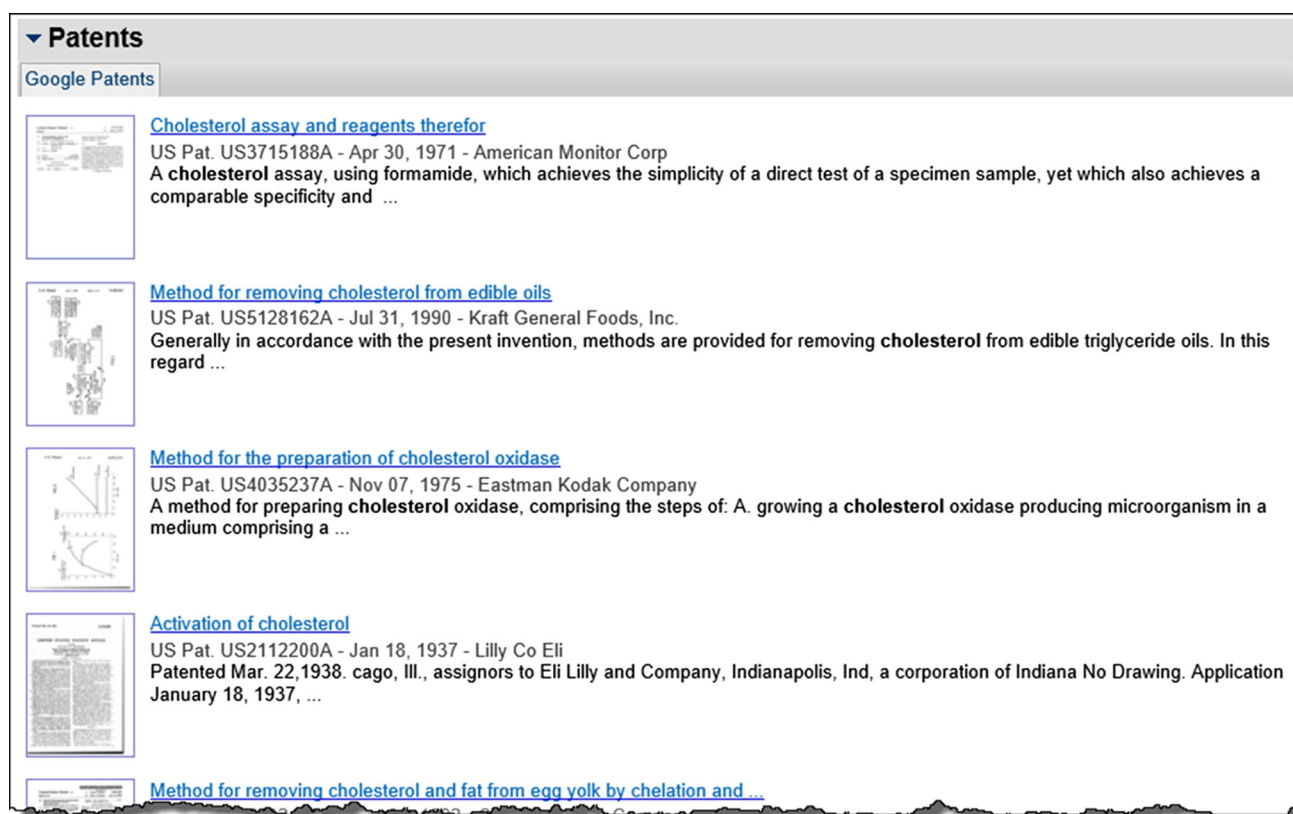


Fig. 2 One of multiple pages of patents retrieved from Google Patents utilizing the series of validated names associated with cholesterol as the basis of a text search against the Google Patents programming interface

sorted by “relevance” (which is calculated based on the length of the synonym as well as its clarity), and then used to call against the API.

