An update on the International Spectroscopic Data Bank Project

Antony N. Davies

External Professor, University of Glamorgan, Wales, UK, and Institute of Spectrochemistry and Applied Spectroscopy, Dortmund, Germany

As regular readers of this column will know there have been reports at irregular intervals on the ongoing attempt to establish an international spectroscopic data bank. The concept is based on the successful protein crystallographic data bank (PDB) into which the key electronic coordinate data is placed following almost ever peer-reviewed protein crystallographic publication.[1] This databank is then made available to the crystallographic community and is a key resource for those active in this field.

Back in 1996 at the first LISMS (Linking and Interpreting Spectra through Molecular Structures) conference in Warwick, UK, a motion was passed which established the need for a similar resource for depositing spectroscopic data.[2,3] The main driving force behind this motion was the increasingly worrying gaps, which were becoming apparent in the coverage of known chemistry by reference spectroscopic databases [see text box].

This gap can be quite dramatically shown in Figure 1, where the number of chemical substances registered with the Chemical Abstracts Service up to the end of July last year is shown. With the addition of the biosequences to the database the number of substances has rapidly overtaken the number of documents cited and has been the cause of the disproportionately rapid increase in the total number of substances registered. Whether or not these are taken into account, the proportion of substances with reference spectroscopic data available (green curve) is tiny compared with the amount of known chemistry. The up-to-date numbers of chemical substance registrations has now passed 33 million compared to only 19.6 million when the support documentation for the EU dedicated call was prepared last year and 14 million in 1996 when the pharmaceutical and chemical industry representatives were worried enough

"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.

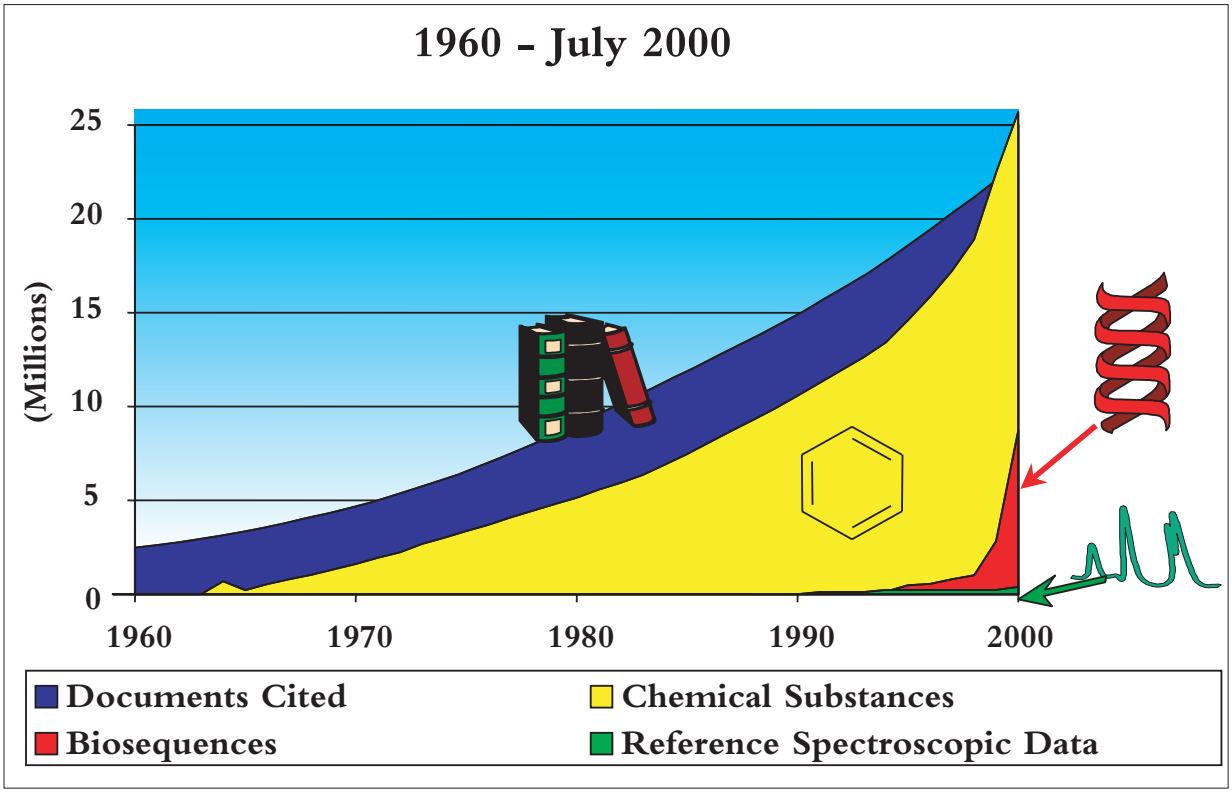

Figure 1. A comparison of the growth in documents, chemical substances, biosequences and reference spectroscopic data.

about the current state-of-the-art to pass the resolution cited in the supporting document!

The motion called on the European Union to make this a priority for the fifth Framework programme. Further discussion meetings were held at various conferences between 1996 and 1998 in Europe as well as in the USA, resulting in two formal so-called Expressions-of-Interest (EOIs) in establishing this spectroscopic databank system in 1999 and in 2000. Both EOIs were positively assessed and resulted in dedicated calls for the establishment of a Thematic Network within the EU Competitive and Sustainable Growth programme (Call identifiers: Growth – Dedicated call 10/99: M&T, Infrastructure and Growth Dedicated call October 2000: M & T, infrastructure).

