TONY DAVIES COLUMN

Keeping the dream alive

Antony N. Davies,[a,b] Peter Lampen,[c] Stephen R. Heller[d] and Evan Bolton[e]
[a]Strategic Research Group – Measurement and Analytical Science, Akzo Nobel, Deventer, the Netherlands
[b]SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of Glamorgan, UK
[c]Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany
[d]InChI Trust, Silver Spring, Maryland 20902, USA
[e]National Center for Biotechnology Information, US National Library of Medicine, Bethesda, MD 20894, USA

Work on a new home for the spectroscopic data from the International Spectroscopic Data Bank (IS-DB) is currently underway and so we thought it fitting to look back on its inception and think about how far we have come.

During the joint industrial/academic conference on "Linking and Interpreting Spectra through Molecular Structures" (LISMS) hosted by Warwick University in 1996, a motion was carried which was the basis of the initiative to produce a spectroscopic data submission system to support auxiliary data in scientific publications. At the time this vision was strongly influenced by a number of factors. Increasing computing power was starting to allow more complex data analysis which needed to be fed by better documented and more comprehensive reference quality spectroscopic data—the existing collections were, with few exceptions, measured on older technologies and represented only a tiny fraction of known chemistries. Even the biggest reference spectroscopic collection at that time represented less than 1% of the known chemistry as documented by the Chemical Abstract Service Index numbers. An excellent lead had been taken in this area by the Protein Crystallography community where it was expected that serious crystallographers always deposited their reference data sets used to produce their scientific publications. These crystallographic coordinates could then be used to assist all crystallographers in their own work by being made available through the Protein Crystallographic Data Bank. The idea behind the Warwick Challenge was

"The Need for an Analytical Reference Data Archive: A Resolution"
3 September 1996

The chemical, pharmaceutical and materials industries are a major economic force and job provider in Europe. Keeping research and development abreast of the rest of the world is important to the scientific and economic success of Europe.

Confirming and elucidating chemical structures are major tasks in the discovery and development of new products, in quality control and in environmental analyses. There are currently in excess of 14,000,000 registered chemical compounds, and more than 500,000 new ones are added each year. Although analytical data (from separation science and spectroscopic methods) are used during the synthesis, purification and identification of all of these compounds, few of the data are available, with the chemical structures, in a form useful to the academic or industrial analytical community.

The largest electronic collections of analytical data represent 1% or less of the known chemical structures. It is estimated that as many spectra are recorded in industrial and academic laboratories in a single day as are contained in the largest electronic analytical databases. Nearly all of these spectra are discarded or are unavailable, even to those who acquired them.

Access to large electronically stored collections of spectroscopic and separations data stimulates significant progress in chemical research and in automated methods for structure/spectrum and structure/biological-activity correlation. This has wide implications for human health, new materials, environmental protection, sustainable development and educational progress.

Combinatorial chemistry is a major advance for discovering new materials and new chemical compounds for human health, crop protection or other uses. Rapid methods for confirming chemical identity depend critically on access to large analytical data sets. Time and money are often spent duplicating analyses of known compounds simply because archival data are not available. The efficiencies gained by enabling access to analytical data archives will contribute to maintaining competitive European industrial and academic research and development.

It is unlikely that any single company or institute could take on the effort of building such an electronic repository. It is more appropriate that the initial funding stimulus for this project come from international public sources. Eventually the repository would become self-funding through fees for access to the data. However, without EU support, the project will not begin or achieve enough momentum to sustain itself.

Because such a collection of analytical data would be an important European scientific resource, members of the European analytical chemistry community strongly encourage the European Union to include an electronic analytical data repository as a priority area in the forthcoming Fifth Programme.

Text Box 1. The Warwick Challenge.

TONY DAVIES COLUMN

to replicate the infrastructure used to collate, quality control and publish the PDB for spectroscopic data.[1,2]

Meeting the challenge

With a small team based around the conference organisers committed to following up on the Challenge the fight for funding began. The details of how this system should work were hammered out at a number of conference events both in Europe and the USA in the following two years. The support of the major scientific publishers, both academic and purely commercial, was sought; as their backing was seen as key to the success of the project.

One of the goals of the project was to establish a workflow which would allow for a form of enhanced peer review process to be put in place. Submitted spectroscopic data would initially be made available to the publishers and through a secure link to the reviewers of the paper to which the spectra belonged. This concept would enable an enhanced peer review though access to the full high-resolution spectroscopic data to ease the difficult job reviewers have to carry out, by confirming or allaying fears about misinterpretation of spectroscopic data often represented only as graphical images. The plan was that, once the article had been accepted for publication, the spectroscopic data and associated chemical information would be made available through links in the electronic versions of the particular publication. There was also a Publishers' Consultative Committee planned to make sure that developments in the industry which might affect the project were monitored, so that when the system finally went live the modus operandi had not been made obsolete by some development in the publishing world.