For many users of the ChemSpider database users have informed us that the primary value of the analytical data is as *reference* data. The users utilize the data on ChemSpider to compare with their own lab-generated data. ChemSpider therefore provided the ability to upload spectral data of various forms against a chemical record such that an individual chemical can have an aggregated set of analytical data to assist in structure verification. Over 3,000 spectra have been added to ChemSpider by users, with additional data being added regularly. These data include infrared, Raman, mass spectrometric and NMR spectra with the majority being ^1H and ^{13}C spectra. Spectral data can be submitted in JCAMP [8] format and displayed in an interactive spectral display widget allowing zooming and expansion.

We can ask the question “Wouldn’t a search of the web suffice for retrieving chemistry data?” Certainly standard search engines can return a lot of hits based on a text-based search for a particular chemical, a search on ChemSpider has a number of advantages over a simple Google search. The variety of information about a compound provided at

ChemSpider is hard to match on any other free Web site and new data continues to be validated and updated by practicing chemists, and in many cases they have been reviewed for accuracy by the user community.

Using ChemSpider as a building block

ChemSpider “web services” provide programmatic access to ChemSpider and this allows for other platforms to utilize the services and harvest data content for use inside their own systems. As an example the NMR spectral data contained within the database are the basis for the Spectral Game [9], which has already been used by more than 10,000 students in nearly 100 countries. This game lets students learn how to interpret NMR spectra by validating either ^1H or ^{13}C spectra against two or more structures. The game increases in complexity as it progresses; ultimately, students must choose a spectrum match from among five structures. By using the game players are also participating in data curation as we review patterns in the data to see if we can catch errors. For example, when a specific spectrum shows a lot of users are incorrectly matching the compound to spectrum we examine the data

to check for issues. As a result we have been able to remove spectra with very intense water peaks and in two cases the JCAMP spectrum was reversed and needed to be reprocessed.

The ChemSpider web services also allow for instrument vendors to utilize the chemical data for compound identification by mass spectrometry. As an example the programming interface is used to query specific data sets using a set of parent ion masses extracted from an LC–MS spectrum to identify the various components. Examples include the identification of metabolites by using masses extracted from an LC–MS analysis of urine and searching against a limited number of datasets on ChemSpider containing human metabolite data.

The data are also available to the Open PHACTS project [10], a project funded by the Innovative Medicines Initiative [11]. The RSC is one of the key participants in the project providing a chemical registration service derived from some of the approaches implemented on ChemSpider and presently being worked on to release as an Open Source platform to the community. Work is also underway to support the PharmaSea [12] project for helping to identify new natural products from the ocean and, ultimately, to host some of the data associated with the resulting chemical compounds.

Serving up commercial databases to the community

RSC has also deployed the UK-centric Chemical Database Service (CDS) [13] offering access for academic scientists to a suite of commercial databases and services. The project is a 5 year grant-funded project with the development of a chemistry data repository as one of the deliverables (vide infra). The CDS acts as a hub to a number of databases including reaction databases, thermophysical and crystallographic data and various other forms of information including available chemicals from vendors. The services include prediction algorithms for physicochemical properties, NMR spectroscopy prediction, systematic nomenclature generation and various other facilities.

Ongoing cheminformatics projects at the Royal Society of Chemistry

While there are numerous commercial reaction databases, there is no free database of chemical syntheses that the community can contribute to or comment on. In order to extend the crowdsourcing capabilities to syntheses a new database was established known as ChemSpider SyntheticPages (CSSP) [14]. The community is fully responsible for populating the database with their contributions, as

CSSP is essentially a publishing platform (an example article is shown in Fig. 3). The system can host multimedia content, spectral data and links to the ChemSpider database.