The response to the first call failed as the reviewers added an additional proviso requesting details of the envisaged project structure beyond the lifetime of the EU funding period, which had not been foreseen in the initial call. This deficit was corrected in time for the second call and the consortium succeeded in putting together a proposal called EUROSPEC which achieved the highest ranking of all proposals in that particular review round. This should be seen not only as a great success for those involved in the consortium but also as a major political success for spectroscopy as a whole as it clearly establishes the essential nature of our science within the general sphere of EU priorities.

So where do we go from here? The project will begin on 1 January 2002 and will obviously take some time to get the infrastructure established for data submission. However, as various projects in the recent past have already developed essential building blocks this will hopefully lead to a rapid development time. Providing no unforeseen problems appear you should be able to begin depositing spectroscopic data associated with articles in the journals pioneering this development during 2002.

The Publishers

One of the most encouraging aspects of the work involved in setting up this project has been the universal encouragement and support from the scientific publishers, both commercial publishers

Date	**Fri Sep 21 10:06:45 EDT 2001**
Count	**18,683,587** organic and inorganic substances
	14,349,381 sequences
Total	**33,032,968** chemical substance registrations
CAS RN	**357913-38-3** is the most recent CAS Registry Number

Figure 2. Chemical substance registrations as of 21 September 2001.

and academic societies. I have been amazed at their willingness to advance the aims of the project and at a time when many have come under attack for their commercial attitude to publishing scientific results this has to be acknowledged and applauded!

There is currently a list of some 18 European and American scientific journals suggested by their publishers for inclusion in the start-up phase of the project. One of the first steps will be to bring the publishers' representatives together to decide on the exact wording of changes to their respective "Instructions for Authors" to request or require electronic deposition of the spectroscopic data associated with their articles.

Enhanced peer-review

Initially the submitted spectroscopic data will only be made available to the publishers and through a secure link to the reviewers of the paper in question. This is to enable an enhanced peer-review where access to the full spectroscopic data often eases the difficult job reviewers have to carry out by confirming or allaying fears about misinterpretation of spectral data. Once the article has been accepted for publication the spectroscopic data and associated chemical information can then be made available for viewing through links in the electronic versions of the particular publication or, for those using the printed version, by typing in the unique URL into a web browser.

Data availability

For this project to succeed the data authors must quickly see a benefit in their own work of having this data available. One of the valid worries initially expressed by the EU was the need to ensure that the benefits gained through their investment in the project during the first three years didn't evaporate on 1 January of year four! This, as well as the desire to make the data available under license to the companies currently supplying the spectroscopic community with reference spectroscopic data and some excellent manipulation software, has led to compromise being proposed and agreed in principle with the database suppliers. The spectroscopic data in the databank will be accessible as single files linked to their associated chemical data such as names, chemical structures etc. However, the project will not be supplying a spectral search engine interface—preferring to leave this type of utilisation to the organisations in a better position to cover needs in this area. This will allow the limited resources available within the project to be better targeted on the solving the problems of setting up the infrastructure to handle and make available the submitted data.

Checks and balances

One of the important features of the project will be the advisory bodies. The publishers will establish one advisory body to ensure that the project is aware not only of current trends especially in the are of electronic publishing but also of discussions on possible new directions which the industry might be considering taking. This is essential if the project is to avoid making decisions which might result in resources being invested in areas which become irrelevant within a short time.

The second advisory body to be established will be an end-users consultative committee whose primary duty will be to advise on the level of acceptability of the technological solutions put in place. There is always a danger in information technology strong projects of this nature that the participants want to use the latest/fastest/prettiest solution which—although giving the system developers bragging rights in the bar in the evening—tend to alienate the end-users who are often not in a position to have the latest hardware or software on their desktops!

Both bodies will serve as the conduit through which problems with using the systems can be assessed and fed in into the project for consideration.

Publicity

Finally, the demand for this project has come from academic and industrial spectroscopists alike. In order for it to succeed and for a really useful resource to be generated the community needs to support it. This will require some extra work for the individual spectroscopists but this effort should be repayed many times over in the coming years. As we need as many groups as possible to adopt spectroscopic data deposition to their standard working practises publicising the project will be a major part of the initial stages of the initiative. The first steps will be taken at the IRDG (Infrared and Raman Discussion Group) meeting at the LGC-NW (Laboratory of the Government Chemist – North West) meeting hosted by Andy Brookes on 1 November—see you there!

References

1. The Protein Data Bank is operated by the Research Collaboratory for Structural Bioinformatics (RCSB). For more information see http://www.rcsb.org/pdb/ and the main European partner is the Cambridge Crystallographic Data Centre in the UK, see http://pdb.ccdc.cam.ac.uk/pdb/.
2. A.N. Davies, D.V. Bowen and M.M. Cashyap, R. Hillhouse, J. Hollerton, and K. Taylor (Eds), *Linking and Interpreting Spectra through Molecular Structures*. IM Publications, Chichester, UK, ISBN 1 901019 2 (1997).
3. A.N. Davies, "Halfway up the stairs—The Warwick Challenge", *Spectroscopy Europe* **8(5),** 30–33 (1996).