The planners designing the system were well aware that there was strong support in the community for the initiative—but that if the data which the scientists had taken time and trouble to compile were not made available rapidly after submission so that they could start to see the benefits of collaborating, then this support would quickly wain. To ensure that the needs of the community were kept in the focus of the project team an End-User Consultative Committee was also planned. A compromise was agreed with the spectroscopic software vendors in that the project would not seek to produce any search options by spectral matching but leave that to the specialist vendors, data submitted to the archive would be available as single data sets by indexing through their reference chemical and literature metadata. It is pleasing to see that, what was at the time-quite a ground-breaking concept, has become commonplace for accompanying data in a multitude of scientific publications, and this position will continue in the new access to the archive envisaged.[3]

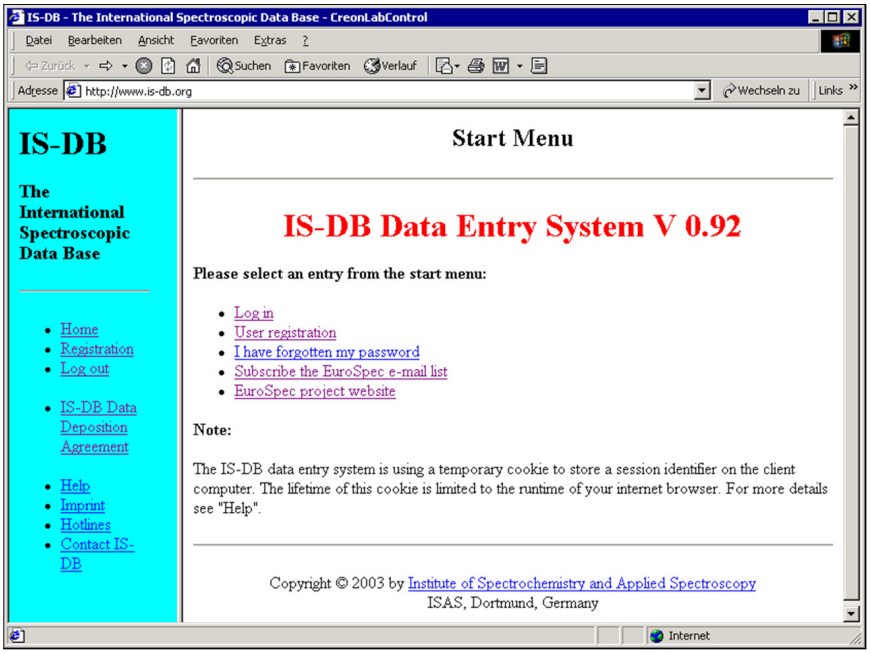

Figure 1. The IS-DB data entry system for spectroscopic data and associated metadata at launch.

Funding

European Union Research Framework funding was sought under the Infrastructure section of the Competitive and Sustainable Growth programme but it took two Expressions of Interests before funding was assigned, as the initial reviewers were keen to see better planning for the survival of the submitted data beyond the requested funding period. With this additional planning in, the Expression of Interest received the highest review score of all applications in the second call and went forward for detailed planning.[4]

Over seven years after the original conference and Challenge was issued, EU funding allowed the project to launch

The EuroSpec Consortium

- ISAS, Institute of Spectrochemistry and Applied Spectroscopy, Dortmund, Germany
- CreonLabControl AG, Frechen, Germany
- LGC, Runcorn, United Kingdom
- INA P-G, Institut National Agronomique Paris – Grignon, Paris, France
- ICT, Institute of Chemical Technology, Prague, Czech Republic
- SPECS and BioSPECS BV, Rijswijk, The Netherlands
- IM Publications, Chichester, United Kingdom
- University of Aveiro, Aveiro, Portugal

Text Box 2. The original EuroSpec Consortium.