After submission, a SyntheticPage is reviewed by one or more members of the editorial board and the comments are made available to the author, who is informed via email. The author then makes edits online and, when accepted by the editorial board, the article is published. This differs from the classical review process as the original review and feedback is only between the editorial board and the author rather than via an anonymous and extended review process through a set of selected reviewers. The process is generally very fast relative to classical review, commonly <48 h. Once published the article can then be commented on by the community. CSSP hosts hundreds of synthetic procedures and new submissions are made regularly. The RSC Reactions database, also presently under development, is intended to be the largest freely accessible database of chemical syntheses and is expected to release over half a million reactions to the community. It will be the host for all reactions extracted from the archive using text-mining approaches (vide infra).

The Chemical Validation and Standardization Platform (CVSP) is a freely available internet-based platform for the community for the processing of chemical compound datasets. The platform both validates and standardizes chemical structure representations according to sets of systematic rules. The chemical validation algorithms detect issues with submitted molecular representations using predefined molecular patterns that are chemically suspicious or potentially requiring manual review. Each identified issue is assigned one of three arbitrary levels of severity (Information, Warning, and Error) in order to inform the user of the need to browse and review subsets of their data. The validation includes validation of atoms, bonds, valences, and stereo. This platform has already been applied to the analysis of a large number of data sets prepared for deposition to our ChemSpider database and in preparation of data for the Open PHACTS project. The CVSP is expected to be released as Open Source software by Summer 2014.

The most active cheminformatics related projects at RSC presently under development will extend the reach and utility of the data for the community as well as providing deep integration into our scientific article archive. The DERA project (Digitally Enabling the RSC Archive) will extract data from the entire RSC archive of over 300,000 articles and will include chemical compounds, reactions, spectral data, property data and so on. The data will then be made available to the community via the appropriate interfaces.

Hydrogenation of Ethyl 3-(1-pyrenyl)acrylate; Ethyl 3-(1-pyrenyl)propanoate

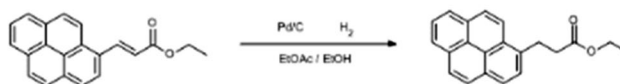
SyntheticPage 510

DOI: 10.1039/SP510

Submitted Sep 29, 2011, published Oct 06, 2011

Anish Mistry (a.mistry@warwick.ac.uk)

A contribution from Fox Group, Warwick University



Chemicals Used

Ethyl 3-(1-pyrenyl)acrylate (1 equiv, prepared)

10% Pd/C (2 mol%, Alfa Aesar)

Ethanol

Ethyl Acetate

Hydrogen gas

Procedure

Ethyl 3-(1-pyrenyl)acrylate (3.02 g, 10.1 mmol) was firstly dissolved in ethyl acetate (70 ml). Ethanol (70 ml) and the 10% Pd/C (2 mol%) were then added to the mixture and the air evacuated out the system and replaced with hydrogen. The reaction was left to stir under a hydrogen balloon at room temperature for 29 hours. The reaction mixture was then filtered through a pad of celite with ethyl acetate and the solvent removed under *vacuo* to yield a dark yellow solid (3.02 g, 99%).

Fig. 3 A screenshot showing an example reaction from ChemSpider SyntheticPages

From ChemSpider to generic data repository

One of our key future directions for which development is well underway is the RSC Data Repository. This repository for research data will utilize our experience in data management for chemistry to create a new flexible architecture that will extend us outside of the limitations of the original ChemSpider model which could only host chemical compounds for which InChIs could be generated. The new repository will include a number of content buckets including chemical compounds, materials, reactions, crystallographic data, spectral data, tables of data, integrated electronic notebook content etc. that will be both segregated yet federated. As an example of the change in capability, while ChemSpider contains spectral data held only as data associated with a chemical compound in the database they are *not* contained in a spectral database such that the data can be searched using spectral searching. These separate content buckets will be developed under the new architecture to provide access via programmatic APIs so that the content, searches and visualization components can be utilized to deliver multiple platforms matched to the needs of the various user communities. The data repository will be open to the community to host their data in the

cloud and they will be able to share data collaboratively, under embargo or publicly, if they choose.

