**Science, Education & Industry**

Thomas Graham House

Science Park, Milton Road

Cambridge CB4 0WF, UK

**Tel** +44 (0) 1223 420066

**Direct line** +44(0) 1223 432655

**Emai**l ileyj@rsc.org

**www.rsc.org**

Rt Hon David Willetts MP

Chair, Research Sector Transparency Board

House of Commons

London

SW1A 0AA

22nd July 2013

Dear Minister,

**Open data: oversight of, and standards for, curation of subject repositories**

I am writing following our brief discussion at the Foundation for Science and Technology meeting at the Royal Society on July 10th, and your suggestion I write to the Board.

The issue the Royal Society of Chemistry (RSC) wishes to raise is this: in the emerging world of open data that results from government-funded research, there is an apparent lack of any co-ordinated national approach to the oversight of data at discipline level; who then is responsible for the data taxonomies and standards that will ensure federated searches will retrieve consistent and coherent results?

In discussion with both HEFCE and EPSRC it is quite clear that both bodies consider the responsibility for enabling access to data resides with individual higher education institutions. This poses a serious problem for those wishing to access data within specific discipline areas as they will be confronted by needing to interrogate potentially some 200 separate institutional repositories. Even if it were possible to overlay a search facility that aggregated the search outputs from these repositories, there is no guarantee that the data in each repository will have been tagged with metadata in a consistent way.

For this reason, the RSC considers it to be essential for there to be some kind of national oversight at a subject level, setting standards for data deposition, and either providing the aggregating facility or taking responsibility for subject data curation.

Alongside our role in administering and developing the EPSRC-funded National Chemical Data Service for the academic community, the RSC itself has a major initiative underway to build what we have given the working title of “the Global Chemistry Network” – networking the world’s chemistry data resources for data sharing, collaboration and open innovation. This will be a ground-breaking, world-leading information resource with high applicability across the broader science domain, and built upon the RSC’s existing freely-available database of 28,000,000 chemical structures called ChemSpider ([www.chemspider.com](http://www.chemspider.com)). This visionary service is intended to meet a number of needs, including:

* *Research data management*

In response to grant-funded research data needing to be openly available to the wider community and on an international scale, we intend to deliver a system that ensures that research data can be effectively managed and not lost, with possible embargos accepted as one of the conditions. Software application programming interfaces are being developed for the community, so that publicly available data can be consumed and reused for the development of local models and integration into internal systems such as Electronic Laboratory Notebooks. Discussions are already underway to initiate streaming of data into the repository from leading universities. The Global Chemistry Network will enable deposition, community-based annotation and sharing of user data, creating a repository of data around which international research collaborations and innovation can be facilitated.

* *Research Collaboration and Innovation*

The Global Chemistry Network is being developed to serve the chemistry community as an environment for Open Innovation. Through this platform, the RSC will connect major companies, SMEs, individual experts, funders, idea generators and academic researchers together in an interdisciplinary environment for crowdsourcing innovation on a global scale. The RSC’s global network of chemical scientists will be combined with its expertise in content delivery via its online platforms to increase the efficiency and effectiveness of the innovation process; for example, by providing access to new technologies and ideas, facilitating technology transfer, and building collaboration with clearly defined outputs.

* *Access to predictive algorithms and data models*

This capability will provide access to predictive algorithms and data models, including those derived from appropriate data hosted within the Global Chemistry Network. This is very useful because database look-ups for experimental data cannot always deliver the desired results whereas predictive algorithms can provide users with estimated values in lieu of experimental data.

* *RSC Data*

The RSC holds a vast amount of content beyond published articles. This is largely in the form of scientific data contained within the research article as well as supplementary information, but also includes other data sources hosted by RSC. The potential this holds for the scientific community is hugely significant. We are now extracting the chemical data from our journal articles, books and other content, including educational resources, to a degree not attempted before by any publisher so that this data can be fully linked, data mined, repurposed and re-presented to different communities of chemical scientists. We are already starting to do this with third-party content (for example, a dataset of 500,000 searchable reactions linked to publicly-available patent literature).

These kinds of initiatives need a broad base of support, and not just from within the scientific communities themselves. Therefore, we urge the Research Sector Transparency Board to address this issue of discipline oversight, and we would be very willing to discuss our initiatives with the Board if that were appropriate.

Yours sincerely



Professor Jim Iley

Executive Director, Science and Education