TONY DAVIES COLUMN

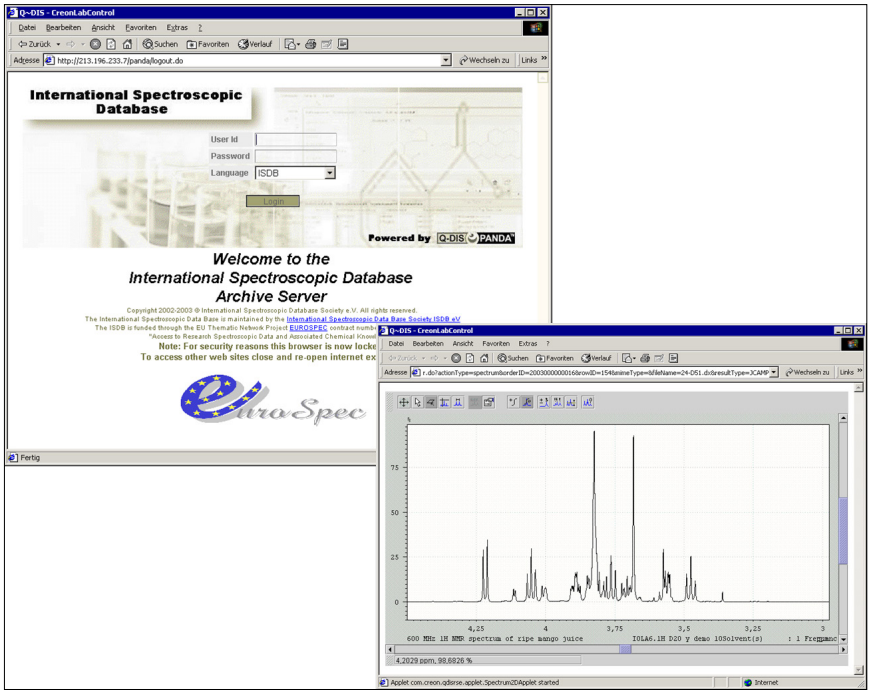

Figure 2. The modified Q~DIS/ Panda web server for the archived spectroscopic data.

in January 2002 with a strong European consortium (Text Box 2).

The project was able to build on earlier work around the collection and handling of spectroscopic data and so was able to hit the ground running. The data deposition front-end had to be built from scratch, but the web-front-end was supplied by CreonLabControl GmbH, Cologne, Germany, based on a customised and re-configured Q~DIS/ Panda electronic scientific data management system. The Achema conference was chosen to launch the IS-DB system on 19 May 2003. The project team was proud to be able to launch the twin IS-DB Data Entry and IS-DB Archive Server systems one month ahead of schedule.[5]

The infrastructure that was required to be set up to handle the rights questions to the deposited spectra saw the founding of a registered charity under German law called the International Spectroscopic Data Base e.V., who were required to receive non-exclusive rights to the use of the deposited spectra and would hopefully be able to be the organ through which the longer term stability of the system could be organised beyond the lifetime of the European Union funding. Strangely this charity has proved more stable than a number of the original participating organisations! The society has been greatly supported by the original ISAS institute in Dortmund (now the "Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V.") which was the main location for the data handling servers and continued to serve as the location for the host servers for many years after the project completed.

Keeping the dream alive

With the original team moving to different roles in various organisations, active support for the systems has fallen essentially on one man. Deposition of fresh data ceased several years ago and with some prompting by the German tax and charity authorities it was decided that the time has come to find a more permanent, non-commercial home for the submitted data in line with the original data deposition guarantees. A number of scientific organisation were considered as potential successors, but the fine print of their systems reserved the right to make the data available in a restricted commercial manner in the future which is in breach of the constitution of the current rights holders. So, the Board of the charity decided, through the good offices of Steve Heller, to contact Evan Bolton at the PubChem Project at the National Center for Biotechnology Information, US National Library of Medicine in Bethesda, USA. The PubChem Project deals in small molecules and they have a fundamental ethos around making quality data available for the general scientific good which is completely aligned to the IS-DB e.V. society.[6] In recent weeks, Peter Lampen and Evan Bolton have been working through the mechanics of transferring the spectra and the associated metadata to PubChem with a view to keeping the data available, initially through the PubChem FTP portal, but with a view to integrating it into the main data archive over time (Figure 3). This will hopefully ensure the continuing availability of the IS-DB collection to the spectroscopic community until long after all of the original participants have retired!

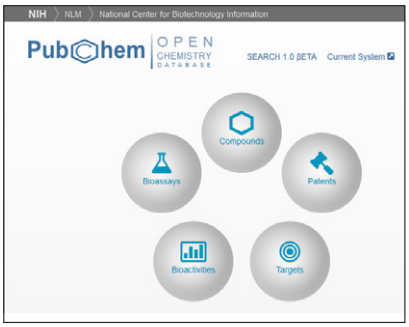

Figure 3. The PubChem search interface.

References

1. A.N. Davies, "Halfway up the stairs—The Warwick Challenge", *Spectrosc. Europe* **8(5),** 30–33 (1996).
2. A.N. Davies, D.V. Bowen, M.M. Cashyap, R. Hillhouse, J. Hollerton and K. Taylor (Eds), *Linking and Interpreting Spectra through Molecular Structures*. IM Publications, Chichester, UK (1997). ISBN: 1-901019-01-2
3. A.N. Davies, "An update on the International Spectroscopic Data Bank Project", *Spectrosc. Europe* **13(5),** 24–26 (2001). bit.ly/1RMaTPU
4. *EuroSpec—Access to Research Spectroscopic Data and Associated Chemical Knowledge*. EU Grant No. G7RT-CT2001–05063.
5. A.N. Davies, P. Lampen and P. Hughes, "EuroSpec goes live at Achema", *Spectrosc. Europe* **15(3),** 25–26 (2003). bit.ly/1QsTolb
6. https://pubchem.ncbi.nlm.nih.gov/