The high-level architecture of the Data Repository (DR) is presented in Fig. 4. Raw structured data in various cheminformatics formats will be derived from a number of sources including web pages, APIs and endpoints, popular file sharing services, Electronic Lab Notebooks (ELNs) and general structured documents which can be text-mined for chemical information. All of this information will be passed through a Deposition Gateway (see Fig. 5) which will segregate out the information by domain type (compounds, reactions, spectra, etc.), will map the data fields to semantically known types (SMILES, InChI, names, properties, etc.), will perform chemical validation and standardization (the CVSP platform is a part of the Deposition Gateway), and then display the results to the end-user and curator. Once all issues are analyzed and fixed, or erroneous records removed from the deposition, then the data is spooled into the appropriate databases. The Deposition System is being constructed in a modular manner that allows for the addition of support for new data types without rebuilding the core pipeline. We hope to add more modules in future to expand our coverage of chemistry.

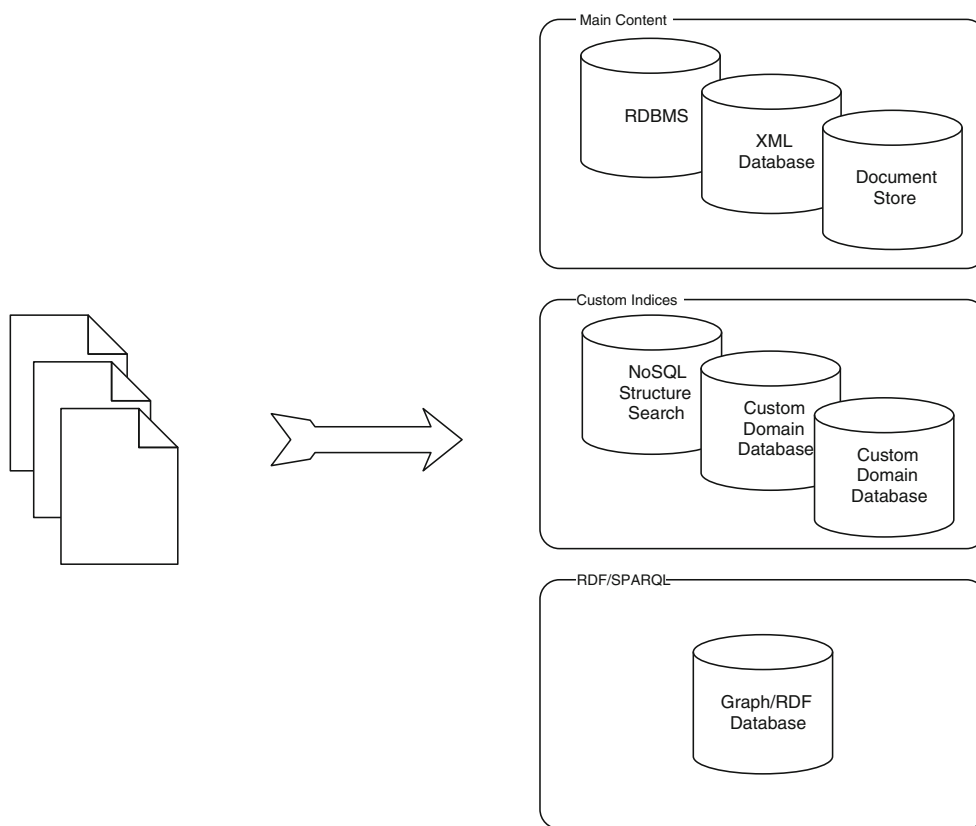


Fig. 4 The data are hosted in the appropriate databases and content stores appropriate to handling both structured and unstructured data

Our experiences in building registrations systems is that only part of information is best handled by relational databases (RDBMS) as they tend to lock the data model, and any change or expansion of that model is a very painful process. An alternative to this approach is well known: the use of a schema-less NoSQL database. Currently we utilize a heterogeneous architecture where the data types are stored in their respective databases based on the best suitability principle: essentially relational data which require real-time changes are hosted in an RDBMS and hierarchical data which are better represented as a whole document are stored in either XML or some other complex document store. Currently the platforms we are investigating are BaseX [15] for XML documents and MongoDB [16] for content that is being actively exchanged with a user interface.

Custom (non-alphanumeric) information from specific domains such as chemical structures, reactions and spectra cannot be handled by either of the available general purpose database management systems and traditionally this information was handled by installing chemically intelligent cartridges into relational databases. With more open source tools available it has become possible to easily build NoSQL database solutions for custom domains.

One of the most interesting perspectives of research data management systems is to allow for reasoning and the discovery of new data relationships in the quickly expanding semantic world of Big Data. While theoretically this can be done in any database management system the most efficient setup necessarily includes general graph databases. In our design graph databases serve as the glue connecting together relevant information from other databases allowing the user to quickly reason and find relevant data and possibly perform a deeper analysis by retrieving respective data from those stores.

Since data volumes are likely to be very large, and research information is by its very nature of a distributed decentralized nature, the system has to be capable of managing large distributed volumes of data. Fortunately all database management systems that we currently use support replication, clustering and data distribution natively in one way or another, although some development work is necessary to make it all work together.

While one potential impression is that the system being built is too heterogeneous and difficult to manage, and in theory could be purchased as a monolithic and likely expensive one-size-fits-all platform, the positive view is that all of these complexities are managed at the lower level of the

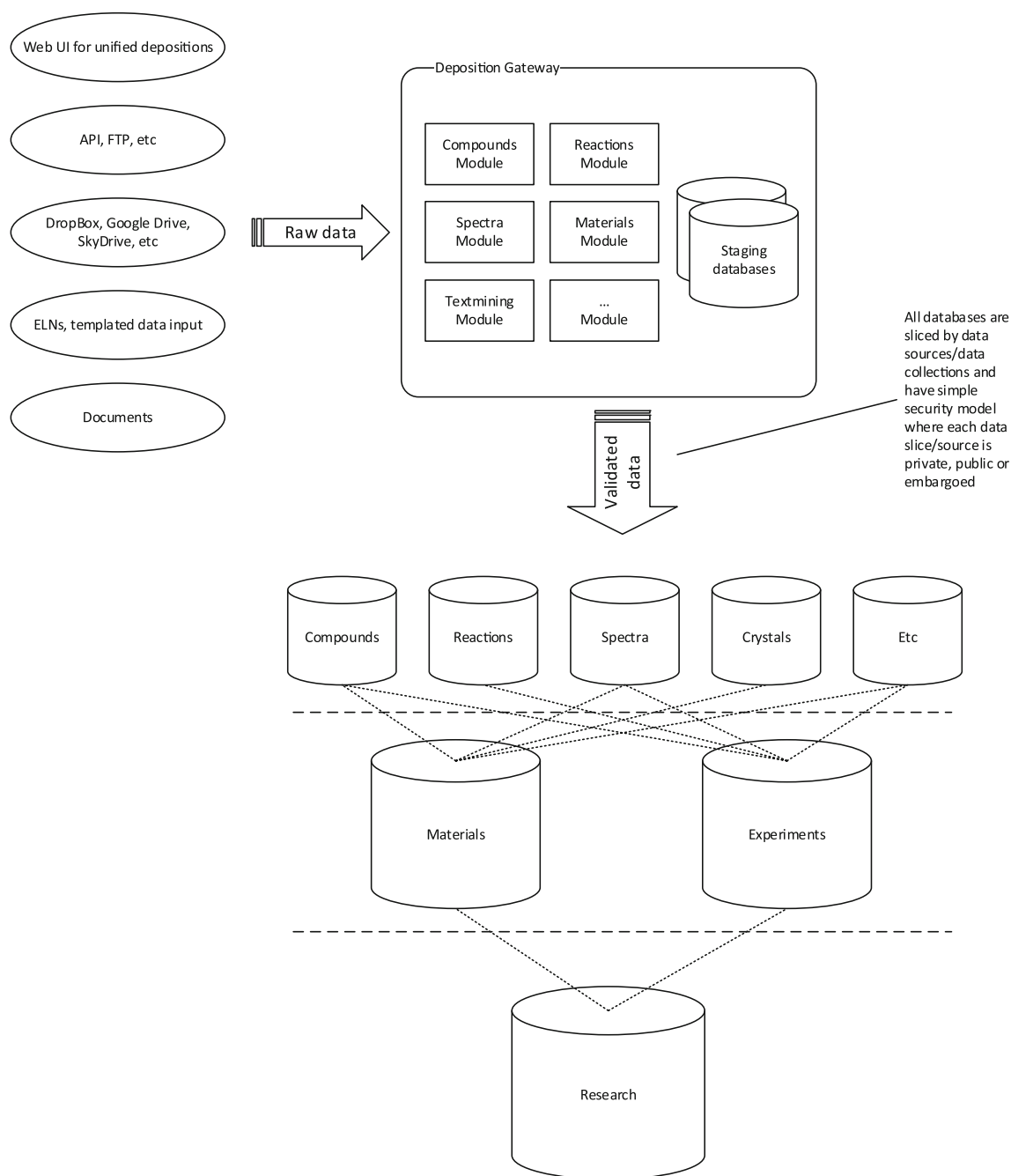


Fig. 5 The deposition gateway processes, validates, segregates and distributes the data into the appropriate containers

system architecture, and even from within the system all operations on objects are performed with a well-defined business model and API. Taking into account that the components of the system we are investigating are all Open Source and in general use by thousands of users around the world, we believe that the inherent flexibility of our approach is far superior to a single dedicated vendor solution.

On the client side the system is built using a set of layers of functionality which allow for delivery of its

services to be exposed as API endpoints (see Fig. 6) as a set of visual user interface elements readily available to build interfaces developed by RSC, for use in third party applications, and a series of completed cheminformatics frameworks.

This multi-year project will require that the challenges of complex data management for the chemical sciences be addressed, that the complexities of data licensing and sharing be navigated, and that the appropriate tools for

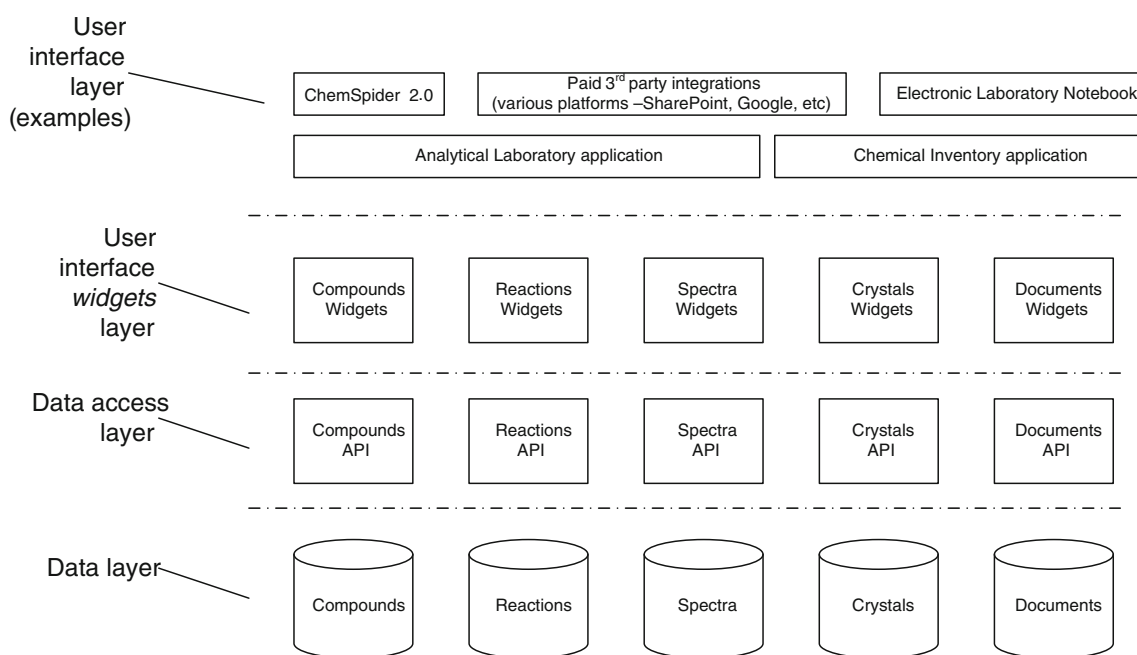


Fig. 6 The multi layer architecture for the data repository—data, data access, collection of user interface widgets and pre-configured user interfaces. In the figure ChemSpider 2.0 refers only to an

searching and displaying the myriad forms of information be developed.

Conclusion

The true collaborative benefits of RSC's cheminformatics platforms such as ChemSpider and the developing data repository will have the greatest impact when they are integrated into federated searches and semantic web linking. Based on the established growth in popularity of our platforms over the past few years we hope that our data repositories and contributions to the open source cheminformatics community will be some of the key building blocks for enabling collaboration and integration for chemistry and will become valuable resources for future generations.

Acknowledgments ChemSpider is the result of the aggregate work of many contributors extending outside of our own team. Our RSC platforms are supported by a dedicated team of IT specialists. The authors acknowledge the support of the Open Source community, the commercial software vendors (specifically Accelrys, ACD/Labs, GGA Software, OpenEye Scientific Software, Dotmatics and many data providers, curators and users for their contributions to the development of the data content in terms of breadth and quality.

References

1. Pence H, Williams AJ (2010) J Chem Educ 87(11):1123
2. Williams AJ (2011) Public compound databases—how ChemSpider changed the rules making molecules on the web free. In: Ekins S, Hupcey MAZ, Williams AJ (eds) Collaborative computational technologies for the life sciences. Wiley, Hoboken, p 363
3. Williams AJ (2010) ChemSpider: integrating structure-based resources distributed across the internet. In: Belford R, Moore JW, Pence HE (eds) Enhancing learning with online resources, social networking, and digital libraries, vol 1060. American Chemical Society, Washington, p 23
4. The IUPAC International Chemical Identifier (InChI). <http://www.iupac.org/inchi/>. Accessed 16 April 2014
5. PubMed. <http://www.ncbi.nlm.nih.gov/pubmed/>. Accessed 16 April 2014
6. Google Scholar. <http://scholar.google.com/>. Accessed 16 April 2014
7. Google Patents. <http://www.google.com/patents>. Accessed 16 April 2014
8. Published JCAMP-DX Protocols. <http://www.jcamp-dx.org/protocols.html>. Accessed 16 April 2014
9. Bradley JC, Lancashire RJ, Lang AS, Williams AJ (2009) J Cheminform 1(1):9
10. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B (2012) Drug Discov Today 17(21–22):1188
11. Hunter AJ (2008) Drug Discov Today 13(9–10):371
12. PharmaSea. <http://www.pharma-sea.eu/>. Accessed 16 April 2014
13. Chemical Database Service. <http://cds.rsc.org>. Accessed 16 April 2014
14. ChemSpider Synthetic Pages. <http://cssp.chemspider.com>. Accessed 16 April 2014
15. BaseX: The XML Database. <http://www.basex.org>. Accessed 16 April 2014
16. MongoDB. <http://www.mongodb.org/>. Accessed 16 April 2014